VBS VVH All-Hadronic

New ABCDNet trails

April 27th, 2023

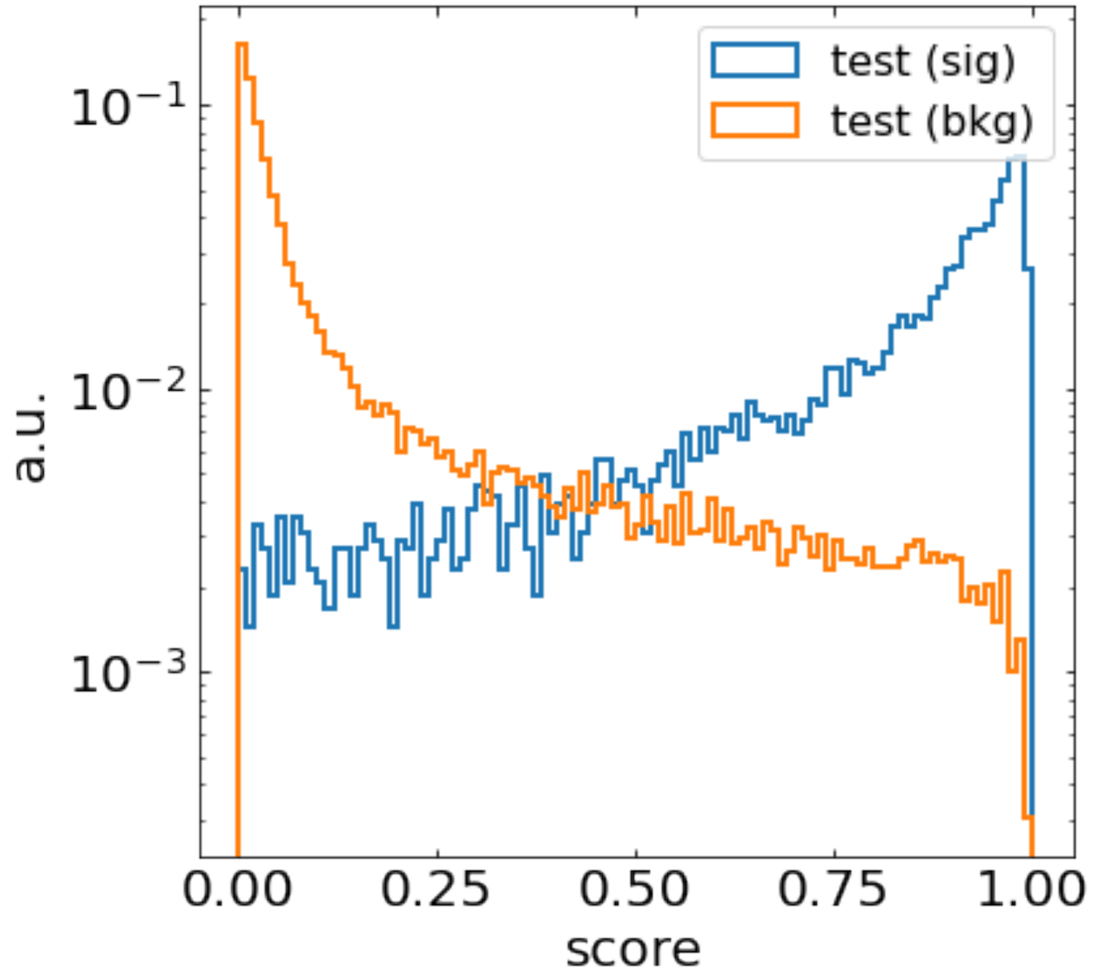
*P. Chang, L. Giannini, **J. Guiang**, Y. Xiang, E. Zenhom*

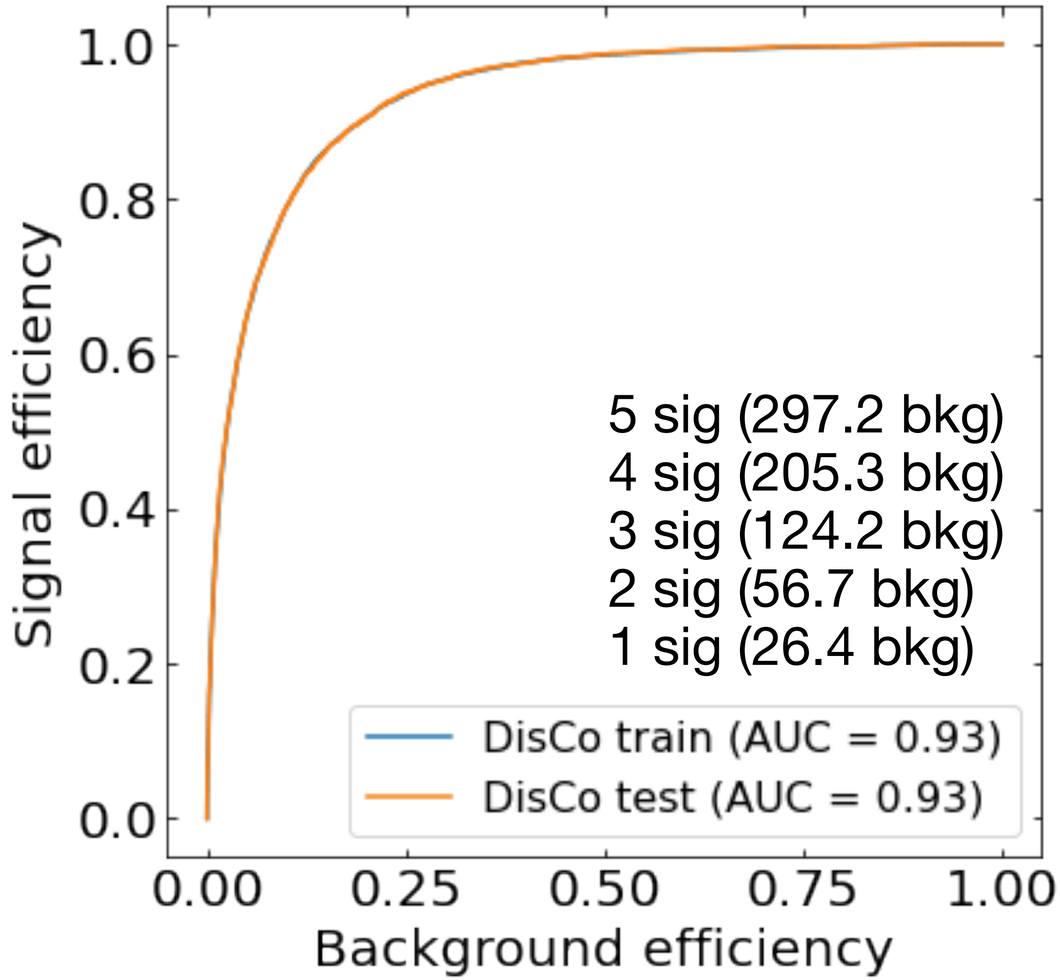
Overview

- Fixed some bugs while performing sanity check
- Re-trained ABCDNet:
  - Same learning rate, etc. as sanity check
  - Larger NN (4 hidden layers, 128 nodes per layer)
- Normalize inputs as follows:
  - $p_T \rightarrow \log(p_T)$
  - Mass → (mass - min)/(max - min)
- Loose presel: Xbb > 0.3 && XWqq > 0.2 && XWqq > 0.2

ABCDNet: λ = 100 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \text{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$

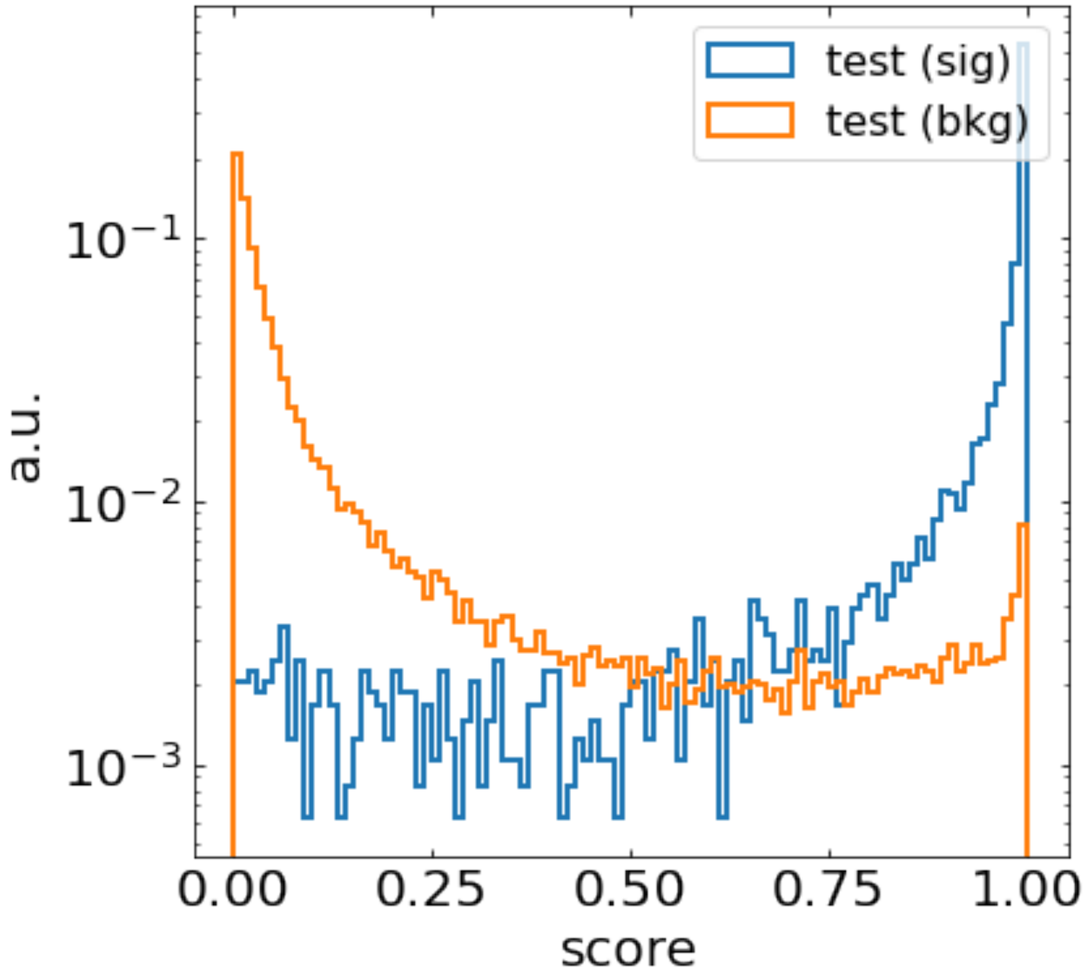


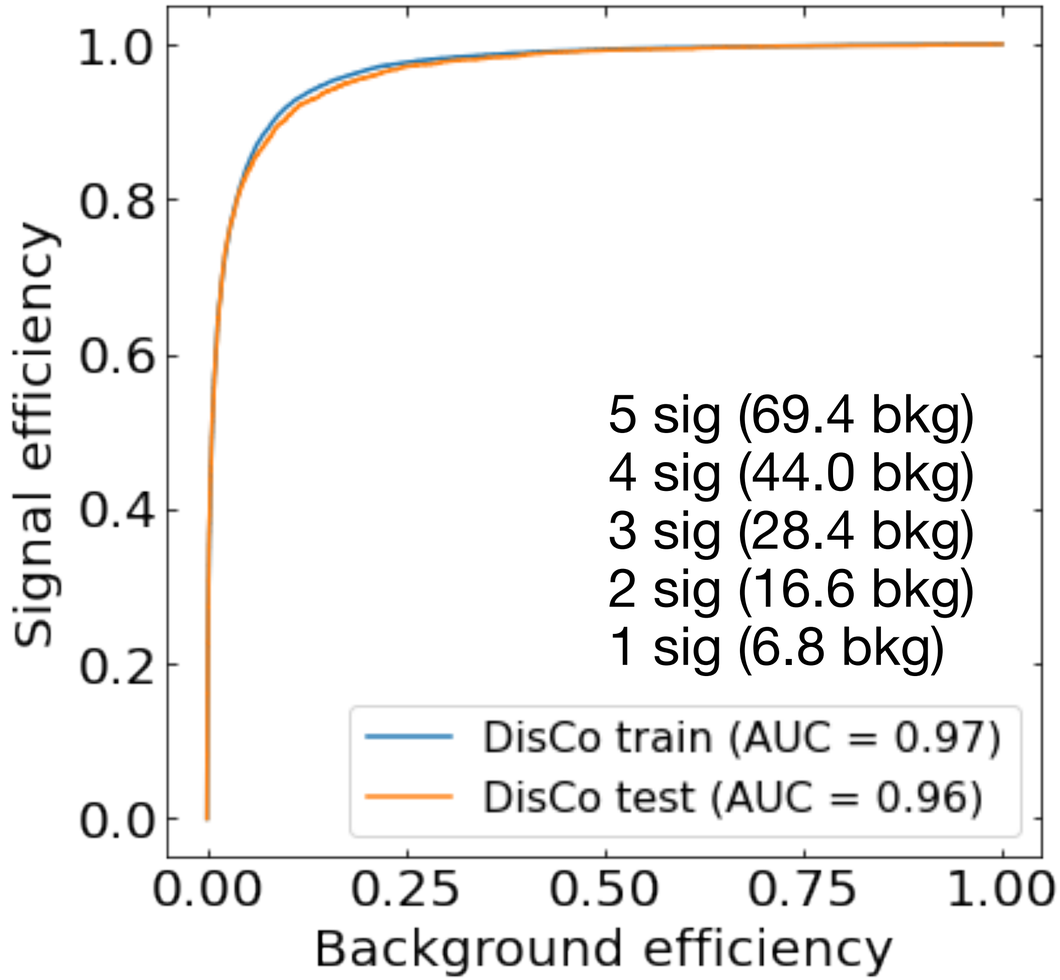


Epoch = 100 | LR = 0.001 (constant) | λ = 100

ABCDNet: λ = 100 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \text{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$



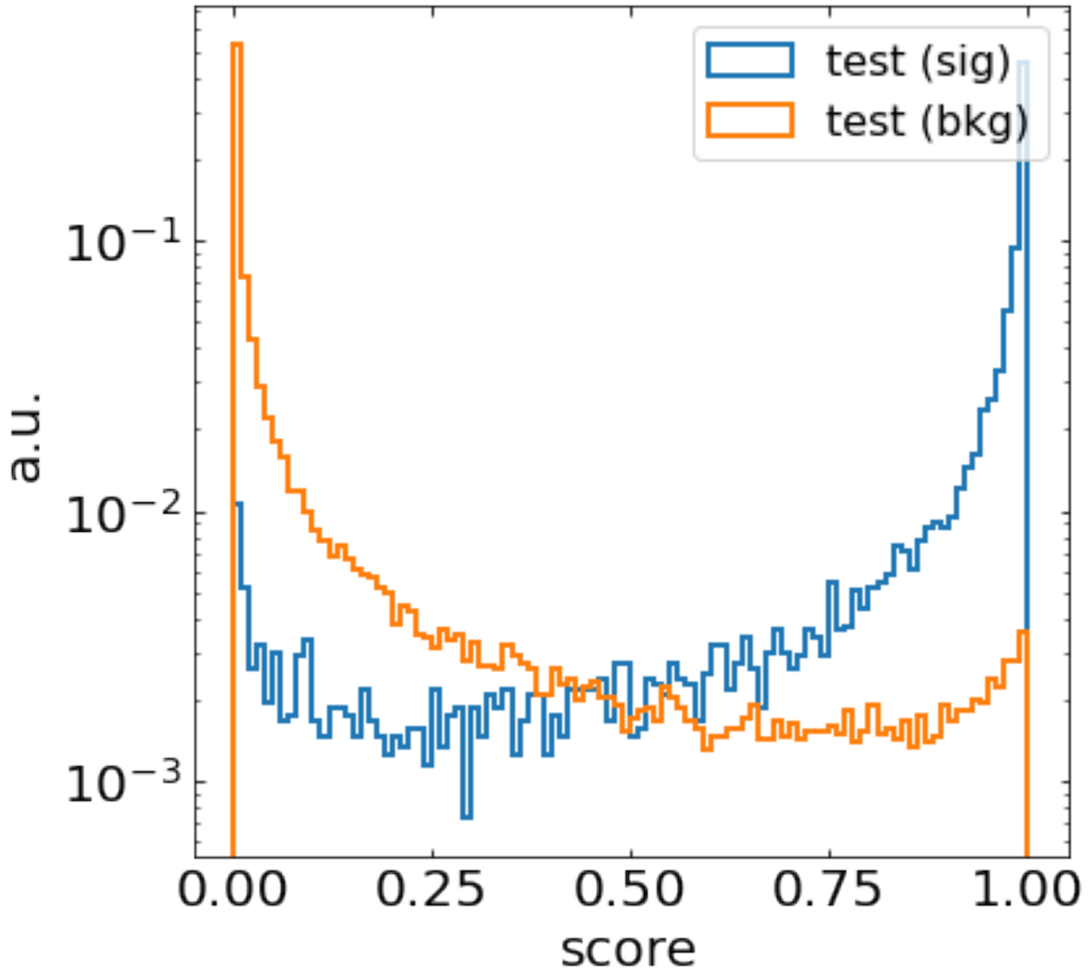


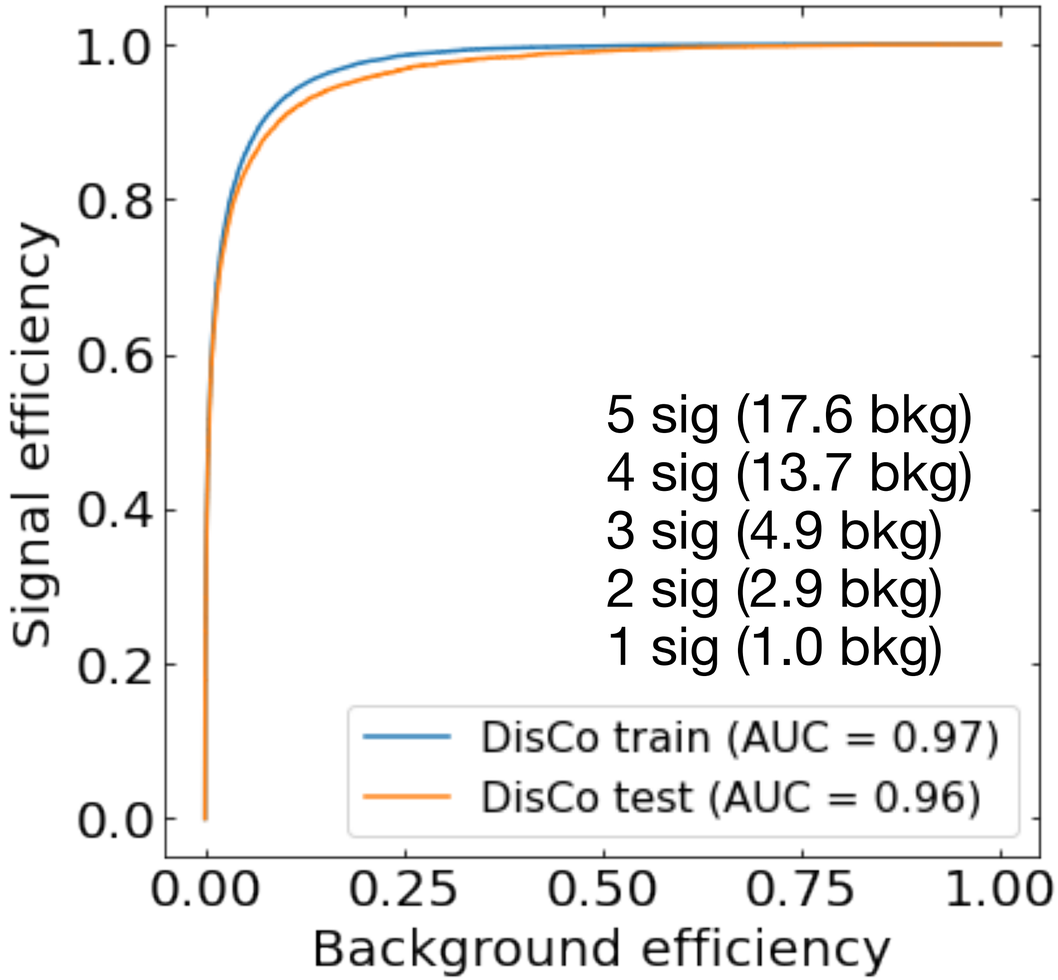
Epoch = 1000 | LR = 0.001 (constant) | λ = 100

