Domain Adaptation in Colliders: Reconcile Simulation and Data

Alexander Karlberg[1] Cheng-Wei Chiang[2] David Shih[3]
Stephane Cooperstein[1] Shang-Fu Wei[4]

[1]CERN [2]National Taiwan University [3]Rutgers University [4]The University of Tokyo

The 2nd AI+HEP in East Asia,
Jan. 2026

東京大学
THE UNIVERSITY OF TOKYO

Introduction

There is always an uncertainty induced by simulation.

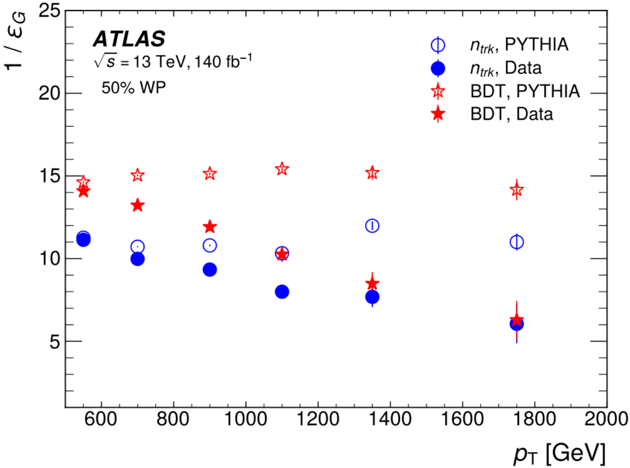
There is always an uncertainty induced by simulation.

Typical working flow. Trained by simulation $\rightarrow$ Inference on data.

Domain difference between simulation and data:

- Parton shower
- Hadronization
- etc.

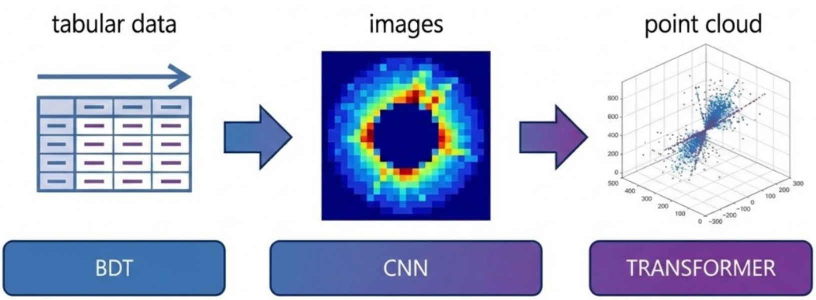
Introduction



[arXiv:2308.00716](https://arxiv.org/abs/2308.00716)

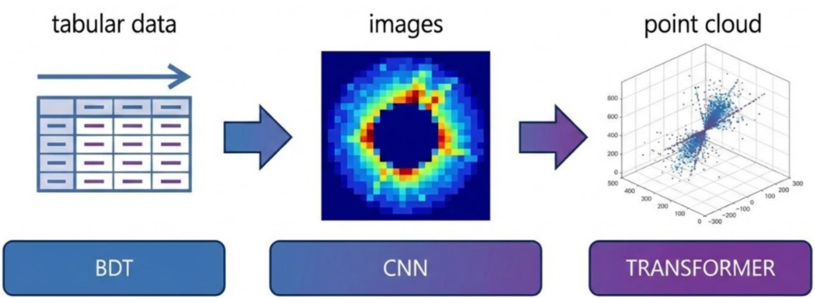
Introduction

Trend. We are accessing ever-increasingly granular information.



Introduction

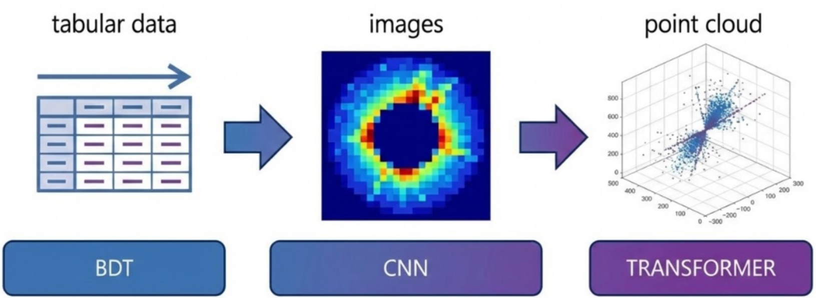
Trend. We are accessing ever-increasingly granular information.



