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

ML4Jets, Vienna
Sep 2026

Introduction

There is always an uncertainty induced by simulation.

Introduction

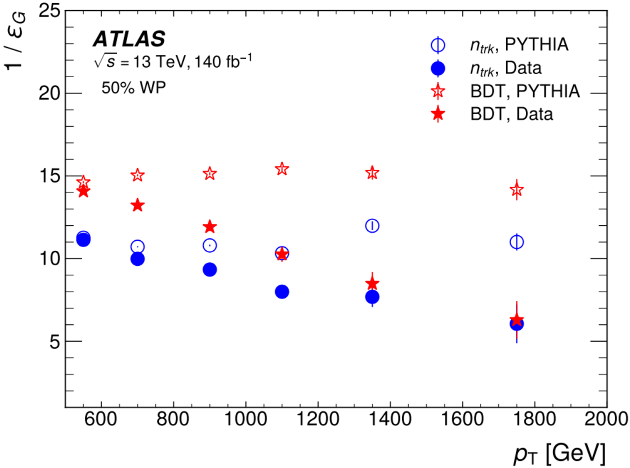
There is always an uncertainty induced by simulation.

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

Domain difference between simulation and data:

- Parton shower
- Hadronization
- etc.

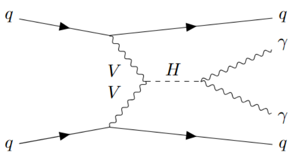
Introduction

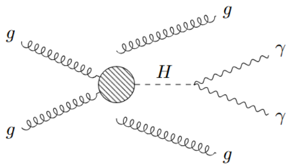


arXiv:2308.00716

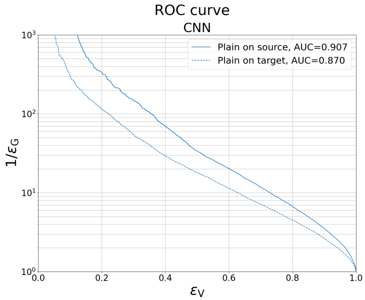
Introduction

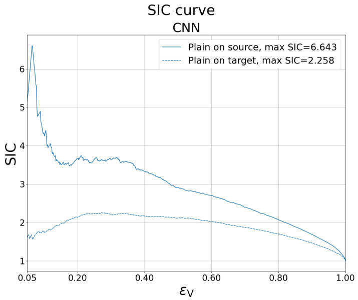
Assume a VBF Higss vs. ggF Higgs classification





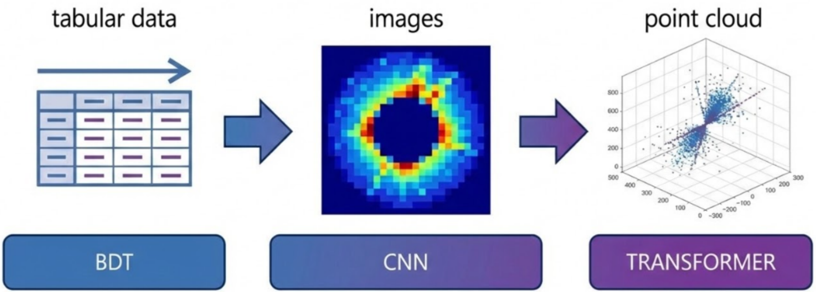
Train on Pythia, test on Pythia/Herwig





Introduction

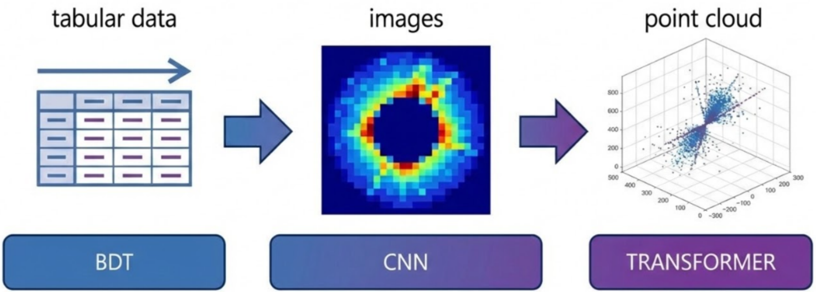
Trend. We are accessing ever-increasingly granular information.



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?