UC San Diego

ABCDNet: λ = 100 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \mathrm{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$



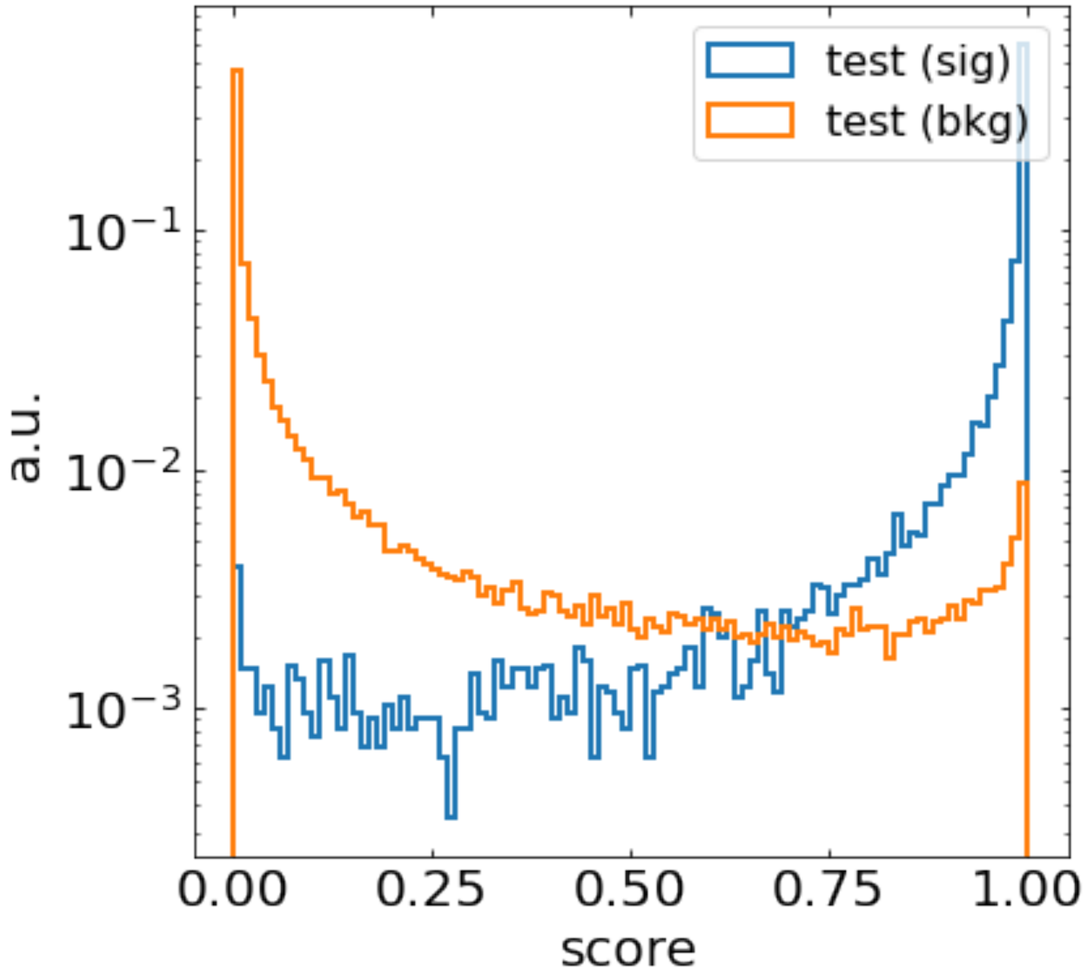


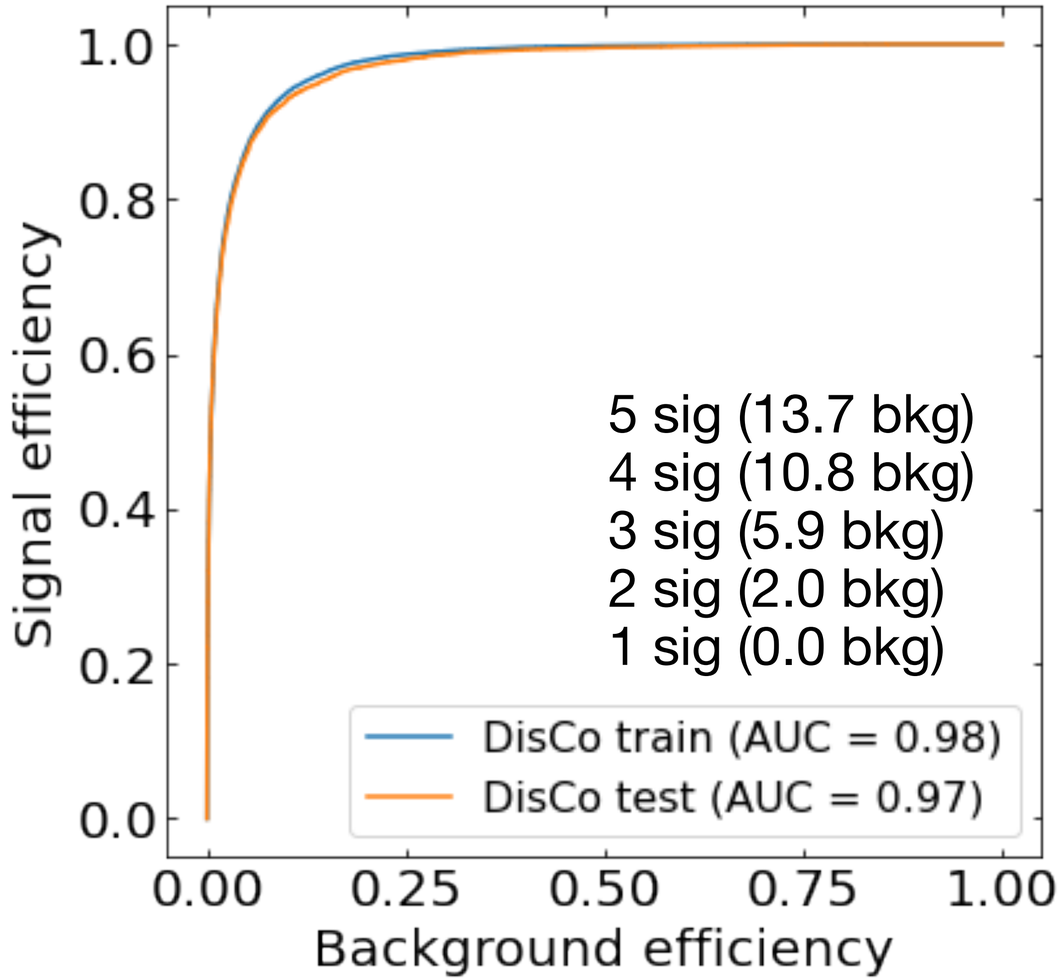
Epoch = 1000 | LR = 0.01 (constant) | λ = 100

UC San Diego

ABCDNet: λ = 100 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \mathrm{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$



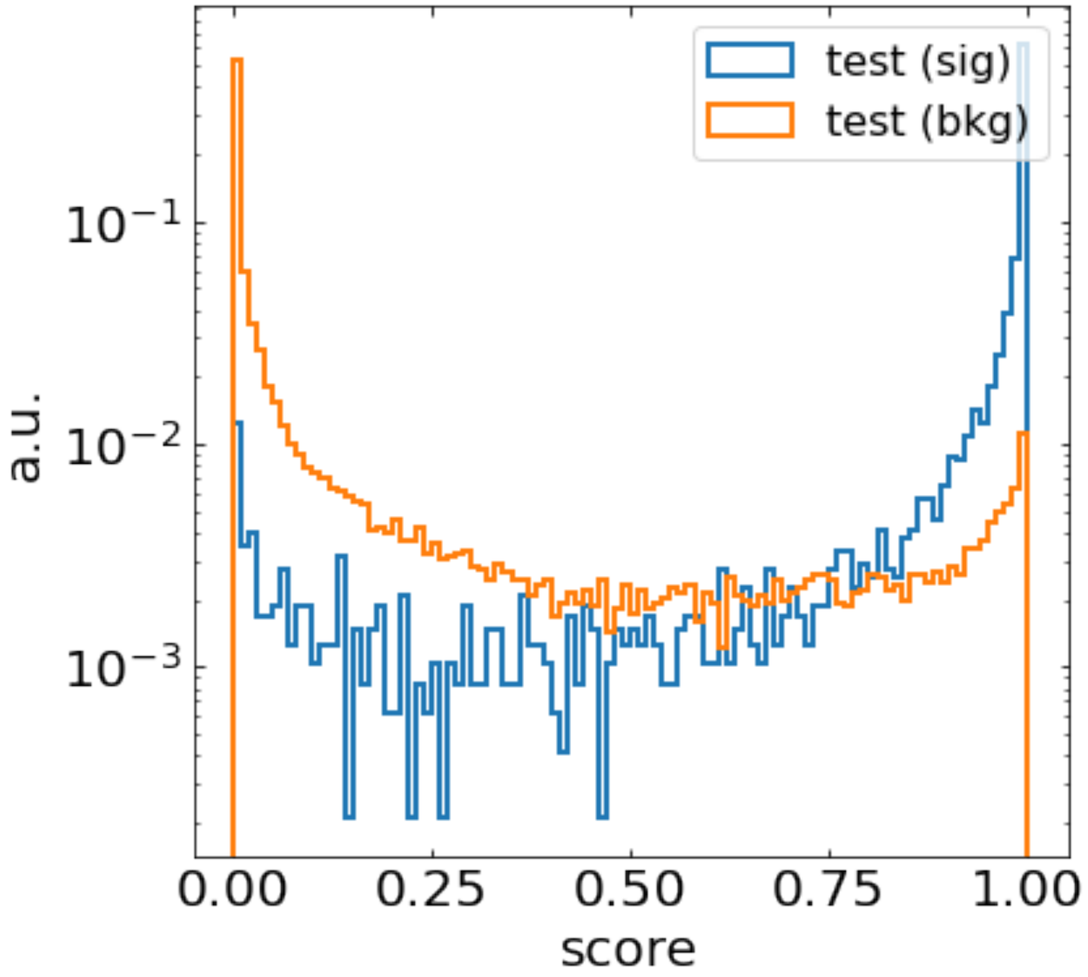


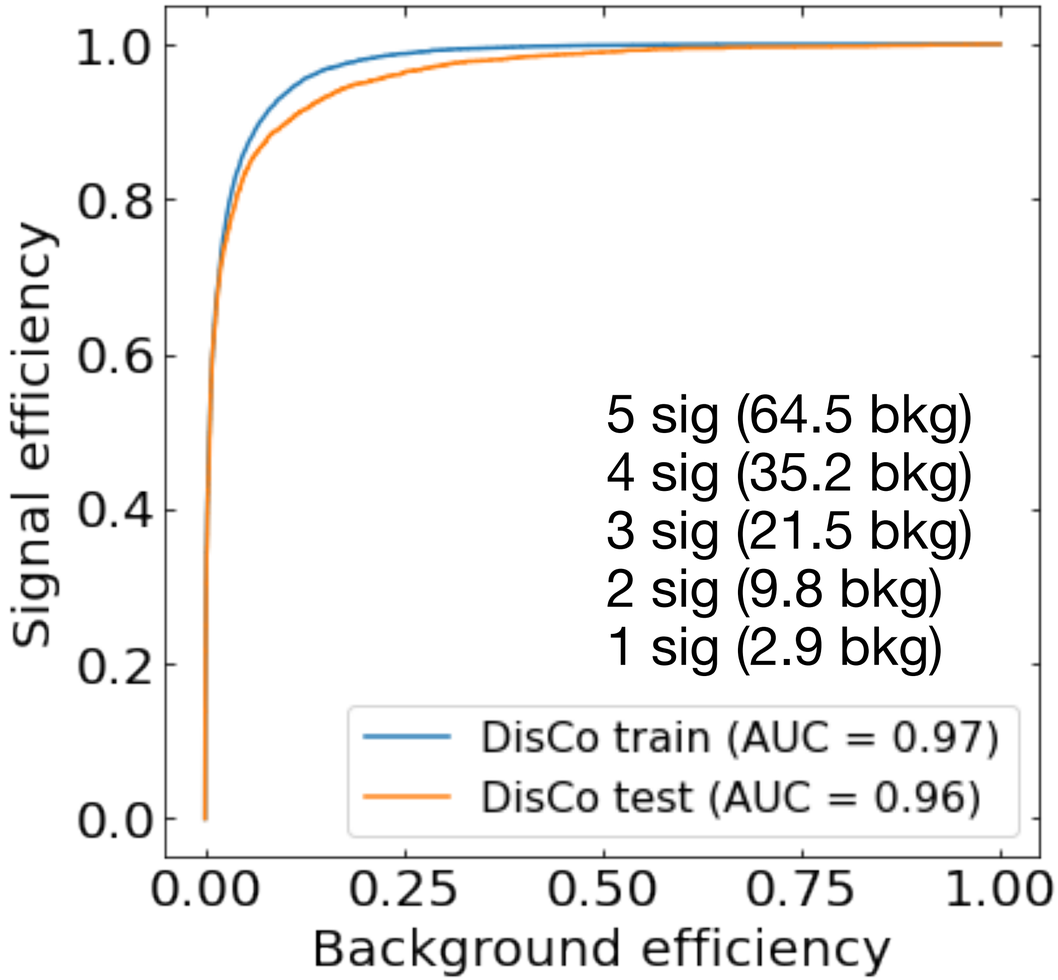
Epoch = 2500 | LR = 0.01 (constant) | λ = 100

UC San Diego

ABCDNet: λ = 200 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 200 \times \text{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$



