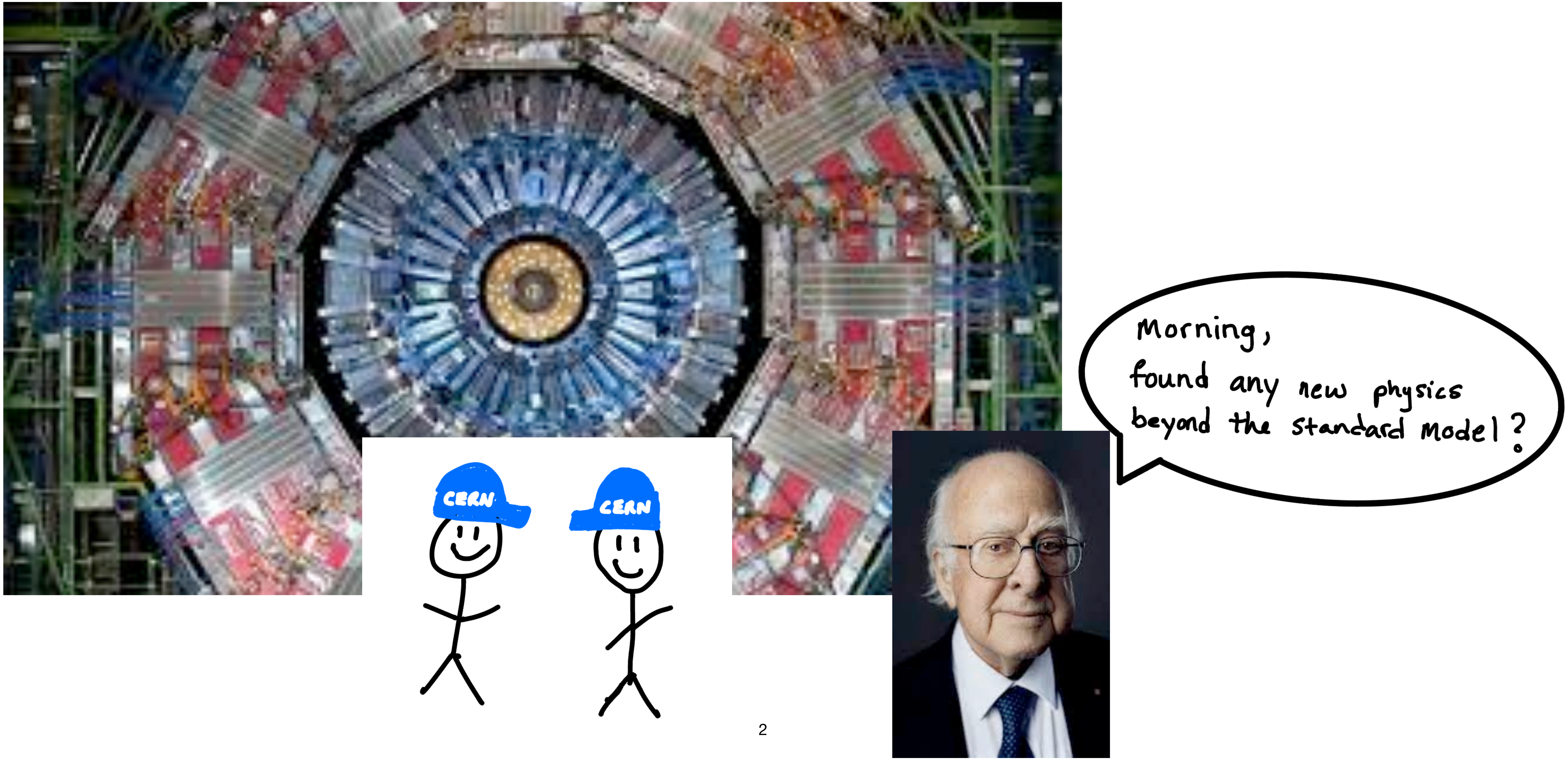
Anomaly Detection at CMS L1 Trigger

Graph Neural Networks, Variational Autoencoders, and Interpreting Anomalous Latent Spaces

Andrew Skivington, A3D3 Postbac, 20/07/2023

What is Anomaly Detection?

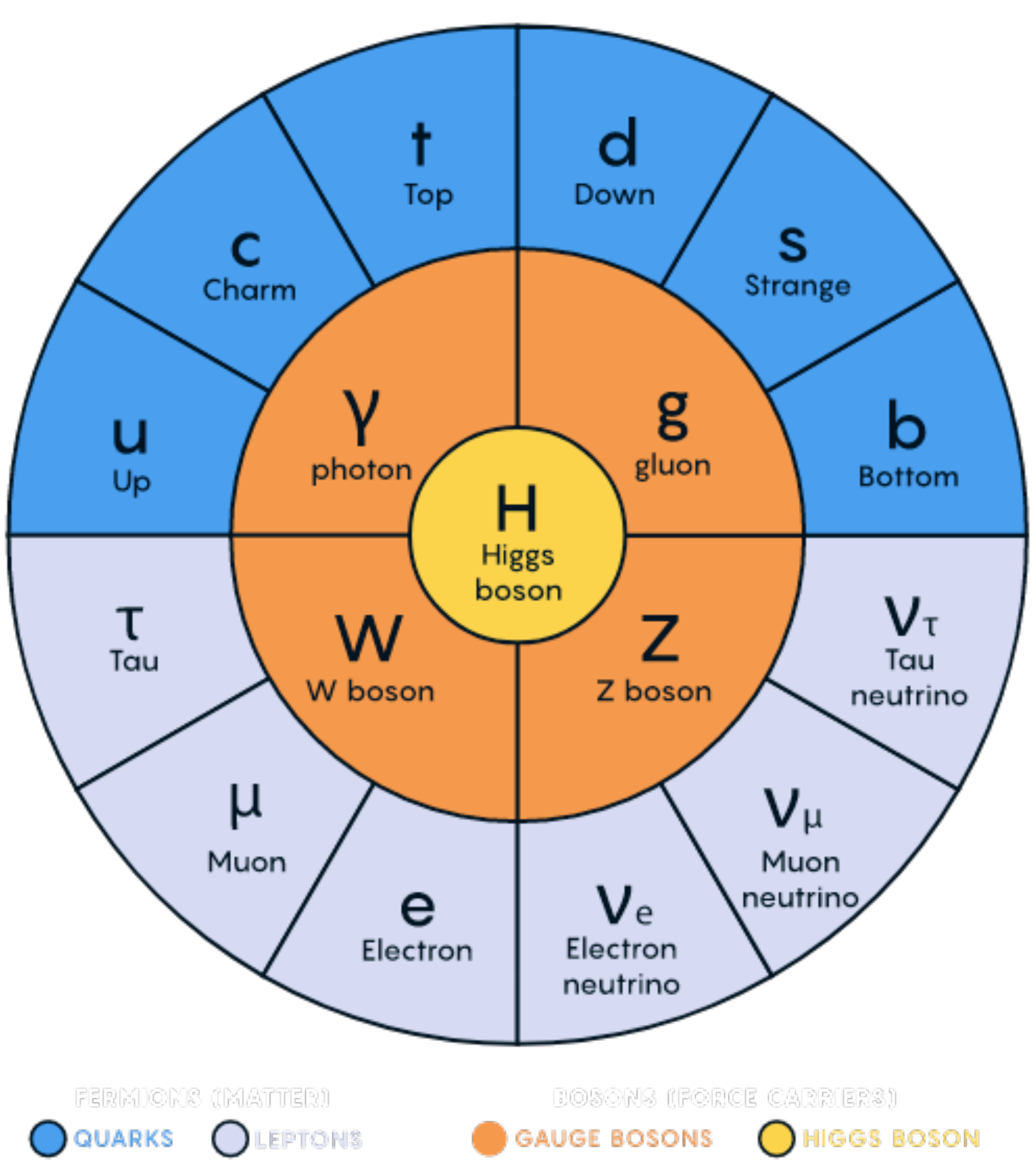


The Big Picture and Goals

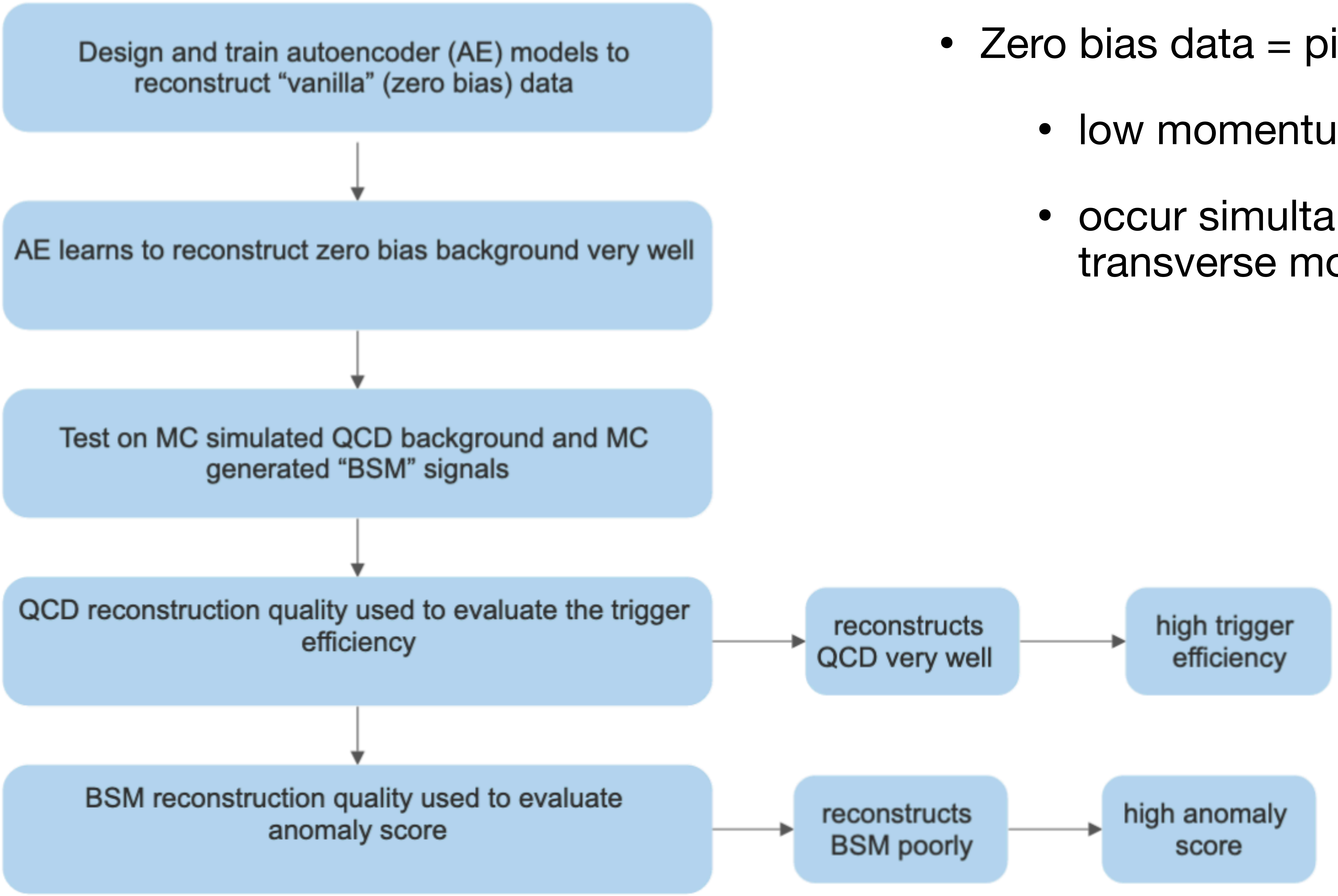
- Traditionally new physics searches are theoretically motivated
- But what if we have been discarding interesting physics?
- We want theory independent algorithms that can detect “anomalous” signals
    - “Anomalous” - anything the algorithm DOESN’T reconstruct well
- These poorly reconstructed signals could be:
    - New Physics :D
    - Detector flaws :/

Preliminary Definitions

- QCD = Quantum chromodynamics
  - Theory used to perform Monte Carlo (MC) background and signal estimations
  - Theory predicts interactions between quarks
    - Mediated by gluons
      - strong force carrying particle
      - Bind quarks to form hadrons
        - Think “Standard Model Physics”
- BSM = Beyond Standard Model Physics
  - All things the standard model can’t explain
  - “New physics”

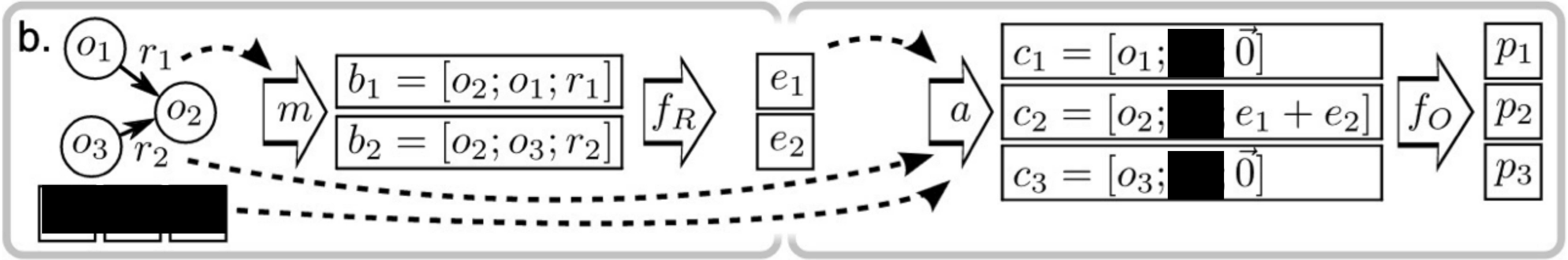


How's It Done?

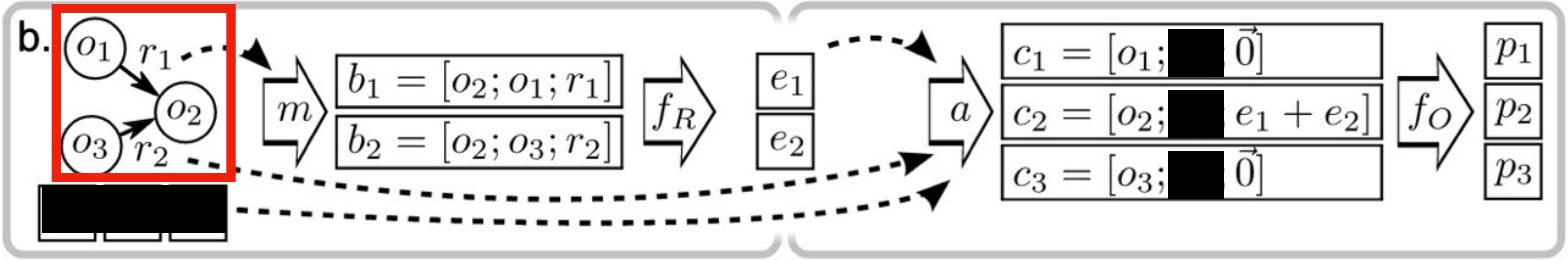


- Zero bias data = pileup data
    - low momentum events
    - occur simultaneously as high transverse momentum events

Network Architecture: Interaction Network (IN)

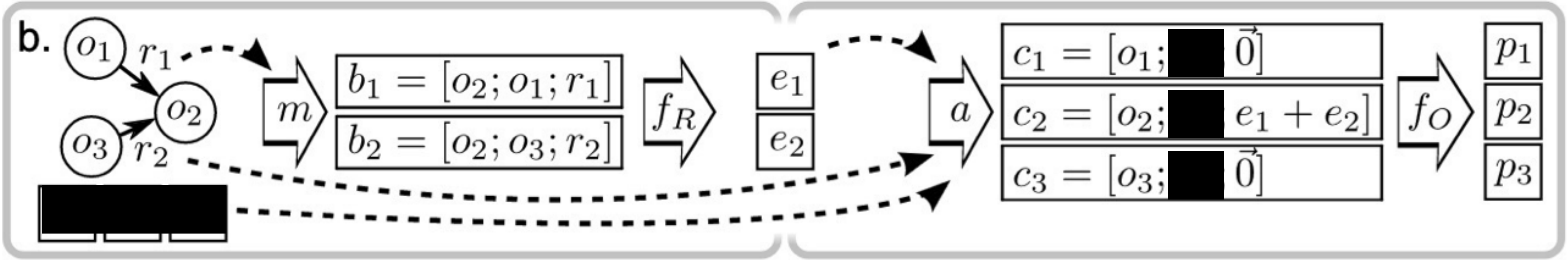


Network Architecture: Interaction Network (IN)

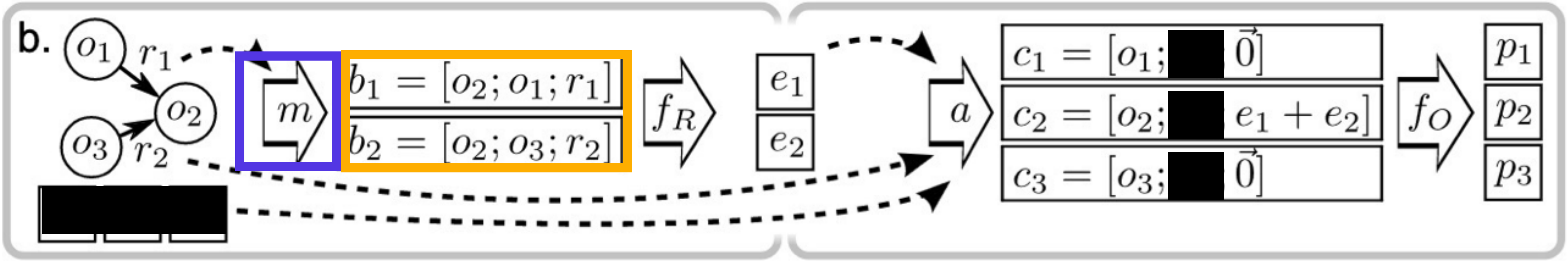


- Graph input data: nodes represent physics objects w/ respective continuous features (i.e. px, py, pz)
    - objects are connected by directional edges
    - IN's have two adjacency matrices: sender and receiver matrices

Network Architecture: Interaction Network (IN)

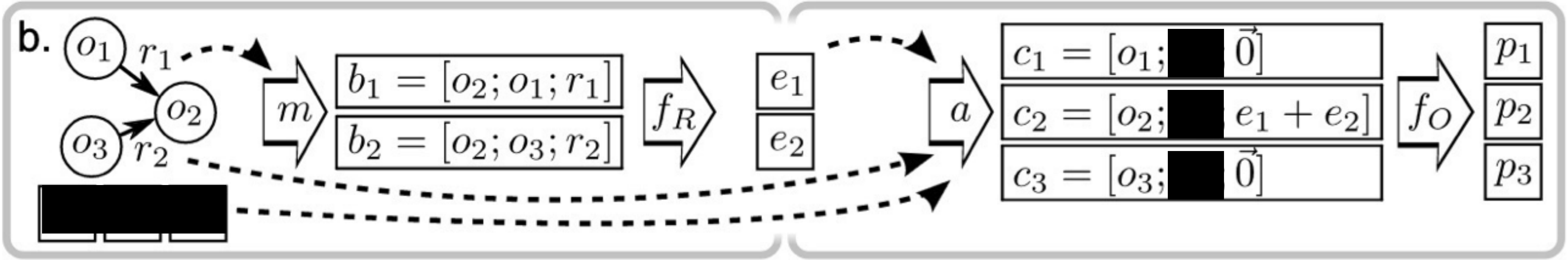


Network Architecture: Interaction Network (IN)

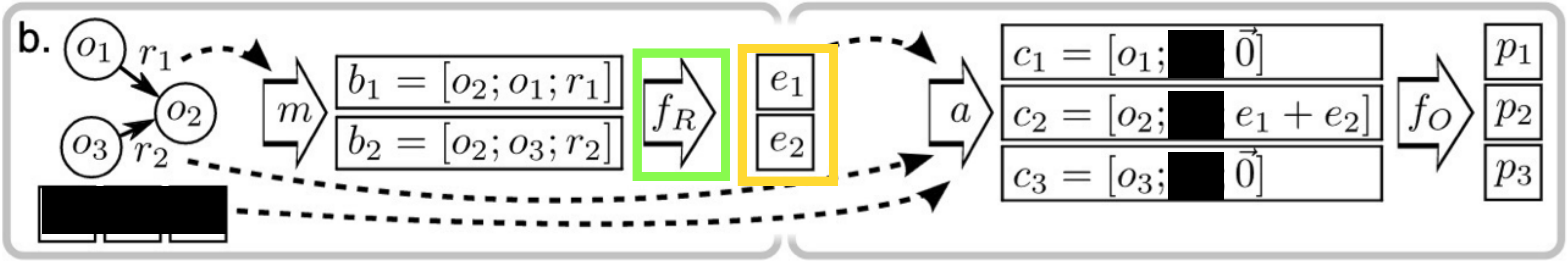


- Marshaling function: message passing step where node-edge projections take place
    - returns interaction terms

Network Architecture: Interaction Network (IN)