Unsupervised domain adaptation
Learn from simulation (with label) + data (without label)

Outline

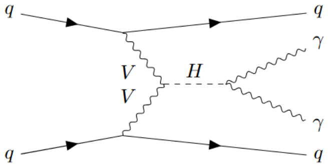
Table of Contents

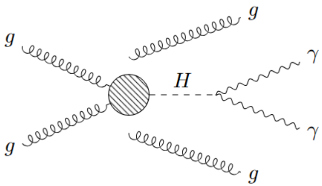
Dataset

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

Two domains are needed to be adapted:

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

Dataset

Pre-selections.

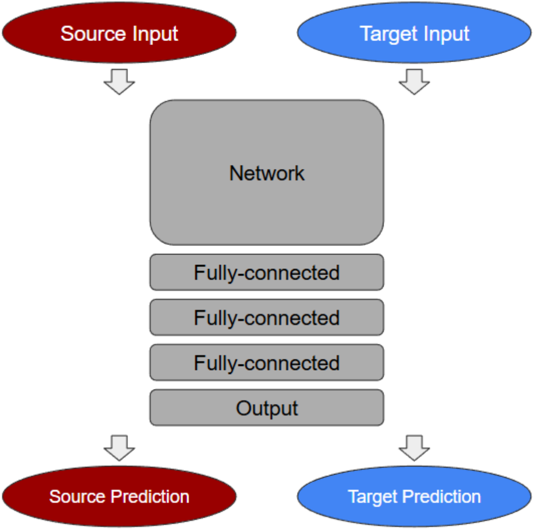
- $N_\gamma \geq 2$
- $N_j \geq 2$ with $p_{T,j} \geq 25$ GeV
- $\Delta\eta_{jj} \geq 2.5$
- $p_{T,H} < 200$ GeV

Pre-processing. Weight rescaling for each event to respect the xs

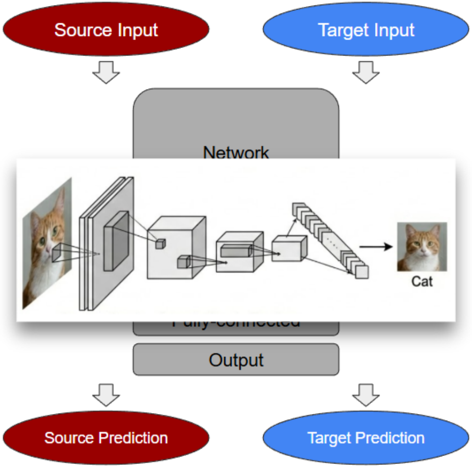
Pythia	N_{evt}	σ (pb)
VBF	242k	1.100(3)
ggF	97k	0.960(1)

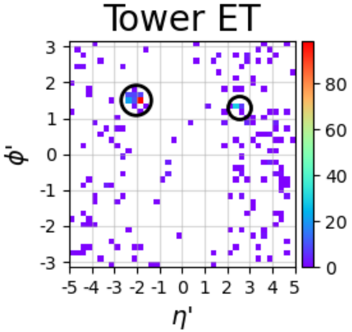
Herwig	N_{evt}	σ (pb)
VBF	247k	1.123(1)
ggF	107k	1.072(3)

Unsupervised Domain Adaptation (UDA)



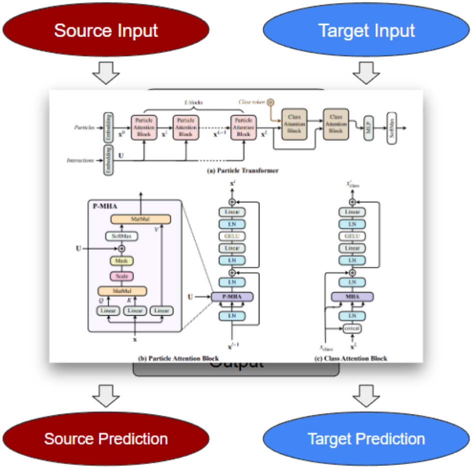
Unsupervised Domain Adaptation (UDA)



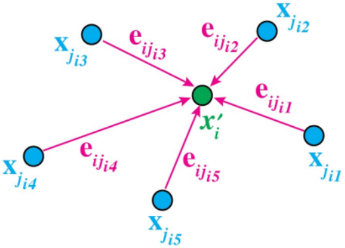


arXiv:2209.05518

Unsupervised Domain Adaptation (UDA)