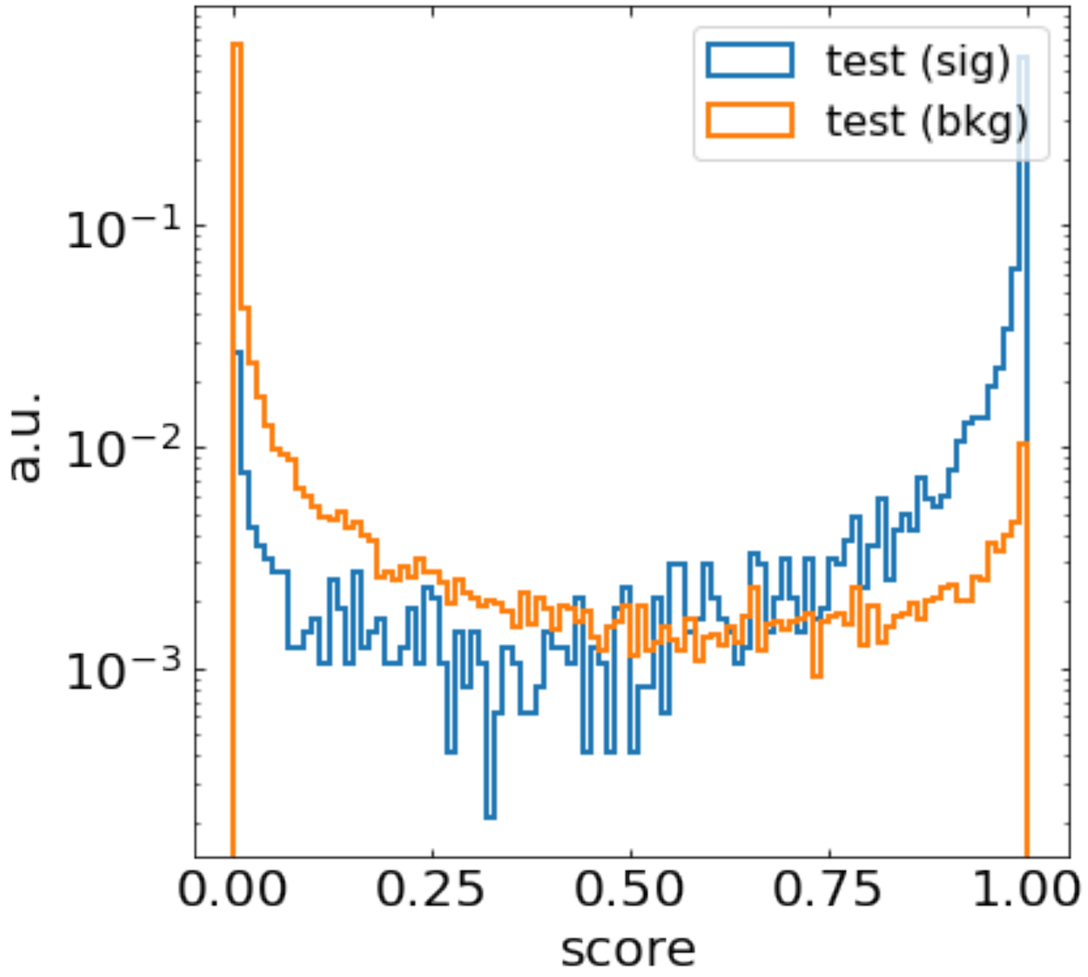


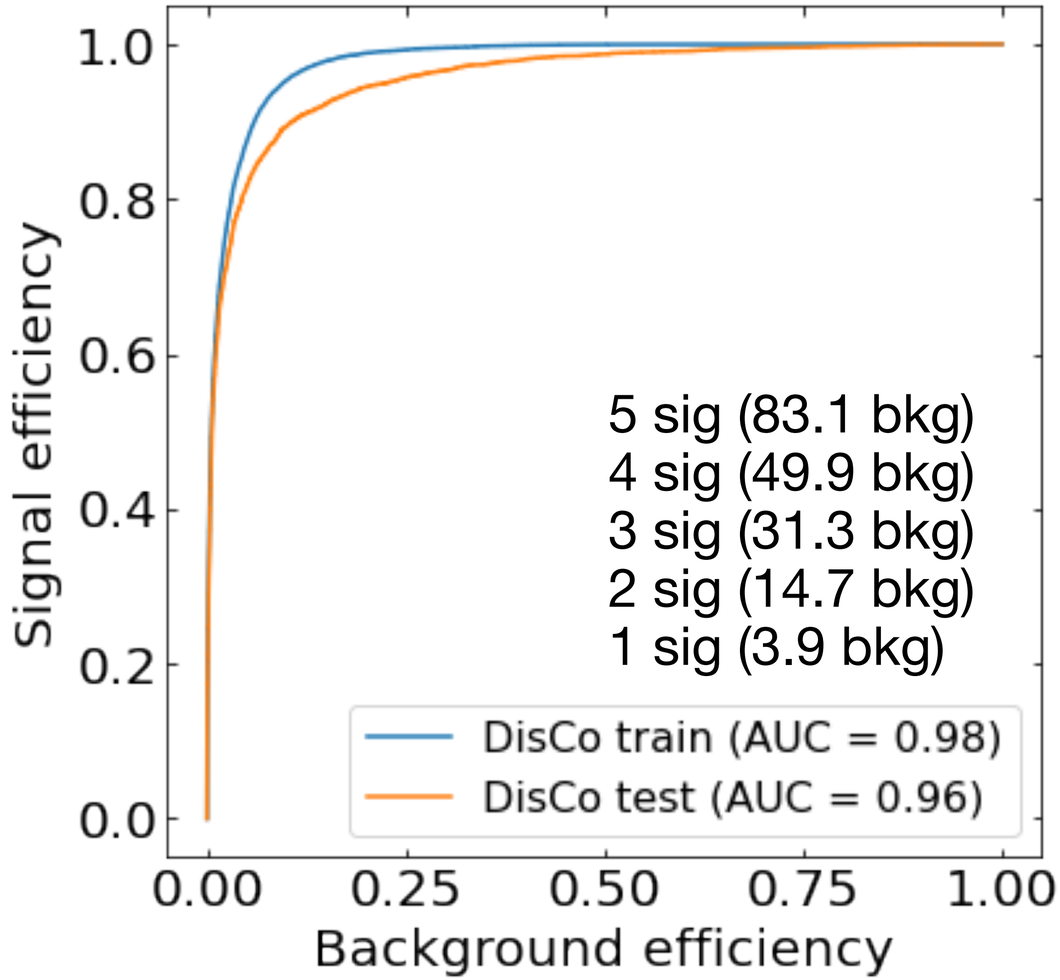
Epoch = 1000 | LR = 0.01 (constant) | λ = 100

UC San Diego

ABCDNet: λ = 200 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 200 \times \mathrm{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$



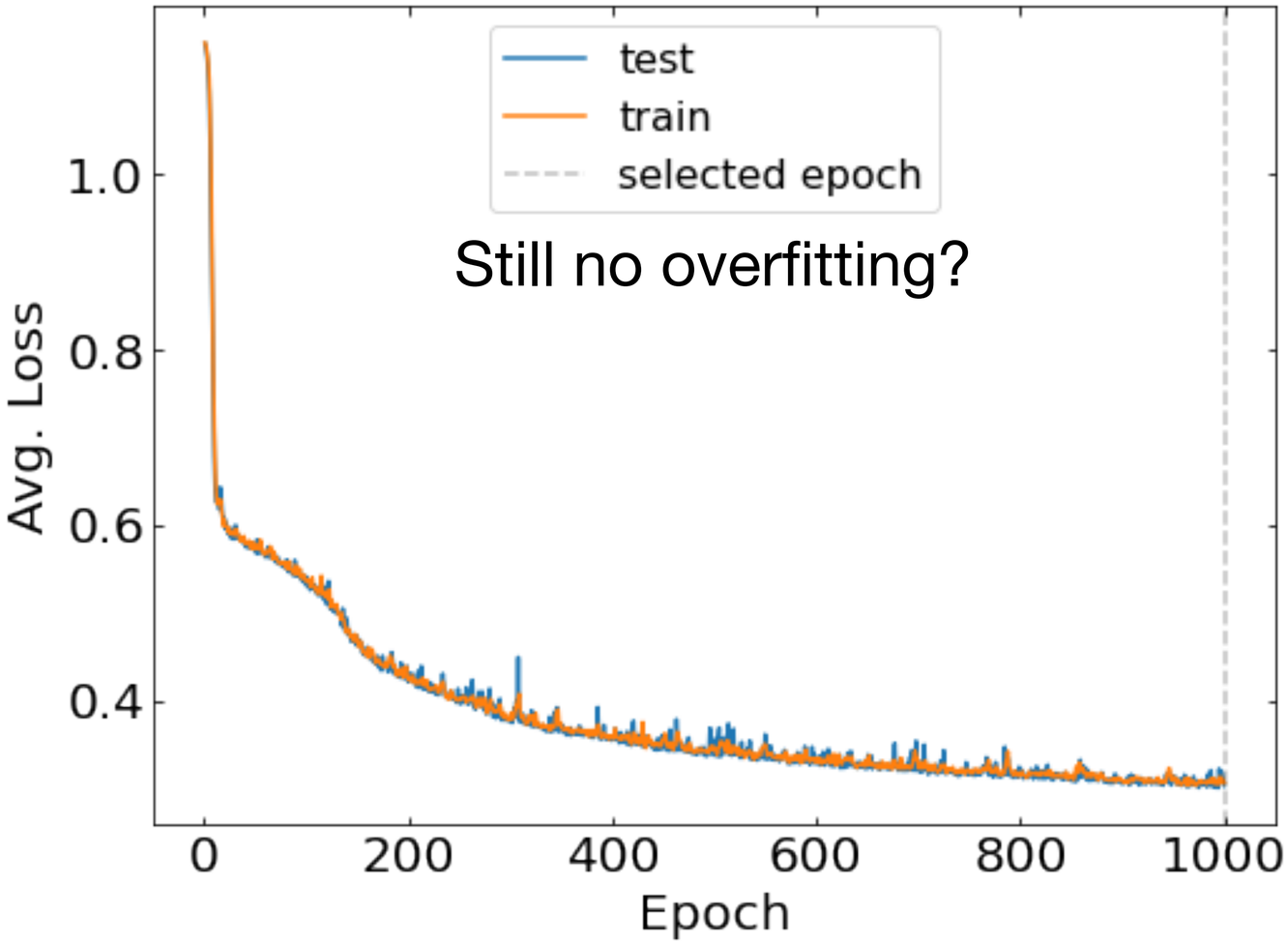


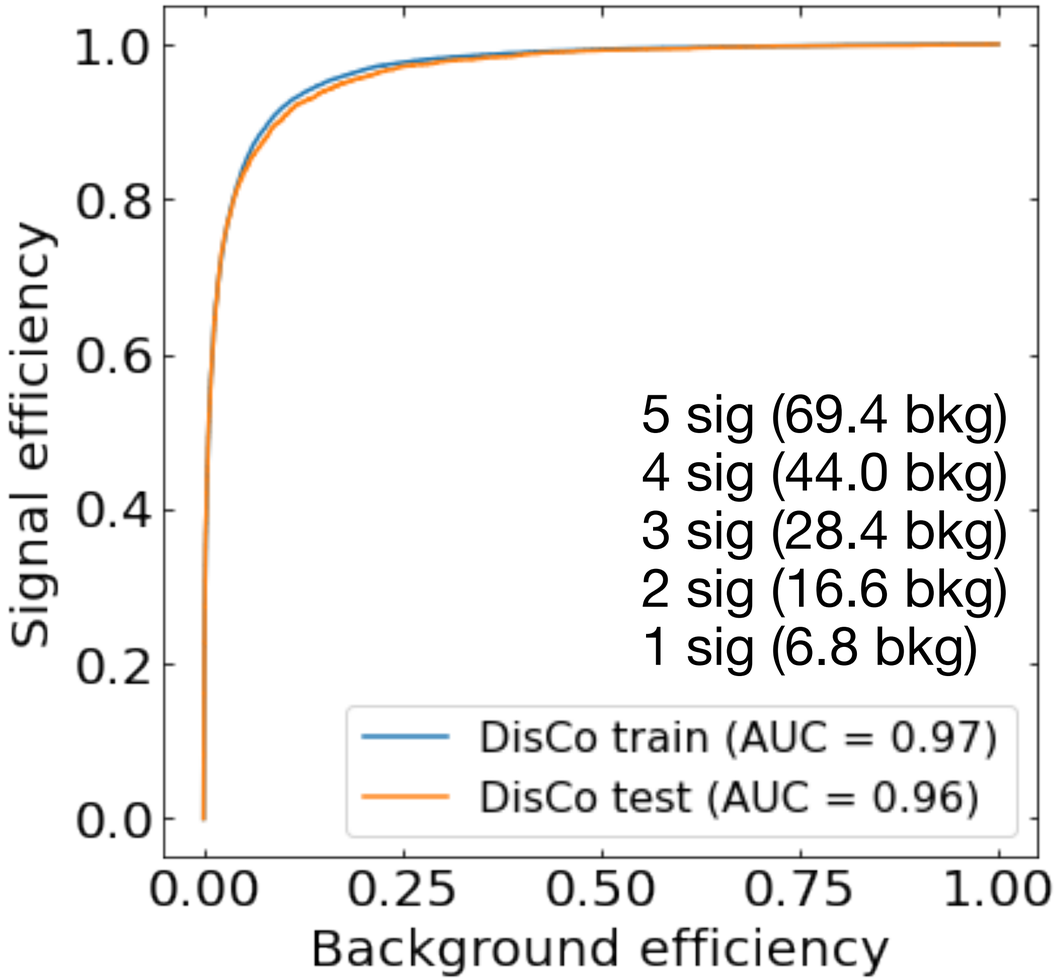
Epoch = 2500 | LR = 0.01 (constant) | λ = 100

UC San Diego

ABCDNet: λ = 100 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$



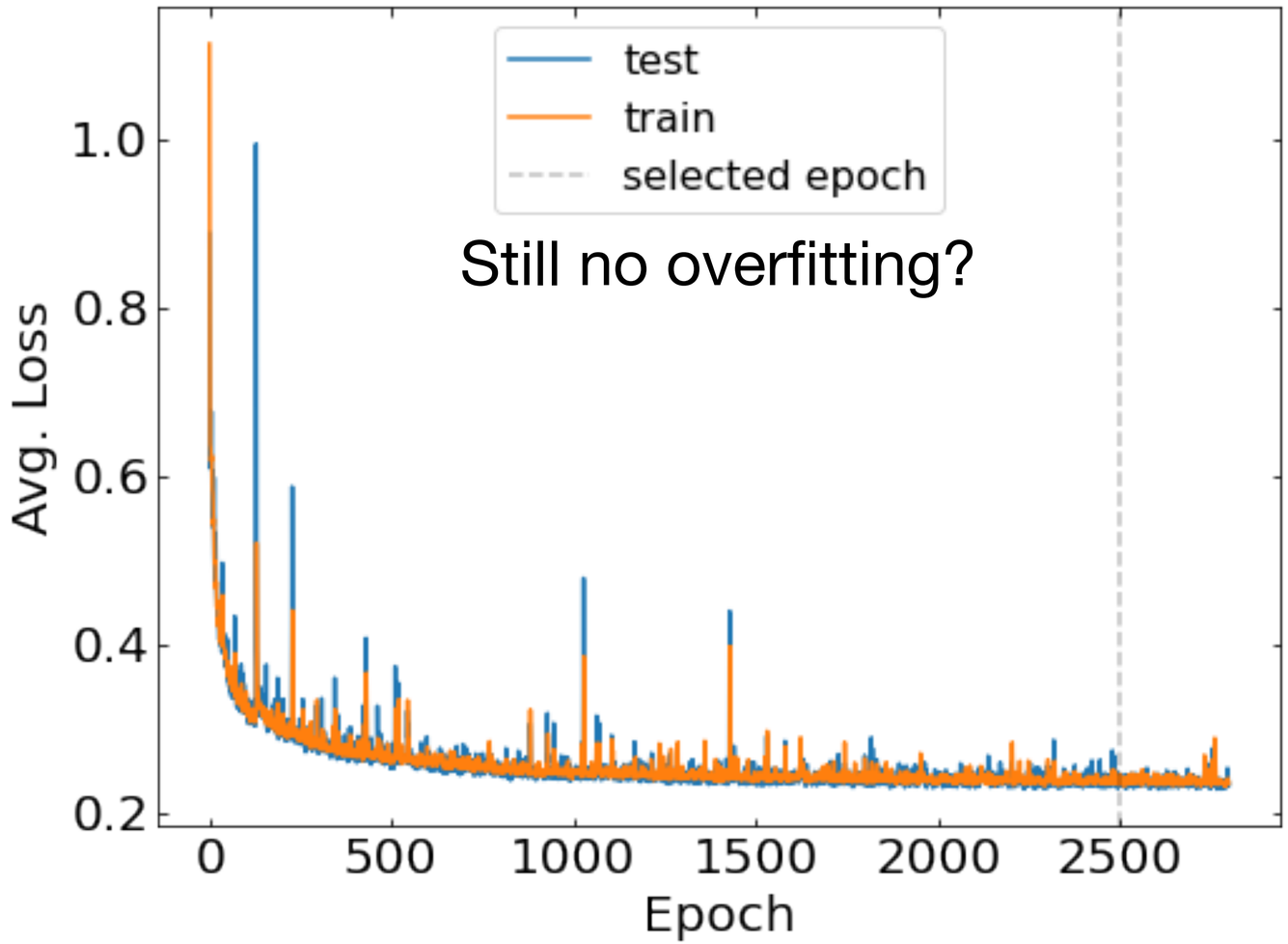


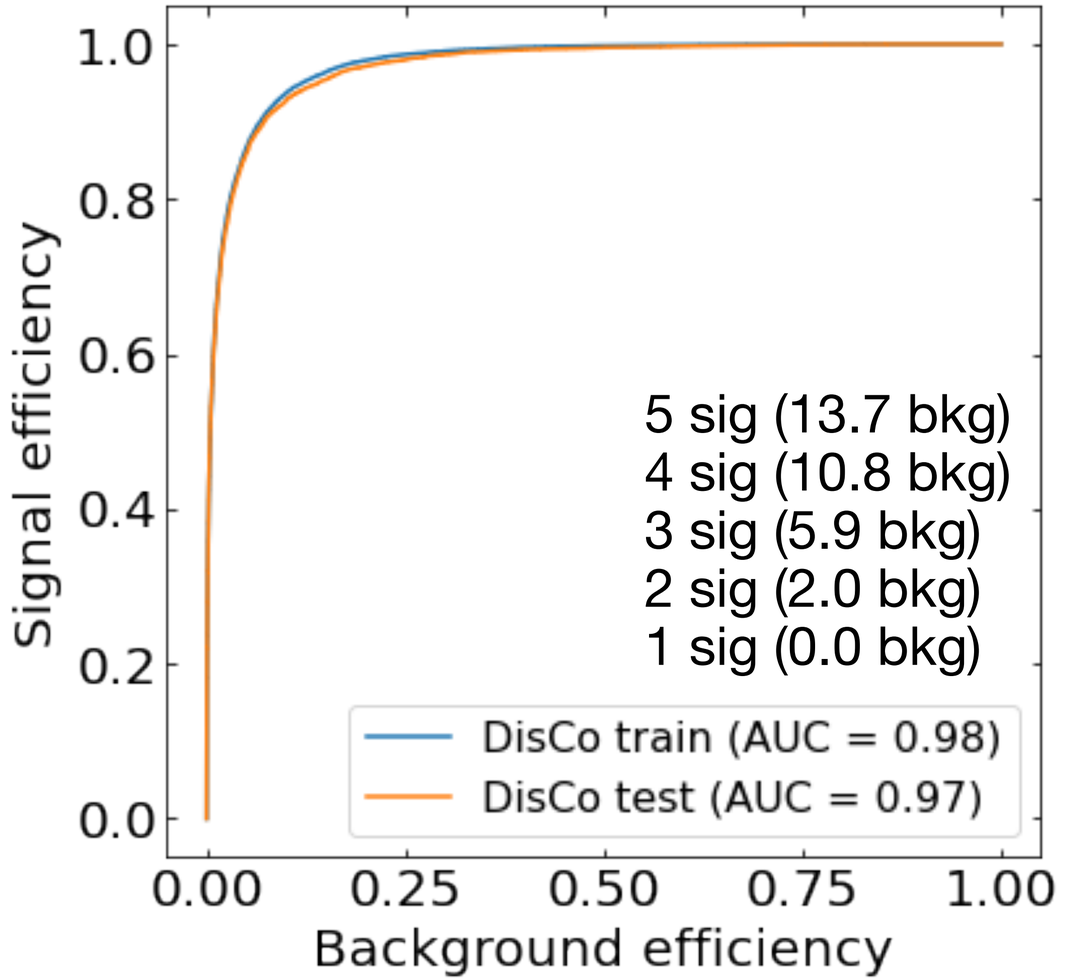
Epoch = 1000 | LR = 0.001 (constant) | λ = 100