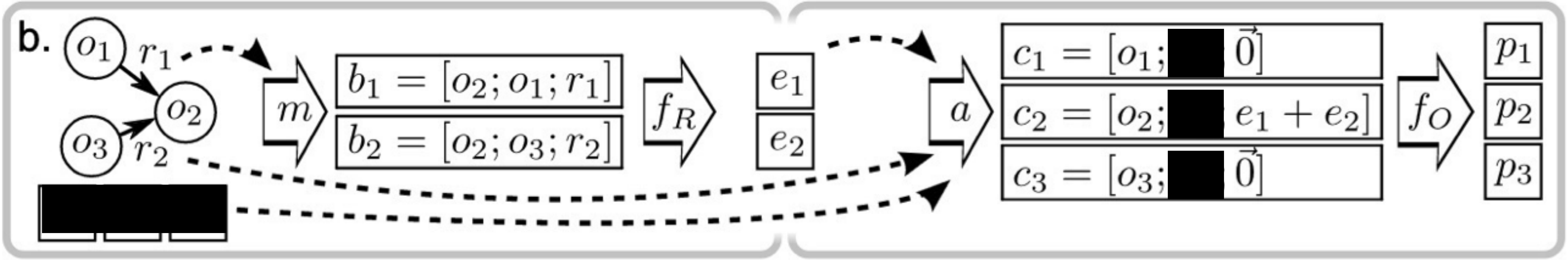


Network Architecture: Interaction Network (IN)

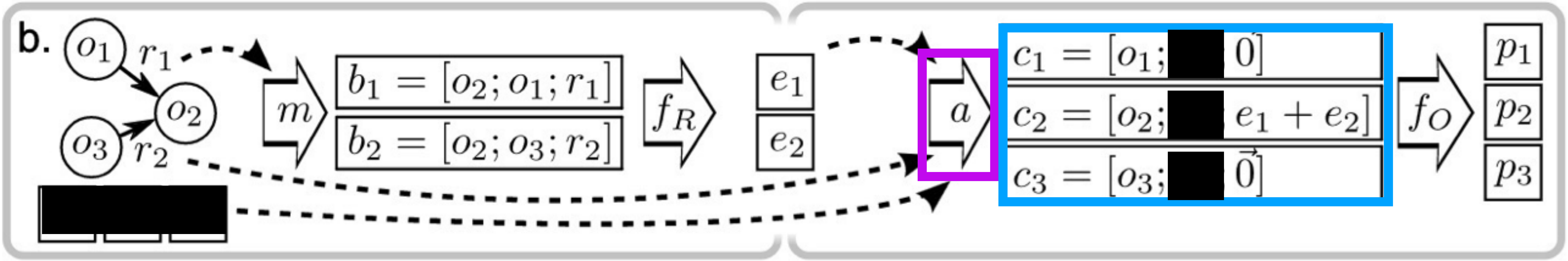


- Relation function (Edge MLP): inputs interaction array into MLP
    - returns the predicted effects of the interaction

Network Architecture: Interaction Network (IN)

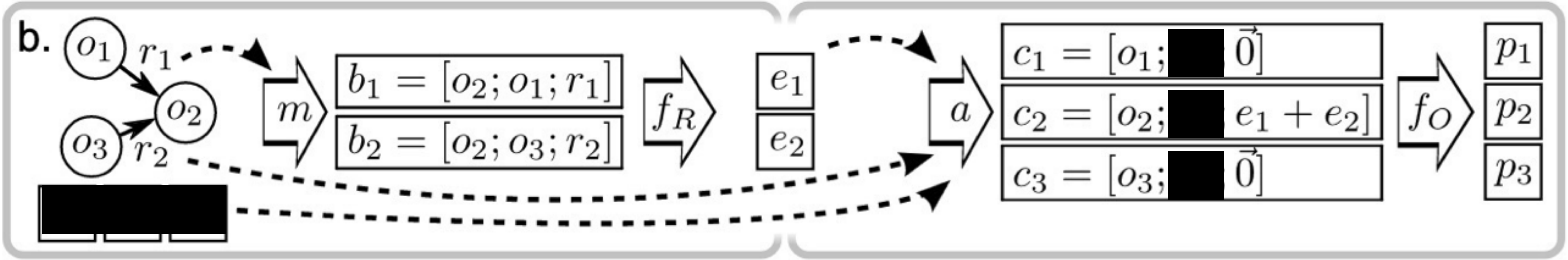


Network Architecture: Interaction Network (IN)

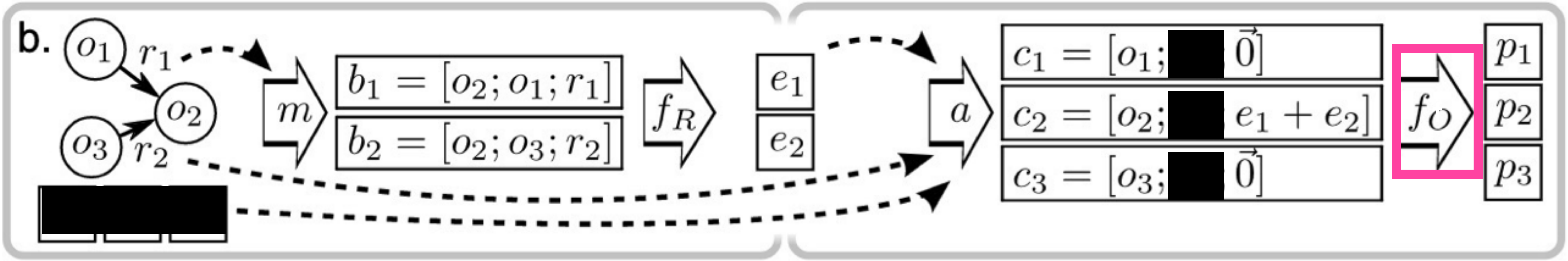


- Aggregation function: concatenates physics object input to the predicted effects
    - analogous to residual connection in ResNet
    - returns aggregator

Network Architecture: Interaction Network (IN)

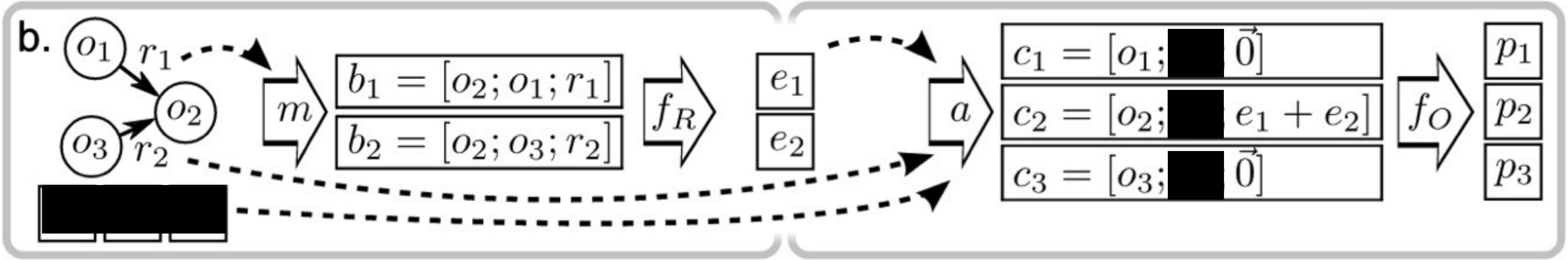


Network Architecture: Interaction Network (IN)

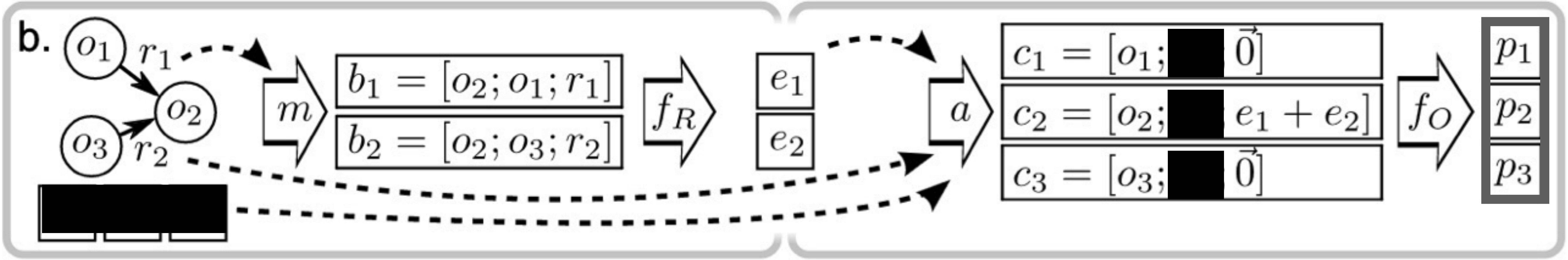


- Object function (Node MLP): inputs aggregated array into MLP
    - returns predicted future state in the example of simulation

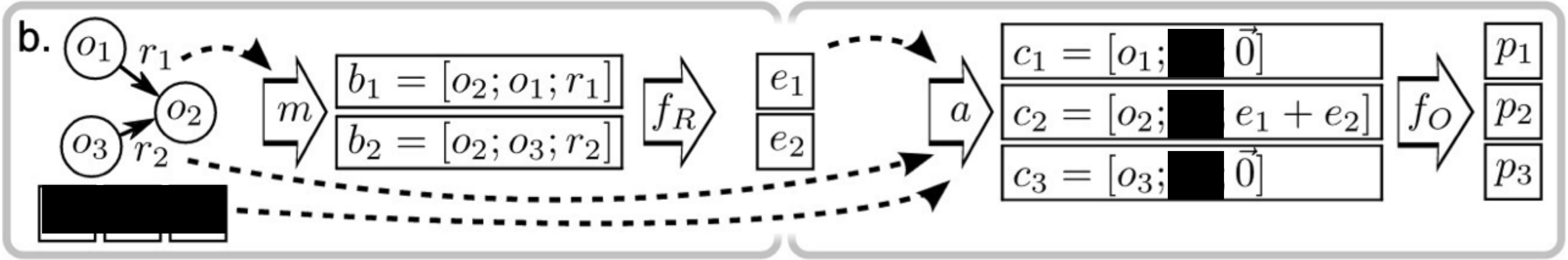
Network Architecture: Interaction Network (IN)



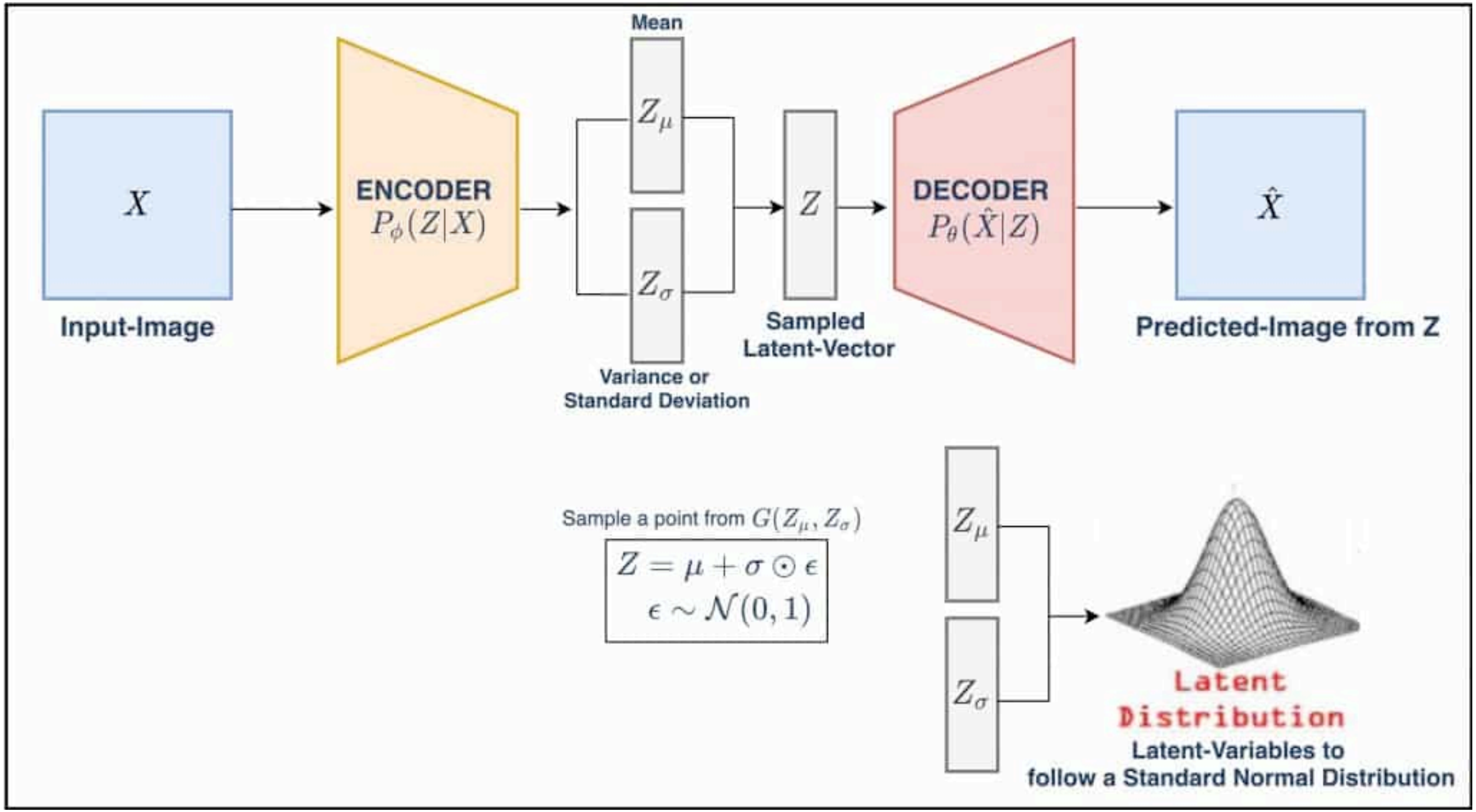
Network Architecture: Interaction Network (IN)



Network Architecture: Interaction Network (IN)

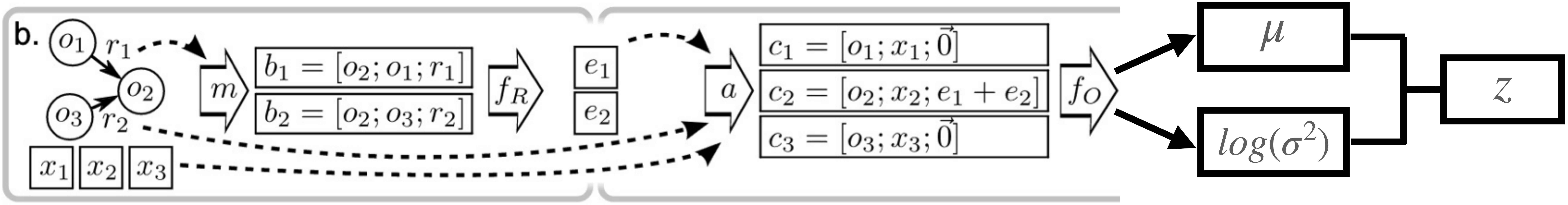


DNN VAE

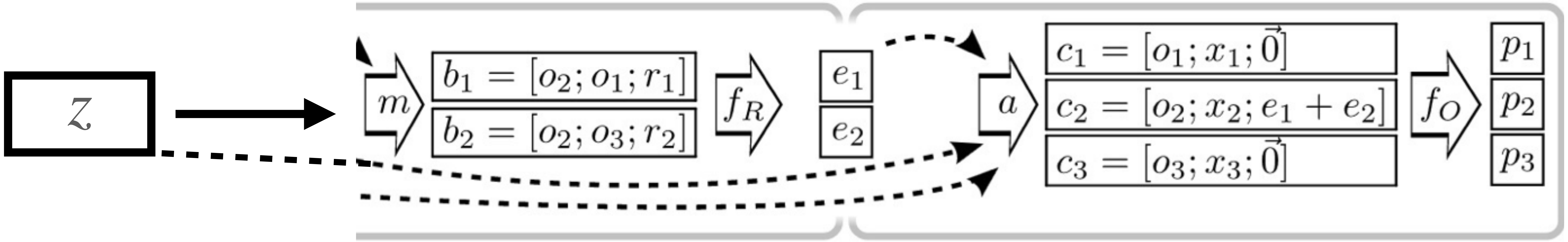


Mean
X
Input-Image
ENCODER
$P_\phi(Z|X)$
Z_μ
Z_σ
Variance or
Standard Deviation
Z
Sampled
Latent-Vector
DECODER
$P_\theta(\hat{X}|Z)$
$\hat{X}$
Predicted-Image from Z
Sample a point from $G(Z_\mu, Z_\sigma)$
$Z = \mu + \sigma \odot \epsilon$
$\epsilon \sim \mathcal{N}(0, 1)$
Z_μ
Z_σ
Latent
Distribution
Latent-Variables to
follow a Standard Normal Distribution

Learnable INVAE (Encoder) Implementation



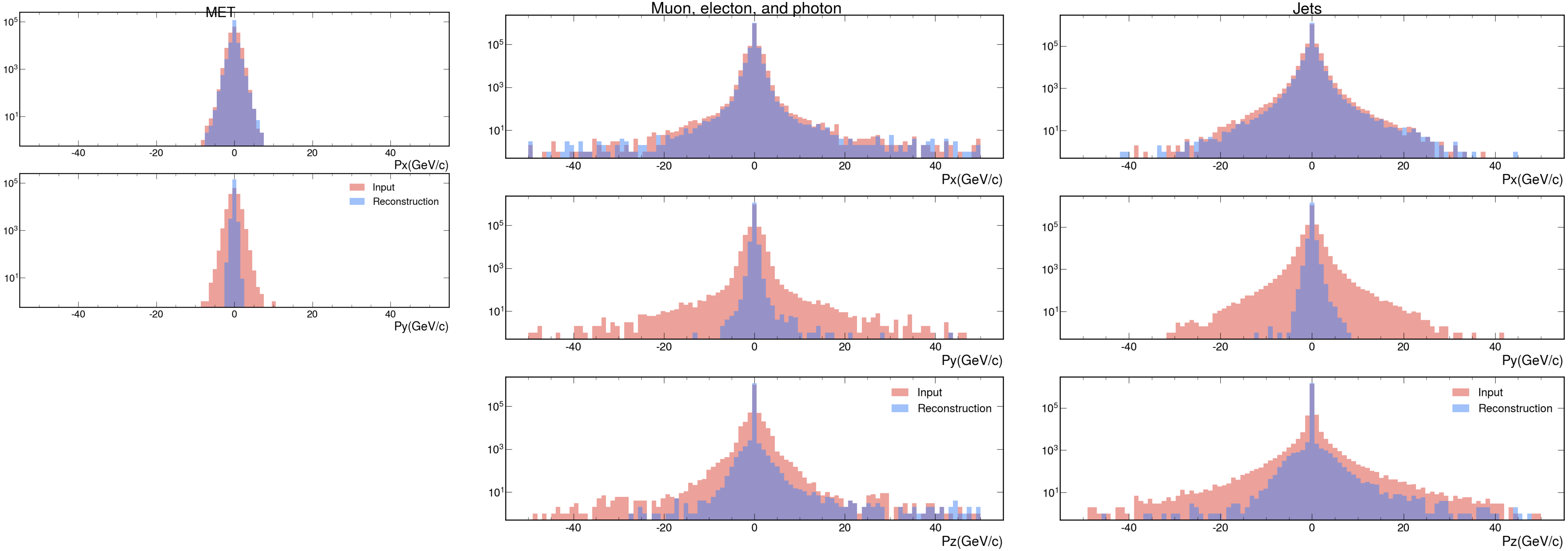
Learnable INVAE (Decoder) Implementation



- The decoder outputs (i.e. p1, p2, p3) are the reconstructed physics objects and their respective momenta (px, py, pz)

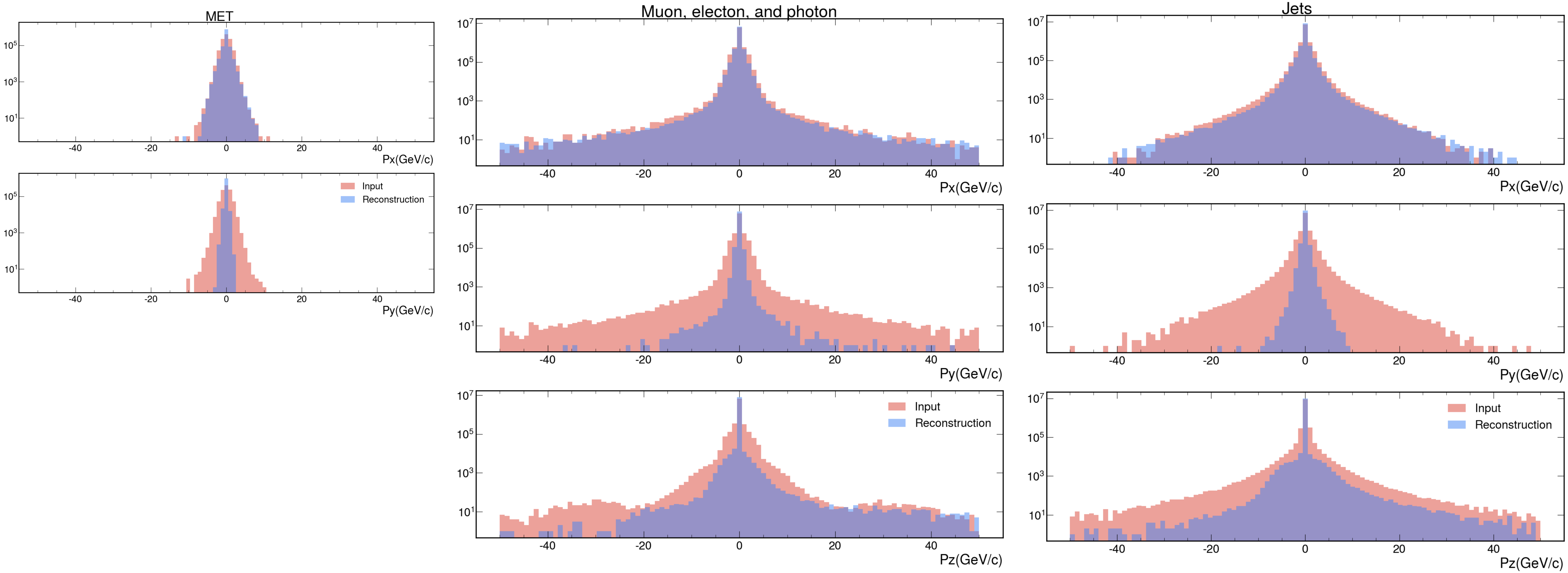
INVAE Results

Zero Bias Reconstruction



- **Expectation**: reconstruction loss should be marginal for zero bias data
- **Summary**: reconstruction is good across all physics objects for px but not py, or pz