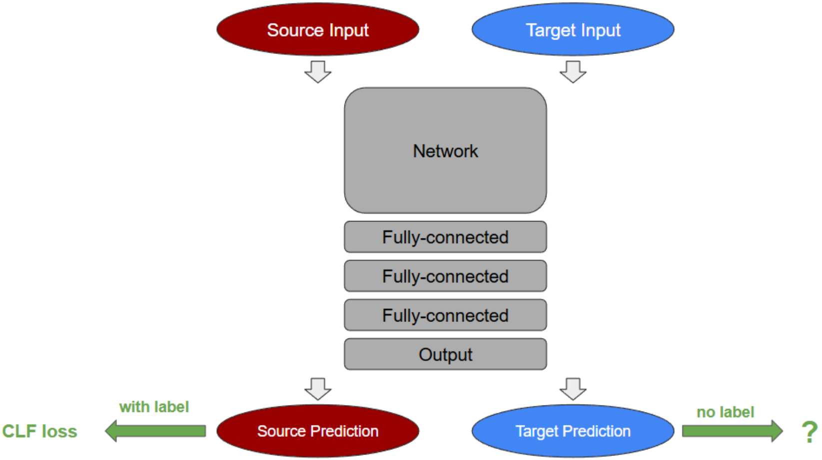


arXiv:2202.03772

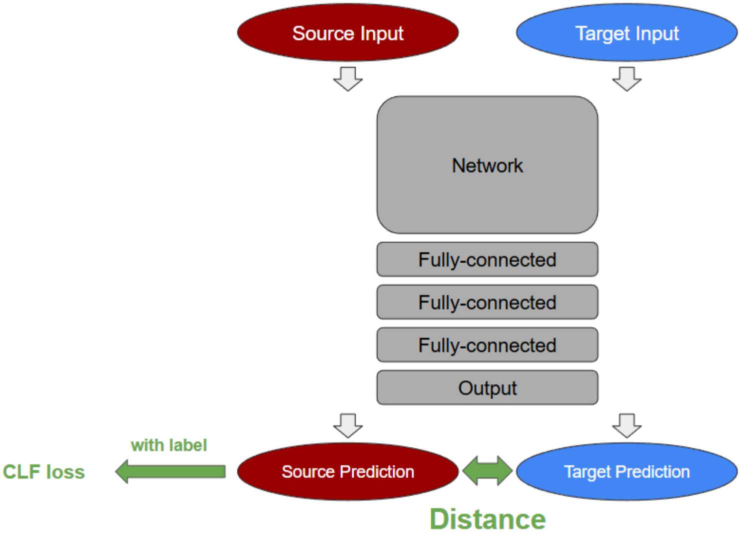


arXiv:1801.07829

Unsupervised Domain Adaptation (UDA)



Unsupervised Domain Adaptation (UDA)



Unsupervised Domain Adaptation (UDA)

Idea. Align the representation of simulation and data.

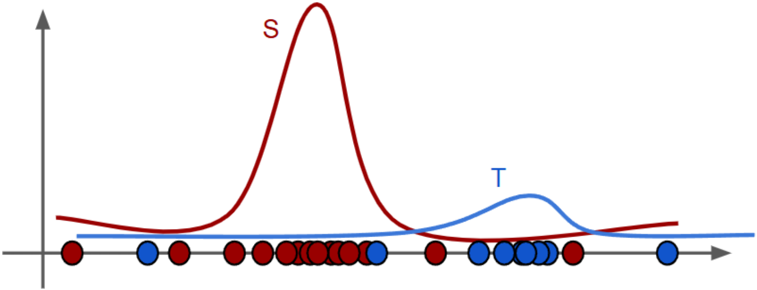
Introduce the **Maximum Mean Discrepancy (MMD)** as a distance of distributions. arXiv:1912.00320

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

with the Gaussian kernel

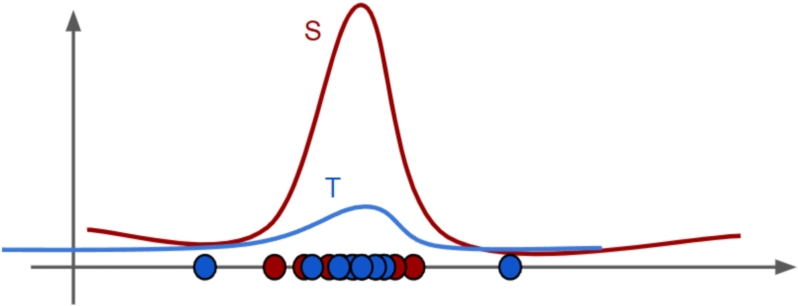
$$k(X,X') \equiv \exp\left(-\frac{\|X-X'\|^2}{\gamma}\right). \tag{2}$$