UC San Diego

ABCDNet: λ = 100 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$



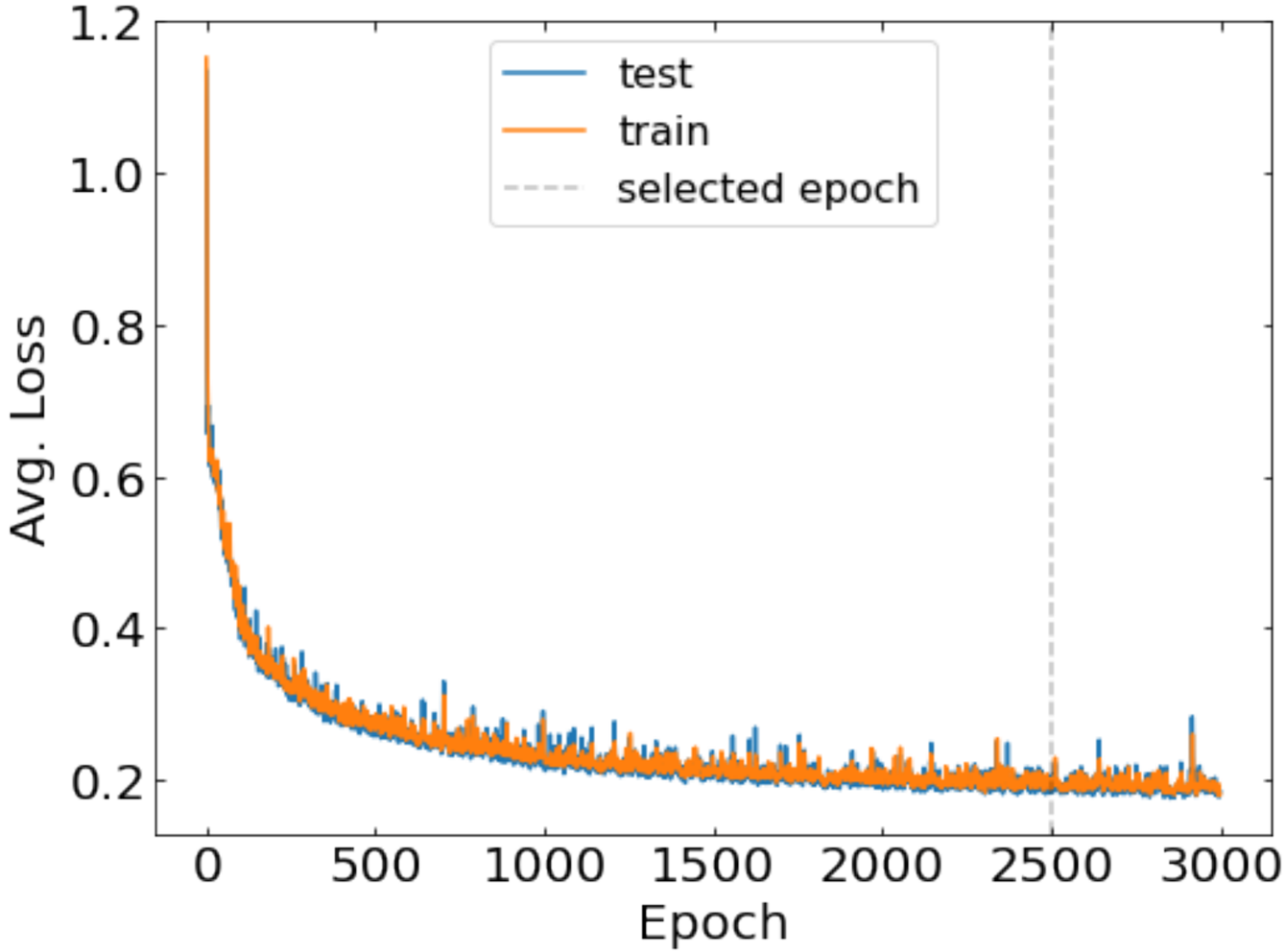


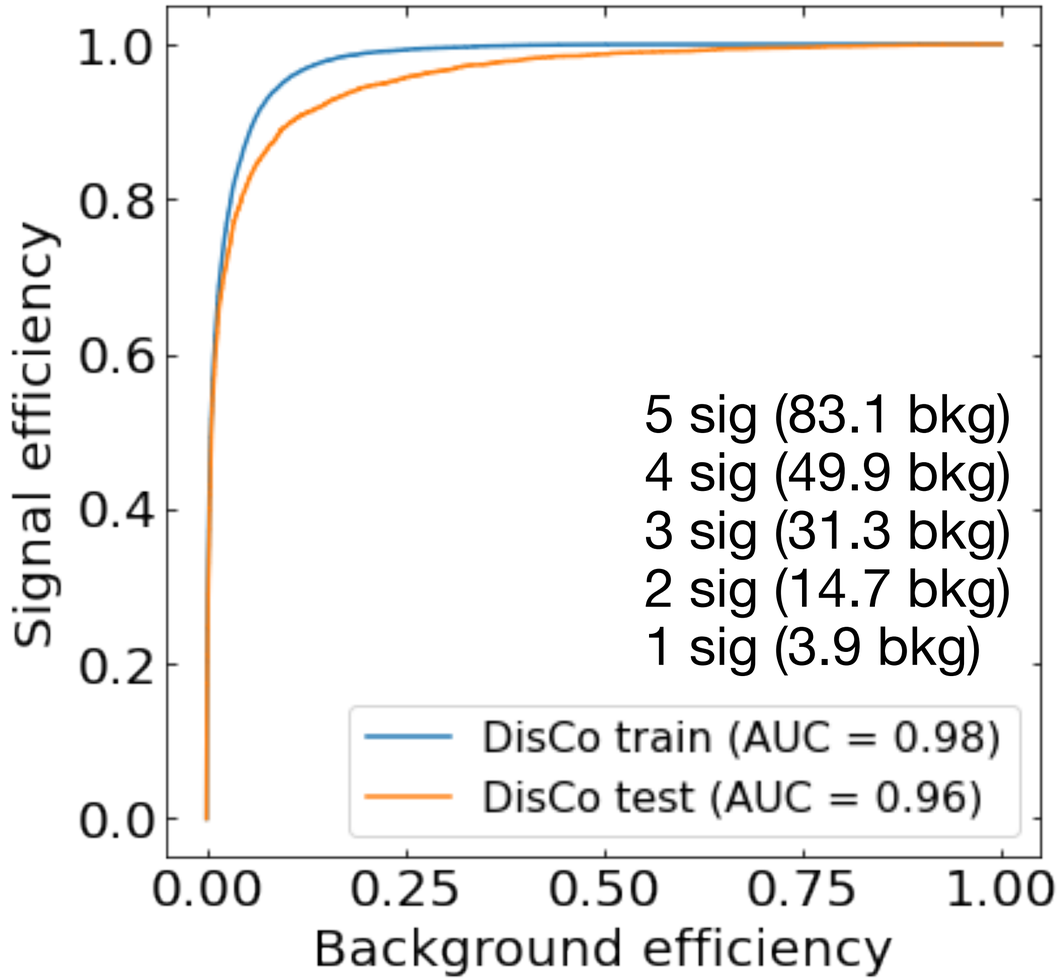
Epoch = 2500 | LR = 0.01 (constant) | λ = 100

UC San Diego

ABCDNet: λ = 200 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 200 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$



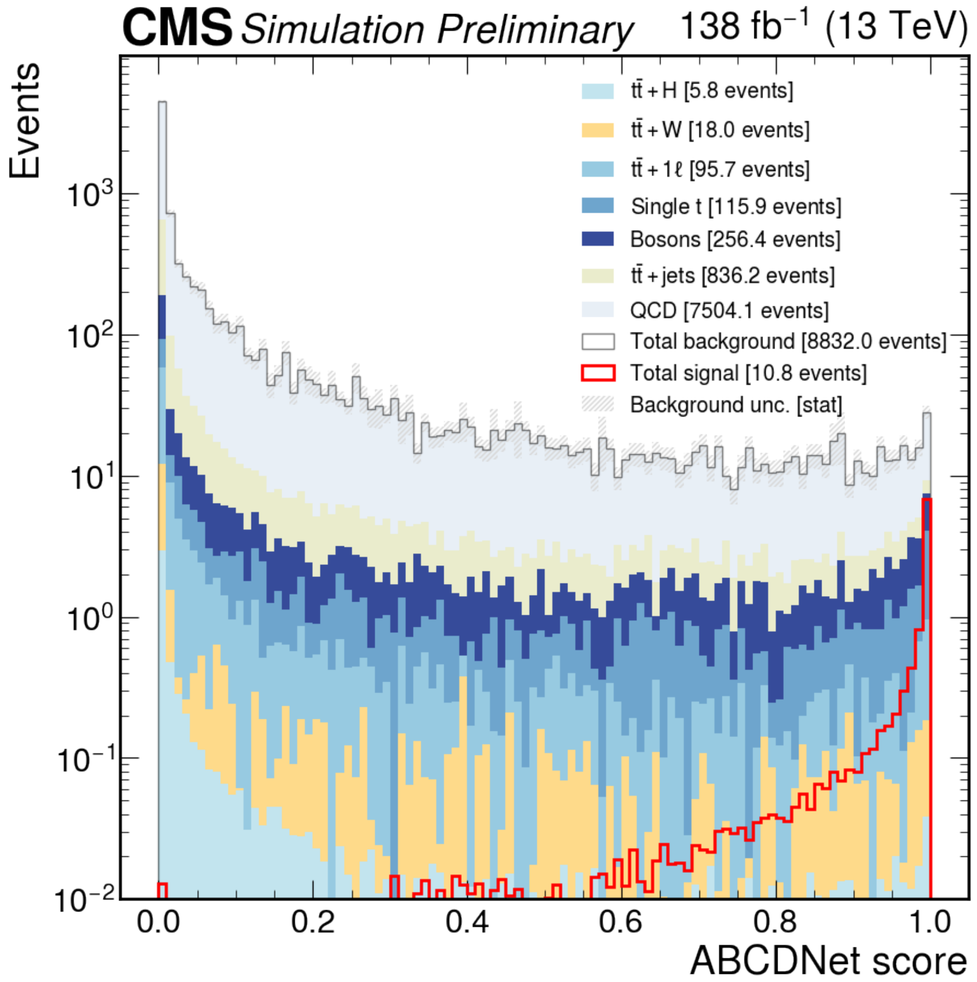


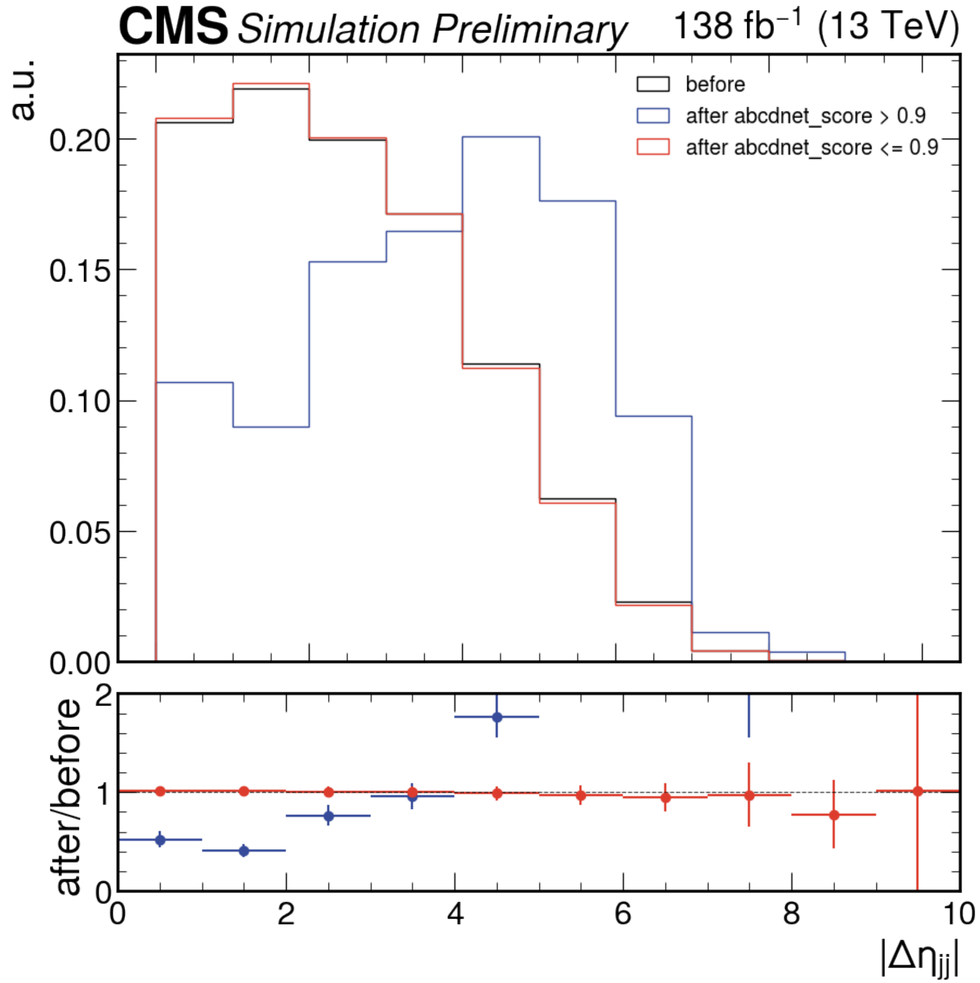
Epoch = 2500 | LR = 0.01 (constant) | λ = 200

UC San Diego

ABCDNet: λ = 100 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \mathrm{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$





Epoch = 2500 | LR = 0.01 (constant) | λ = 200

UC San Diego

ABCDNet: λ = 100 DisCo: ABCD

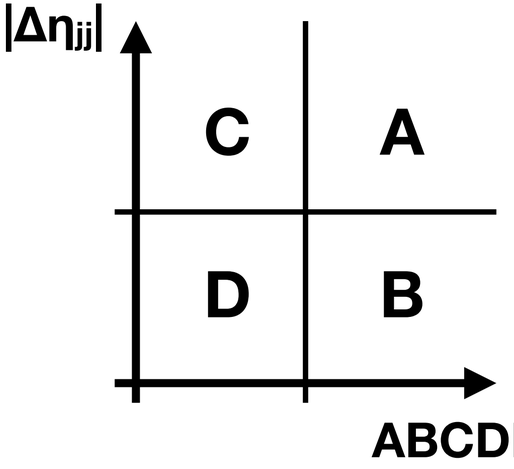
$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

Borrowed from BDT study

Xbb > 0.6 and XWqq > 0.6|0.65 (ld|tr)
and ST > 1300 and M_{jj} > 600 GeV and Hbb fatjet PNet mass < 150 GeV
and Vqq fatjet mass < 120|120 GeV (ld|tr)

Selection	Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
$\|\Delta\eta_{jj}\|$ > 4 and ABCDNet ≤ 0.99	C	81.18	13.32	1.18	0.03	63	7.94
$\|\Delta\eta_{jj}\|$ ≤ 4 and ABCDNet ≤ 0.99	D	35.96	4.61	0.06	0.01	27	5.20
$\|\Delta\eta_{jj}\|$ ≤ 4 and ABCDNet > 0.99	B	2.46	0.82	0.06	0.01	6	2.45
$\|\Delta\eta_{jj}\|$ > 4 and ABCDNet > 0.99	A	**3.97**	0.81	4.96	0.06	—	—

NOT an optimal SR, but ABCD closure is better than before



$A_{MC} = B_{MC} \times C_{MC} / D_{MC} =$ **5.56**

Epoch = 2500 | LR = 0.01 (constant) | λ = 100

UC San Diego

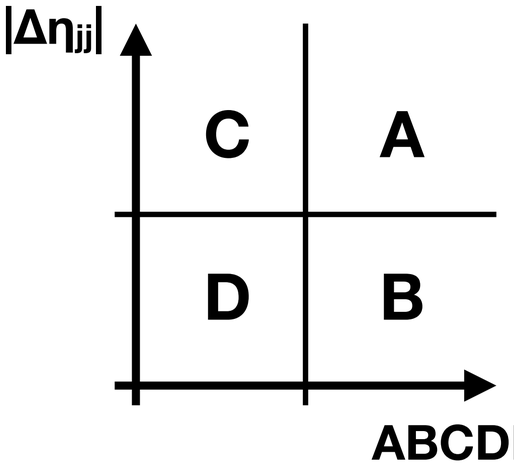
ABCDNet: λ = 100 DisCo: ABCD

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 100 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

Xbb > 0.3 and XWqq > 0.5|0.5 (ld|tr)

Selection	Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
$\|\Delta\eta_{jj}\| > 3.5$ and ABCDNet ≤ 0.99	C	229.14	25.70	1.44	0.03	216	14.70
$\|\Delta\eta_{jj}\| \leq 3.5$ and ABCDNet ≤ 0.99	D	531.40	33.12	0.68	0.02	544	23.32
$\|\Delta\eta_{jj}\| \leq 3.5$ and ABCDNet > 0.99	B	0.09	0.05	0.17	0.01	—	—
$\|\Delta\eta_{jj}\| > 3.5$ and ABCDNet > 0.99	A	**1.46**	0.34	4.49	0.06	—	—

Optimal SR, but ABCD closure is very bad



$$A_{MC} = B_{MC} \times C_{MC} / D_{MC} = \mathbf{0.04}$$

Epoch = 2500 | LR = 0.01 (constant) | λ = 100

UC San Diego

Summary

- Still can’t get model to overfit (is this not an issue?)
- Performant models still do not seem to be giving good ABCD closure

Backup

Sanity Check

- **Goal:** repeat the first example in the PRL paper (3D gaussian variables)
- **(1)** and **(2)** define the 3D gaussians
- **(3)** and **(4)** give the rest:
  - Input: X_1, X_2 (DisCo target: X_3)
  - NN architecture: 3 hidden layers; 128 nodes per layer; ReLU between layers; sigmoid output
  - λ = 1000, Adam optimizer
  - 2M sig, 2M bkg

IV. APPLICATIONS

This section explores the efficacy of single and double DisCo in some applications of the ABCD method.

A. Simple example: Three-dimensional Gaussian random variables

We begin with a simple example to build some intuition and validate our methods. Consider a three-dimensional space (X_0, X_1, X_2), where the signal and background are both multivariate Gaussian distributions. We choose the means $\vec{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ for background and signal as

1

$$\vec{\mu}_b = (0,0,0), \qquad \boldsymbol{\Sigma}_b = \sigma_b^2 \begin{pmatrix} 1 & \rho_b & 0 \\ \rho_b & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\sigma_b = 1.5, \qquad \rho_b = -0.8, \tag{4.1}$$

and

2

$$\vec{\mu}_s = (2.5, 2.5, 2), \qquad \boldsymbol{\Sigma}_s = \sigma_s^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \sigma_s = 1.5. \tag{4.2}$$

3

So for the background, all three features are centered at the origin and features X_0 and X_1 are correlated with each other but independent of X_2. For the signal, all three features are independent but are centered away from the origin. The first feature X_0 will play the role of the known feature for single DisCo in Sec. III.

All of the neural networks presented in this section use three hidden layers with 128 nodes per layer. The rectified linear unit (ReLU) activation function is used for the intermediate layers and the output is a sigmoid function. A hyperparameter of $\lambda = 1000$ is used for both single and double DisCo to ensure total decorrelation. The single DisCo training converged after 100 epochs while the double DisCo training required 200 epochs. Other networks only needed ten epochs. The double DisCo networks

4

were trained using a single neural network with a two-dimensional output. All models were trained using Tensorflow [89] through Keras [90] with Adam [91] for optimization. Two million examples were generated with 15% used for testing. A batch size of 1% of the total was used for all networks to ensure an accurate calculation of the DisCo term in the relevant loss functions.