QCD Reconstruction



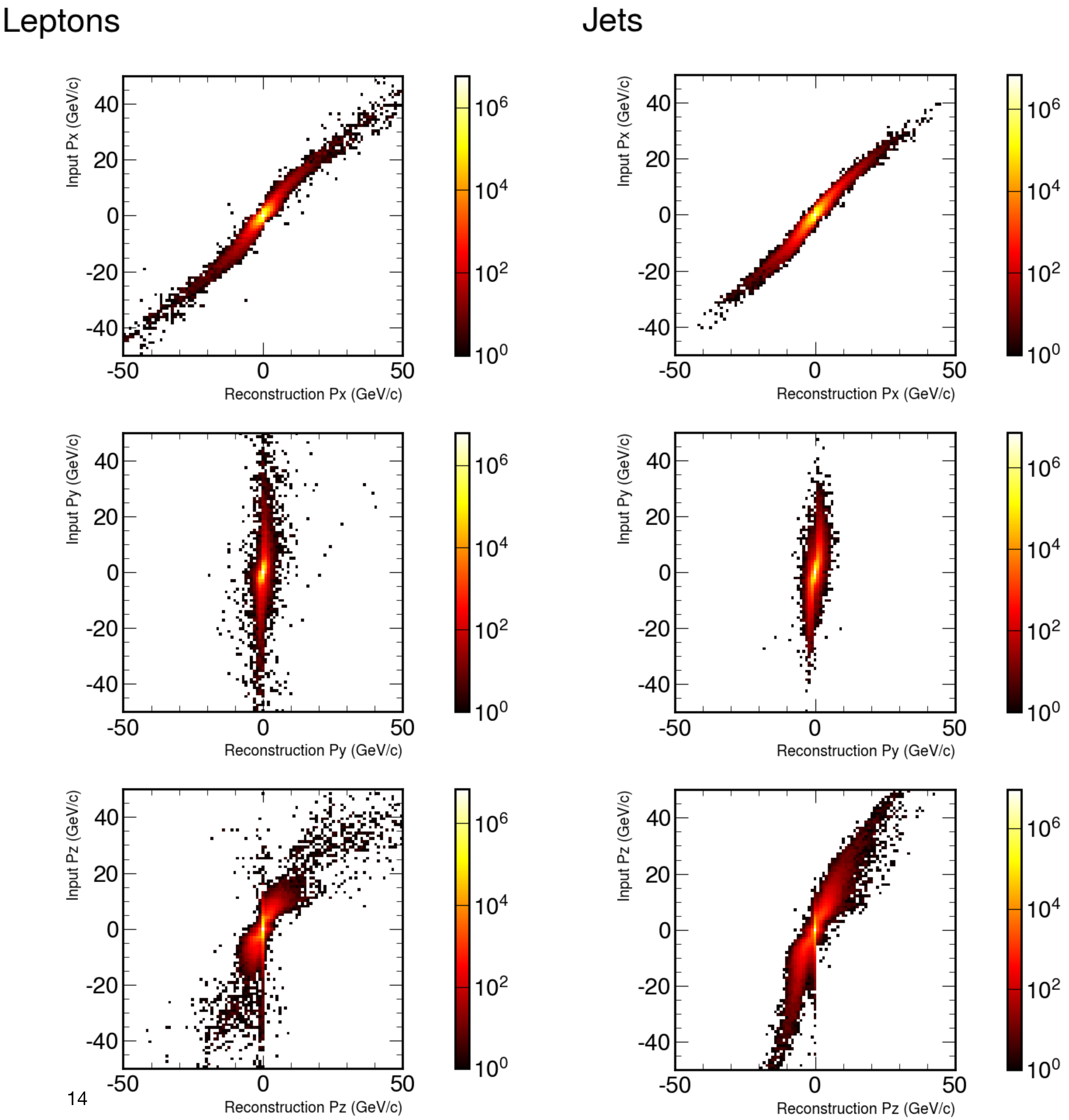
- **Expectation**: reconstruction loss to be comparable to Zero Bias data
- **Result:** similar quality but px and py should be symmetric across background and reconstruction. This is not the case for QCD or zero Bias data.

H -> aa -> 4b Reconstruction

- **Expectation**: reconstruction loss is worse than SM data reconstruction

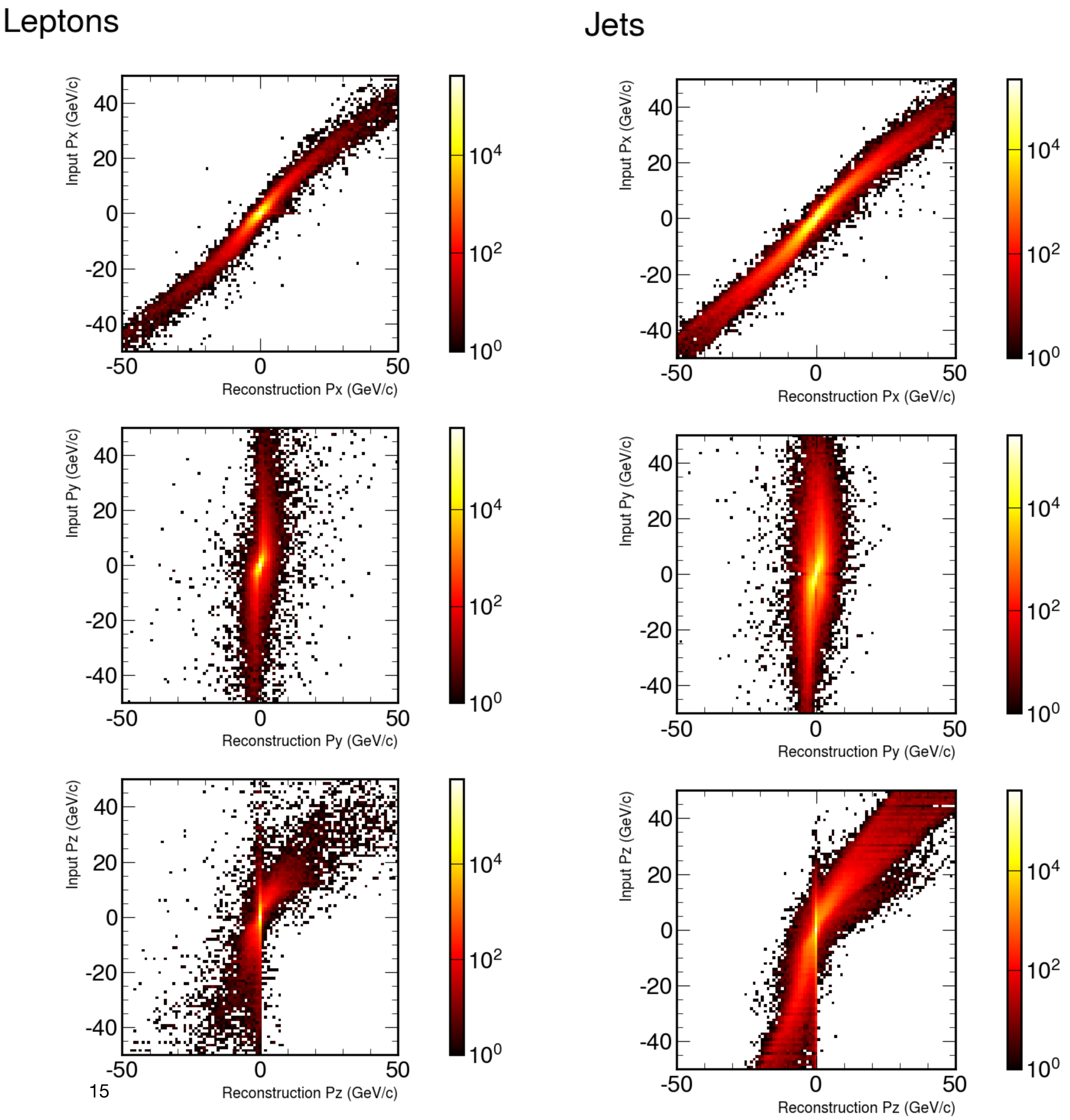
QCD Heat Maps

- Px reconstruction is strongly correlated
- Py is almost independent of input
- **Expectation**: If Px is strongly correlated to its input, so should Py
    - Consequence of cylindrical detector symmetry



H -> aa -> 4b Heat Maps

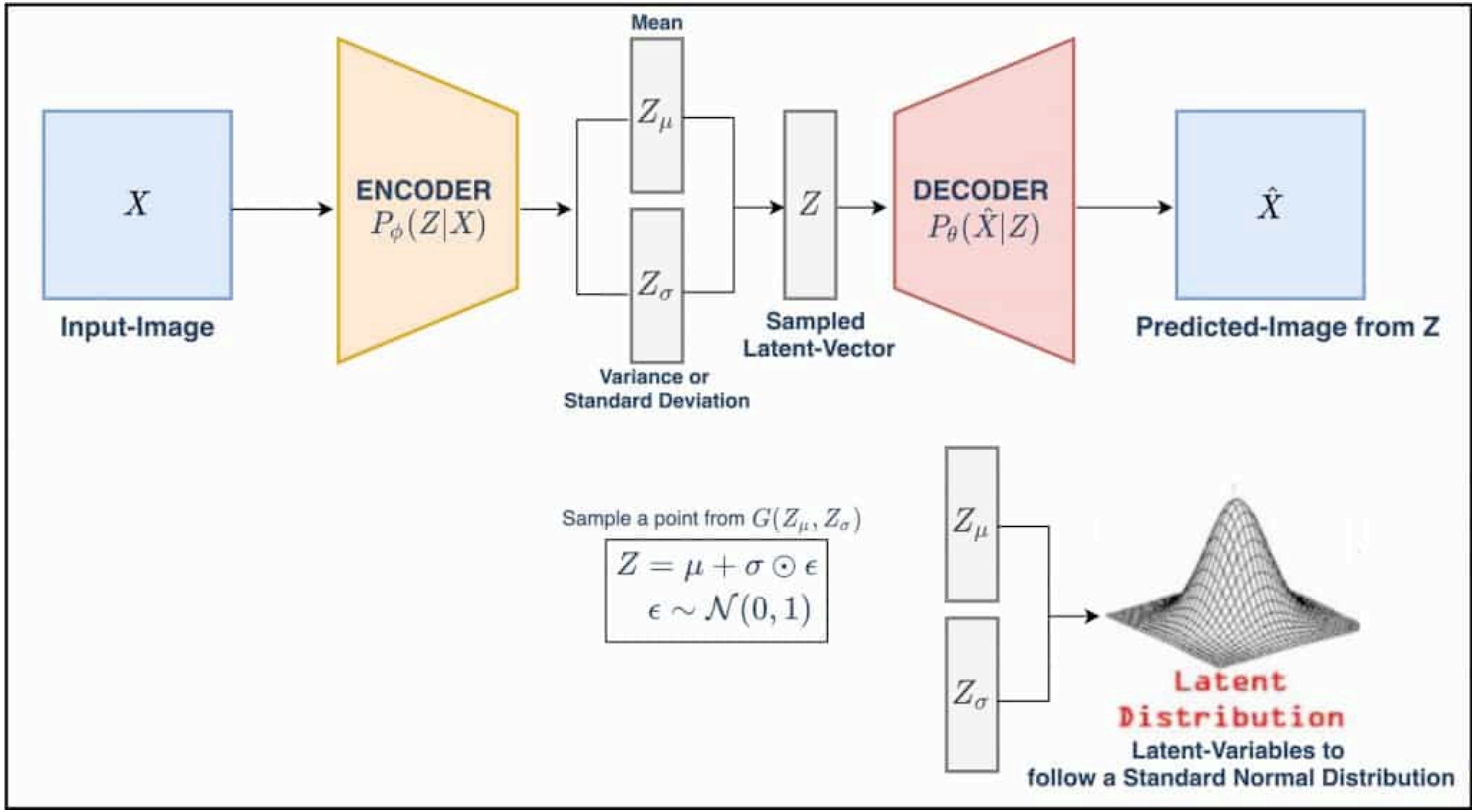
- The same asymmetry between Px and Py reconstruction occurs
- **Expectation:** components for BSM signal would be less correlated than QCD



INVAE ROC Curves

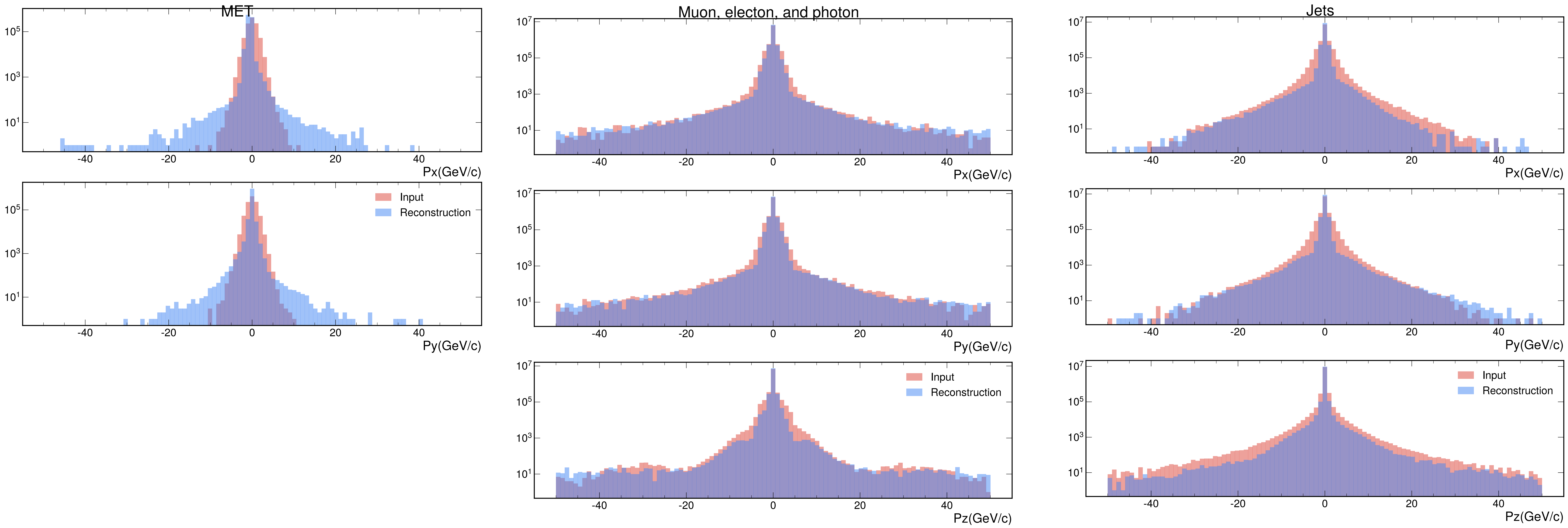
DNN VAE Results

DNN VAE



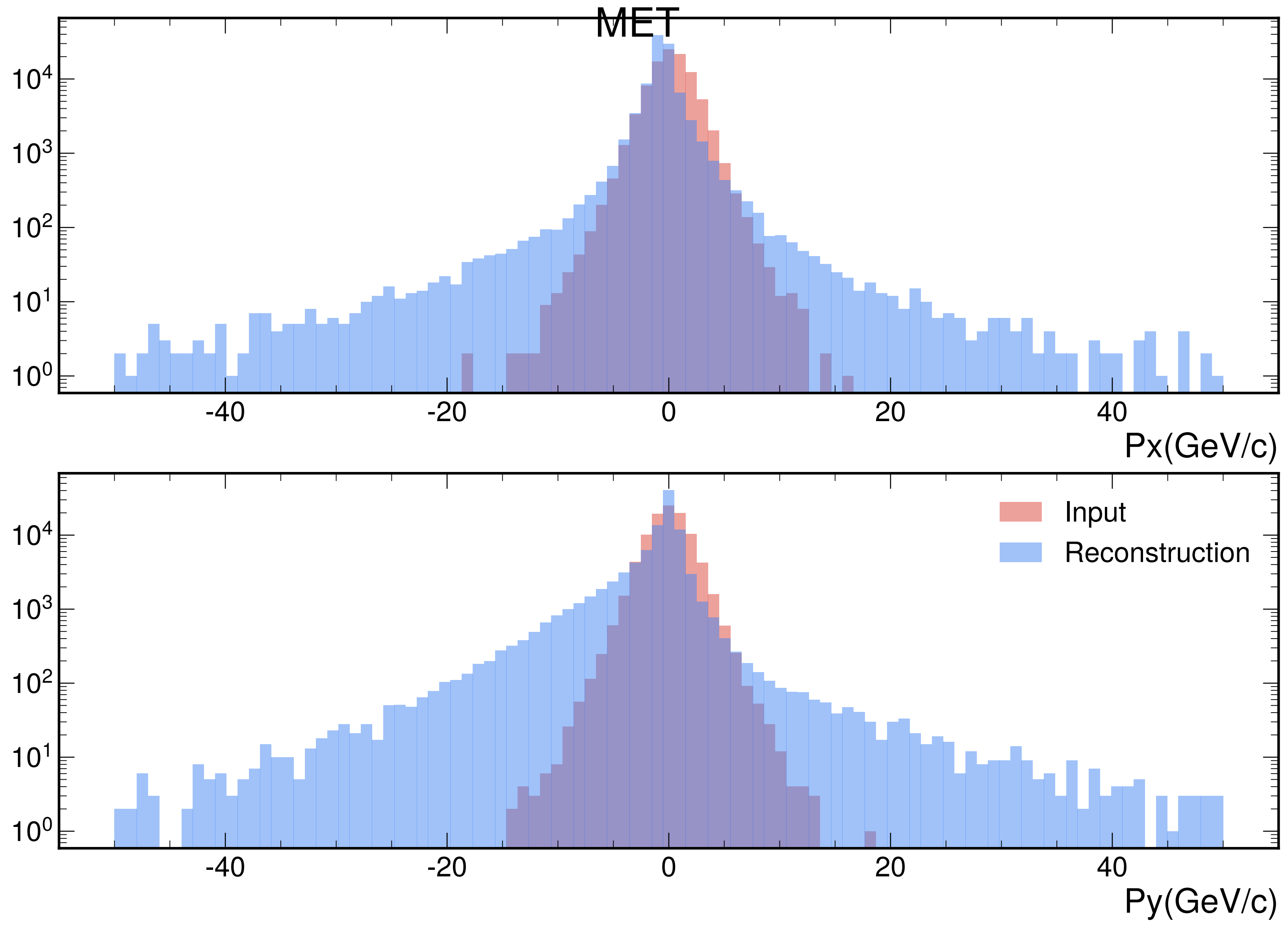
Mean
X
ENCODER
$P_\phi(Z|X)$
Z_μ
Z_σ
Z
DECODER
$P_\theta(\hat{X}|Z)$
$\hat{X}$
Input-Image
Sampled
Latent-Vector
Predicted-Image from Z
Variance or
Standard Deviation
Sample a point from $G(Z_\mu, Z_\sigma)$
$Z = \mu + \sigma \odot \epsilon$
$\epsilon \sim \mathcal{N}(0, 1)$
Z_μ
Z_σ
Latent
Distribution
Latent-Variables to
follow a Standard Normal Distribution