Unsupervised Domain Adaptation (UDA)



$$\text{distance} \sim ||\mathbb{E}_{X\sim S}\left[X\right] - \mathbb{E}_{Y\sim T}\left[Y\right]|| \quad (3)$$

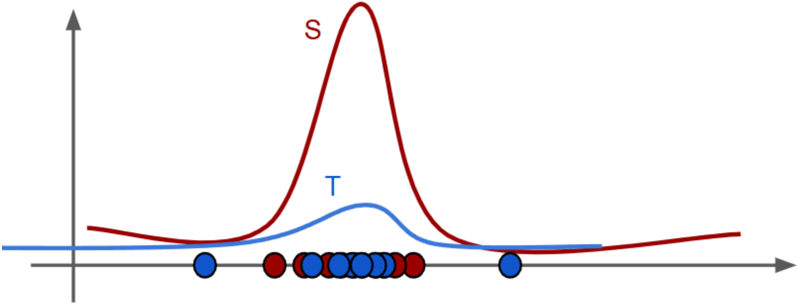
Unsupervised Domain Adaptation (UDA)



$$\text{distance} \sim \left|\left|\mathbb{E}_{X\sim S}\left[\left(X, X^2\right)\right] - \mathbb{E}_{Y\sim T}\left[\left(Y, Y^2\right)\right]\right|\right| \quad (4)$$

(5)

Unsupervised Domain Adaptation (UDA)

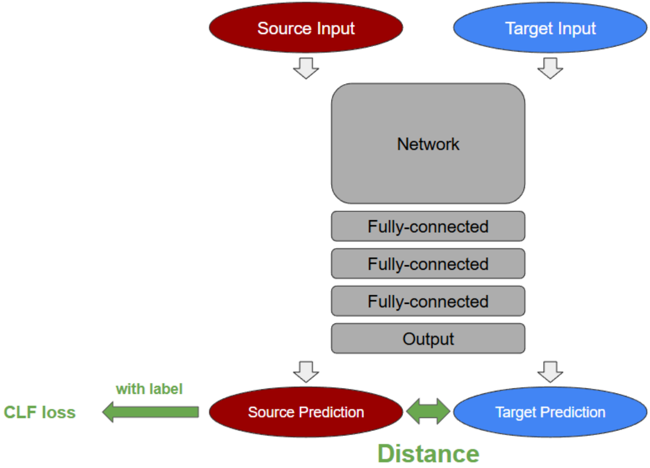


$$\text{distance} \sim \left|\left|\mathbb{E}_{X\sim S}\left[\left(X, X^2\right)\right] - \mathbb{E}_{Y\sim T}\left[\left(Y, Y^2\right)\right]\right|\right| \tag{4}$$

$\downarrow$ higher moment

$$\sim \left|\left|\mathbb{E}_{X\sim S}\left[e^X\right] - \mathbb{E}_{Y\sim T}\left[e^Y\right]\right|\right| \tag{5}$$

Unsupervised Domain Adaptation (UDA)



$$\mathcal{L} = \mathcal{L}_{\mathsf{CLF}}(\text{source}) + \lambda \cdot \mathrm{MMD}^2(\text{source}, \text{target}), \tag{6}$$

Table of Contents

Uncertainty

Toy measurement.

$$N_{\mathrm{V-like}} = \mathcal{L}(\sigma_{\mathrm{V}} \cdot \epsilon_{\mathrm{V}} + \sigma_{\mathrm{G}} \cdot \epsilon_{\mathrm{G}}) \tag{7}$$

$$N_{\mathrm{G-like}} = \mathcal{L}(\sigma_{\mathrm{V}} \cdot (1 - \epsilon_{\mathrm{V}}) + \sigma_{\mathrm{G}} \cdot (1 - \epsilon_{\mathrm{G}})). \tag{8}$$

$$\begin{pmatrix} \sigma_{\mathrm{V}} \\ \sigma_{\mathrm{G}} \end{pmatrix} = \frac{1}{\mathcal{L}} \frac{1}{\epsilon_{\mathrm{G}} - \epsilon_{\mathrm{V}}} \begin{pmatrix} \epsilon_{\mathrm{G}} - 1 & \epsilon_{\mathrm{G}} \\ 1 - \epsilon_{\mathrm{V}} & -\epsilon_{\mathrm{V}} \end{pmatrix} \begin{pmatrix} N_{\mathrm{V-like}} \\ N_{\mathrm{G-like}} \end{pmatrix}. \tag{9}$$