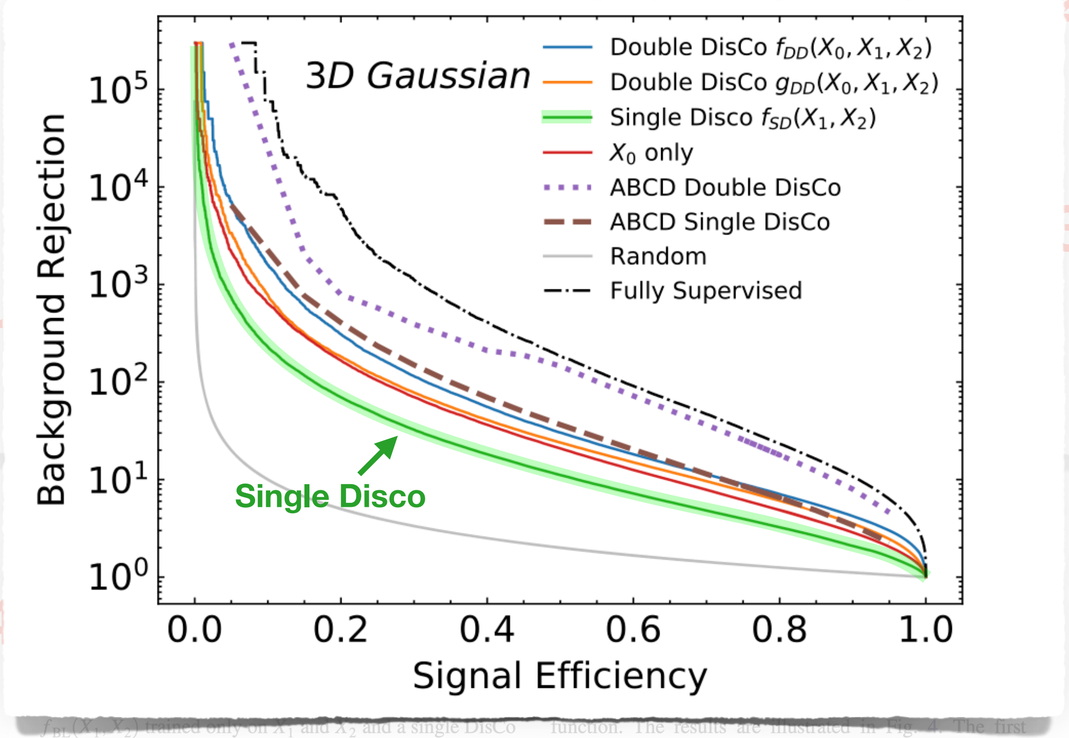
We first consider two classifiers: a baseline classifier $f_{\mathrm{BL}}(X_1, X_2)$ trained only on X_1 and X_2 and a single DisCo classifier $f_{\mathrm{SD}}(X_1, X_2)$ which includes a penalty for correlations between f_{SD} and X_0. The values of these classifiers for events drawn from the distributions are plotted in Fig. 3 against the X_0, X_1, or X_2 values of these events. We see that even though X_0 was not used in the training of the baseline, the classifier output is still correlated with X_0 because of the correlations between X_0 and X_1. In contrast to the baseline classifier, the single DisCo classifier is independent of both X_0 and X_1 and is simply a function of X_2. Intuitively, it makes sense that a classifier that must be independent of X_0 must also be independent of X_1. This is justified rigorously in Appendix B.

For double DisCo, we train two classifiers $f_{\mathrm{DD}}(X, Y, Z)$ and $g_{\mathrm{DD}}(X, Y, Z)$ according to the double DisCo loss function. The results are illustrated in Fig. 4. The first classifier depends mostly on Z and the second classifier depends mostly on X and Y. However, the residual dependence on all three observables is not a deficit of the training procedure: even though the three random variables are separable into two independent subsets (X, Y) and Z, the two classifiers learned by double DisCo

Sanity Check

- **Goal:** repeat the first example in the PRL paper (3D gaussian variables)
- **(1)** and **(2)** define the 3D gaussians
- **(3)** and **(4)** give the rest:
  - Input: X_1, X_2 (DisCo target: X_3)
  - NN architecture: 3 hidden layers; 128 nodes per layer; ReLU between layers; sigmoid output
  - λ = 1000, Adam optimizer
  - 2M sig, 2M bkg

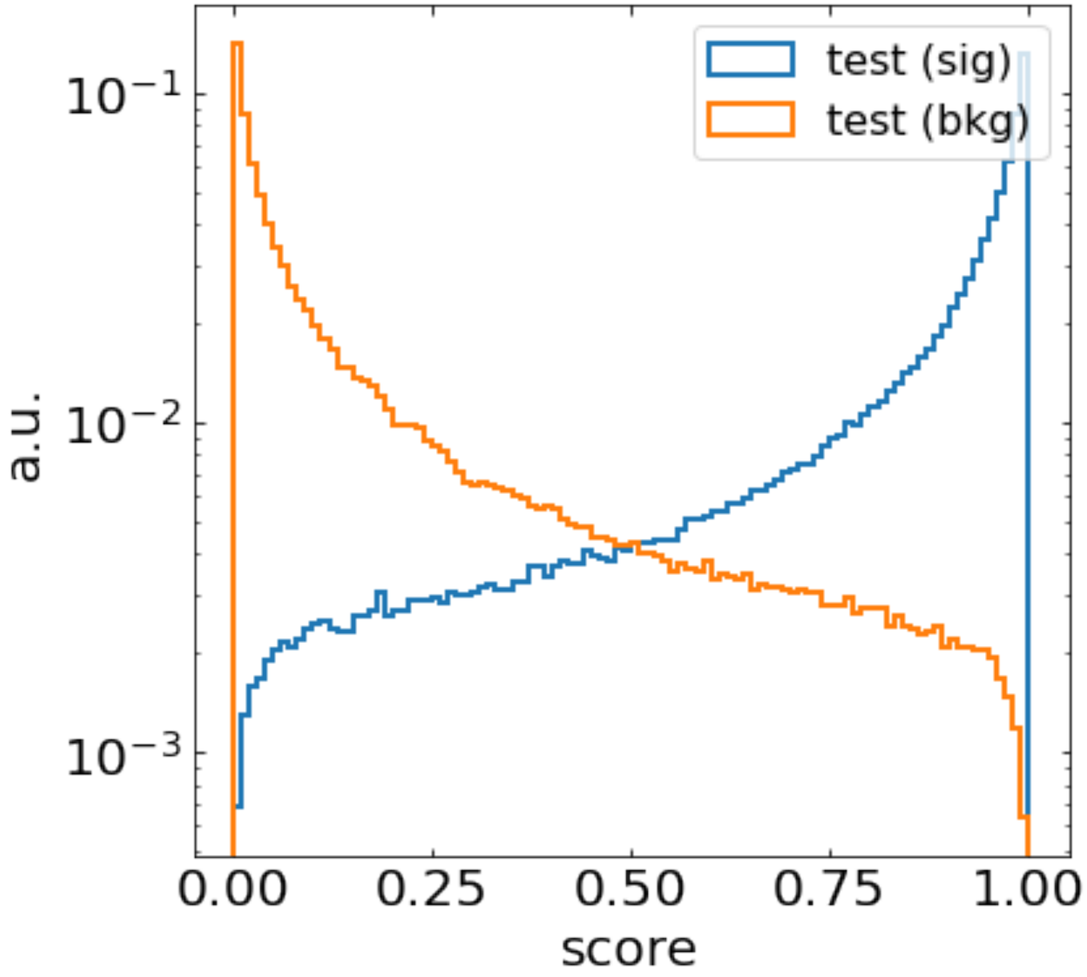
Target: recreate this ROC curve!

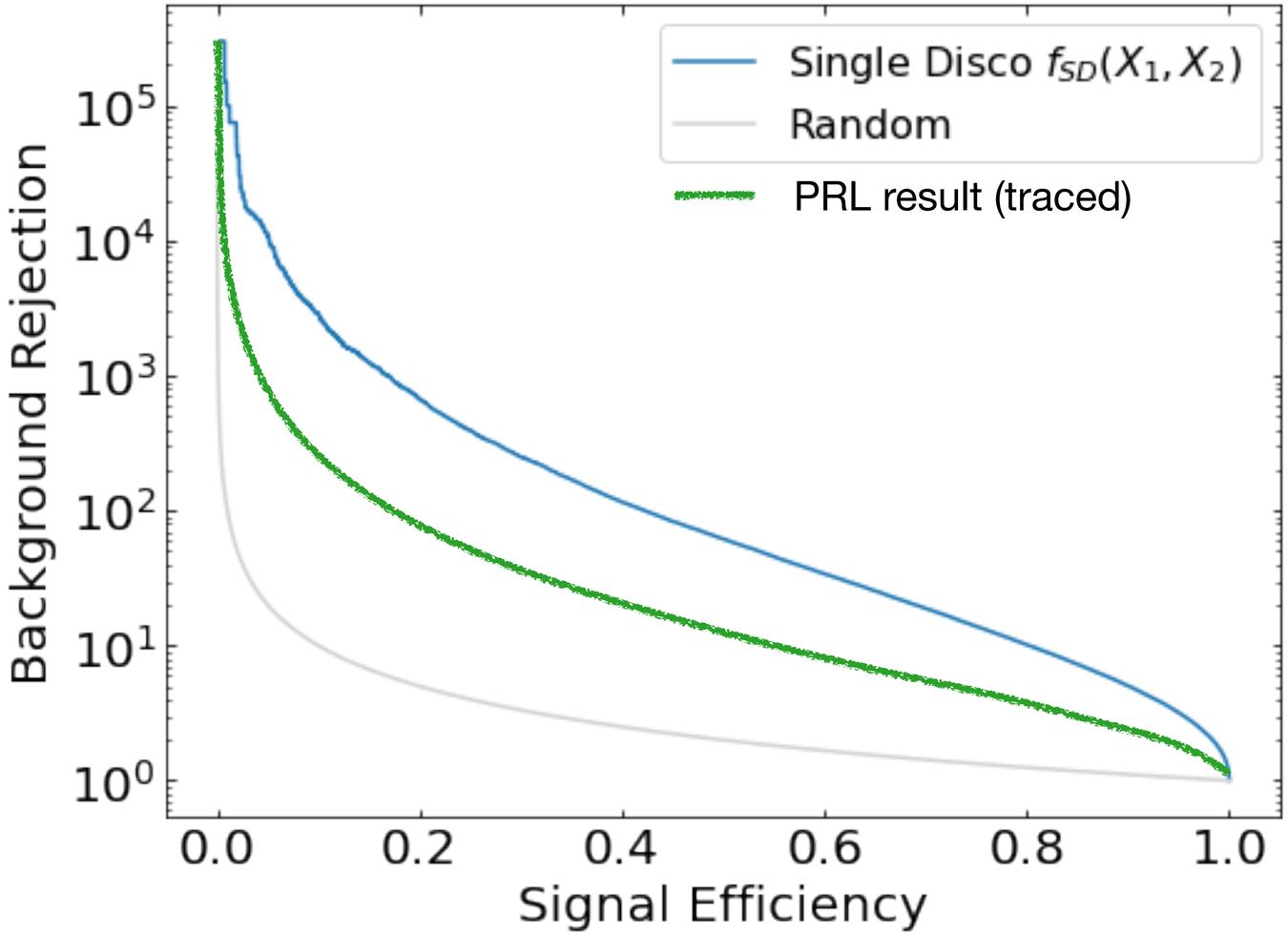


TPR = TP/P = (true positives)/(positives)
FPR = FP/N = (false positives)/(negatives)

Sanity Check: No DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 0 \times \text{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$

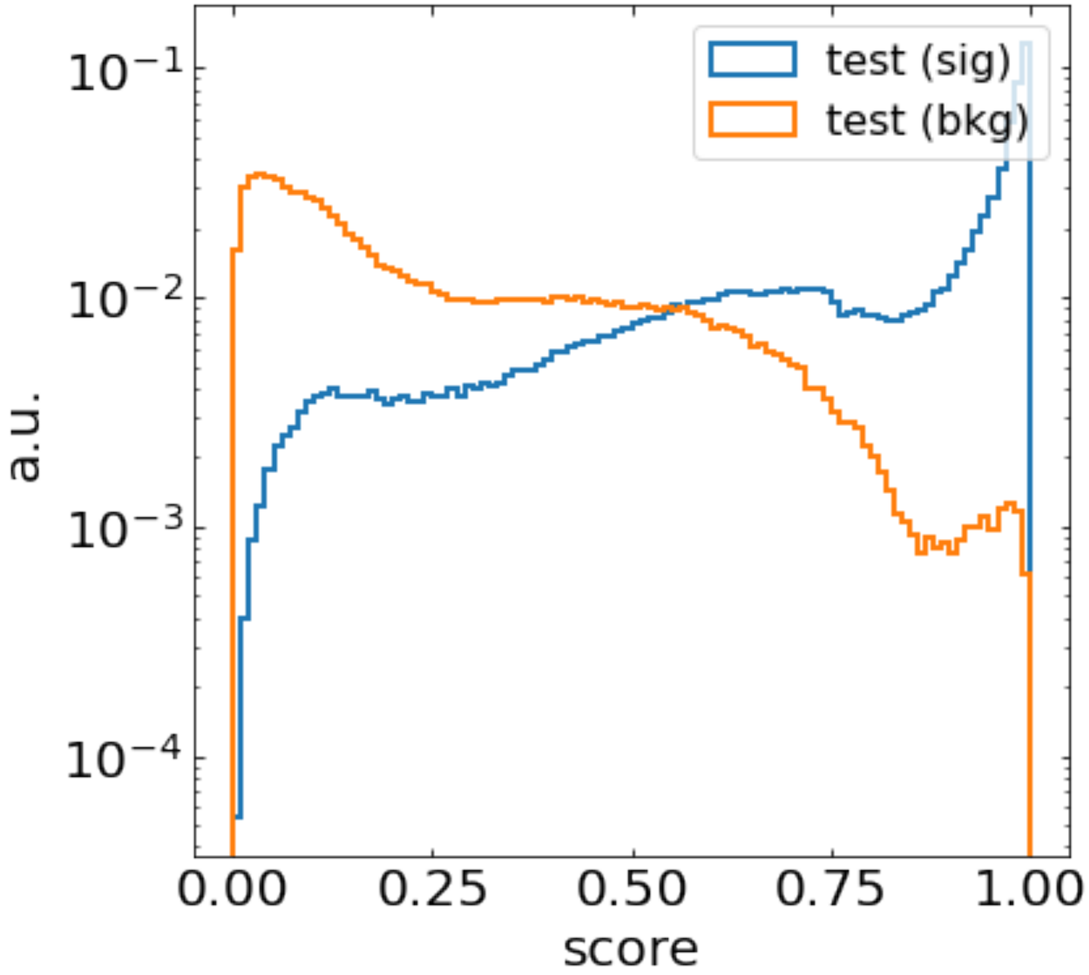


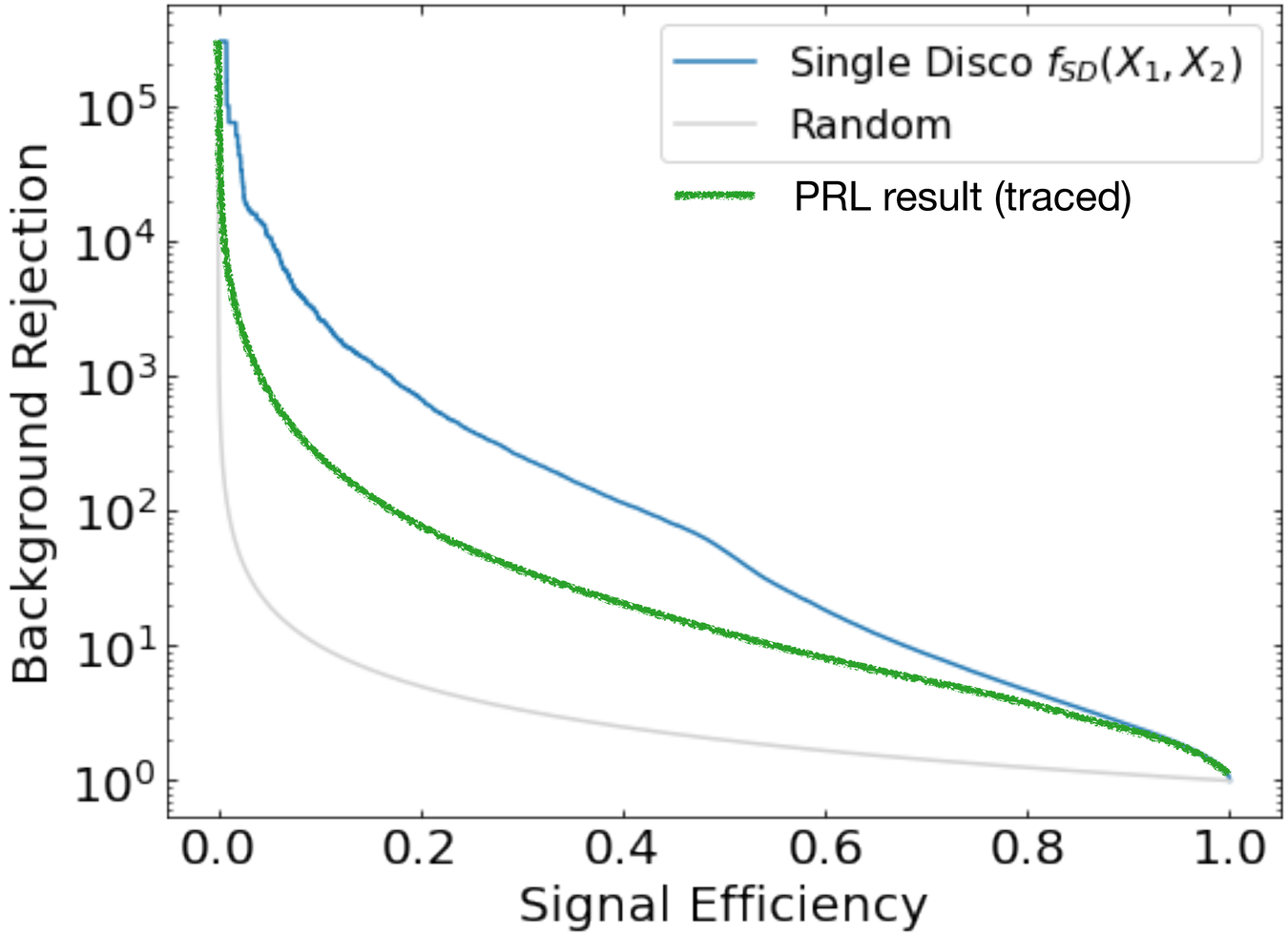


UC San Diego

Sanity Check: PRL DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \text{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$