How to ensure the model does NOT learn the artifact in the simulation?

Introduction

Trend. We are accessing ever-increasingly granular information.



How to ensure the model does NOT learn the artifact in the simulation?

Domain adaptation
Learn feature from simulation while aligning data shape

Outline

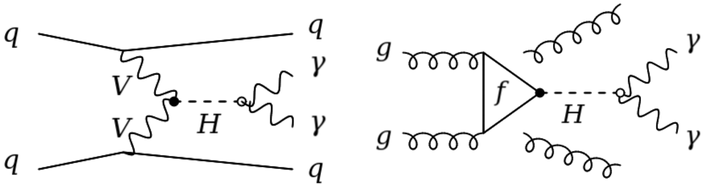
Table of Contents

Dataset

Processes. Higgs production in pp collisions

- Signal: VBF Higgs with up to three jets
- Background: ggF Higgs with up to three jets

Assume the Higgs further decay to diphoton, $H \to \gamma\gamma$.



Dataset

Generation.

Domain	Generator		Shower & Hadronization		Detector
Pythia	NLO PWG	$\longrightarrow$	**Pythia**	$\longrightarrow$	ATLAS card
Herwig	NLO PWG	$\longrightarrow$	**Herwig**	$\longrightarrow$	ATLAS card

This yields four sets of data:

	sig	bkg
Pythia	Pythia VBF	Pythia ggF
Herwig	Herwig VBF	Herwig ggF

Dataset

Domains.

	sig	bkg
Pythia	Pythia VBF	Pythia ggF
Herwig	Herwig VBF	Herwig ggF

Two domains are needed to be adapted:

- Source/simulation/labeled: **Pythia VBF** + Pythia ggF
- Target/data/unlabeled: **Herwig VBF** + Pythia ggF

Pre-selections.

- $N_\gamma \geq 2$
- $N_j \geq 2$ with $p_{T,j} \geq 25\,\text{GeV}$
- $\Delta\eta_{jj} \geq 2.5$

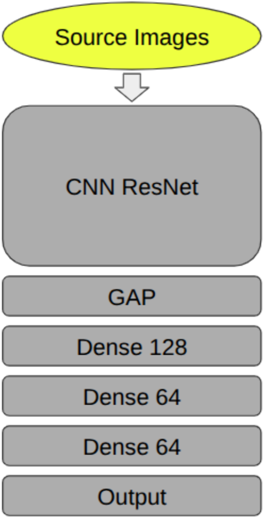
Pre-processing. crop dataset $\rightarrow$ rescale weights.

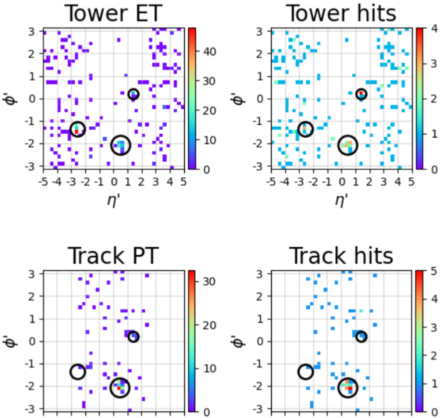
Pythia	Raw N_{evt}	Sum of weight	XS (pb)
VBF	77k	780k	1.150
GGF	77k	780k	1.046

Herwig	Raw N_{evt}	Sum of weight	XS (pb)
VBF	85k	868k	1.177
GGF	85k	868k	1.165

Model

A very simple ResNet CNN as a prototype.





Model

Idea. Align the latent representation of simulation and data.

Model

Idea. Align the latent representation of simulation and data.