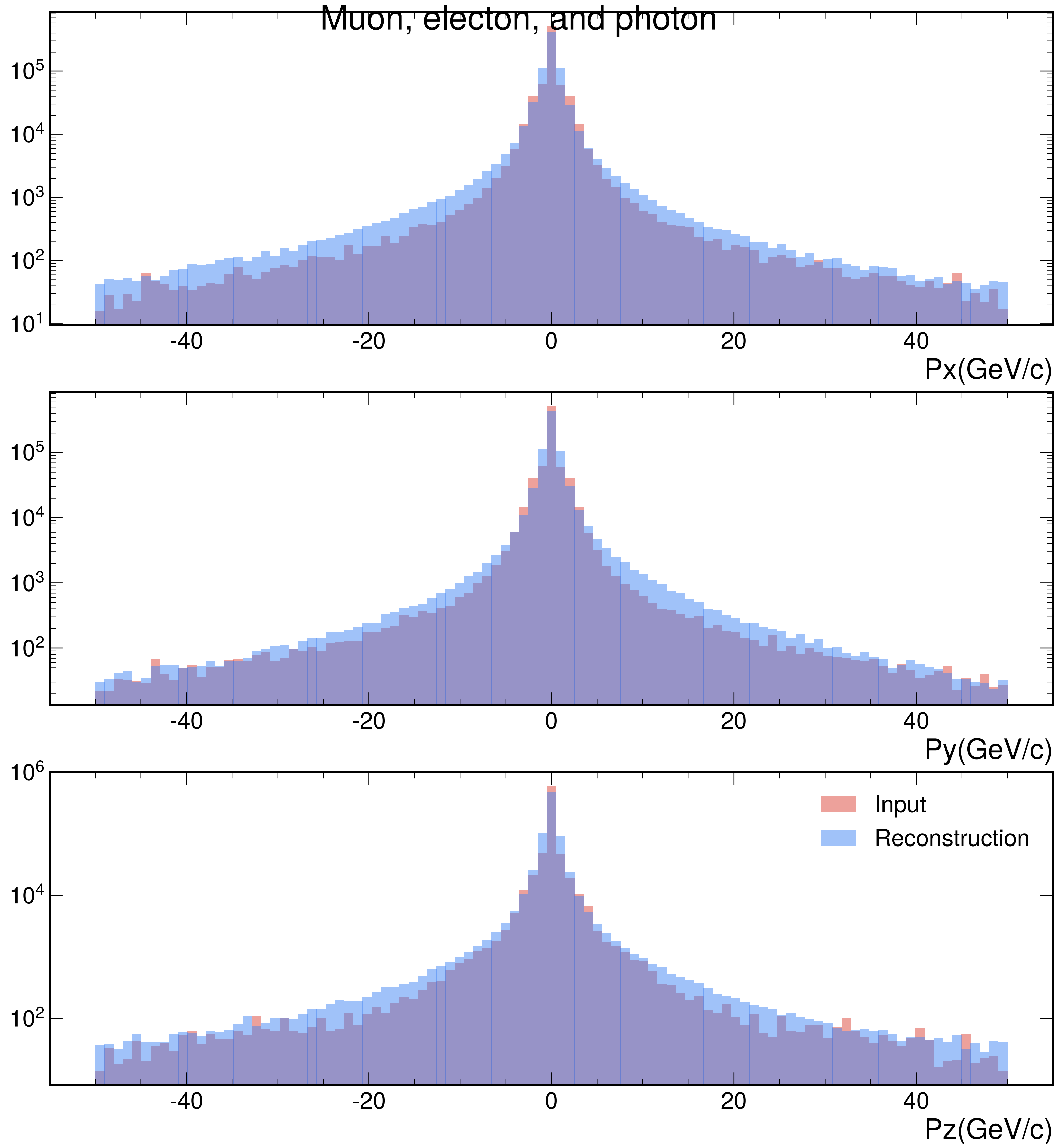
QCD Object Reconstruction

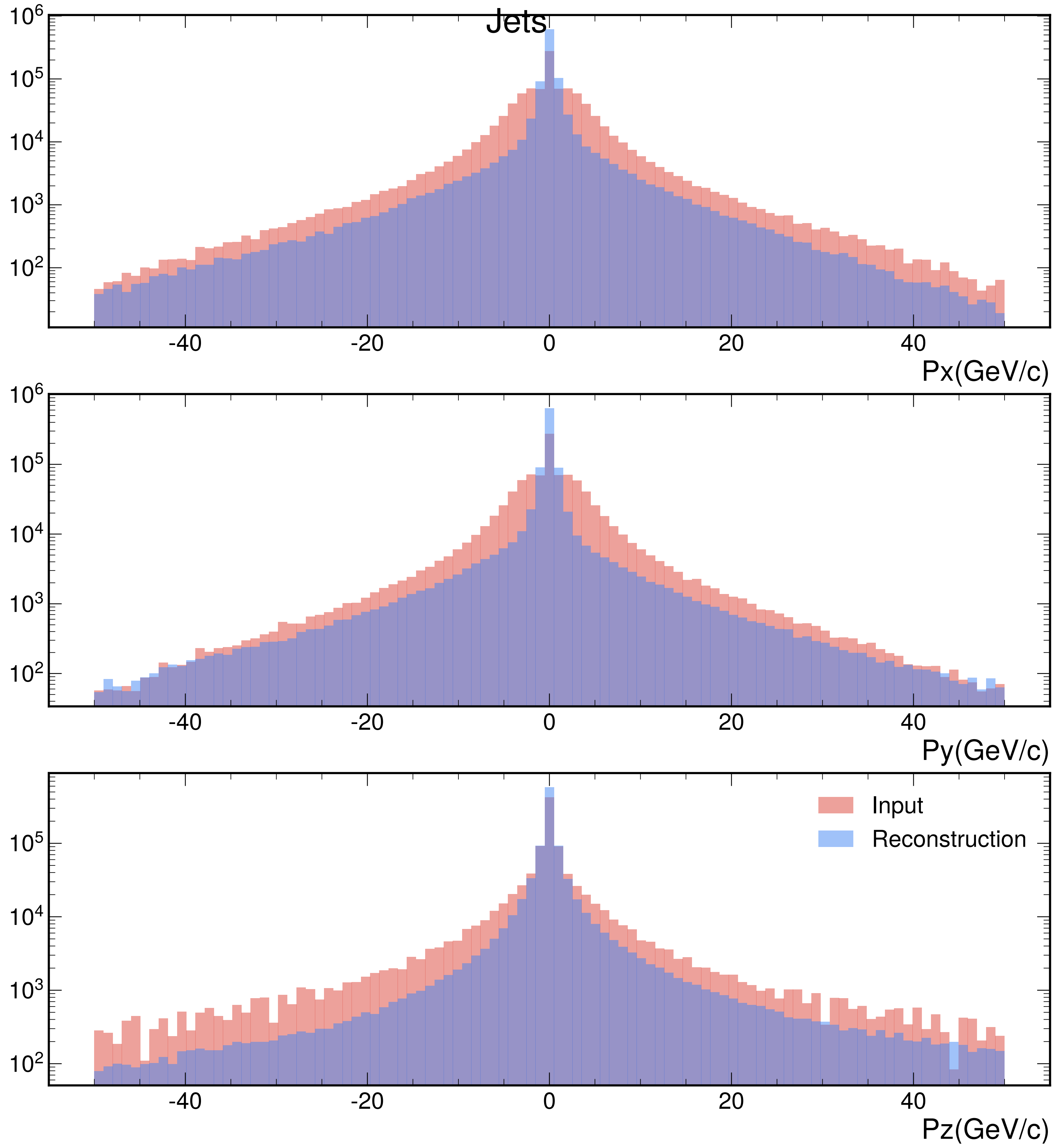


- Better reconstruction across all components
- Minimal loss for QCD events as expected

H -> aa -> 4b Reconstruction





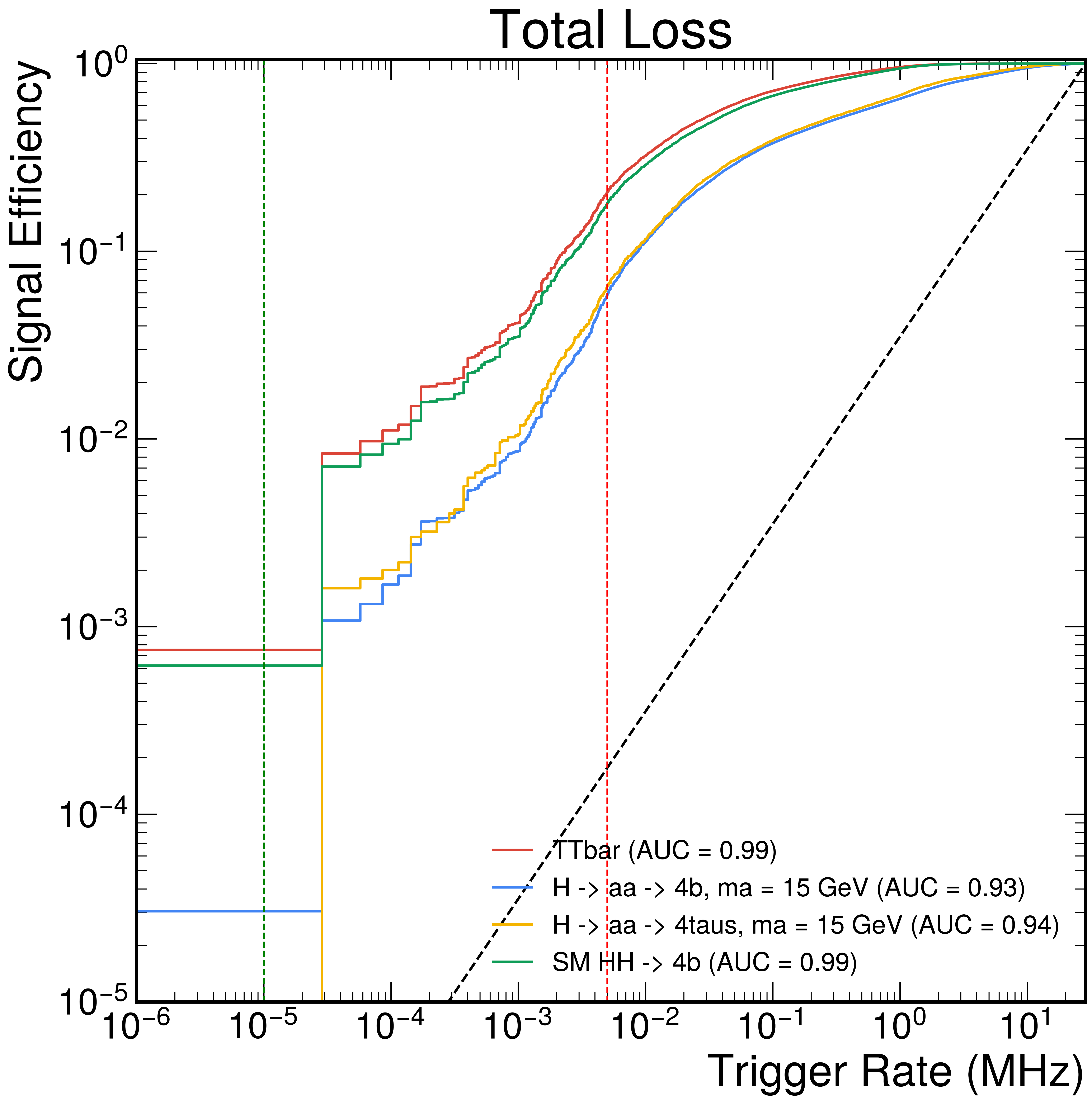


- greater reconstruction loss
- especially for MET
- This was the expected behavior

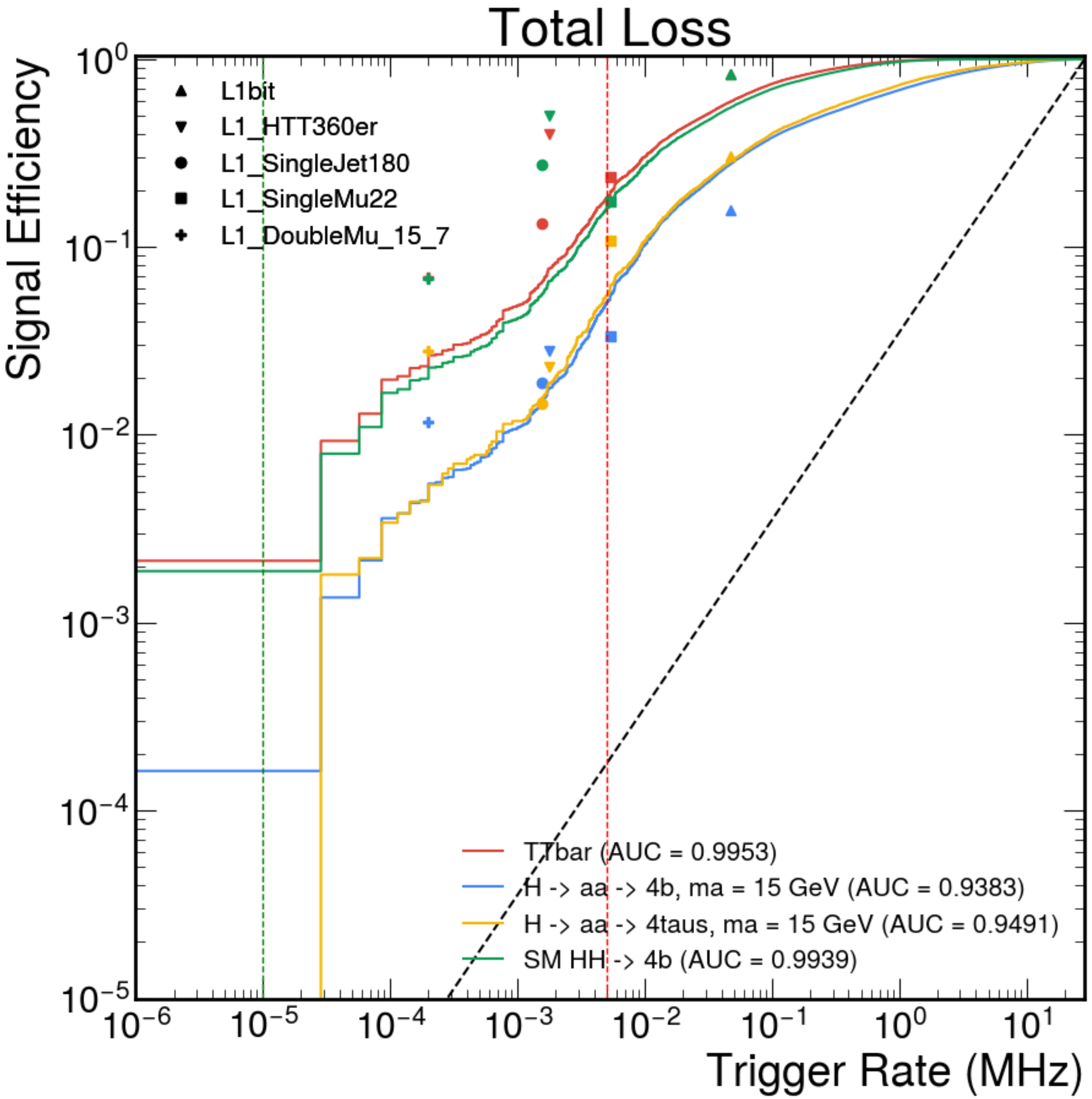
VAE ROC Curve

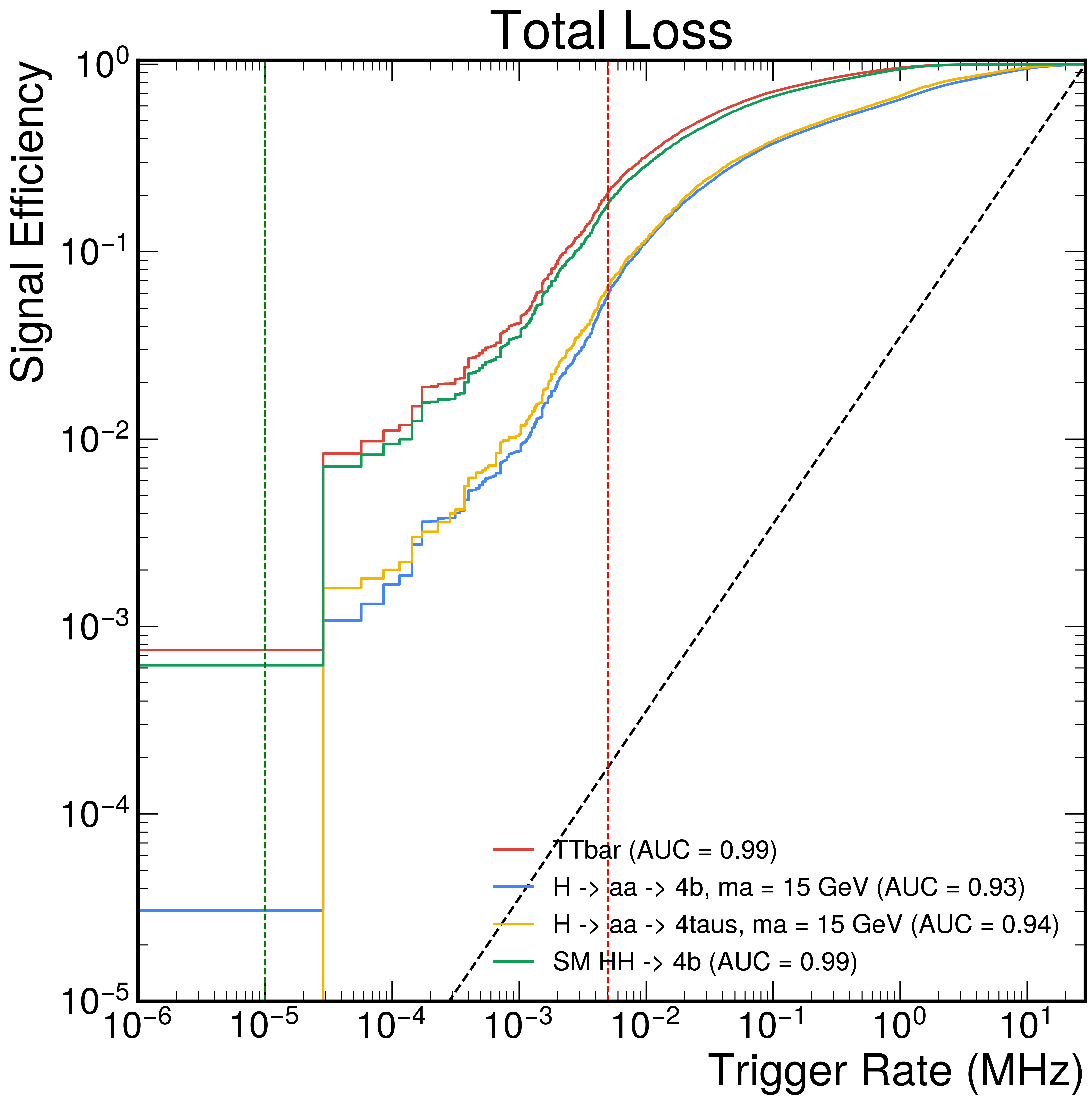
- Total Loss = MSE + KLD
- $$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\mathrm{KLD}(q_\phi(z|x)||p(z)) = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2\right)$$



INVAE (left) vs DNN (right) ROC Curve Comparisons





- **Result:** AUC is comparable for INVAE and DNNVAE, but the quality of the reconstruction is better for the DNN

Next Steps

- Train/test INVAE on the ADC2021 datasets
- Do one last hyperparameter search
    - See if network can generalize to all components in reconstruction
- Based on results:
    - write external paper on performance of INVAE and INAE compared to DNNAE and VAE benchmarks for AD

Interpreting Anomalous Latent Space

Current Project

- We have a VAE encoder algo we are putting on L1 trigger menu!
- However, we have no way of interpreting these encodings…
    - That's where I come in:
        - mapping anomalous latent space
        - Interpret our results
        - Distinguish between anomalous signal candidates and potential detector flaws
- Questions to answer:
    - Where do these anomalous signal candidates 'live' in latent space?
    - Is there clustering? Can it be interpreted/explained?
    - Are components of latent space correlated?
    - What are the most important components in latent space corresponding to higher anomaly score?

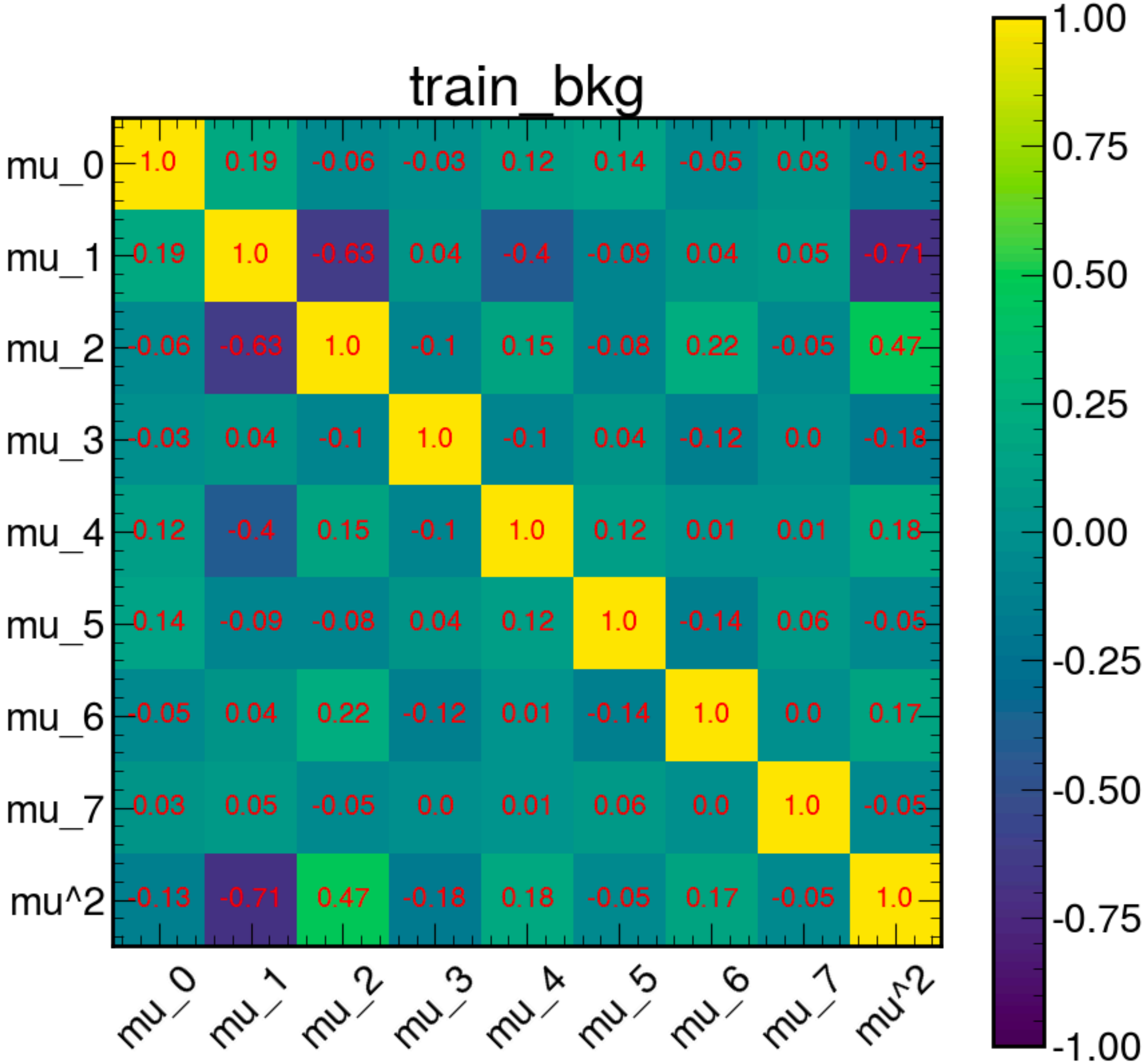
Background Info

- How the anomaly score is quantified:
    - mu
        - Encoded latent vector
        - mu = < mu_0, mu_1, …, mu_7 >
        - 8-D vector
        - used to calculate the anomaly score:
            - mu^2 :
                - Squared sum of all components of mu-vector
                - 1-D vector
                - mu^2 >> 0 —> larger anomaly score

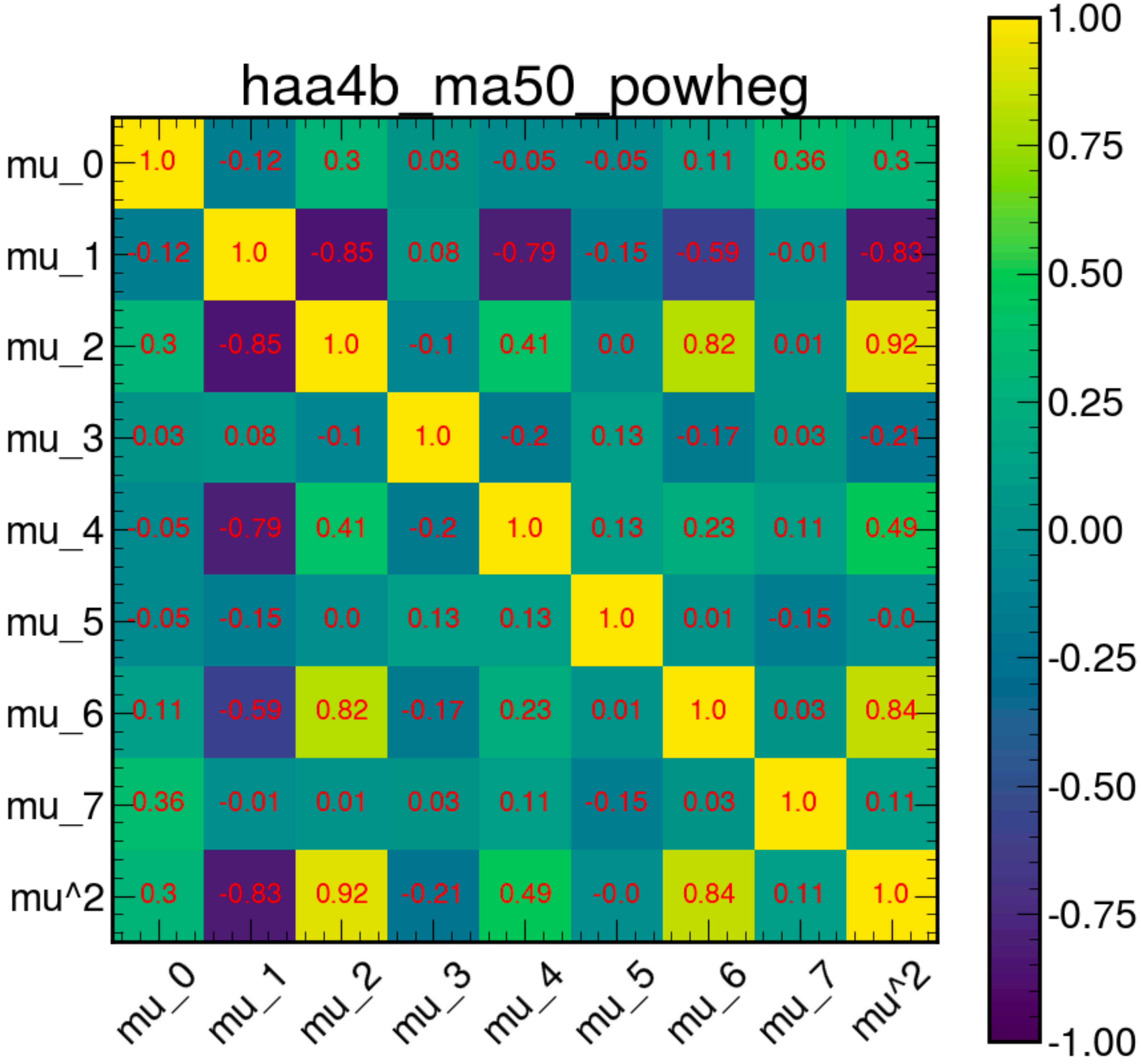
Background Info

- How the anomaly score is quantified:
  - mu
    - Encoded latent vector
    - mu = < mu_0, mu_1, …, mu_7 >
    - 8-D vector
    - used to calculate the anomaly score:
      - mu^2 :
        - Squared sum of all components of mu-vector
        - 1-D vector
        - mu^2 >> 0 —> larger anomaly score

Exploring Latent Variable Correlations



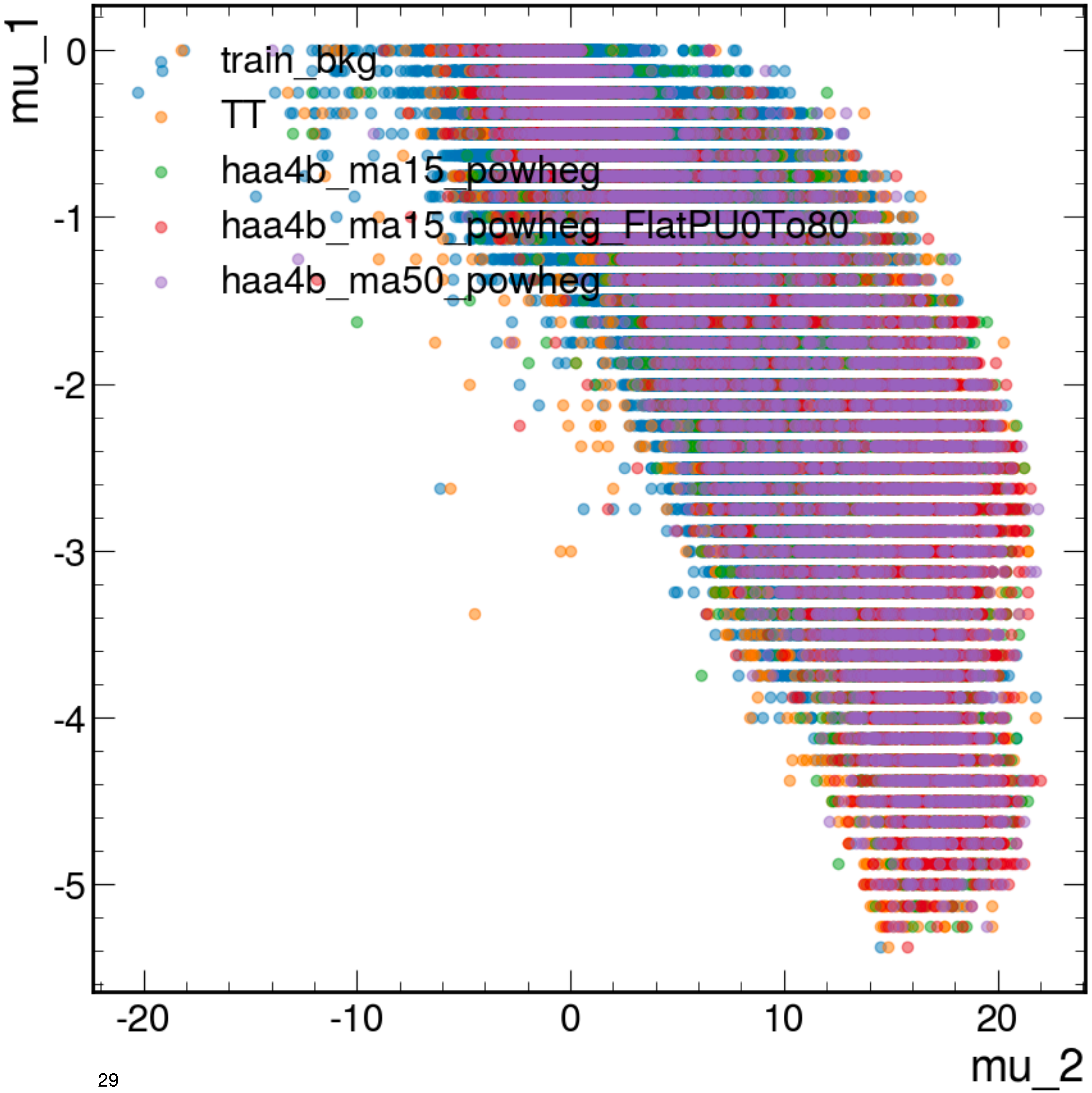
train_bkg
mu_0 1.0 0.19 -0.06 -0.03 0.12 0.14 -0.05 0.03 -0.13
mu_1 0.19 1.0 -0.63 0.04 -0.4 -0.09 0.04 0.05 -0.71
mu_2 -0.06 -0.63 1.0 -0.1 0.15 -0.08 0.22 -0.05 0.47
mu_3 -0.03 0.04 -0.1 1.0 -0.1 0.04 -0.12 0.0 -0.18
mu_4 0.12 -0.4 0.15 -0.1 1.0 0.12 0.01 0.01 0.18
mu_5 0.14 -0.09 -0.08 0.04 0.12 1.0 -0.14 0.06 -0.05
mu_6 -0.05 0.04 0.22 -0.12 0.01 -0.14 1.0 0.0 0.17
mu_7 0.03 0.05 -0.05 0.0 0.01 0.06 0.0 1.0 -0.05
mu^2 -0.13 -0.71 0.47 -0.18 0.18 -0.05 0.17 -0.05 1.0
mu_0 mu_1 mu_2 mu_3 mu_4 mu_5 mu_6 mu_7 mu^2
1.00 0.75 0.50 0.25 0.00 -0.25 -0.50 -0.75 -1.00



haa4b_ma50_powheg
mu_0 1.0 -0.12 0.3 0.03 -0.05 -0.05 0.11 0.36 0.3
mu_1 -0.12 1.0 -0.85 0.08 -0.79 -0.15 -0.59 -0.01 -0.83
mu_2 0.3 -0.85 1.0 -0.1 0.41 0.0 0.82 0.01 0.92
mu_3 0.03 0.08 -0.1 1.0 -0.2 0.13 -0.17 0.03 -0.21
mu_4 -0.05 -0.79 0.41 -0.2 1.0 0.13 0.23 0.11 0.49
mu_5 -0.05 -0.15 0.0 0.13 0.13 1.0 0.01 -0.15 -0.0
mu_6 0.11 -0.59 0.82 -0.17 0.23 0.01 1.0 0.03 0.84
mu_7 0.36 -0.01 0.01 0.03 0.11 -0.15 0.03 1.0 0.11
mu^2 0.3 -0.83 0.92 -0.21 0.49 -0.0 0.84 0.11 1.0
mu_0 mu_1 mu_2 mu_3 mu_4 mu_5 mu_6 mu_7 mu^2
1.00 0.75 0.50 0.25 0.00 -0.25 -0.50 -0.75 -1.00

mu_1 vs mu_2 Plot

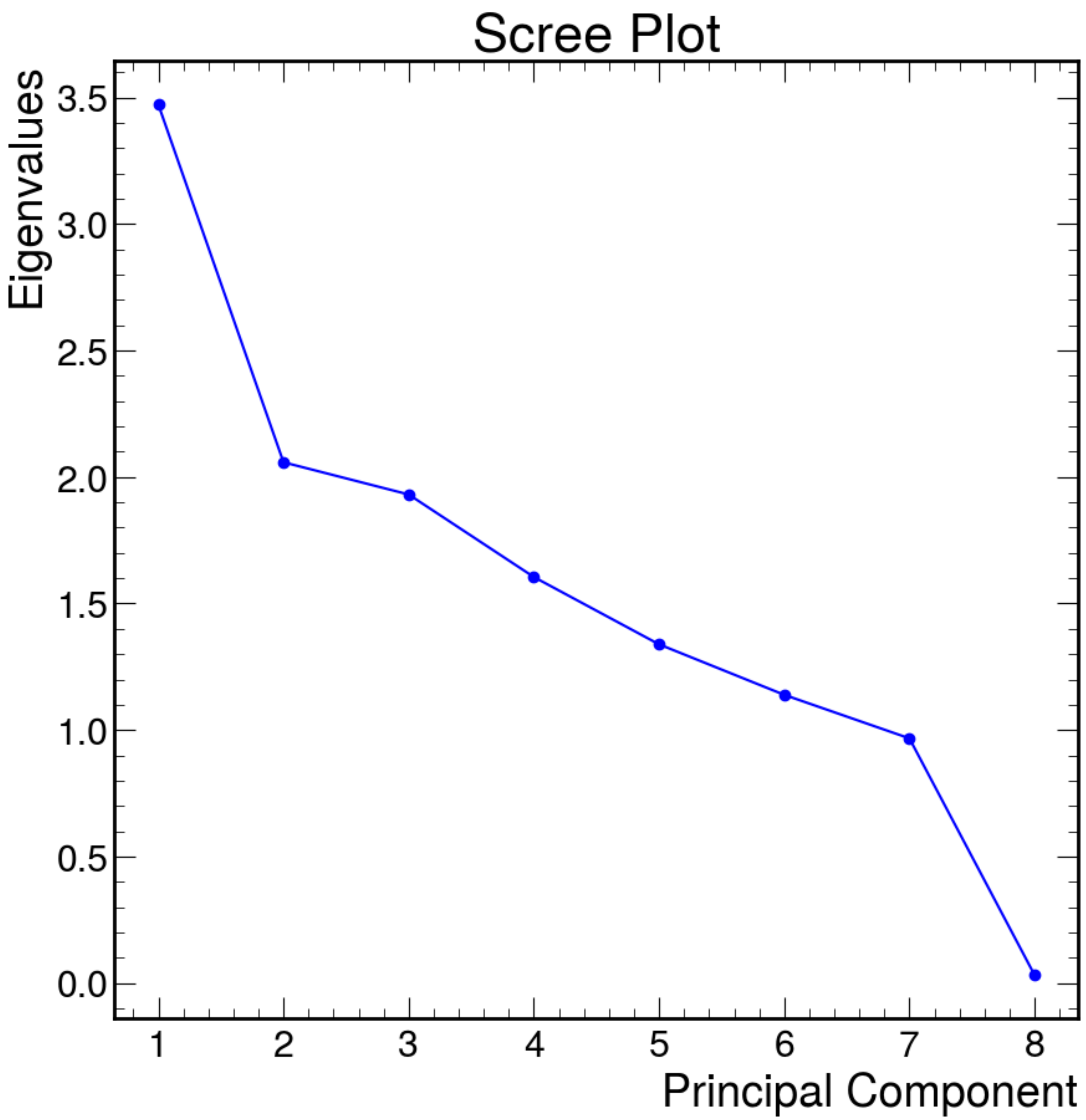
- Visualizes the strong negative correlation
- Note discrete values across all signals are visible because smaller plot domain
- There are some other interesting correlations to visualize…
  - But won't provide more information
  - Hence, PCA is needed
- Conclusion:
  - There are variables that are strongly and weakly correlated to mu_2 and other comps.



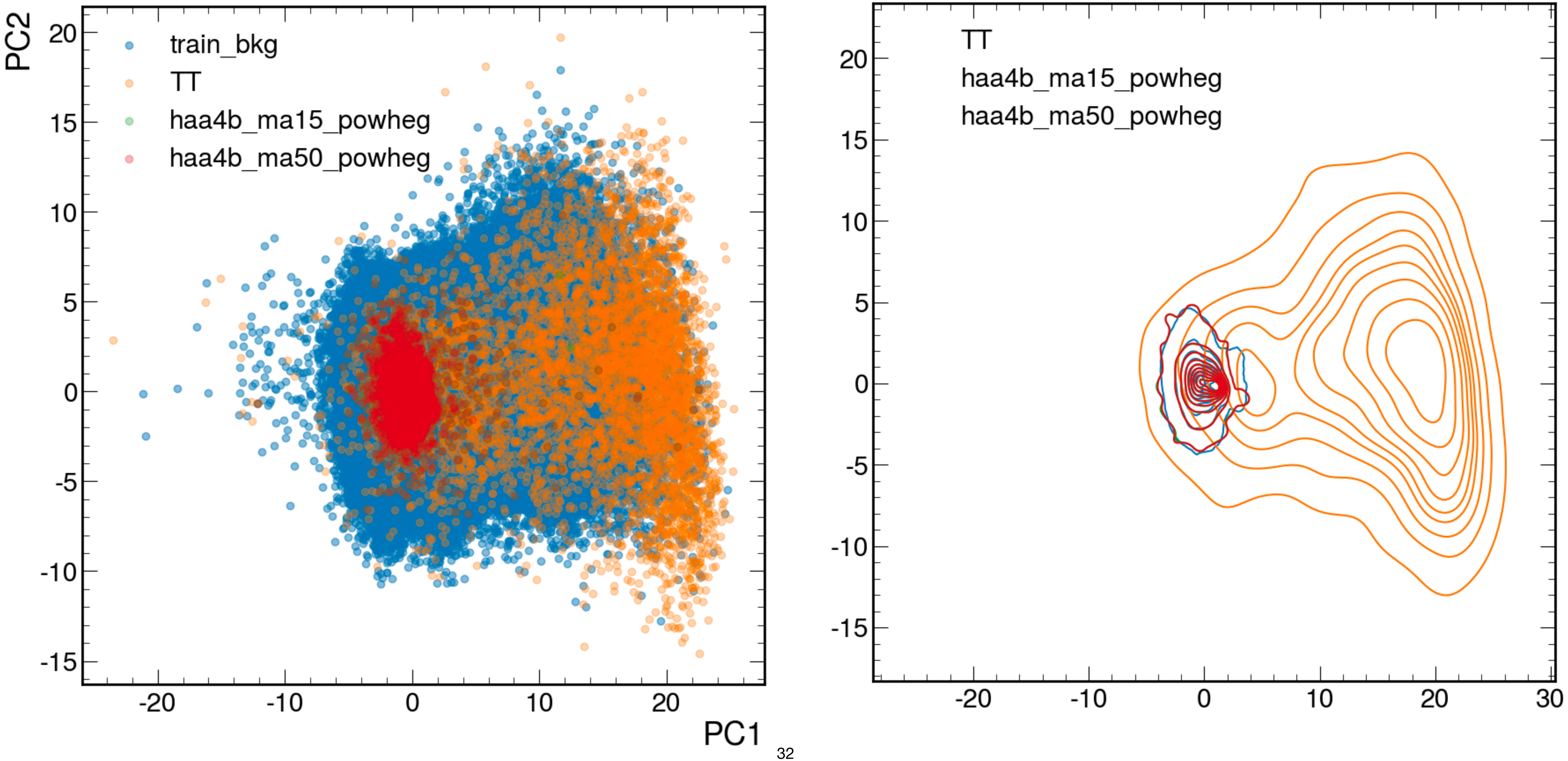
Principal Component Analysis (PCA)

Scree Plot

- Eigenvalues correspond to variances of principal comps.(PC)
- PCA's goal is to determine:
    - What PC correspond to the highest variance
    - Helps determine how many PC to retain during dim. reduction



First k Principle Components (PCs) Scatter and KDE Plots



Summary

- The current DNNVAE out performs the INVAE at the task of anomaly detection
- DNNVAE encoder is being put on L1 trigger
- Final results will be written and publish externally
- Mapping latent space of DNNVAE encoder is in progress
- The results of PCA are being interpreted to:
  - Determine optimal amount of PC
  - Interpret what variables contribute to which PC most
  - Interpret/explain clustering in reduced dim. space