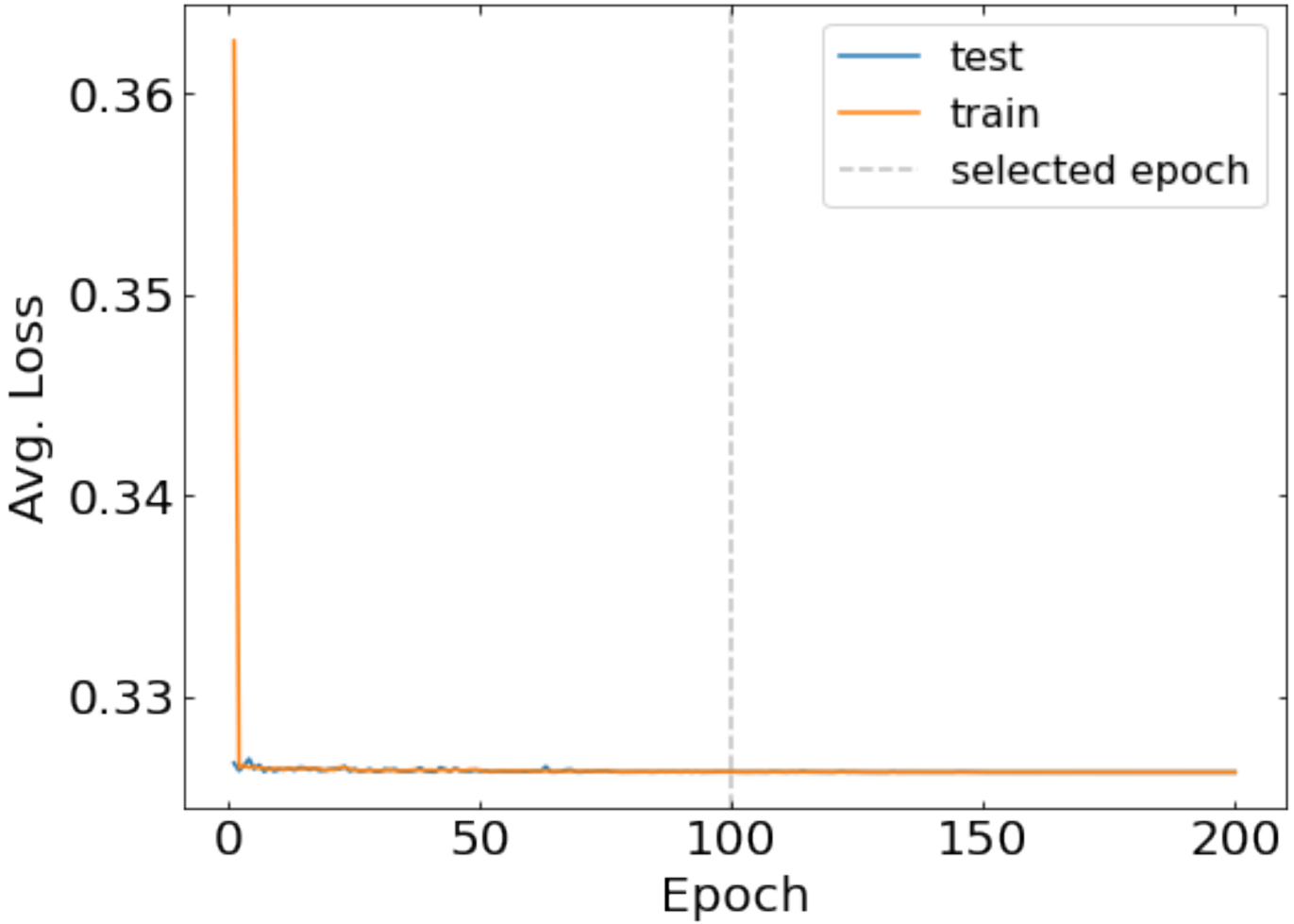


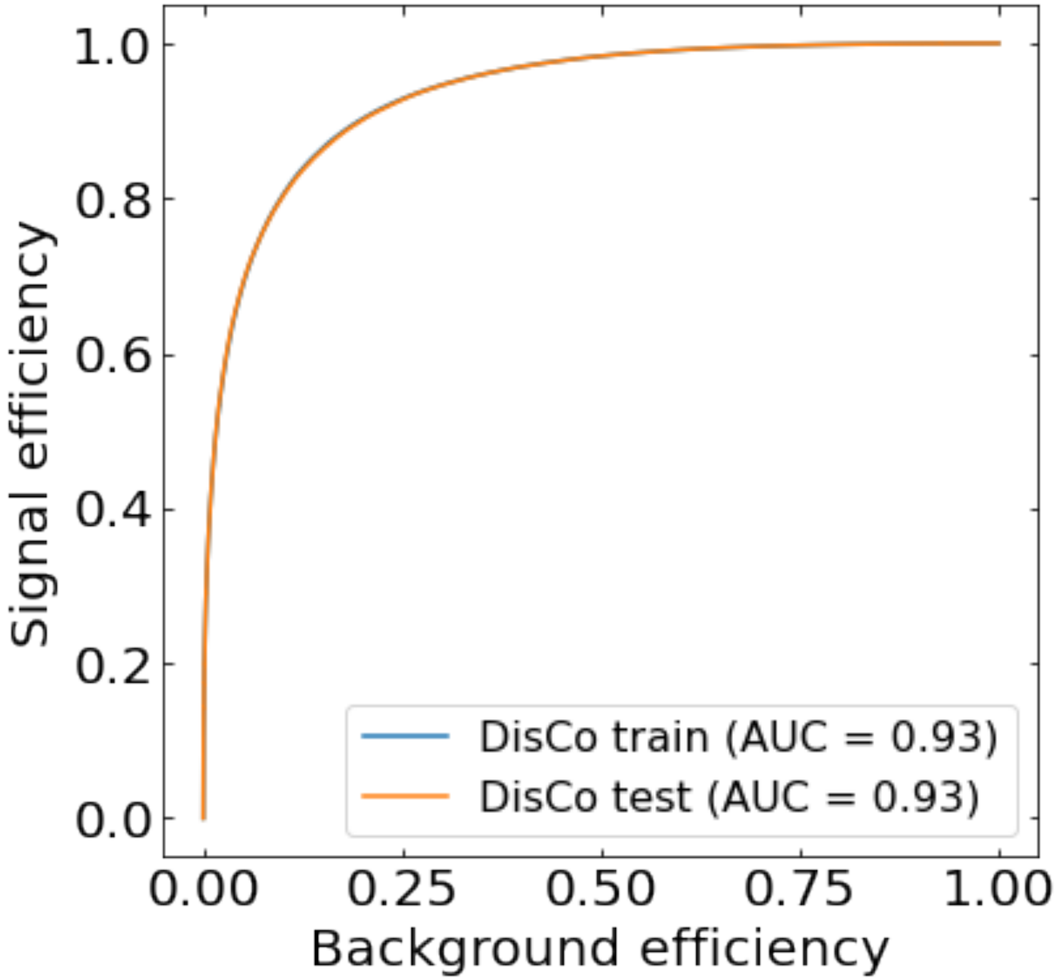


UC San Diego

Sanity Check: No DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 0 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

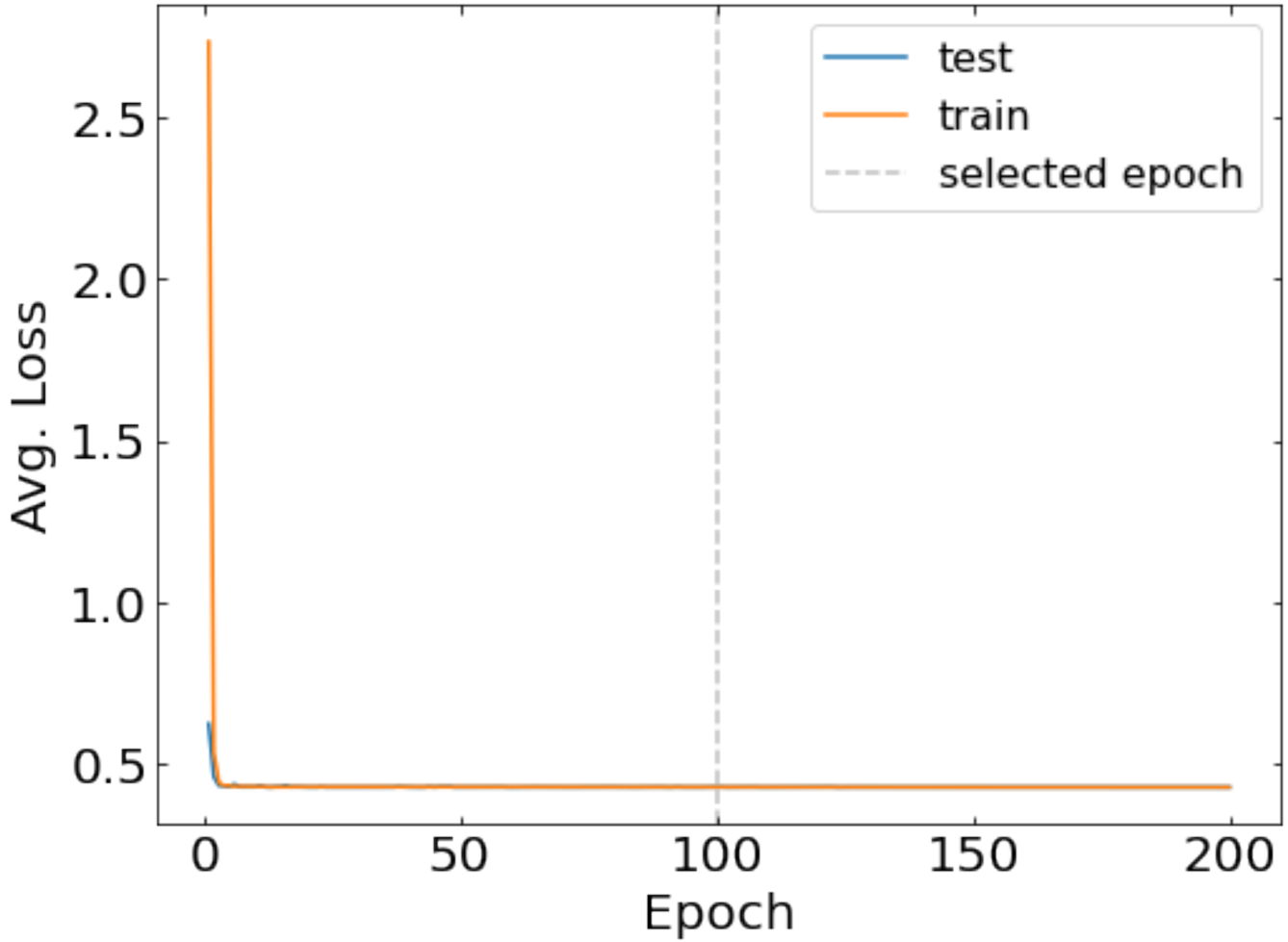


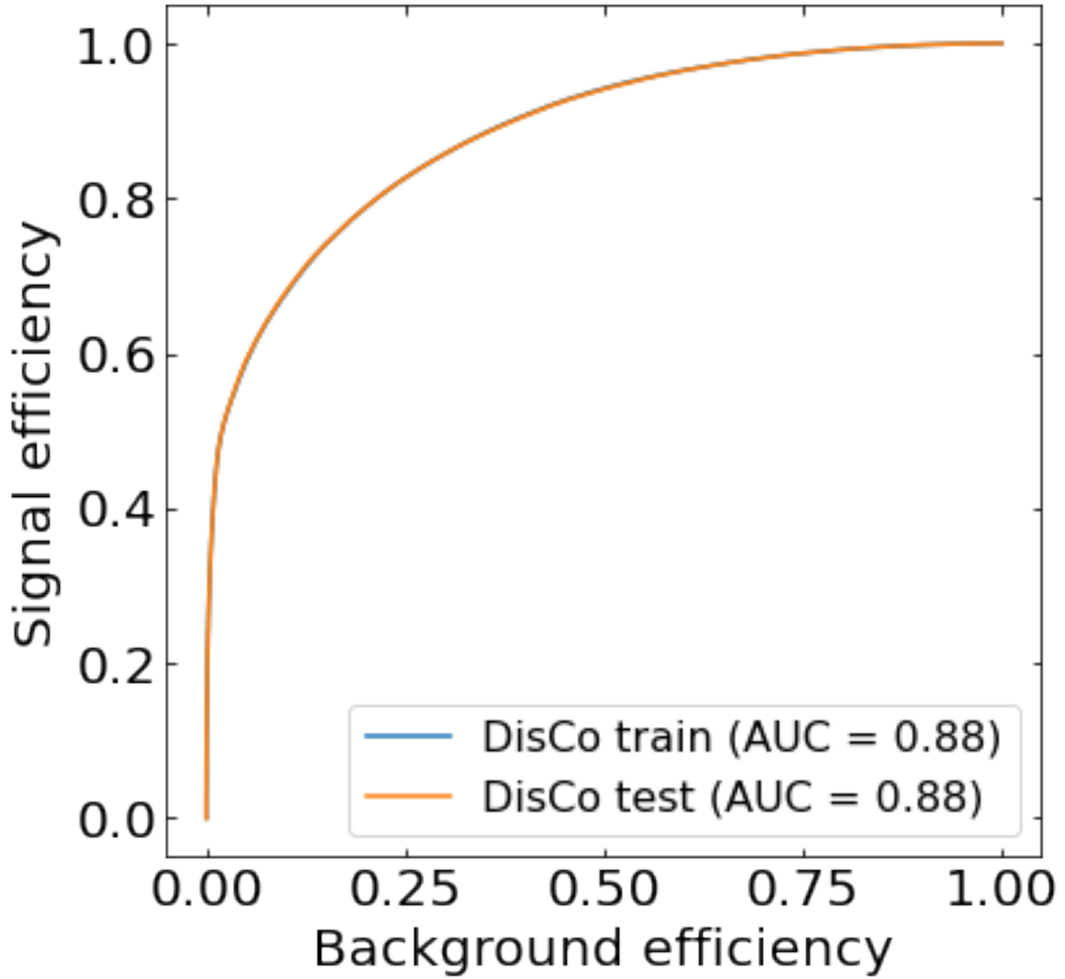


UC San Diego

Sanity Check: PRL DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

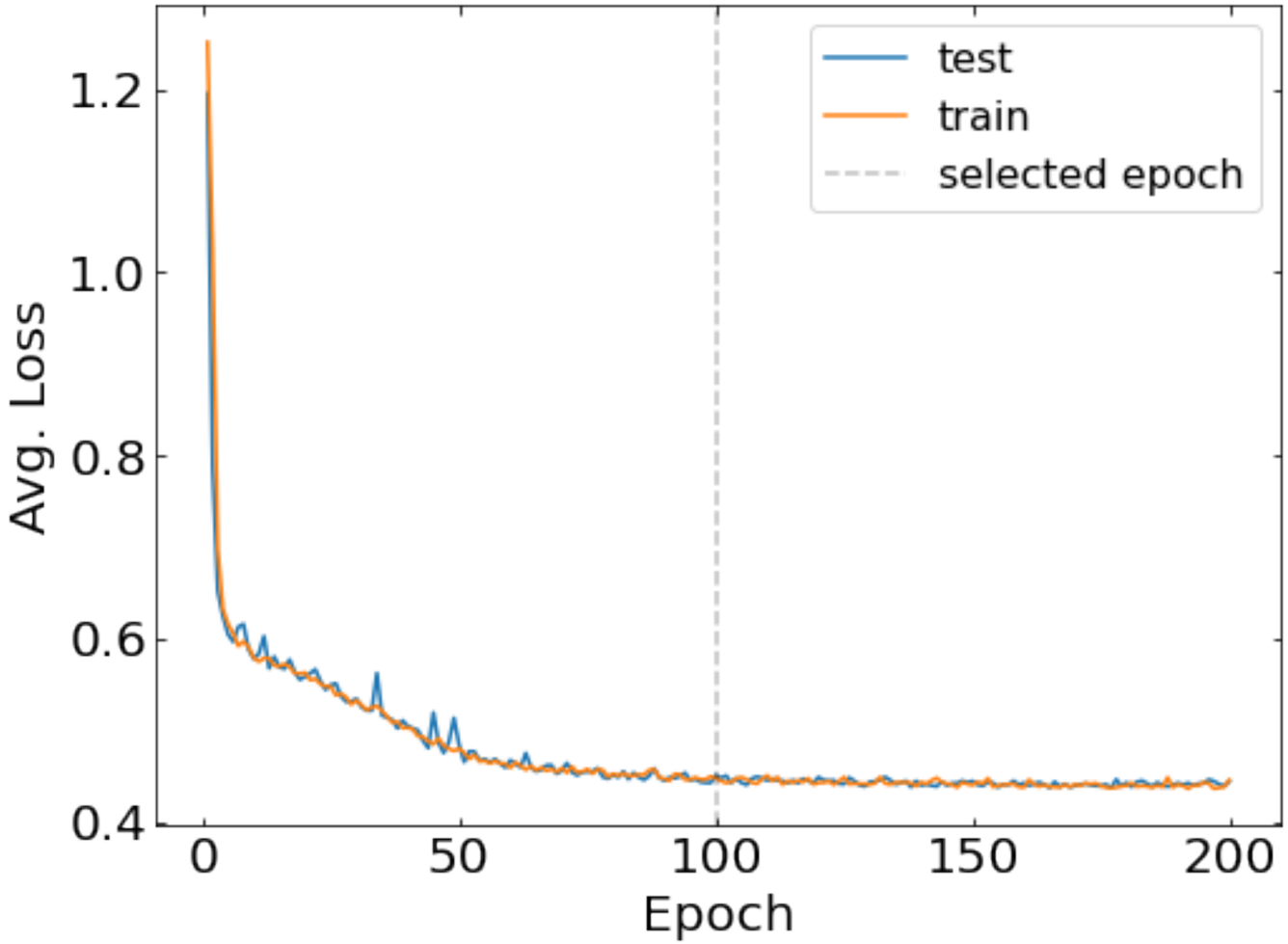


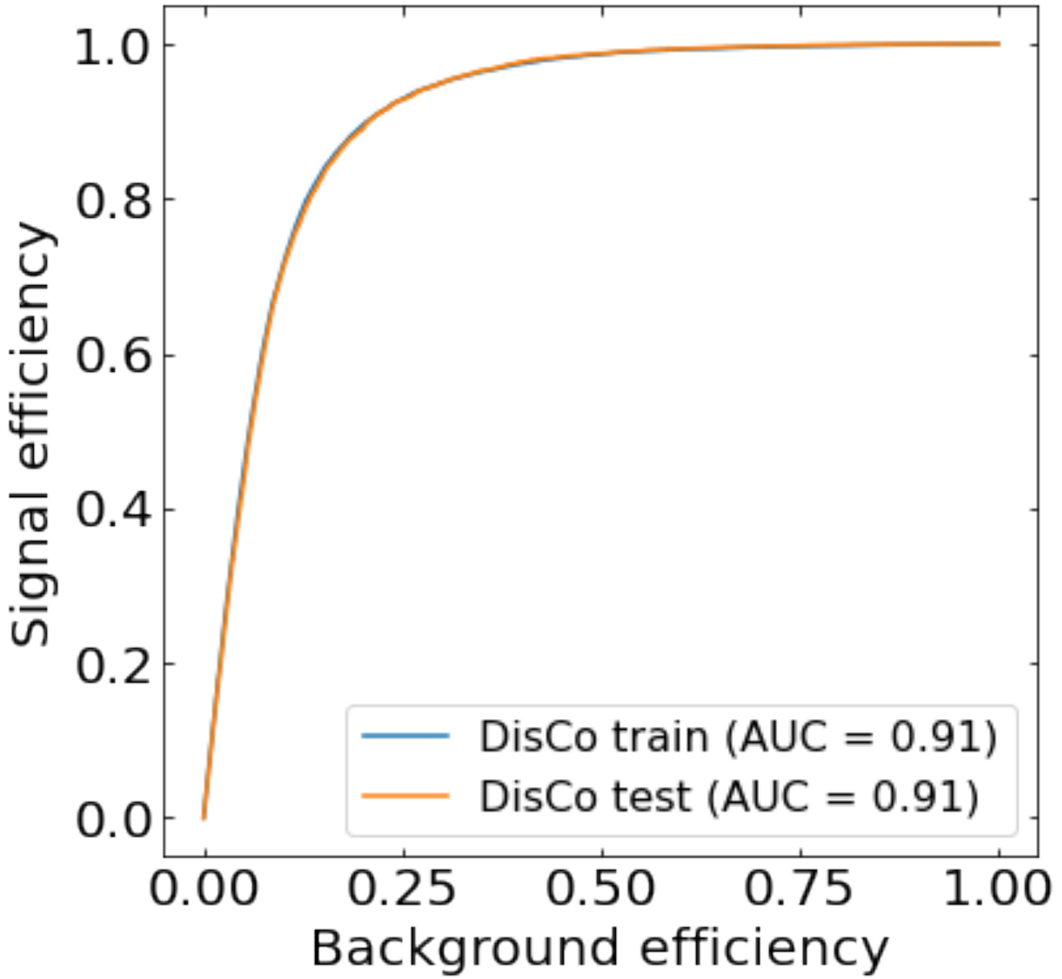


UC San Diego

ABCDNet: No DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 0 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