Uncertainty

Toy measurement.

$$N_{\mathrm{V-like}} = \mathcal{L}(\sigma_{\mathrm{V}} \cdot \epsilon_{\mathrm{V}} + \sigma_{\mathrm{G}} \cdot \epsilon_{\mathrm{G}}) \tag{7}$$

$$N_{\mathrm{G-like}} = \mathcal{L}(\sigma_{\mathrm{V}} \cdot (1 - \epsilon_{\mathrm{V}}) + \sigma_{\mathrm{G}} \cdot (1 - \epsilon_{\mathrm{G}})). \tag{8}$$

$$\begin{pmatrix} \sigma_{\mathrm{V}} \\ \sigma_{\mathrm{G}} \end{pmatrix} = \frac{1}{\mathcal{L}} \frac{1}{\epsilon_{\mathrm{G}} - \epsilon_{\mathrm{V}}} \begin{pmatrix} \epsilon_{\mathrm{G}} - 1 & \epsilon_{\mathrm{G}} \\ 1 - \epsilon_{\mathrm{V}} & -\epsilon_{\mathrm{V}} \end{pmatrix} \begin{pmatrix} N_{\mathrm{V-like}} \\ N_{\mathrm{G-like}} \end{pmatrix}. \tag{9}$$

Sources of the uncertainty:

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

  $\rightsquigarrow$ Estimated by domain difference: $\Delta\epsilon_{\mathrm{V}}$, $\Delta\epsilon_{\mathrm{G}}$.

  ($\Delta\epsilon = \epsilon$ on source $- \epsilon$ on target)

Uncertainty

The variance of the cross section measurement:

$$(\Delta\sigma_{\rm V})^2 = \underbrace{\left(\frac{\partial\sigma_{\rm V}}{\partial N_{\rm V}}\right)^2 (\Delta N_{\rm V})^2 + \left(\frac{\partial\sigma_{\rm V}}{\partial N_{\rm G}}\right)^2 (\Delta N_{\rm G})^2}_{\text{statistical}}$$

$$+ \underbrace{\left(\frac{\partial\sigma_{\rm V}}{\partial \epsilon_{\rm V}}\right)^2 (\Delta \epsilon_{\rm V})^2 + \left(\frac{\partial\sigma_{\rm V}}{\partial \epsilon_{\rm G}}\right)^2 (\Delta \epsilon_{\rm G})^2}_{\text{systematic}} \quad (10)$$

$$= \frac{1}{(\epsilon_{\rm V} - \epsilon_{\rm G})^2} \left[\underbrace{\frac{(1-\epsilon_{\rm G})^2 N_{\rm V} + \epsilon_{\rm G}^2 N_{\rm G}}{\mathcal{L}^2}}_{\text{statistical}} + \underbrace{\sigma_{\rm V}^2 (\Delta\epsilon_{\rm V})^2 + \sigma_{\rm G}^2 (\Delta\epsilon_{\rm G})^2}_{\text{systematic}} \right], \quad (11)$$

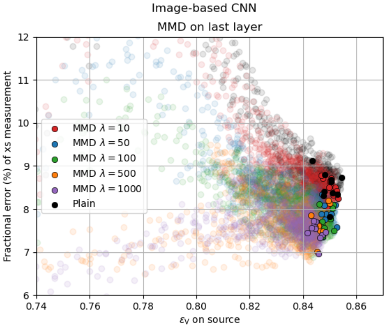
- $\mathcal{L} = 3\,\text{ab}^{-1}$,
- the working point with $\epsilon_{\rm G} = 0.2$.

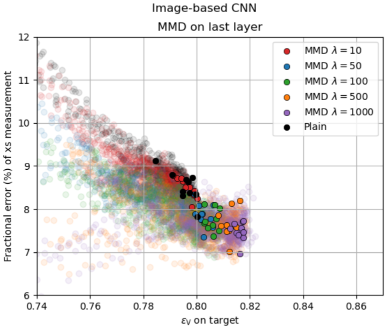
Uncertainty

Compare 10 runs for each of the following models:

- Plain
- MMD $\lambda = 10$
- MMD $\lambda = 50$
- MMD $\lambda = 100$
- MMD $\lambda = 500$
- MMD $\lambda = 1000$

Image-based CNN.