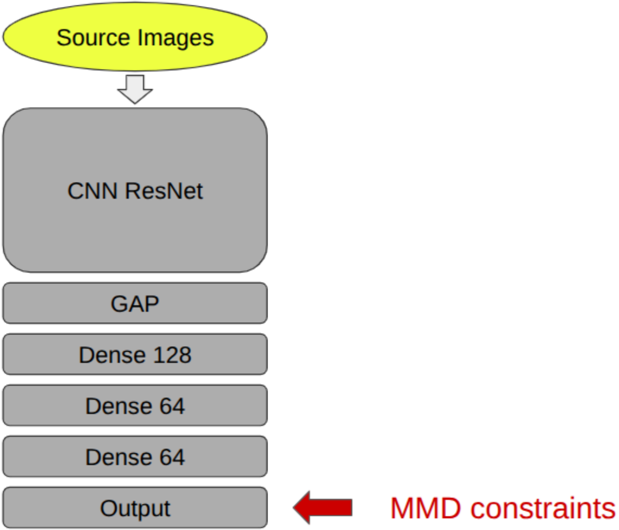
Introduce the Maximum Mean Discrepancy (MMD) as a distance of distributions.

$$\begin{aligned} \mathrm{MMD}^2(P,Q) \equiv \mathbb{E}_{X\sim P}[k(X,X)] + \mathbb{E}_{Y\sim Q}[k(Y,Y)] \\ - 2\mathbb{E}_{X\sim P, Y\sim Q}[k(X,Y)], \end{aligned} \tag{1}$$

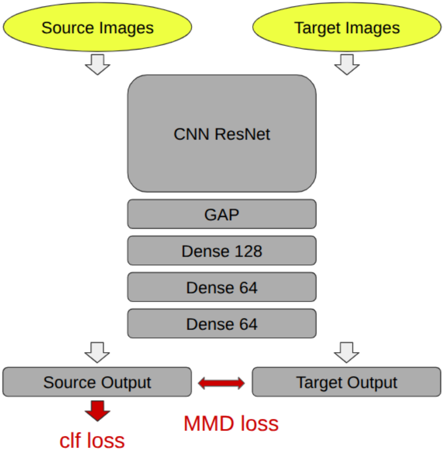
with the Gaussian kernel

$$k(X,X') \equiv \exp\left(-\frac{\|X-X'\|^2}{\gamma}\right). \tag{2}$$

Model


Source Images
CNN ResNet
GAP
Dense 128
Dense 64
Dense 64
Output
MMD constraints

Model



$$\text{Loss} = \text{C.E.}[\text{simulation}] + \lambda \cdot \text{MMD}[\text{simulation}, \text{data}] \tag{3}$$

Table of Contents

TPR history

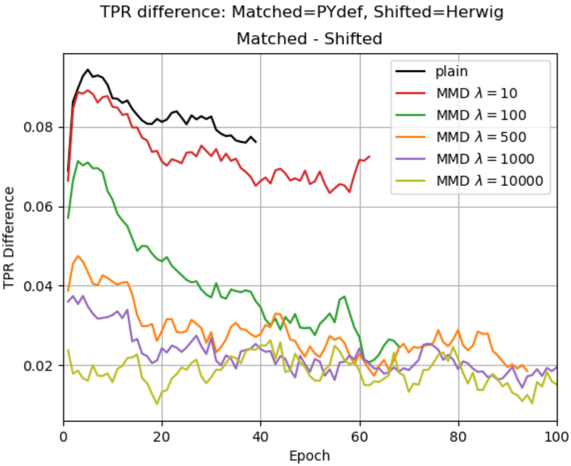
Compare 10 runs for each of the following models:

- Plain
- MMD $\lambda = 10$
- MMD $\lambda = 100$
- MMD $\lambda = 500$
- MMD $\lambda = 1000$
- MMD $\lambda = 10000$

Early stop for validation loss with patience of 50 epochs.