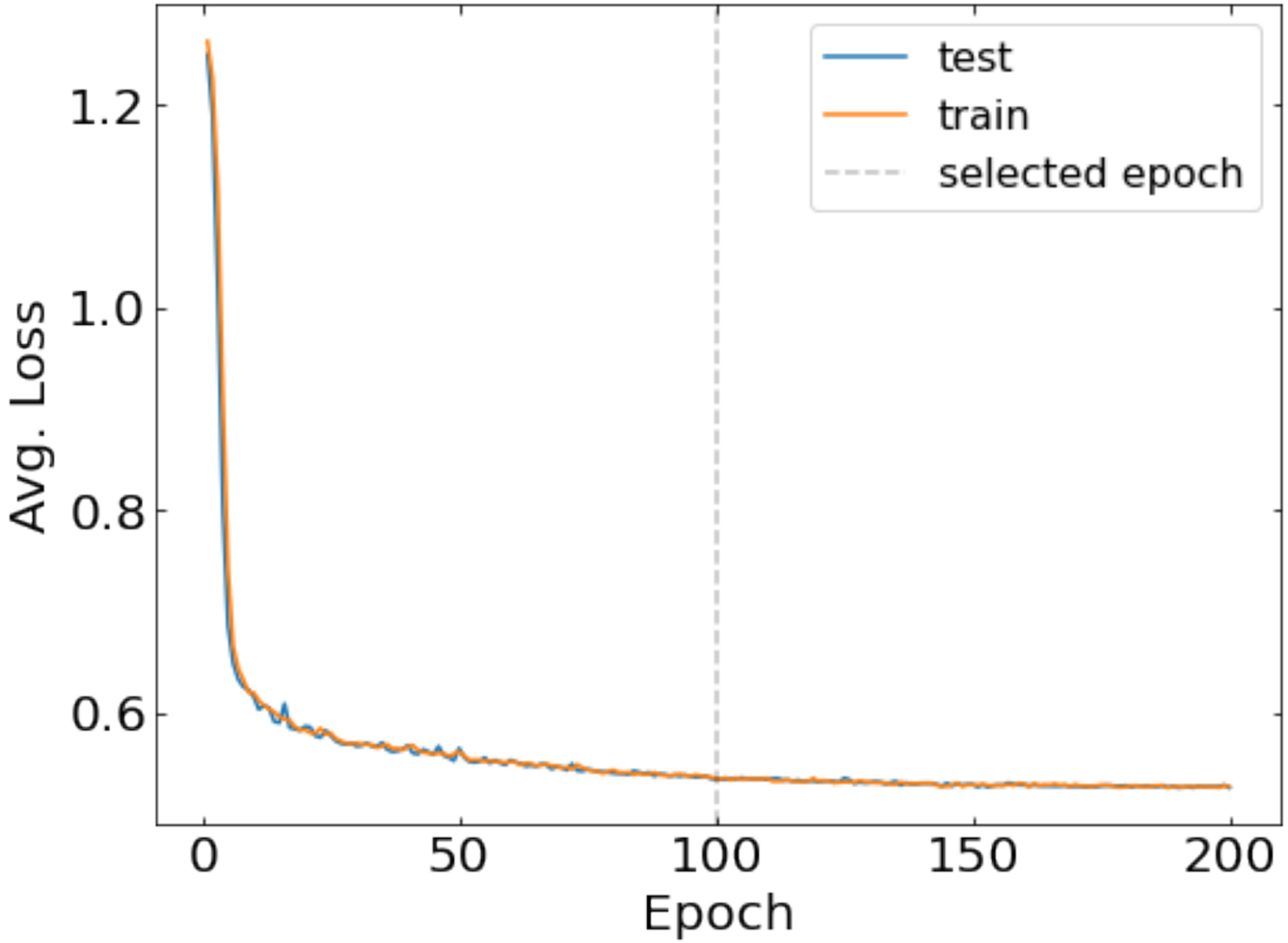


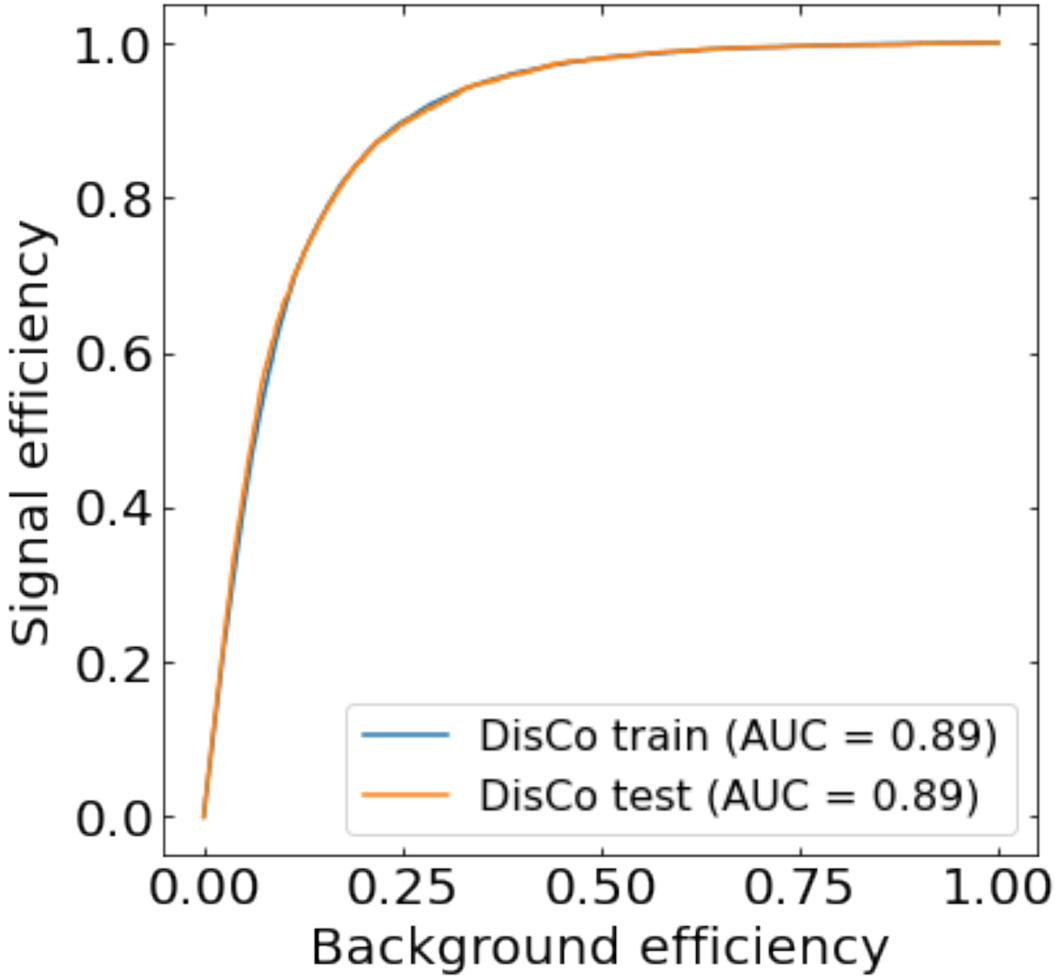


UC San Diego

ABCDNet: λ = 1 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1 \times \mathrm{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

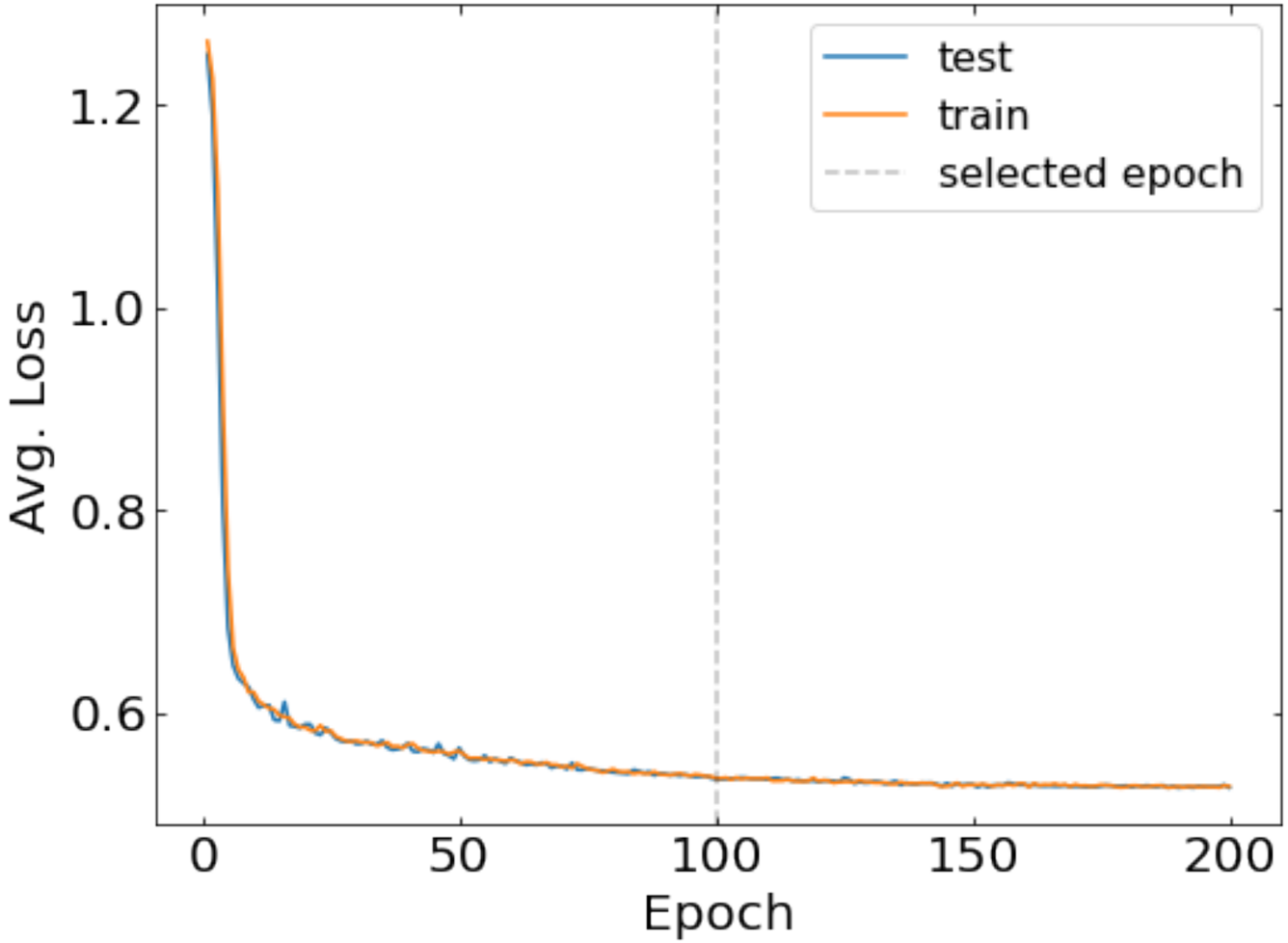


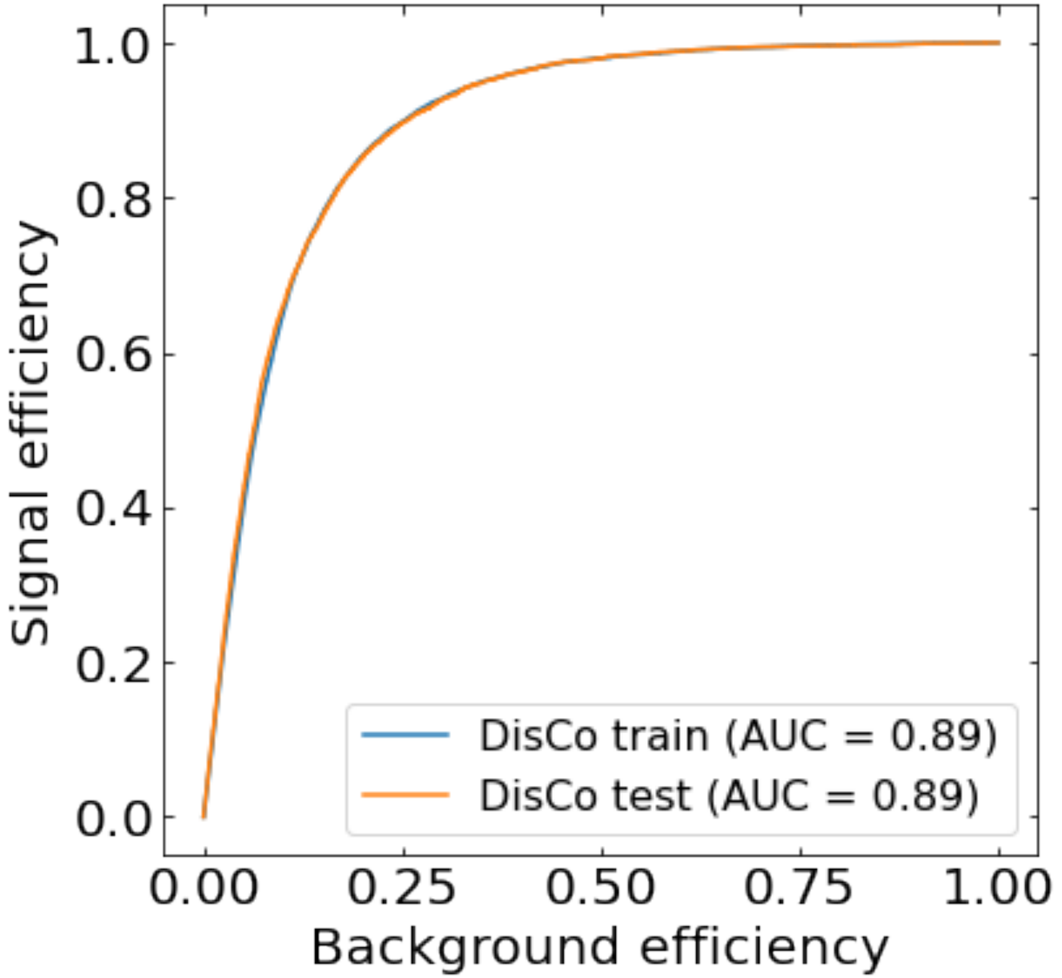


UC San Diego

ABCDNet: λ = 2 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 2 \times \mathrm{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