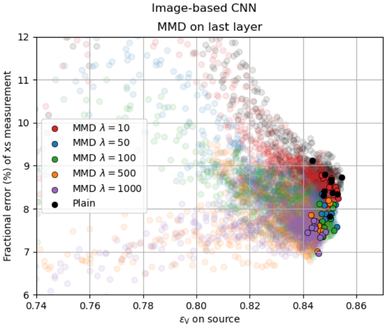


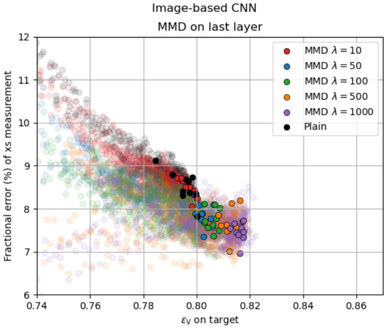
Each epoch in the training history is scattered.

The epoch with the largest source ϵ_V is marked to represent a run.

Uncertainty

Image-based CNN.

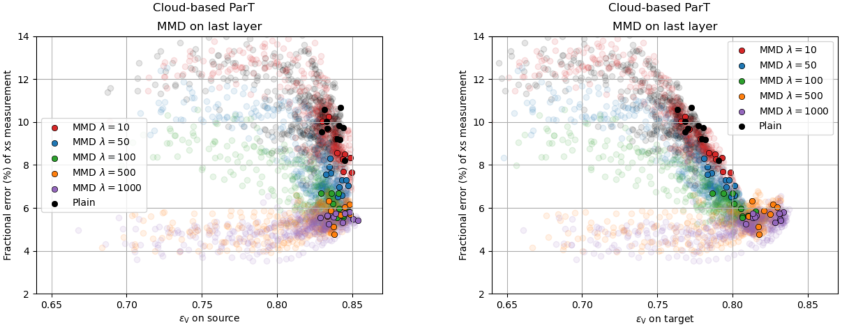




- Source ϵ_V is kept, and the uncertainty is reduced
- Target ϵ_V of MMD is even larger than the plain case.
- Uncertainty: $8\,\% - 9\,\%$ (plain) $\rightarrow$ $7\,\% - 8\,\%$ (MMD $\lambda = 1000$)

Uncertainty

Cloud-based ParT.



- Uncertainty: $9\,\% - 11\,\%$ (plain) $\rightarrow$ $5\,\% - 6\,\%$ (MMD $\lambda = 1000$)

Summary

- Systematic is an important ingredient in measurement.
- Introducing MMD helps to reduce the domain difference, and hence the systematic.
- The MMD can be readily applied to various types of networks and data representations.

Thank you for listening!

Backup - Epoch Selection

To perform epoch selection, the only viable information are

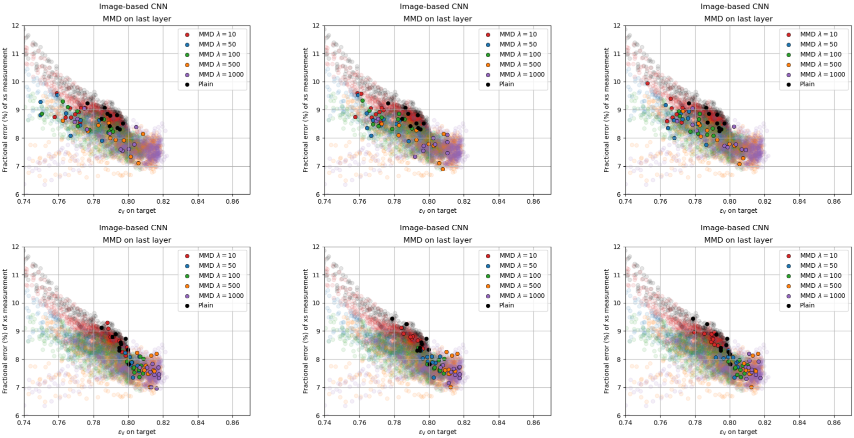
- ϵ_{V} on simulation;
- MMD between simulation and data.