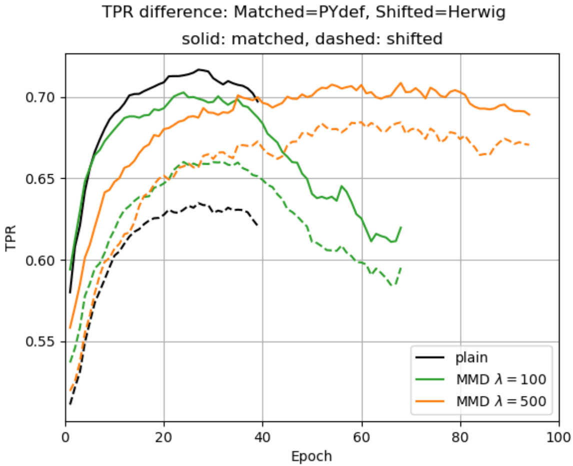
TPR history

The working point is fixed to have $\mathrm{FPR} = 0.1$.



TPR history

The working point is fixed to have $\mathrm{FPR} = 0.1$.



Uncertainty

Toy measurement.

$$\begin{pmatrix} \sigma_V \\ \sigma_G \end{pmatrix} = \frac{1}{\mathcal{L}} \frac{1}{\mathrm{FPR} - \mathrm{TPR}} \begin{pmatrix} \mathrm{FPR} - 1 & \mathrm{FPR} \\ 1 - \mathrm{TPR} & -\mathrm{TPR} \end{pmatrix} \begin{pmatrix} N_{\textsf{V-like}} \\ N_{\textsf{G-like}} \end{pmatrix} \tag{4}$$

Uncertainty

Toy measurement.

$$\begin{pmatrix} \sigma_V \\ \sigma_G \end{pmatrix} = \frac{1}{\mathcal{L}} \frac{1}{\mathrm{FPR} - \mathrm{TPR}} \begin{pmatrix} \mathrm{FPR} - 1 & \mathrm{FPR} \\ 1 - \mathrm{TPR} & -\mathrm{TPR} \end{pmatrix} \begin{pmatrix} N_{\textsf{V-like}} \\ N_{\textsf{G-like}} \end{pmatrix} \tag{4}$$

Sources of the uncertainty:

- Statistical: $N_{\textsf{V-like}}$, $N_{\textsf{G-like}}$

Uncertainty

Toy measurement.

$$\begin{pmatrix} \sigma_V \\ \sigma_G \end{pmatrix} = \frac{1}{\mathcal{L}} \frac{1}{\mathrm{FPR} - \mathrm{TPR}} \begin{pmatrix} \mathrm{FPR} - 1 & \mathrm{FPR} \\ 1 - \mathrm{TPR} & -\mathrm{TPR} \end{pmatrix} \begin{pmatrix} N_{\text{V-like}} \\ N_{\text{G-like}} \end{pmatrix} \tag{4}$$

Sources of the uncertainty:

- Statistical: $N_{\text{V-like}}$, $N_{\text{G-like}} \sim$ Poisson

Uncertainty

Toy measurement.

$$\begin{pmatrix} \sigma_V \\ \sigma_G \end{pmatrix} = \frac{1}{\mathcal{L}} \frac{1}{\mathrm{FPR} - \mathrm{TPR}} \begin{pmatrix} \mathrm{FPR} - 1 & \mathrm{FPR} \\ 1 - \mathrm{TPR} & -\mathrm{TPR} \end{pmatrix} \begin{pmatrix} N_{\text{V-like}} \\ N_{\text{G-like}} \end{pmatrix} \tag{4}$$

Sources of the uncertainty:

- Statistical: $N_{\text{V-like}}$, $N_{\text{G-like}} \sim$ Poisson
- Systematic: TPR, FPR

Uncertainty

Toy measurement.

$$\begin{pmatrix}\sigma_V \\ \sigma_G\end{pmatrix} = \frac{1}{\mathcal{L}} \frac{1}{\mathrm{FPR}-\mathrm{TPR}} \begin{pmatrix}\mathrm{FPR}-1 & \mathrm{FPR} \\ 1-\mathrm{TPR} & -\mathrm{TPR}\end{pmatrix} \begin{pmatrix}N_{\textsf{V-like}} \\ N_{\textsf{G-like}}\end{pmatrix} \tag{4}$$

Sources of the uncertainty:

- Statistical: $N_{\textsf{V-like}}$, $N_{\textsf{G-like}} \sim$ Poisson
- Systematic: TPR, FPR

  $\rightsquigarrow$ Estimate by domain difference: $\Delta\mathrm{TPR}$, $\Delta\mathrm{FPR}$.