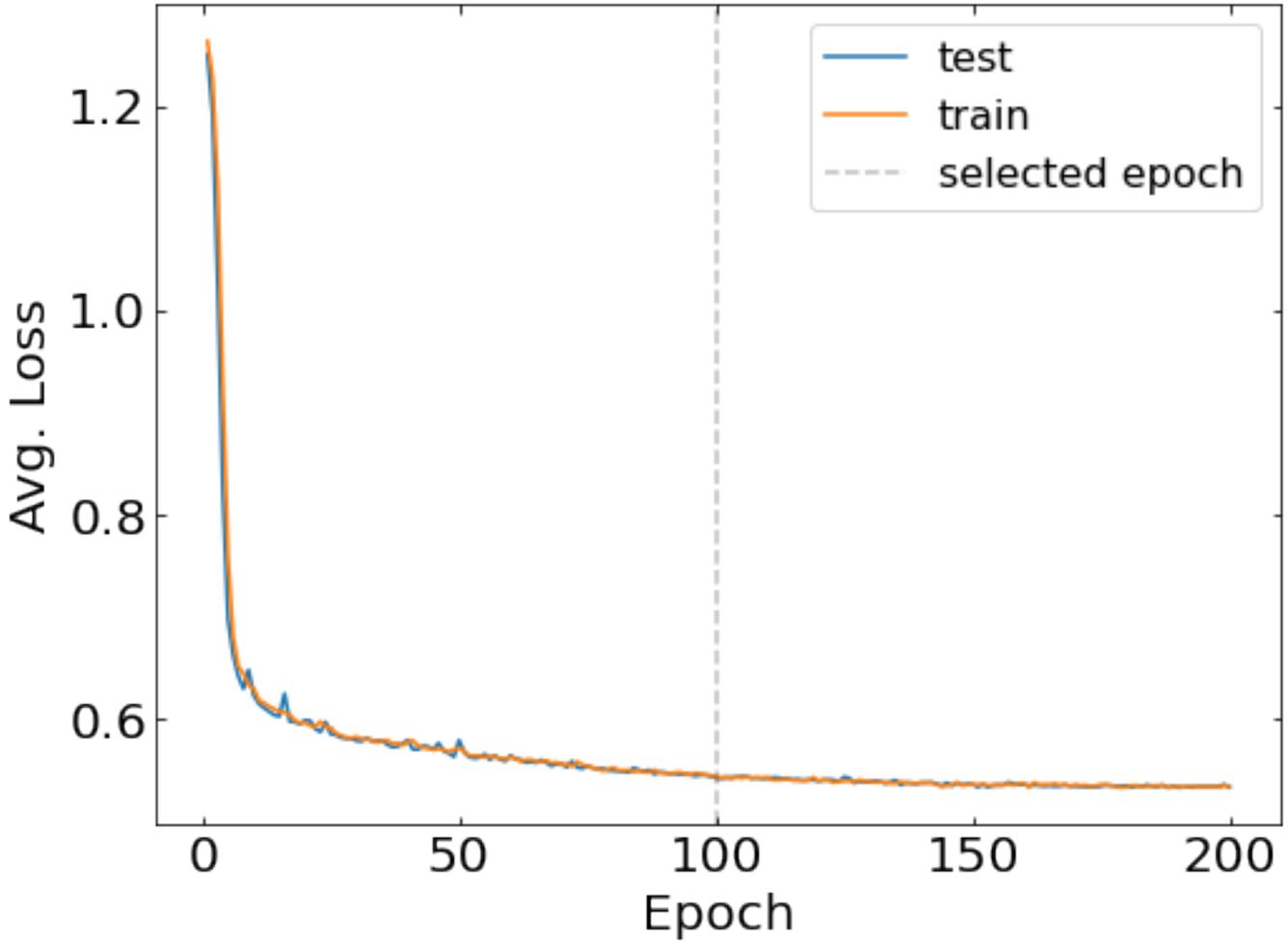


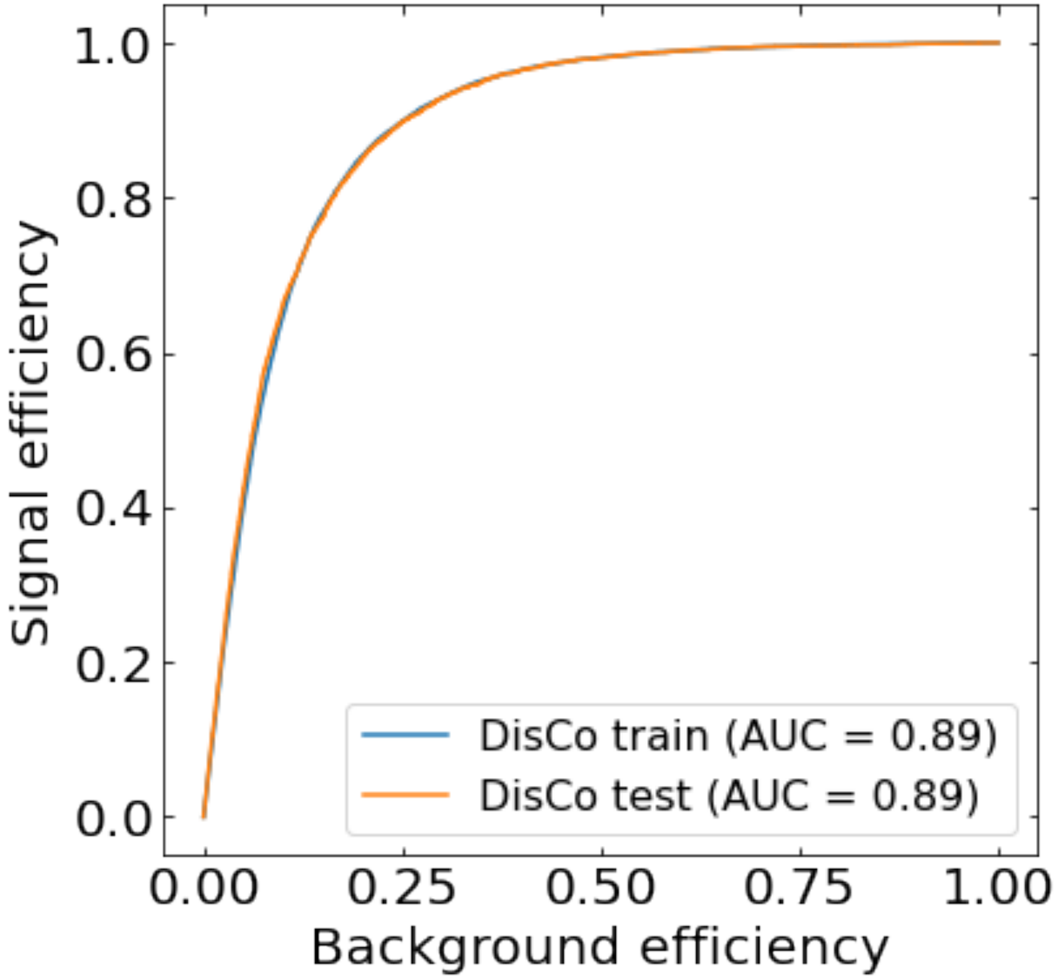


UC San Diego

ABCDNet: λ = 5 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 5 \times \mathrm{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

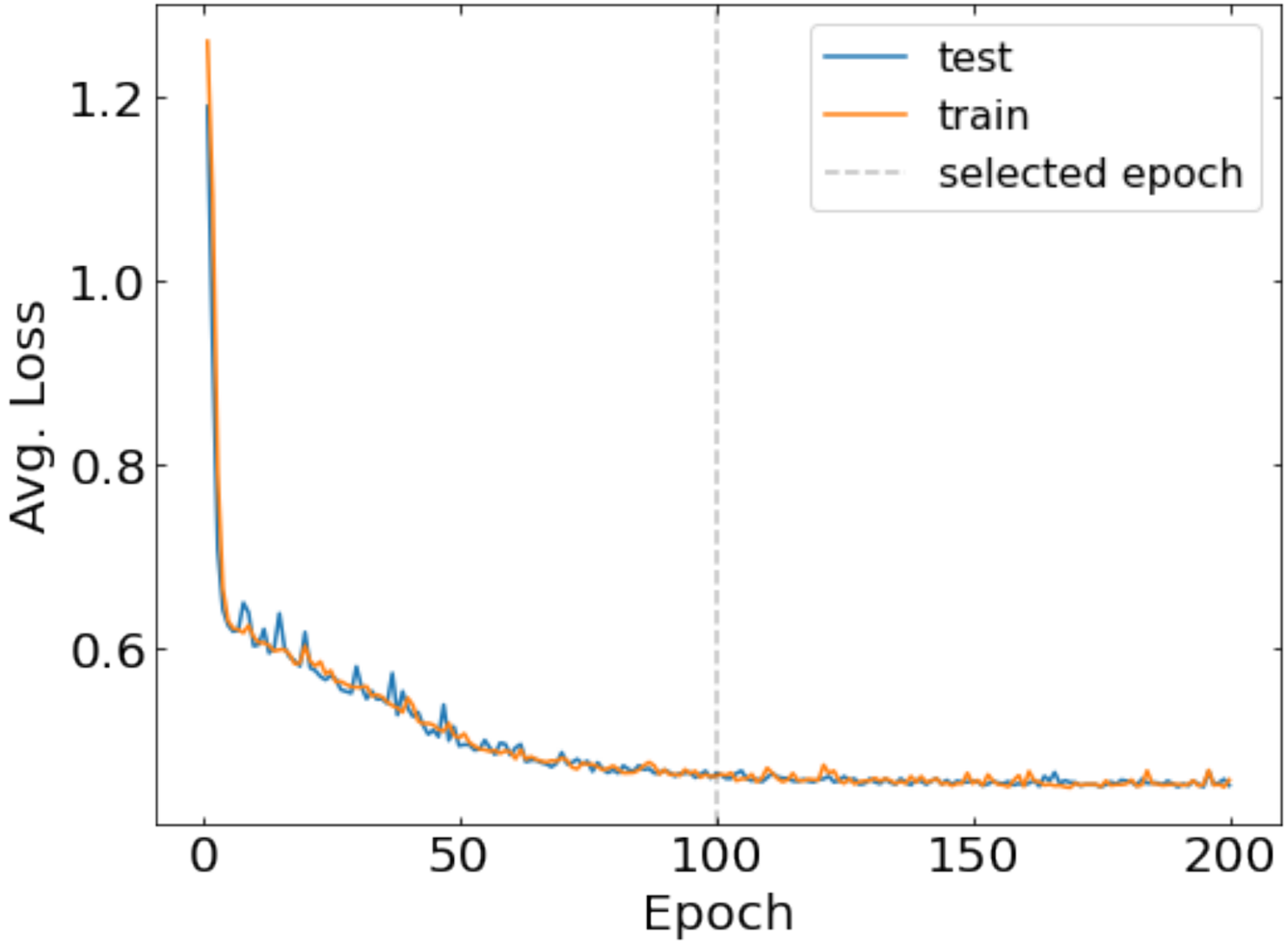


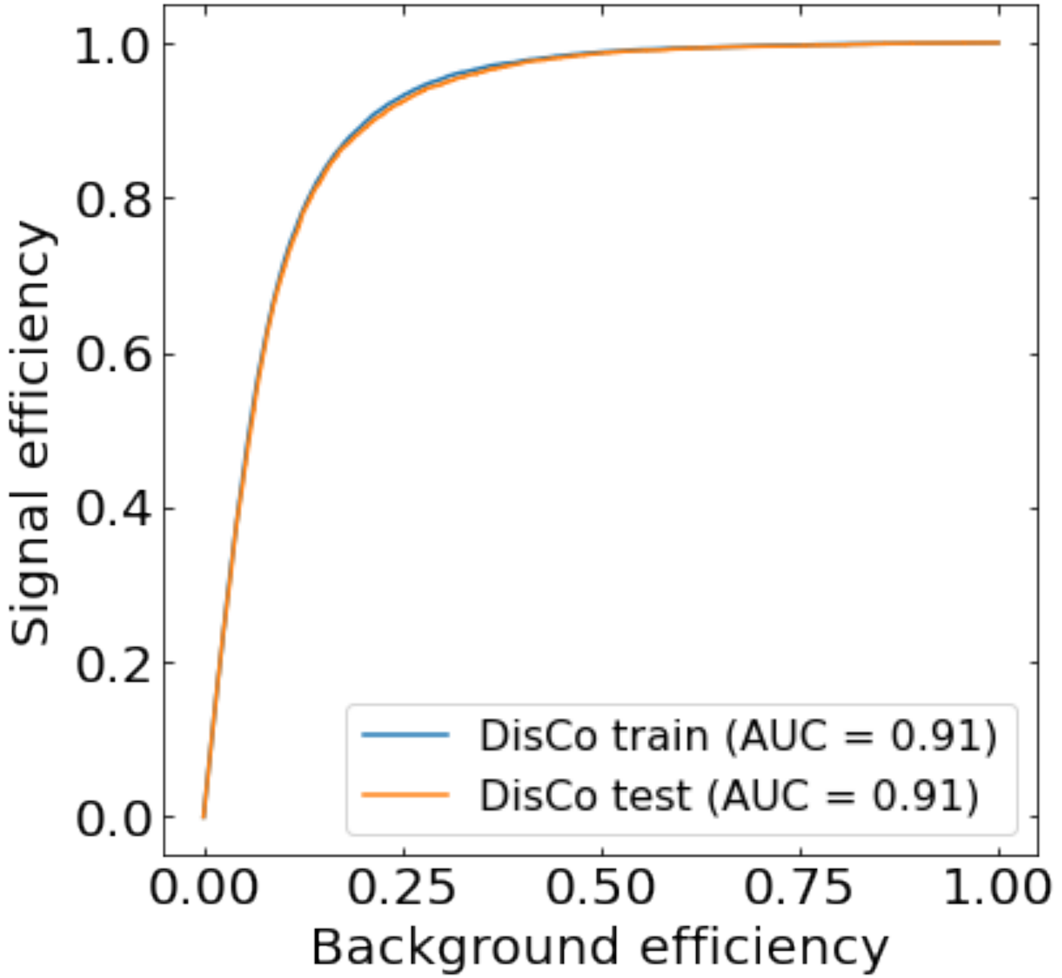


UC San Diego

ABCDNet: λ = 10 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 10 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$

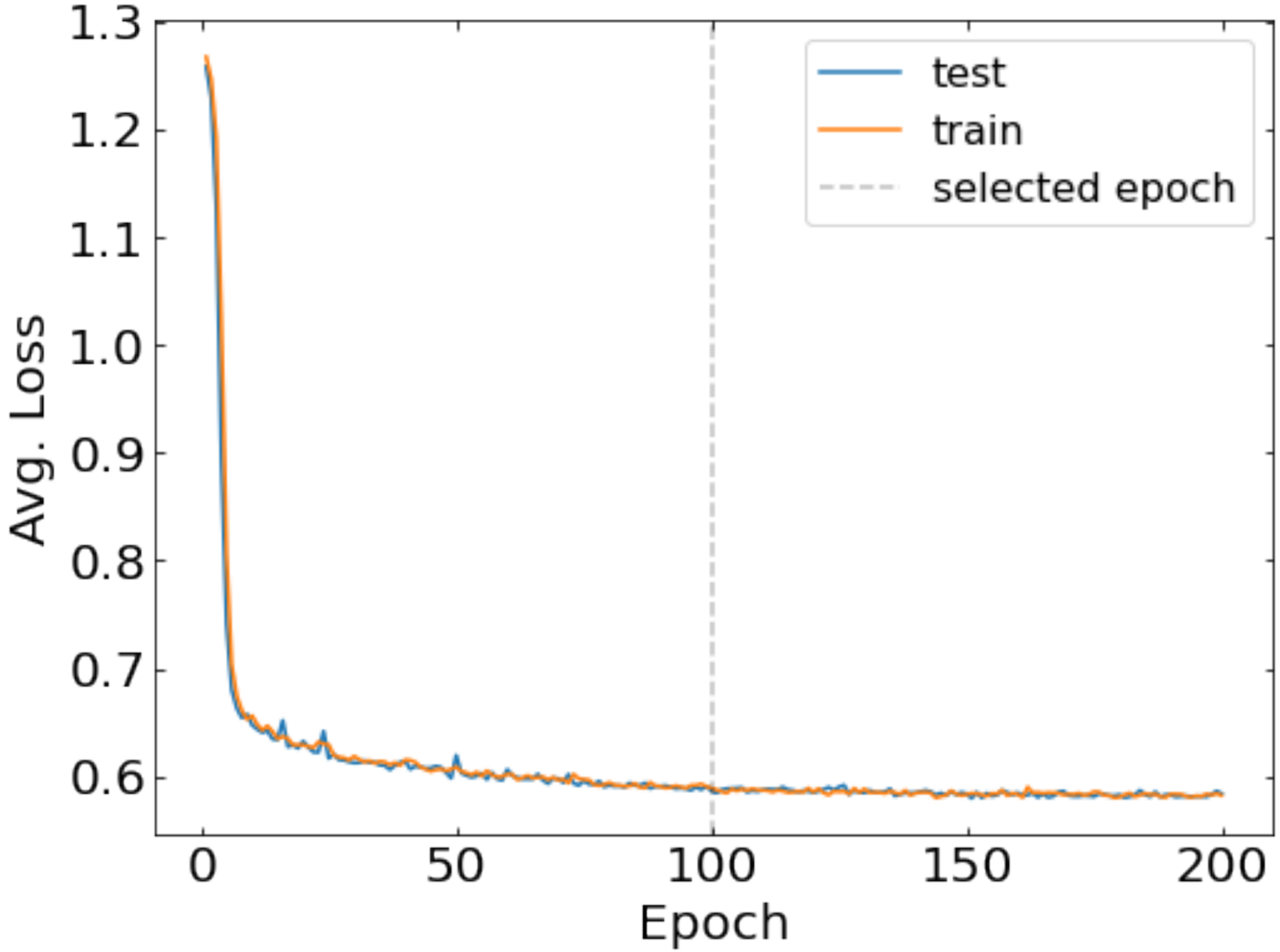


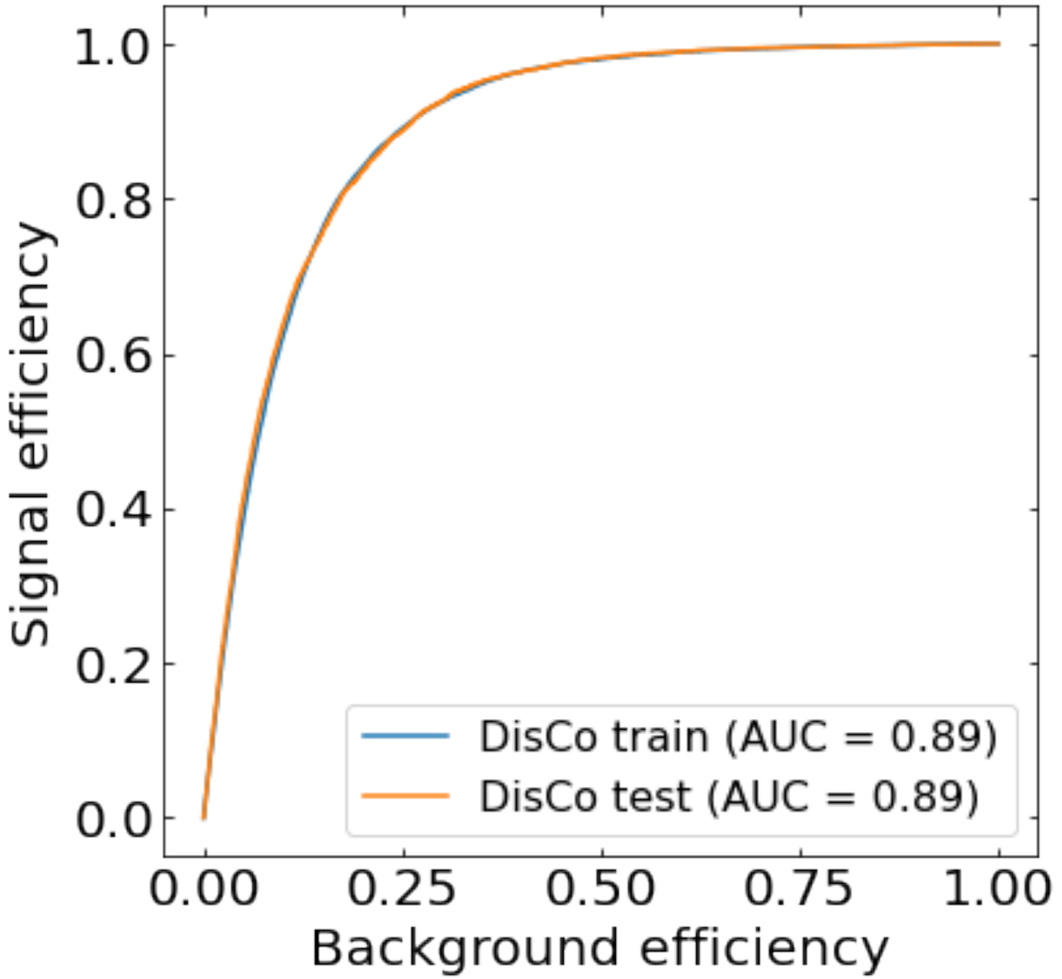


UC San Diego

ABCDNet: λ = 50 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 50 \times \text{dCorr}^2(f_{SD}(X_1, X_2), X_0)$$





UC San Diego