Other possible selection strategies:

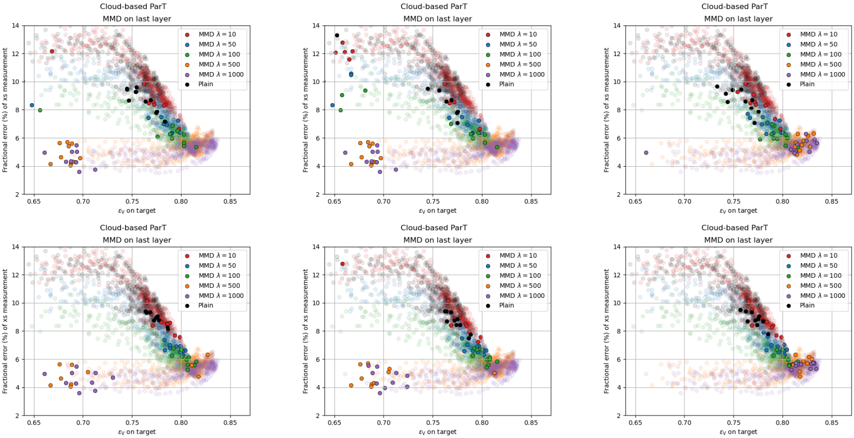
1. Selecting the epoch which has the least MMD on
   1. the third-to-last layer
   2. the second-to-last layer
   3. the last layer
2. First filter out the epochs with MMD less than 0.2 on
   1. the third-to-last layer
   2. the second-to-last layer
   3. the last layer,

   then among the remaining epochs, select the one with the highest ϵ_{V} on the source domain.

Backup - Epoch Selection



CNN: 1a (upper left), 1b (upper middle), 1c (upper right), 2a (lower left), 2b (lower middle), and 2c (lower right).

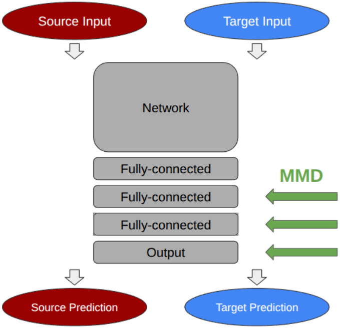


ParT: 1a (upper left), 1b (upper middle), 1c (upper right), 2a (lower left), 2b (lower middle), and 2c (lower right).

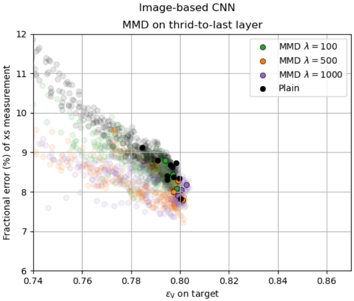
Backup - MMD Layers

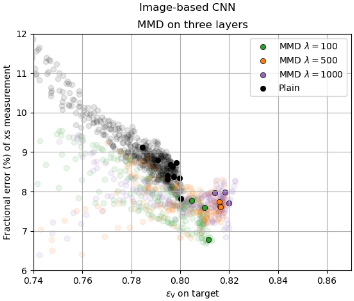
It is possible that the MMD can be applied on the hidden layers of the network.

$$\mathcal{L} = \mathcal{L}_{\text{clf}}(\text{source}) + \sum_{\ell \in \text{layers}} \lambda_\ell \cdot \mathrm{MMD}^2(\text{source}_\ell, \text{target}_\ell), \tag{12}$$



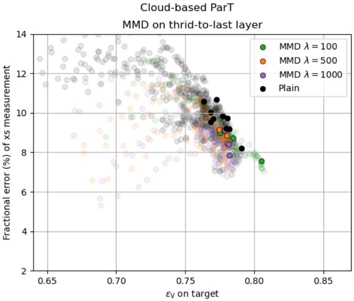
Backup - MMD Layers

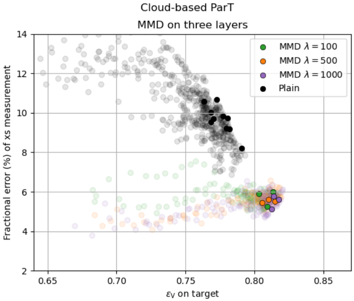




CNN: MMD is applied on the third-to-last (left), and all of the last three layers (right).

Backup - MMD Layers





ParT: MMD is applied on the third-to-last (left), and all of the last three layers (right).