Uncertainty

The variance of the cross section measurement:

$$\mathrm{Var}(\sigma_V) = \left(\frac{\mathrm{FPR}-1}{(\mathrm{FPR}-\mathrm{TPR})^2}\frac{N_{\textsf{V-like}}}{\mathcal{L}} + \frac{\mathrm{FPR}}{(\mathrm{FPR}-\mathrm{TPR})^2}\frac{N_{\textsf{G-like}}}{\mathcal{L}}\right)^2 (\Delta\mathrm{TPR})^2$$
$$+\left(\frac{1-\mathrm{TPR}}{(\mathrm{FPR}-\mathrm{TPR})^2}\frac{N_{\textsf{V-like}}}{\mathcal{L}} + \frac{-\mathrm{TPR}}{(\mathrm{FPR}-\mathrm{TPR})^2}\frac{N_{\textsf{G-like}}}{\mathcal{L}}\right)^2 (\Delta\mathrm{FPR})^2$$
$$+\left(\frac{\mathrm{FPR}-1}{\mathrm{FPR}-\mathrm{TPR}}\frac{1}{\mathcal{L}}\right)^2 N_{\textsf{V-like}} + \left(\frac{\mathrm{FPR}}{\mathrm{FPR}-\mathrm{TPR}}\frac{1}{\mathcal{L}}\right)^2 N_{\textsf{G-like}} \quad (5)$$

$$(6)$$

- $\sigma_V = 1.150\,\text{pb}$, $\sigma_G = 1.046\,\text{pb}$,
- $\mathcal{L} = 3\,\text{ab}^{-1}$,
- the working point with $\mathrm{FPR} = 0.1$ ($\Delta\mathrm{FPR} = 0$).

Uncertainty

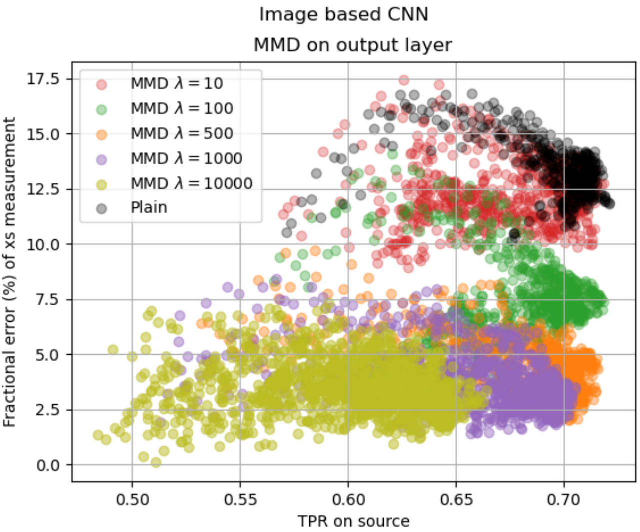
The variance of the cross section measurement:

$$\mathrm{Var}(\sigma_V) = \left(\frac{\mathrm{FPR}-1}{(\mathrm{FPR}-\mathrm{TPR})^2}\frac{N_{\text{V-like}}}{\mathcal{L}} + \frac{\mathrm{FPR}}{(\mathrm{FPR}-\mathrm{TPR})^2}\frac{N_{\text{G-like}}}{\mathcal{L}}\right)^2 (\Delta\mathrm{TPR})^2$$
$$+\left(\frac{1-\mathrm{TPR}}{(\mathrm{FPR}-\mathrm{TPR})^2}\frac{N_{\text{V-like}}}{\mathcal{L}} + \frac{-\mathrm{TPR}}{(\mathrm{FPR}-\mathrm{TPR})^2}\frac{N_{\text{G-like}}}{\mathcal{L}}\right)^2 (\Delta\mathrm{FPR})^2$$
$$+\left(\frac{\mathrm{FPR}-1}{\mathrm{FPR}-\mathrm{TPR}}\frac{1}{\mathcal{L}}\right)^2 N_{\text{V-like}} + \left(\frac{\mathrm{FPR}}{\mathrm{FPR}-\mathrm{TPR}}\frac{1}{\mathcal{L}}\right)^2 N_{\text{G-like}} \quad (5)$$

$$\text{Frac. error} = \frac{\sqrt{\mathrm{Var}(\sigma_V)}}{\sigma_V} \quad (6)$$

- $\sigma_V = 1.150\,\text{pb}$, $\sigma_G = 1.046\,\text{pb}$,
- $\mathcal{L} = 3\,\text{ab}^{-1}$,
- the working point with $\mathrm{FPR} = 0.1$ ($\Delta\mathrm{FPR} = 0$).

Uncertainty



Uncertainty

To perform model selection, the only viable information are

- TPR on simulation;
- MMD between simulation and data.

Uncertainty

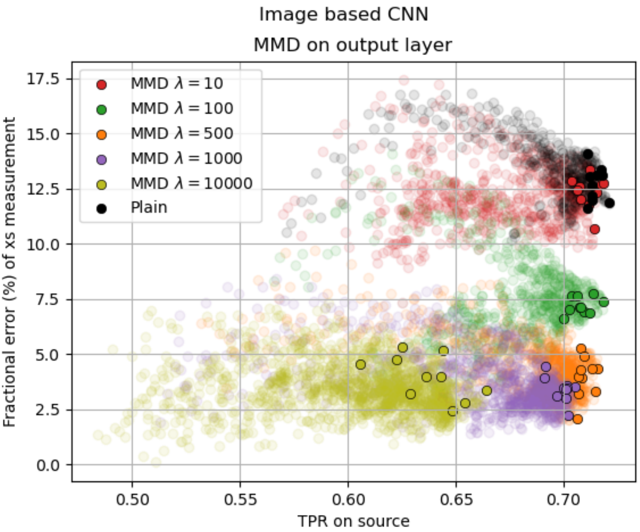
To perform model selection, the only viable information are

- TPR on simulation;
- MMD between simulation and data.

Here we simply choose the epoch with the highest TPR.

Remark. The choice depends on the base model and hyperparameters.

Uncertainty



Summary

Summary

- Systematic is an important ingredient in measurement.
- Introducing MMD helps to reduce the domain difference, and hence the systematic.

To-do list

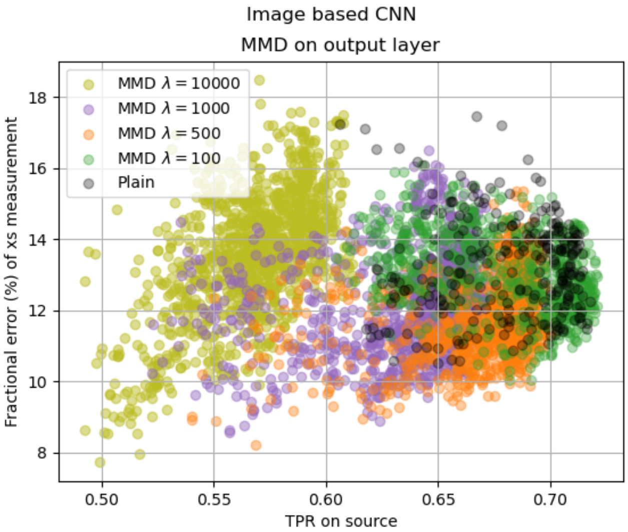
- Refine the uncertainty estimation.
- Apply the MMD plug-in to other models (PartNet, ParT, etc.).

Thank you for listening!