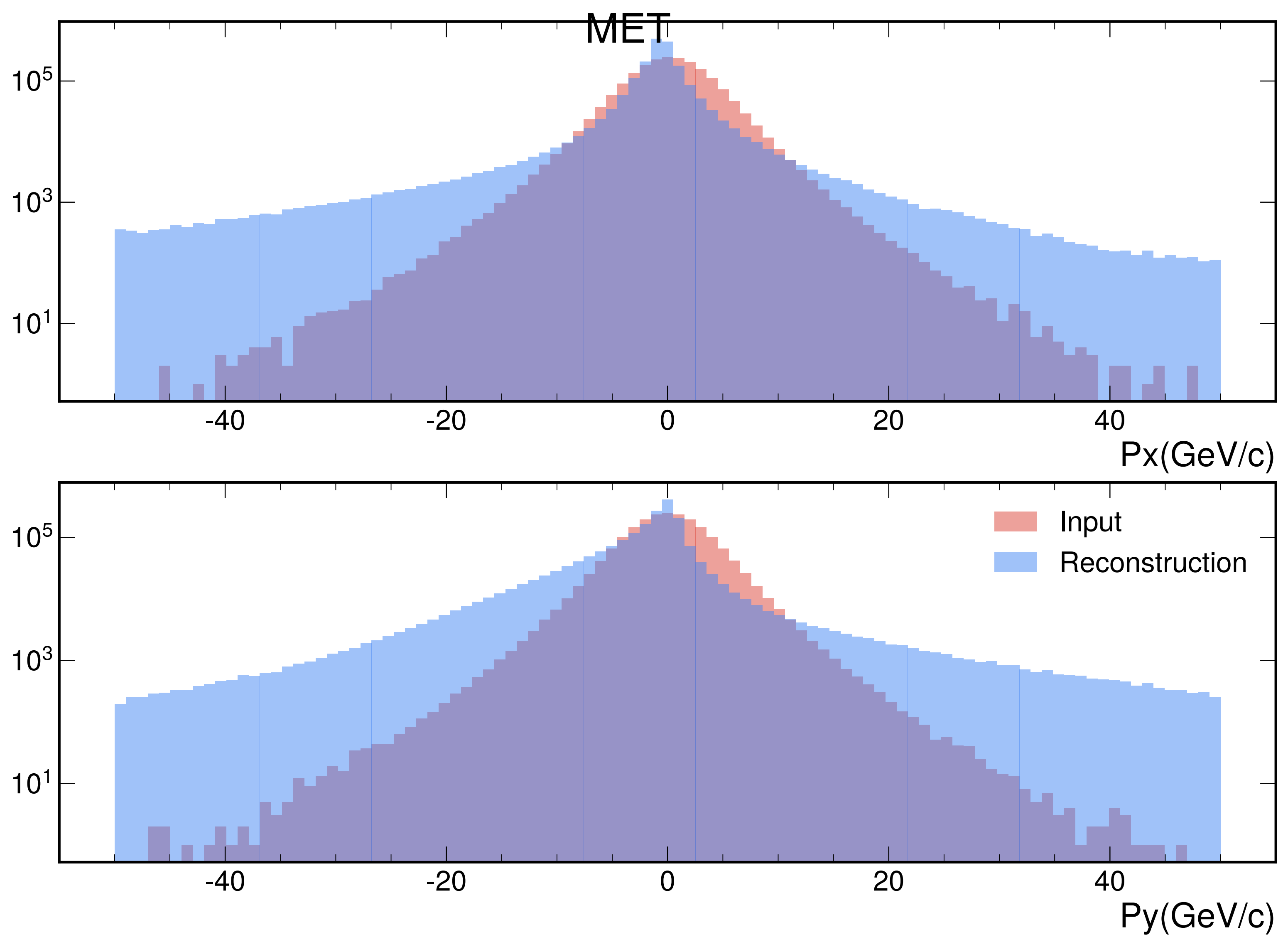
Moving Forward

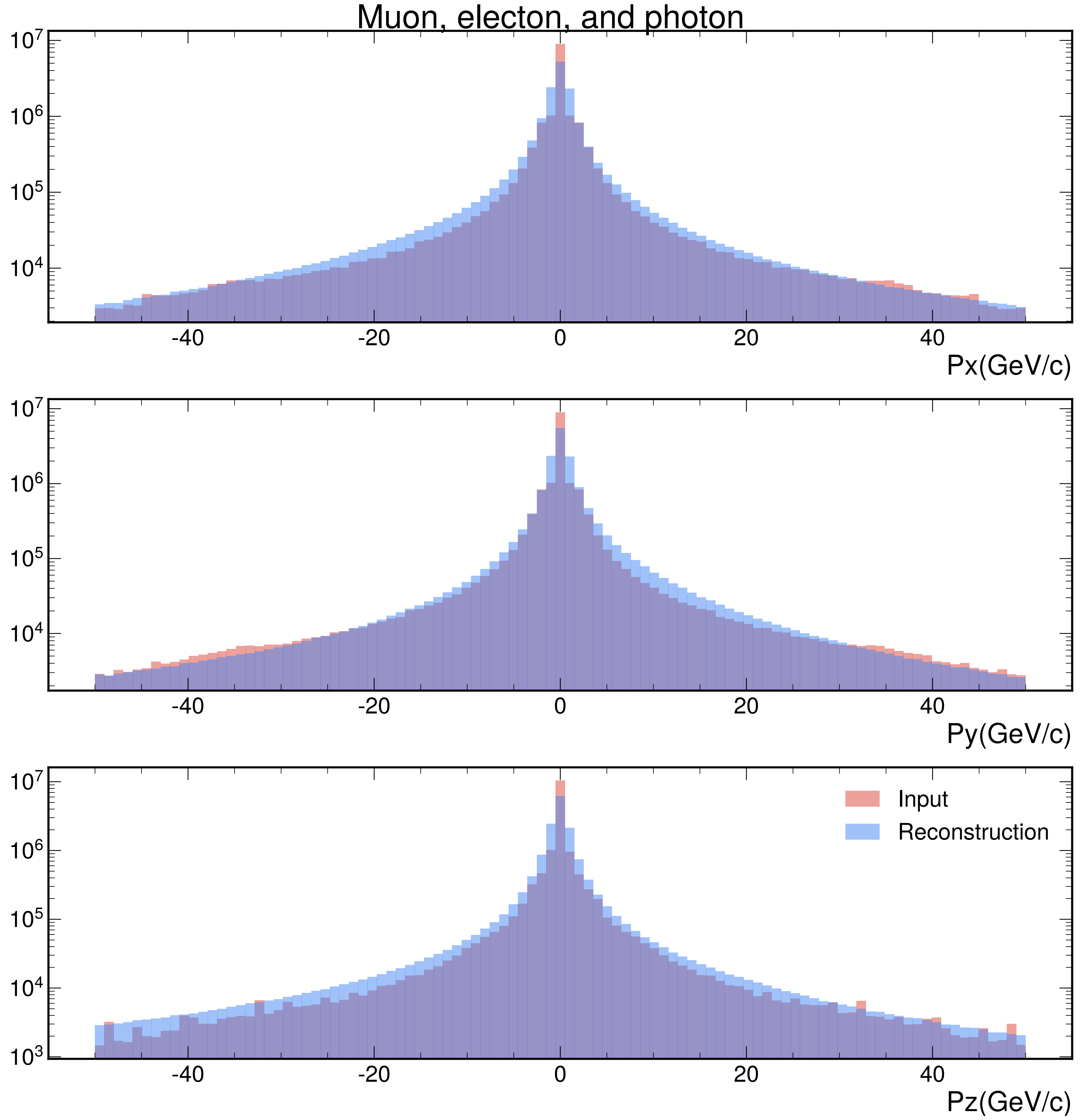
- Use the results from PCA to:
  - Determine optimal amount of PC
  - Interpret what variables contribute to which PC most
  - Interpret/explain clustering in reduced dim. space
- Continue collaborating with CU Boulder graduate and summer students:
  - Complete data characterization

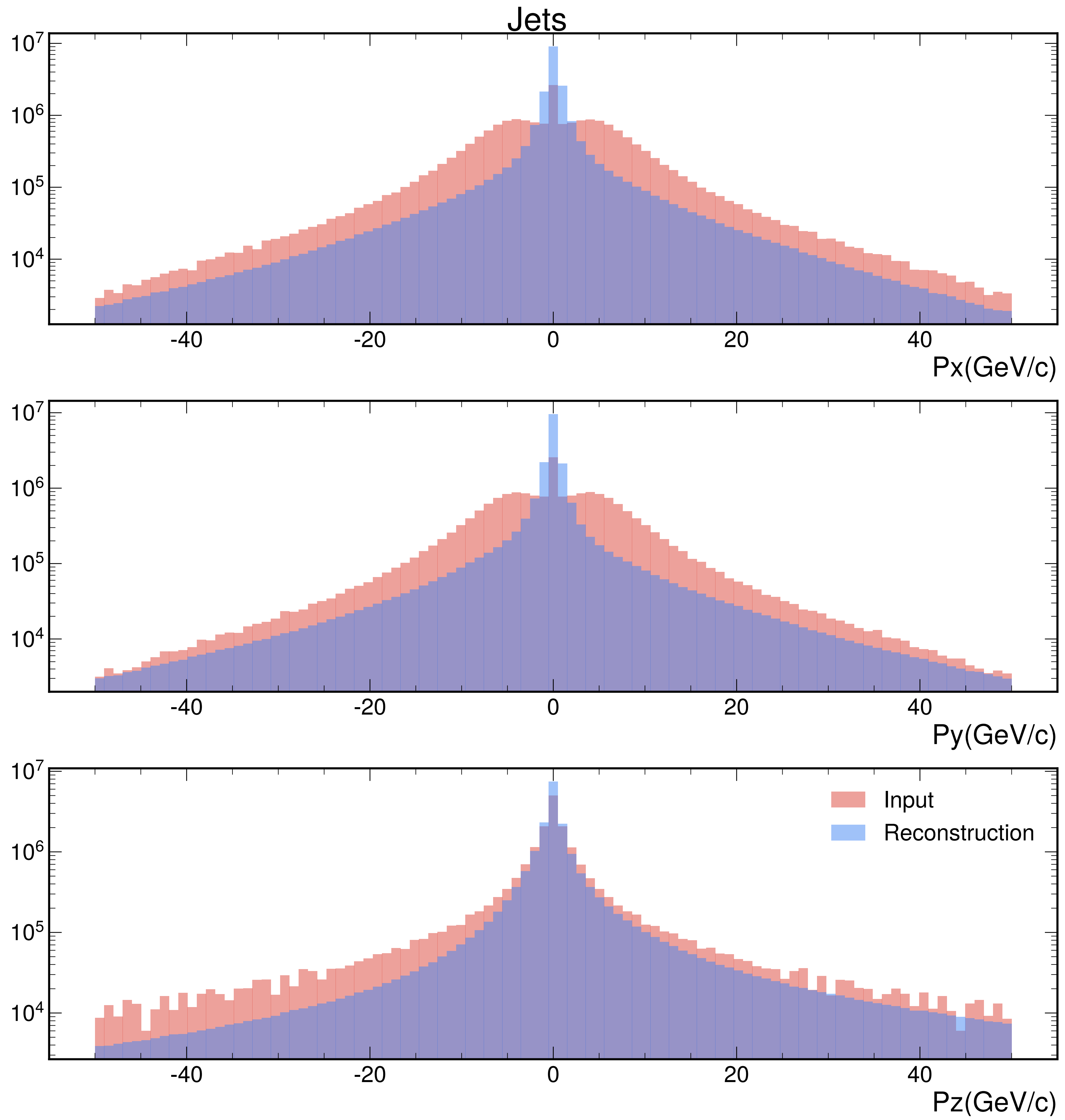
Thank you!

Backup Slides

ttbar Object Reconstruction (DNN)

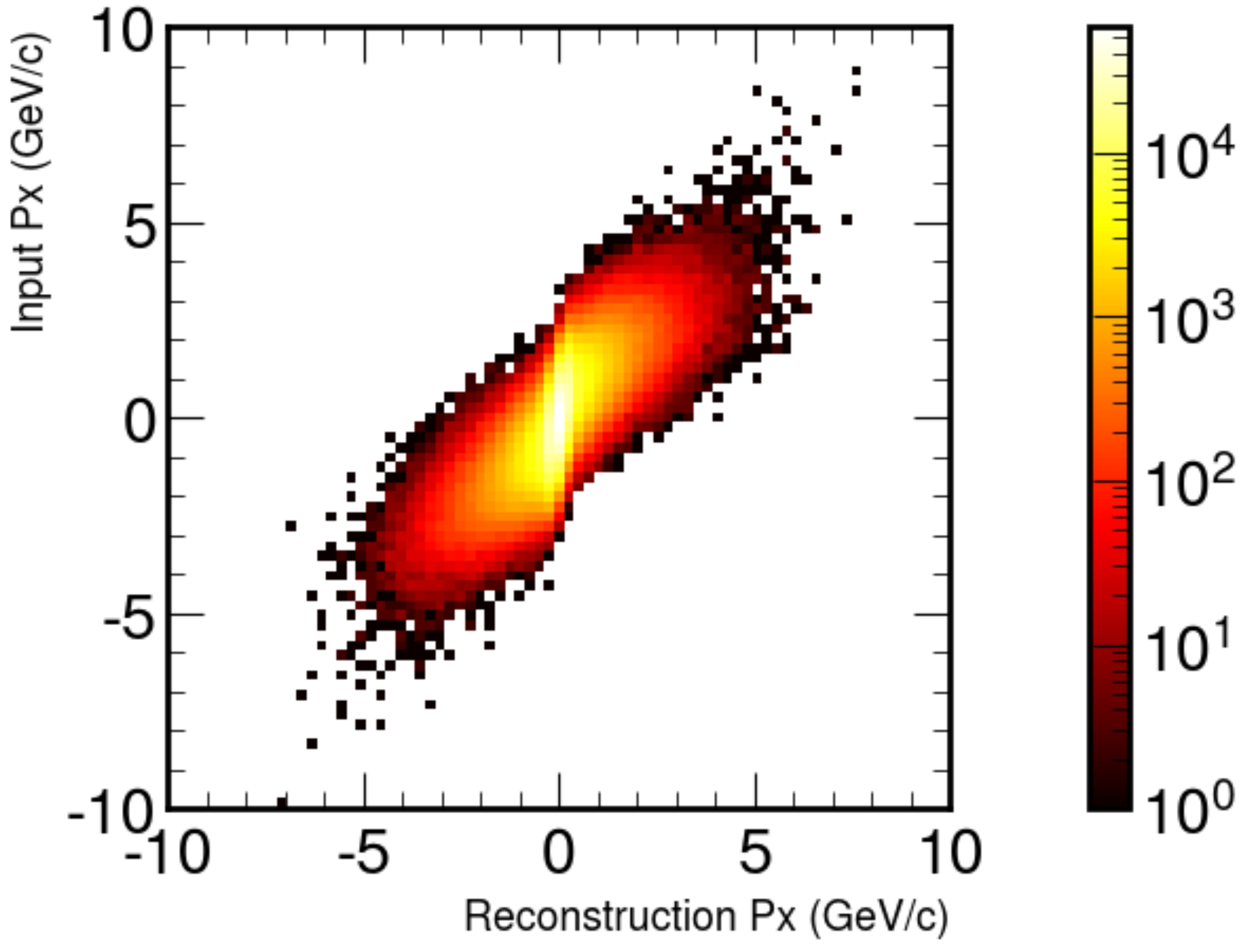


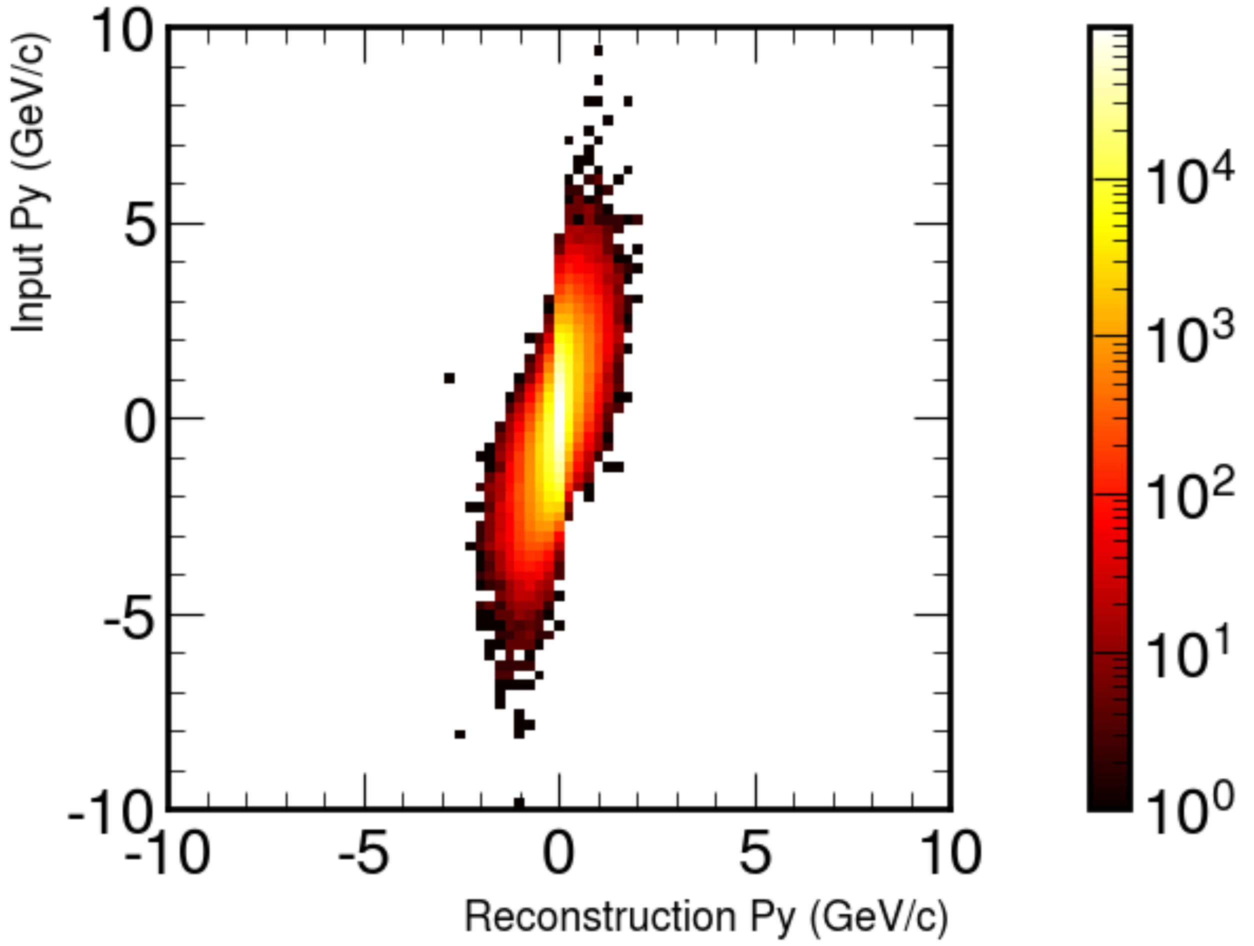




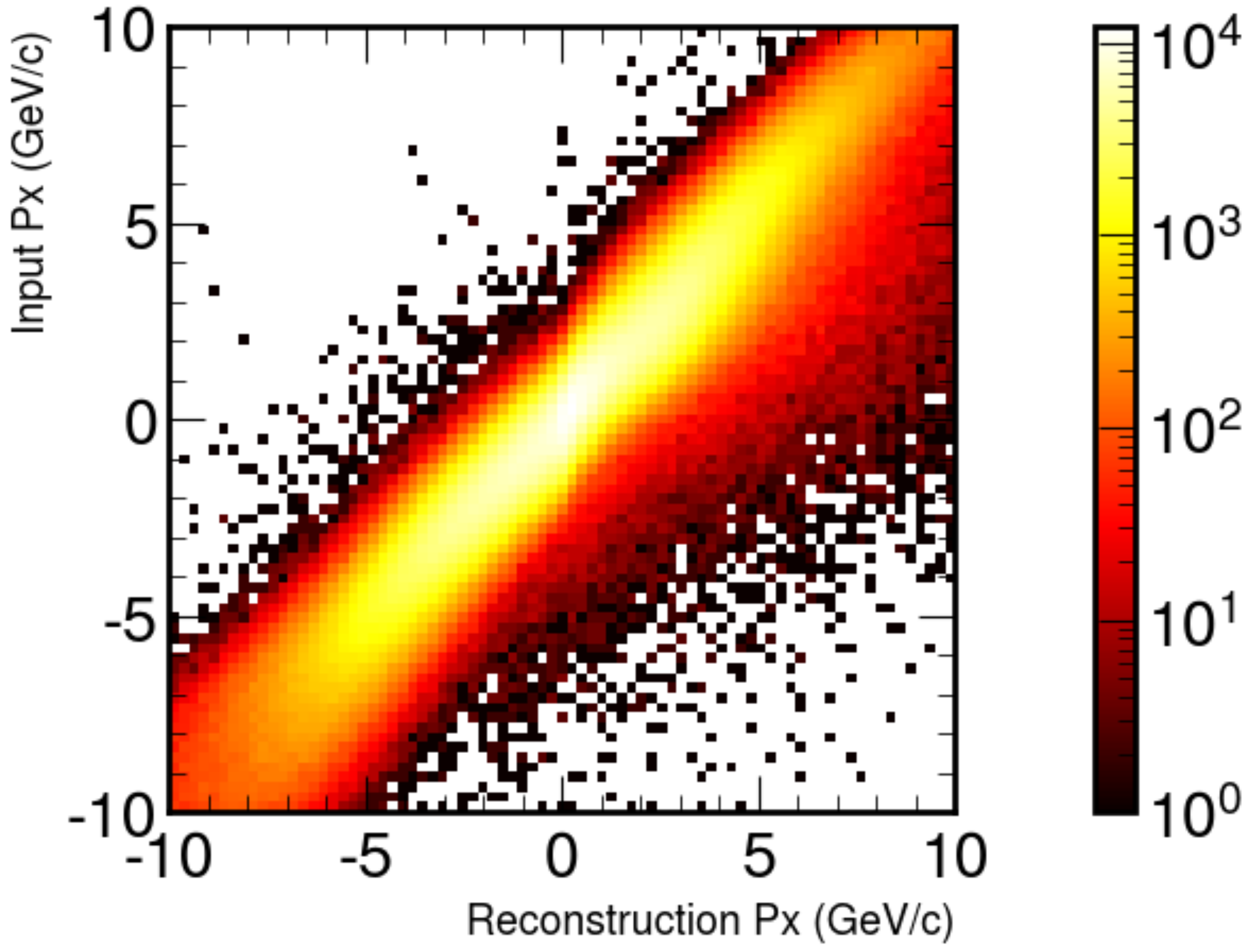
MET Heat Maps

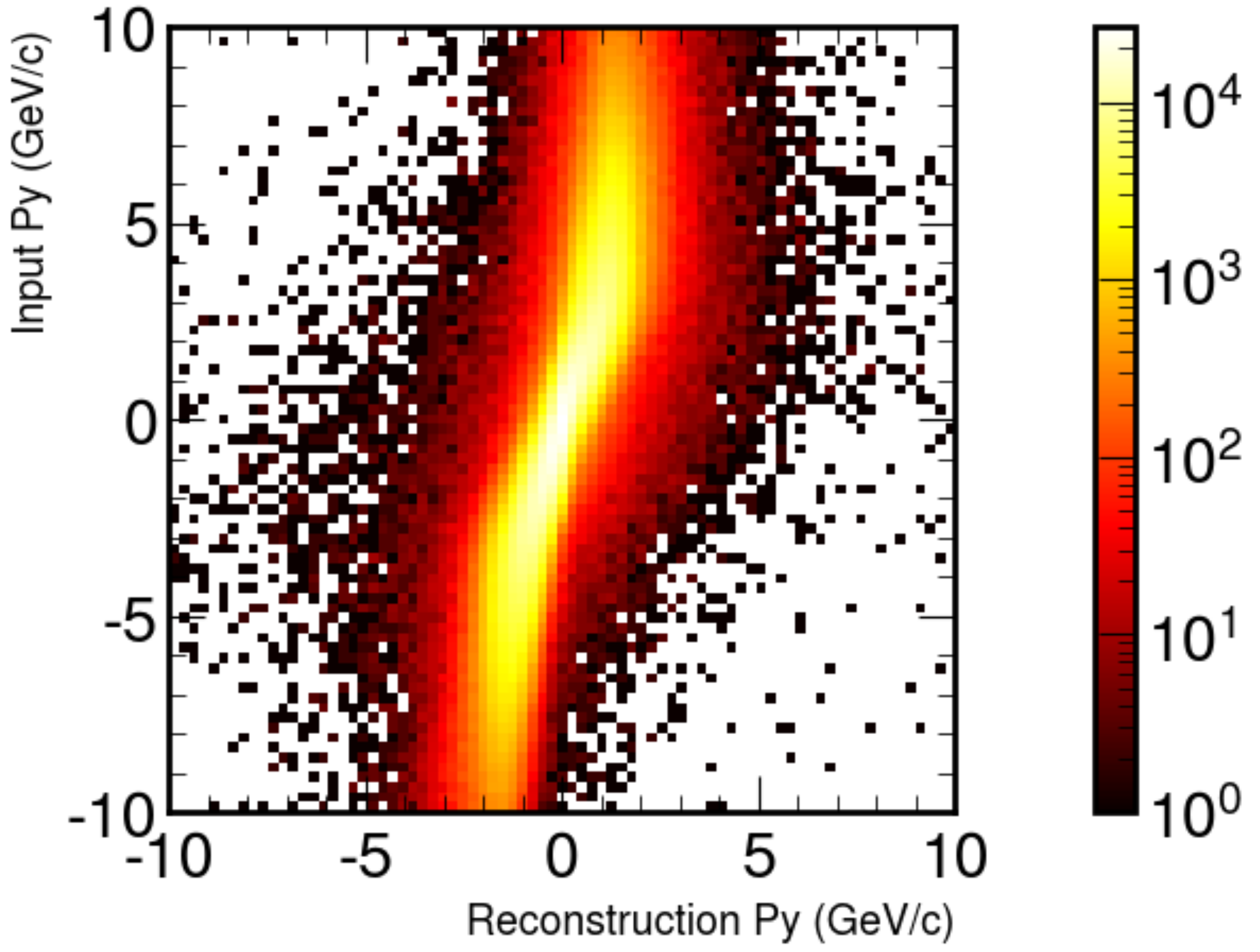
MET



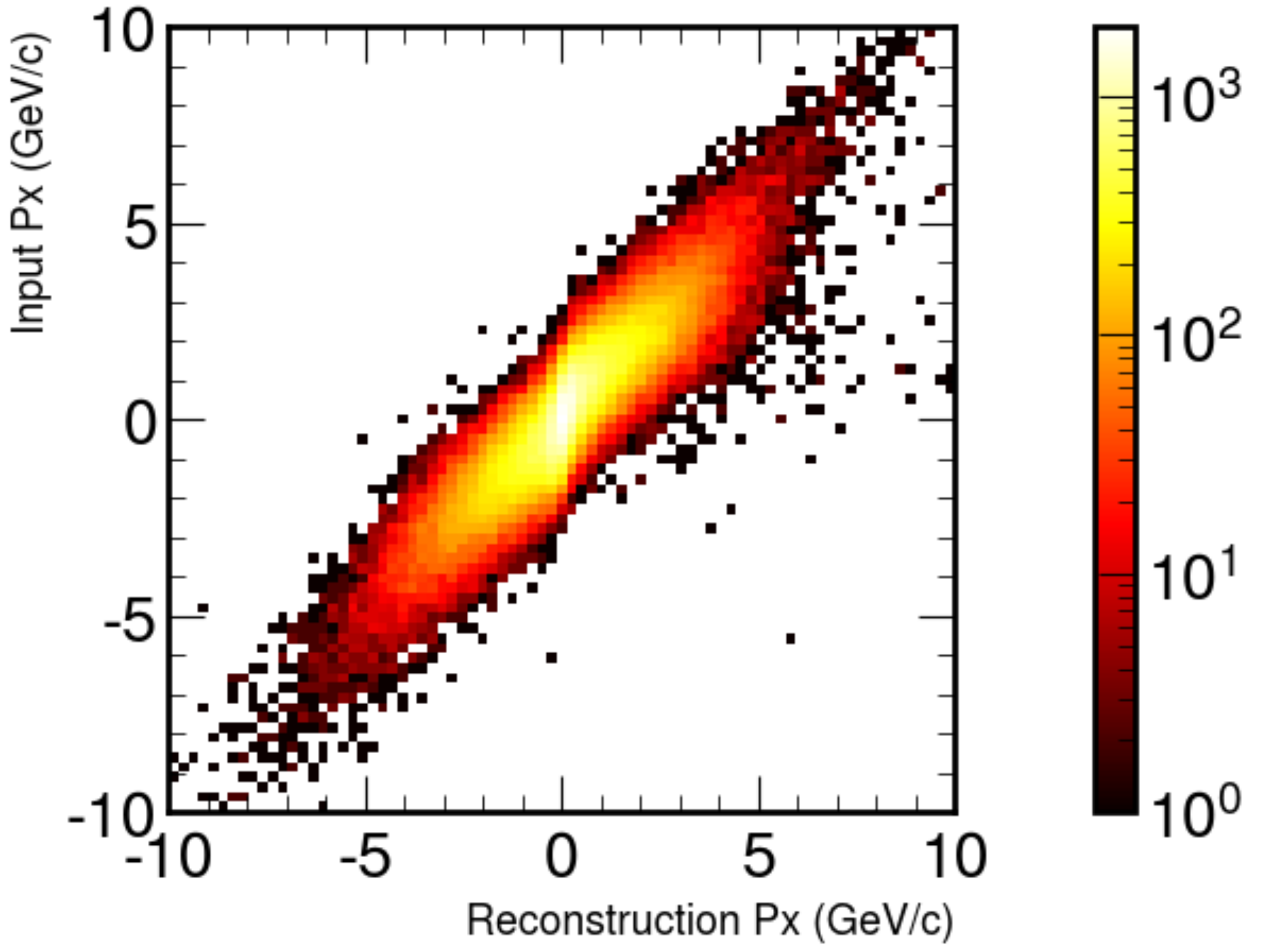


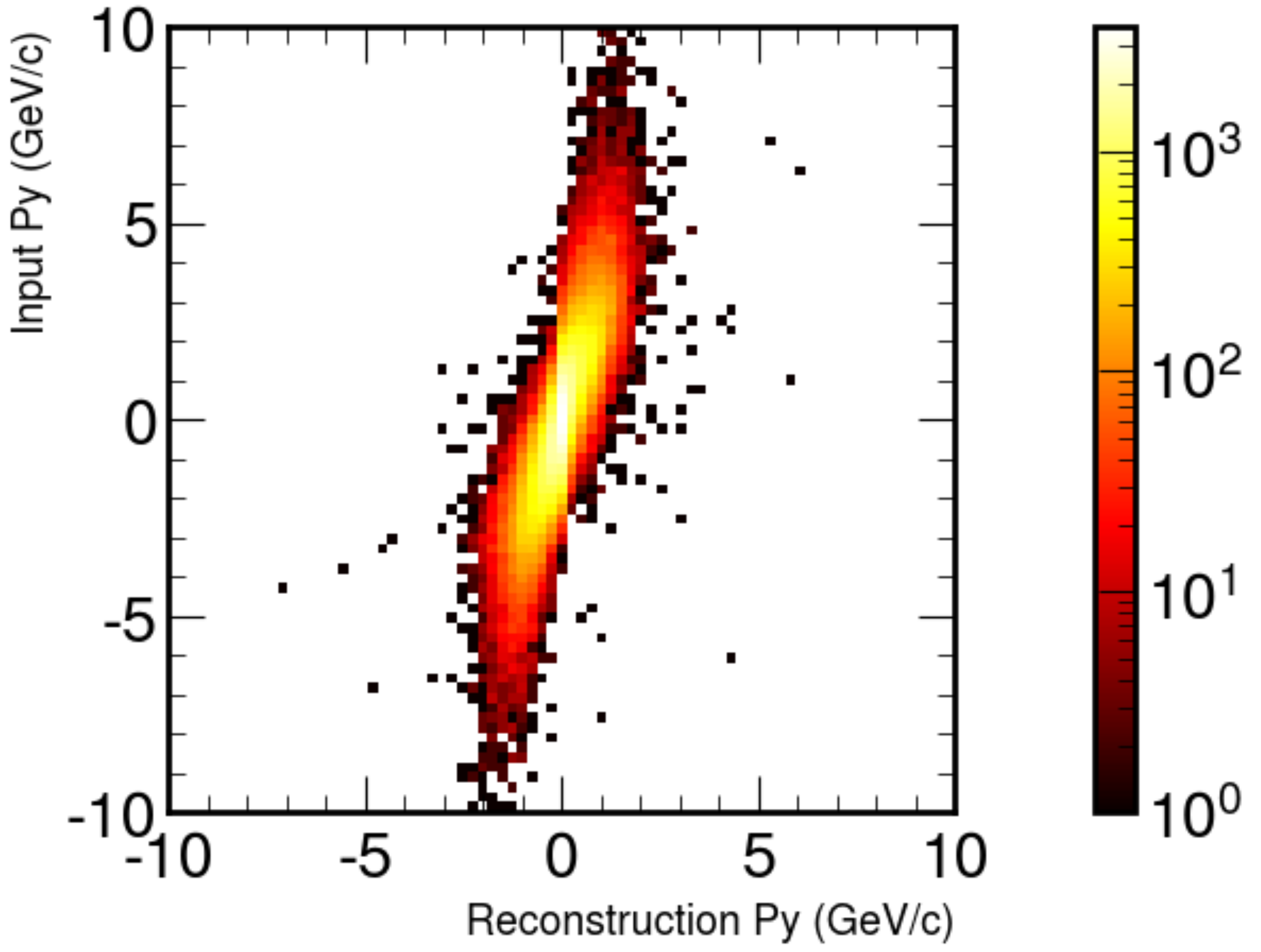
MET





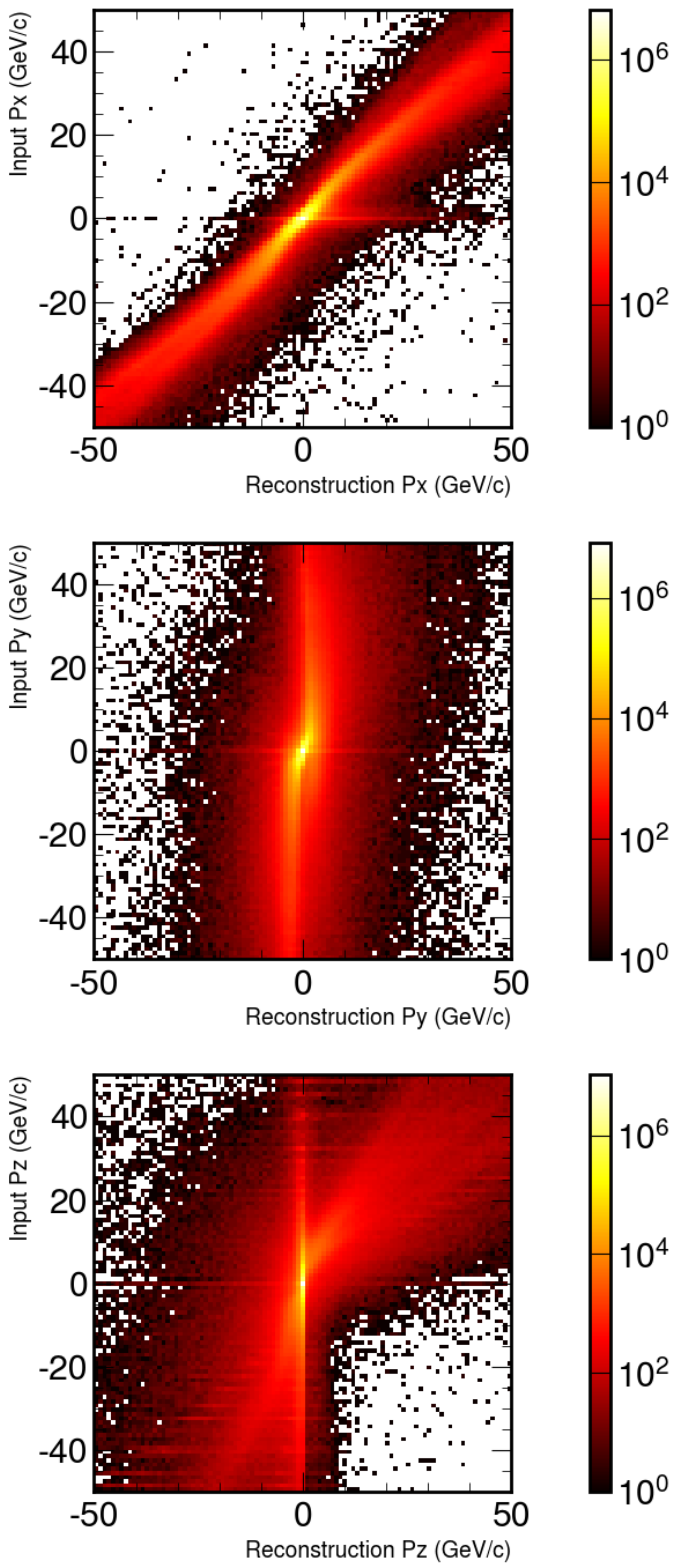
MET



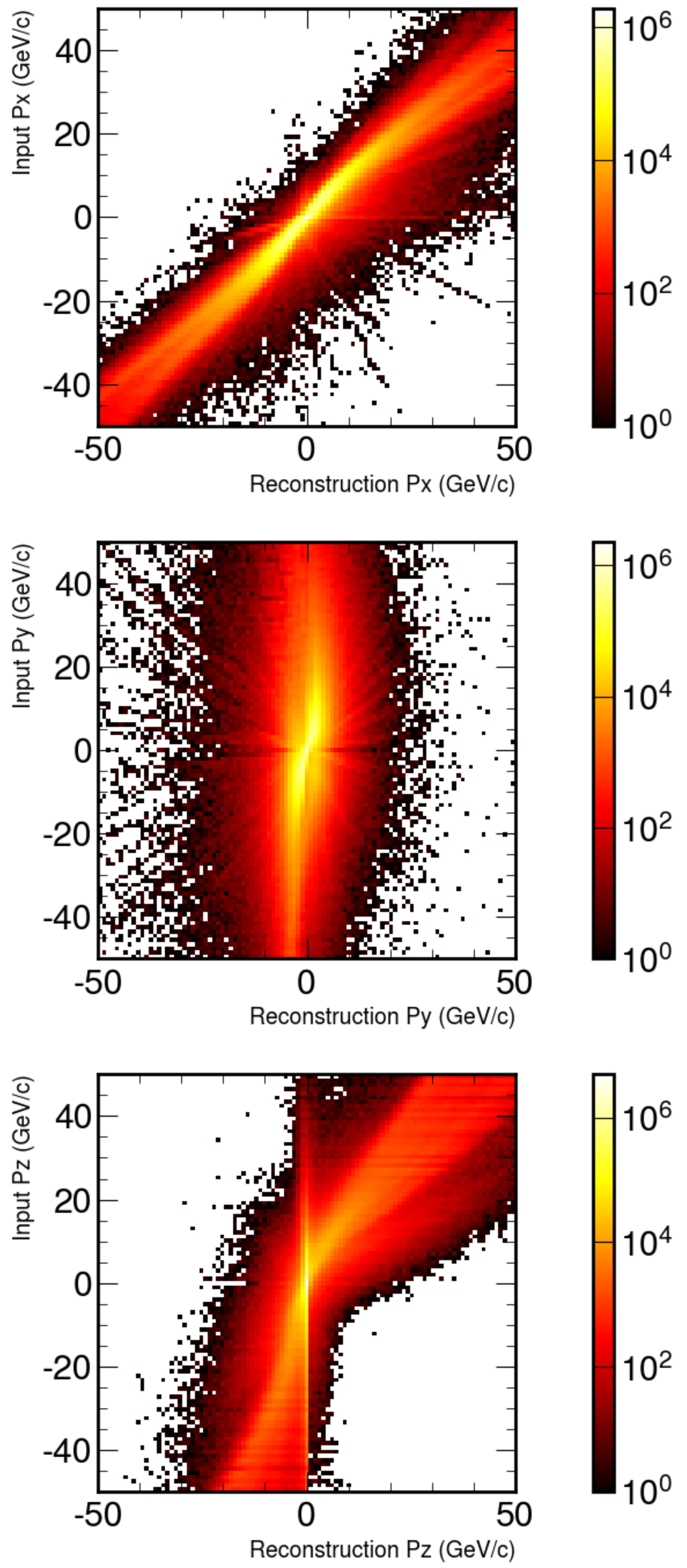


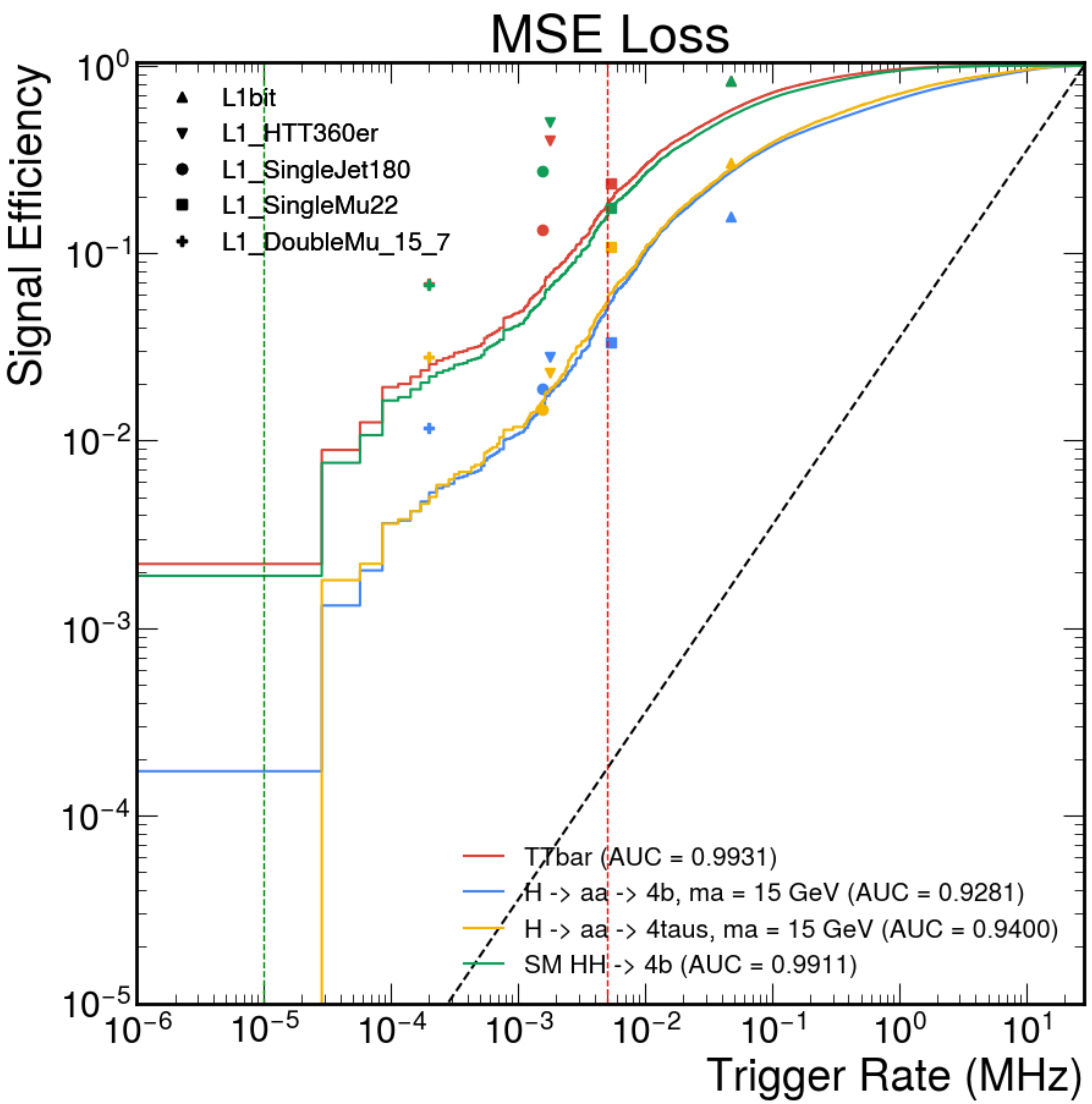
ttbar Heat Maps

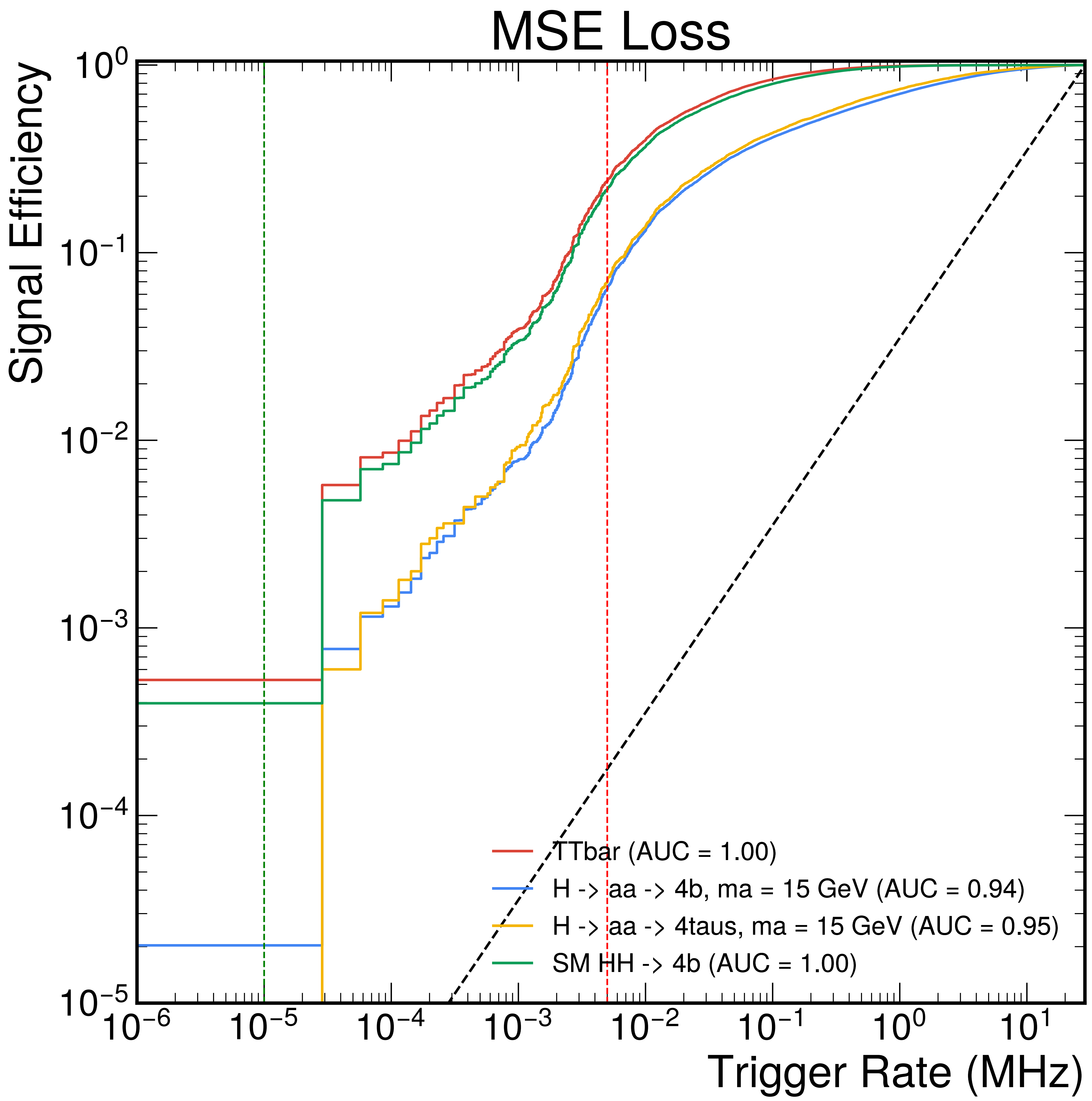
Leptons

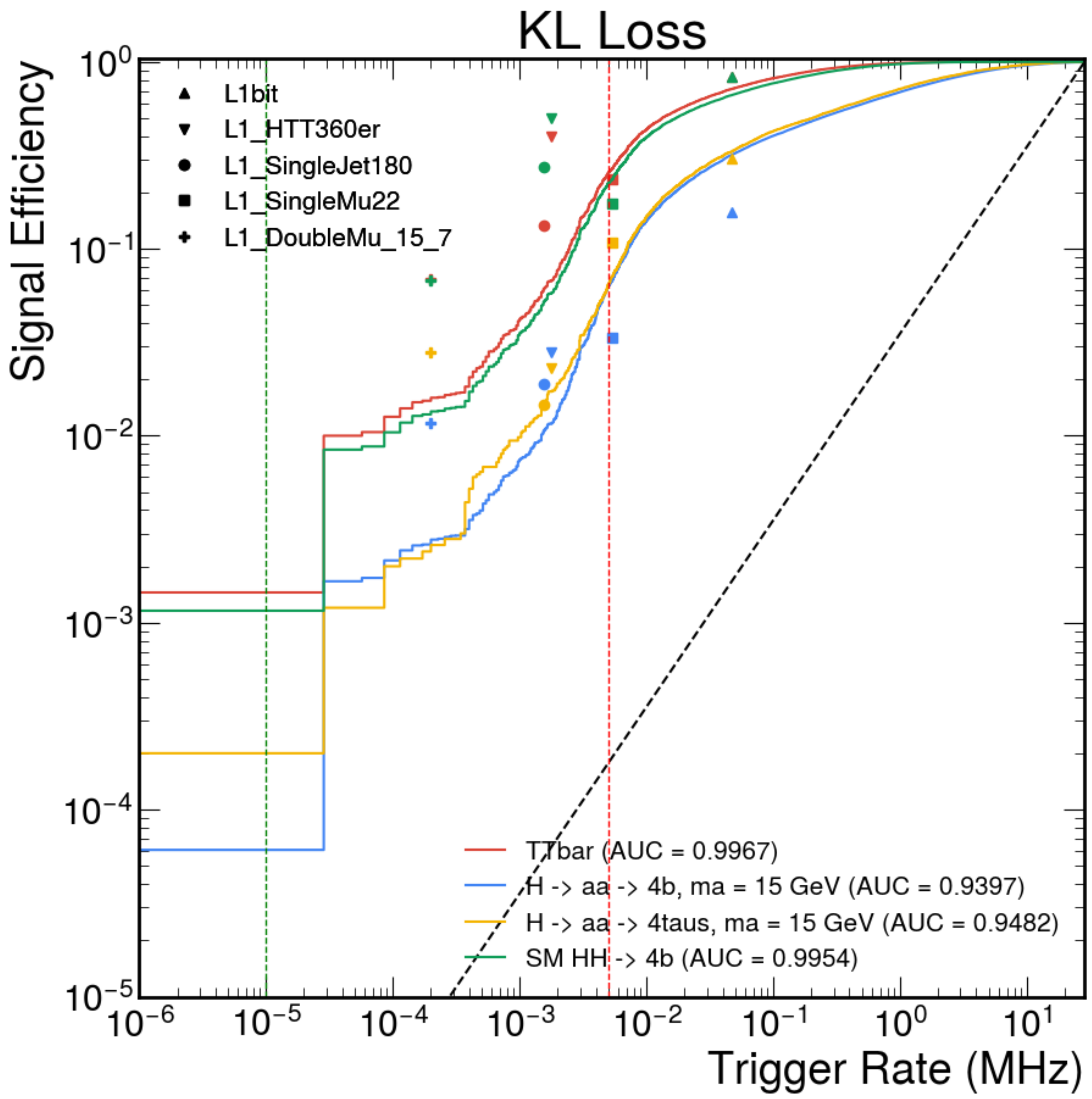


Jets

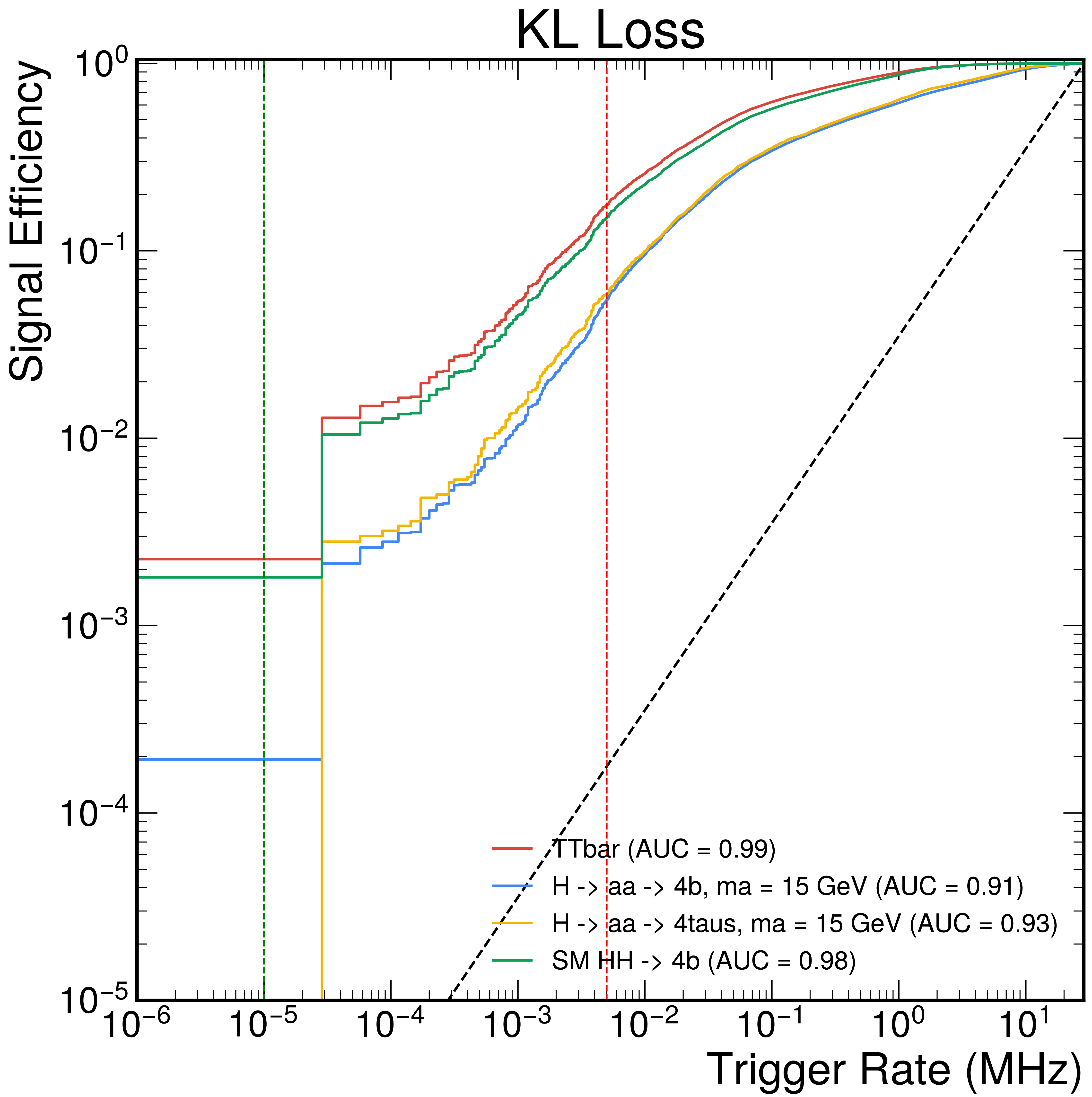



MSE Loss
Signal Efficiency
Trigger Rate (MHz)
L1bit
L1_HTT360er
L1_SingleJet180
L1_SingleMu22
L1_DoubleMu_15_7
TTbar (AUC = 0.9931)
H -> aa -> 4b, ma = 15 GeV (AUC = 0.9281)
H -> aa -> 4taus, ma = 15 GeV (AUC = 0.9400)
SM HH -> 4b (AUC = 0.9911)

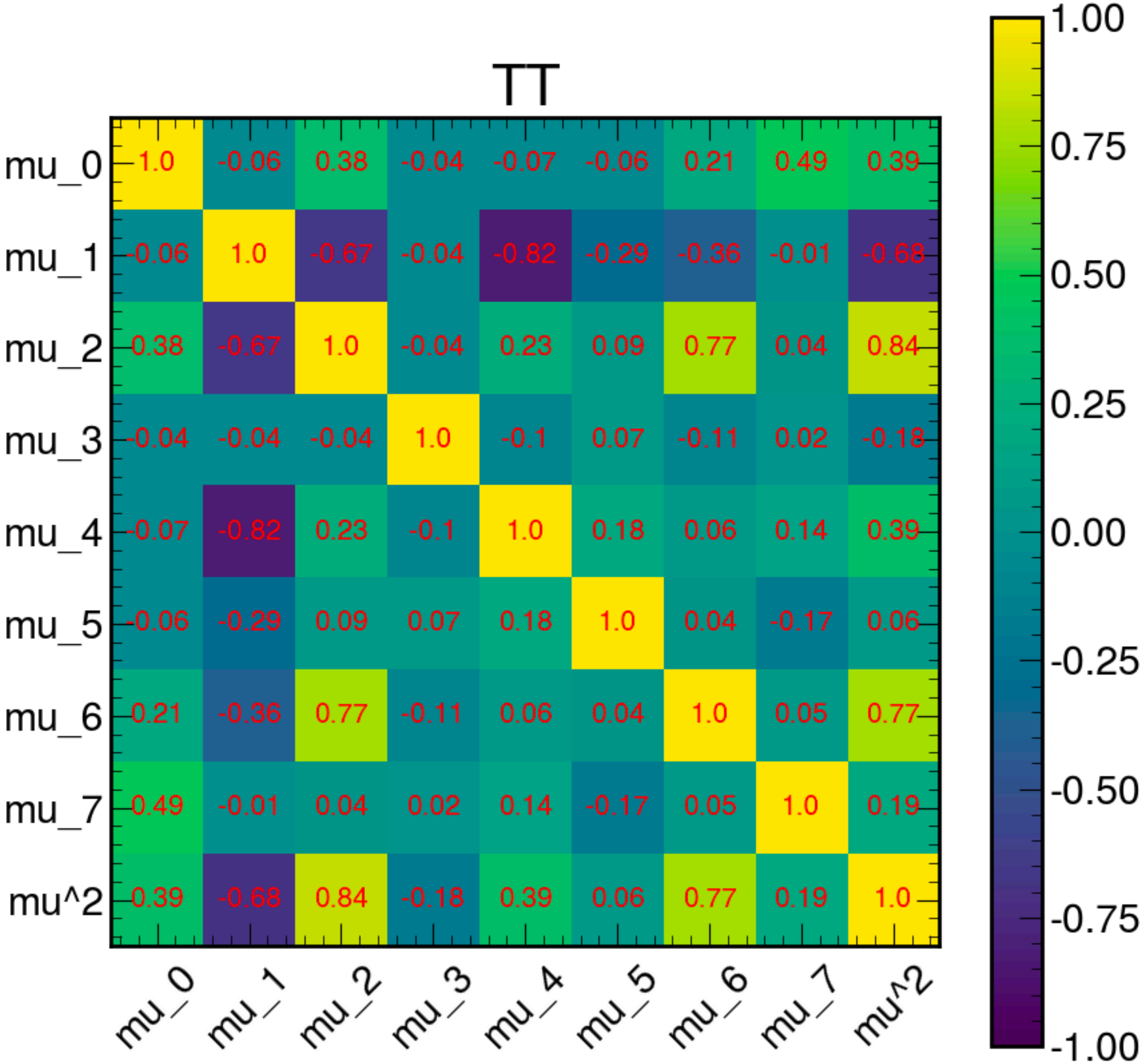

MSE Loss
Signal Efficiency
Trigger Rate (MHz)
TTbar (AUC = 1.00)
H -> aa -> 4b, ma = 15 GeV (AUC = 0.94)
H -> aa -> 4taus, ma = 15 GeV (AUC = 0.95)
SM HH -> 4b (AUC = 1.00)

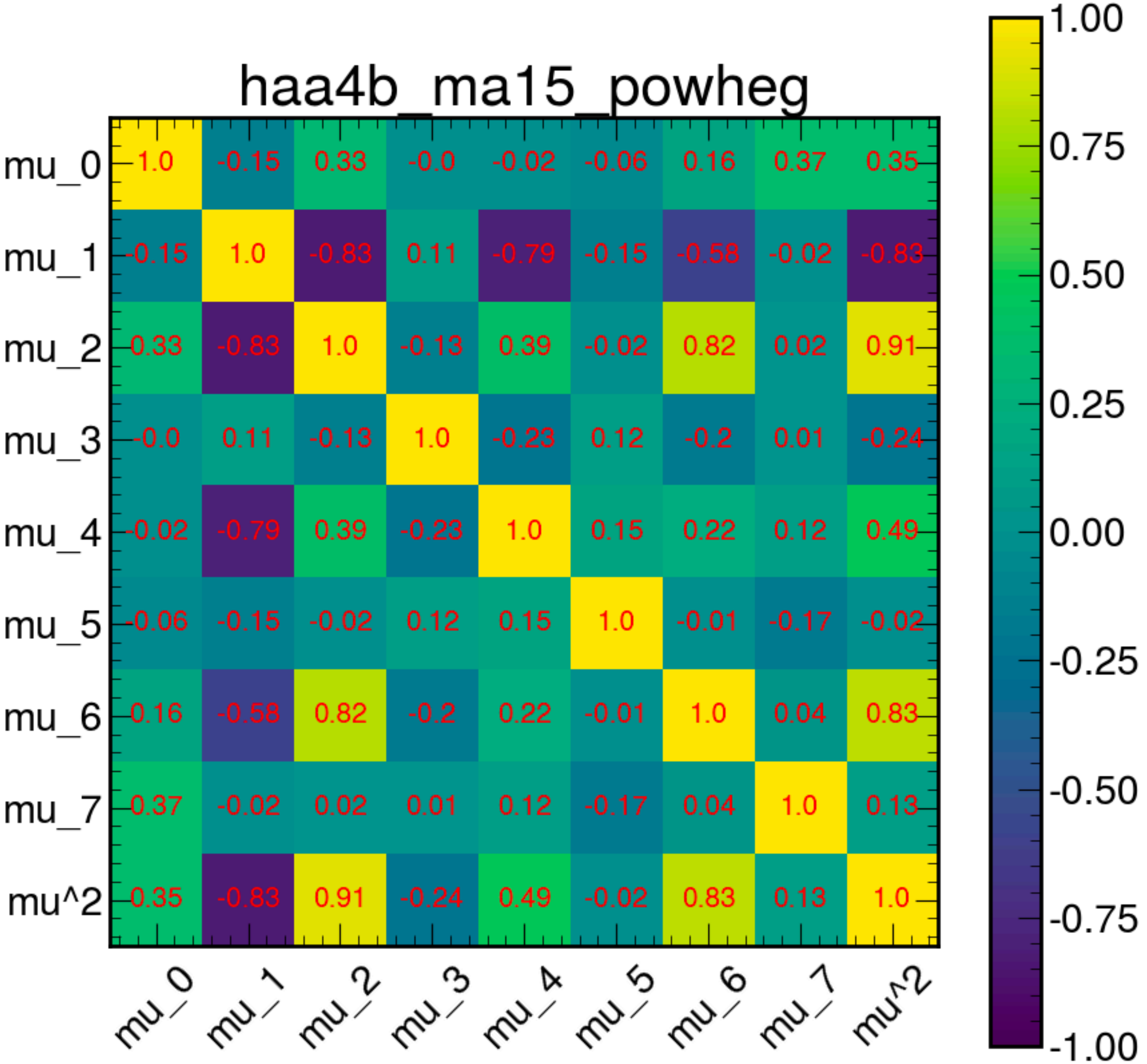


KL Loss
Signal Efficiency
Trigger Rate (MHz)
L1bit
L1_HTT360er
L1_SingleJet180
L1_SingleMu22
L1_DoubleMu_15_7
TTbar (AUC = 0.9967)
H -> aa -> 4b, ma = 15 GeV (AUC = 0.9397)
H -> aa -> 4taus, ma = 15 GeV (AUC = 0.9482)
SM HH -> 4b (AUC = 0.9954)



KL Loss
Signal Efficiency
Trigger Rate (MHz)
TTbar (AUC = 0.99)
H -> aa -> 4b, ma = 15 GeV (AUC = 0.91)
H -> aa -> 4taus, ma = 15 GeV (AUC = 0.93)
SM HH -> 4b (AUC = 0.98)


TT
mu_0 1.0 -0.06 0.38 -0.04 -0.07 -0.06 0.21 0.49 0.39
mu_1 -0.06 1.0 -0.67 -0.04 -0.82 -0.29 -0.36 -0.01 -0.68
mu_2 0.38 -0.67 1.0 -0.04 0.23 0.09 0.77 0.04 0.84
mu_3 -0.04 -0.04 -0.04 1.0 -0.1 0.07 -0.11 0.02 -0.18
mu_4 -0.07 -0.82 0.23 -0.1 1.0 0.18 0.06 0.14 0.39
mu_5 -0.06 -0.29 0.09 0.07 0.18 1.0 0.04 -0.17 0.06
mu_6 0.21 -0.36 0.77 -0.11 0.06 0.04 1.0 0.05 0.77
mu_7 0.49 -0.01 0.04 0.02 0.14 -0.17 0.05 1.0 0.19
mu^2 0.39 -0.68 0.84 -0.18 0.39 0.06 0.77 0.19 1.0
mu_0 mu_1 mu_2 mu_3 mu_4 mu_5 mu_6 mu_7 mu^2
1.00 0.75 0.50 0.25 0.00 -0.25 -0.50 -0.75 -1.00


haa4b_ma15_powheg
mu_0 1.0 -0.15 0.33 -0.0 -0.02 -0.06 0.16 0.37 0.35
mu_1 -0.15 1.0 -0.83 0.11 -0.79 -0.15 -0.58 -0.02 -0.83
mu_2 0.33 -0.83 1.0 -0.13 0.39 -0.02 0.82 0.02 0.91
mu_3 -0.0 0.11 -0.13 1.0 -0.23 0.12 -0.2 0.01 -0.24
mu_4 -0.02 -0.79 0.39 -0.23 1.0 0.15 0.22 0.12 0.49
mu_5 -0.06 -0.15 -0.02 0.12 0.15 1.0 -0.01 -0.17 -0.02
mu_6 0.16 -0.58 0.82 -0.2 0.22 -0.01 1.0 0.04 0.83
mu_7 0.37 -0.02 0.02 0.01 0.12 -0.17 0.04 1.0 0.13
mu^2 0.35 -0.83 0.91 -0.24 0.49 -0.02 0.83 0.13 1.0
mu_0 mu_1 mu_2 mu_3 mu_4 mu_5 mu_6 mu_7 mu^2
1.00 0.75 0.50 0.25 0.00 -0.25 -0.50 -0.75 -1.00