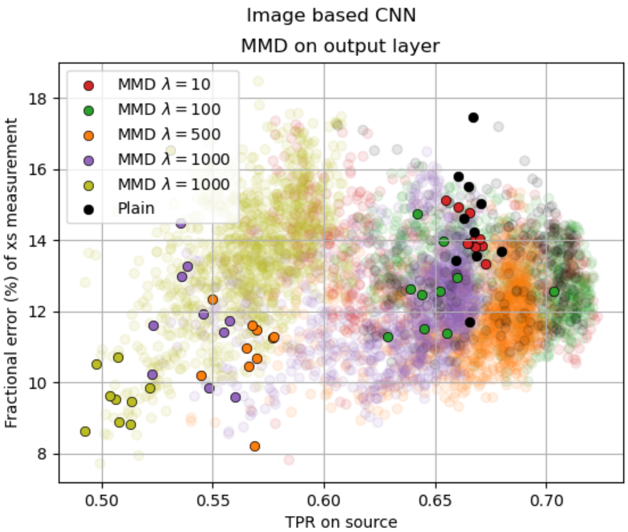
Backup - Full Domain Difference

- Source/simulation/labeled: **Pythia VBF + Pythia ggF**
- Target/data/unlabeled: **Herwig VBF + Herwig ggF**

Backup - Full Domain Difference

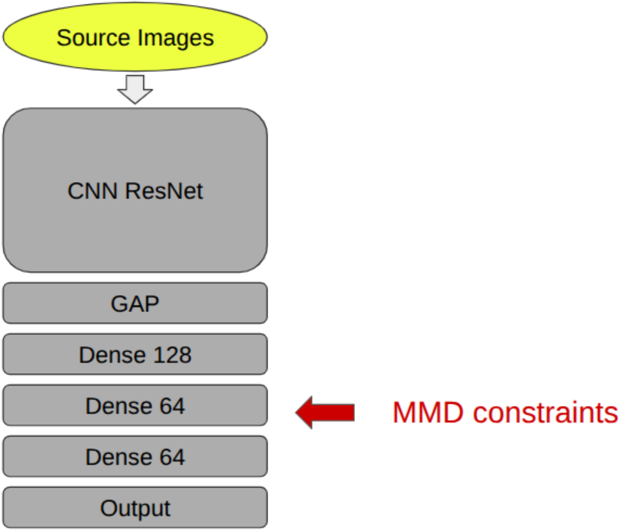


Backup - Full Domain Difference

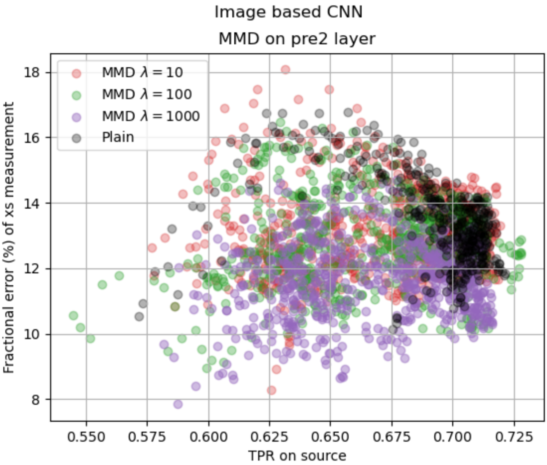


Model selection: Minimal MMD

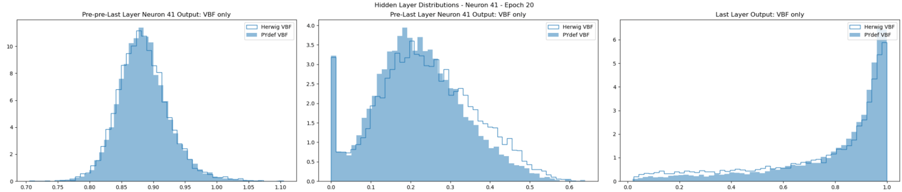
Backup - MMD on Different Layers



Backup - MMD on Different Layers



Pre2 layer:



Last layer: