Decorrelation using Optimal Transport

Malte Algren, Johnny Raine & Tobias Golling

Link to paper

Decorrelation technics

- How to decorrelated a tagger/classifier?

Prior: Reweight the conditions and retrain the classifier

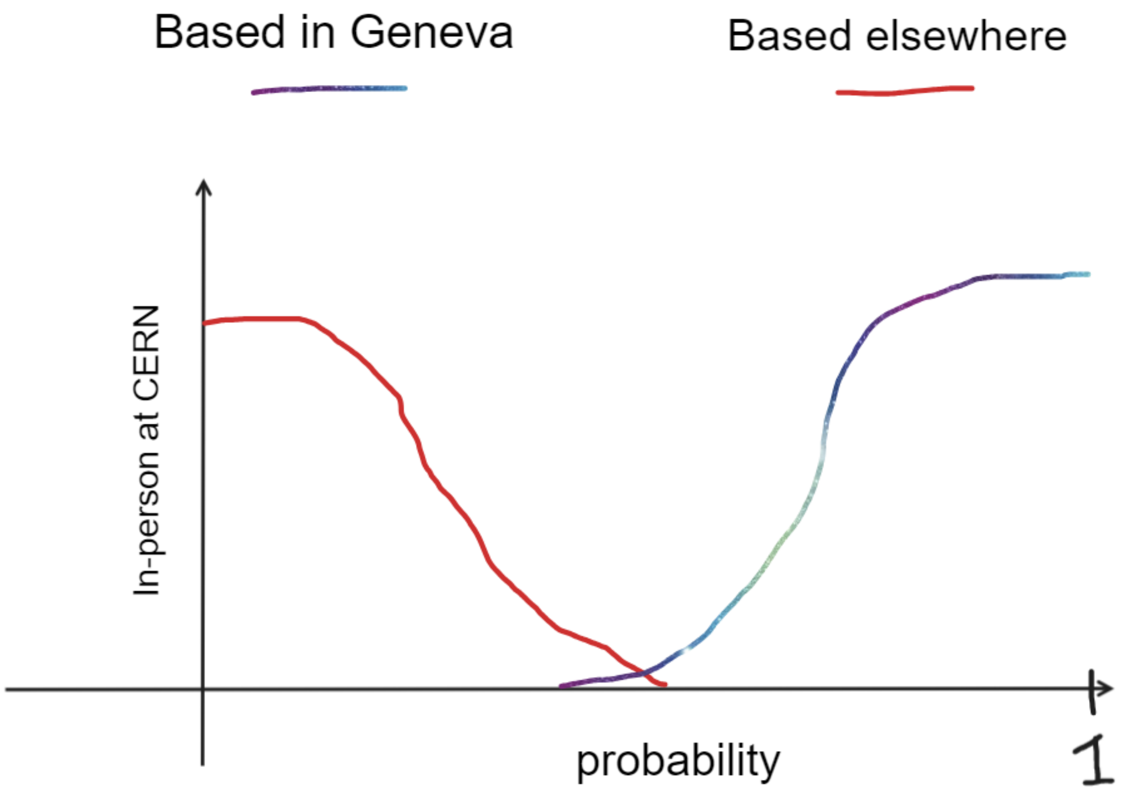
During: Penalize classifier with a correlation metric

  - https://arxiv.org/abs/2001.05310
  - https://arxiv.org/abs/2010.09745
  - Adversarial training

Post: **Monotonic transformation $T(\cdot\,|c)$ given the conditions that breaks the correlations**

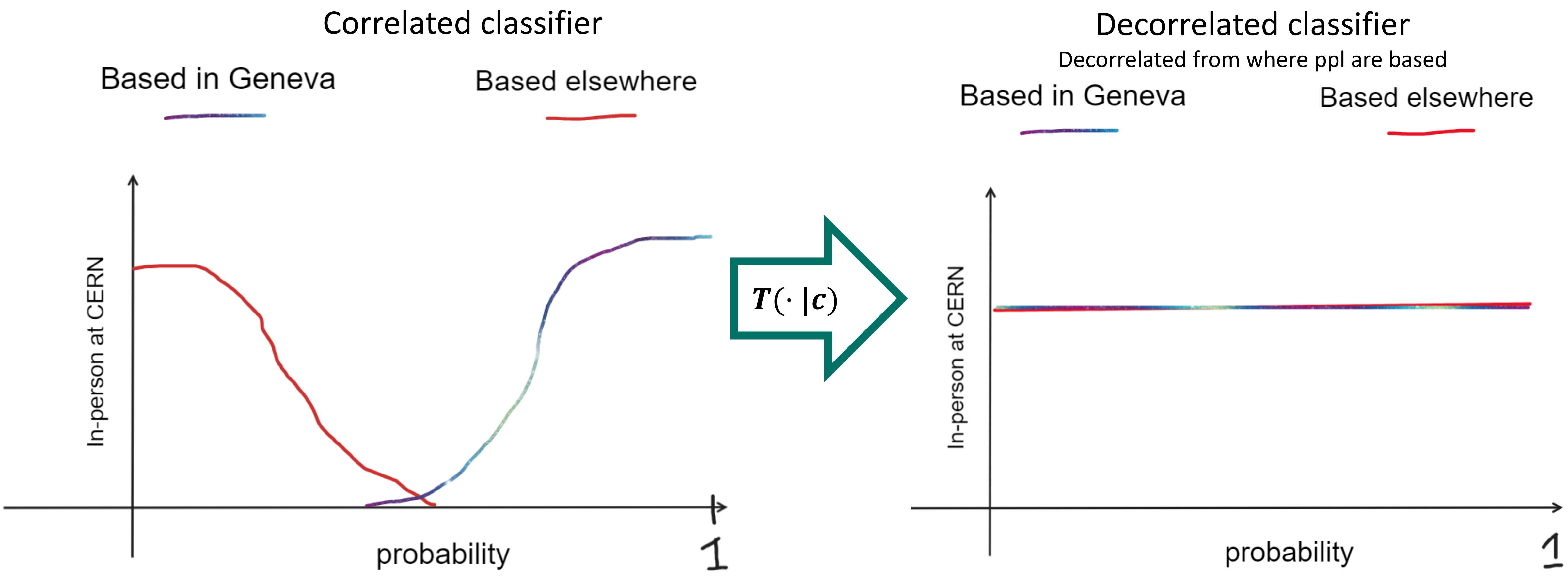
Output of correlated classifier

Correlated if based in Geneva or not



Post training decorrelation

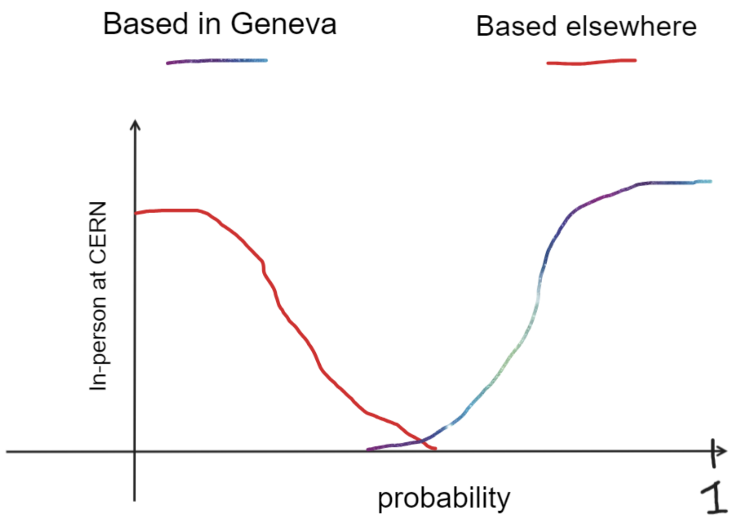
- **Monotonic transformation $T(in - person\ at\ CERN|c)$ can break the correlations**



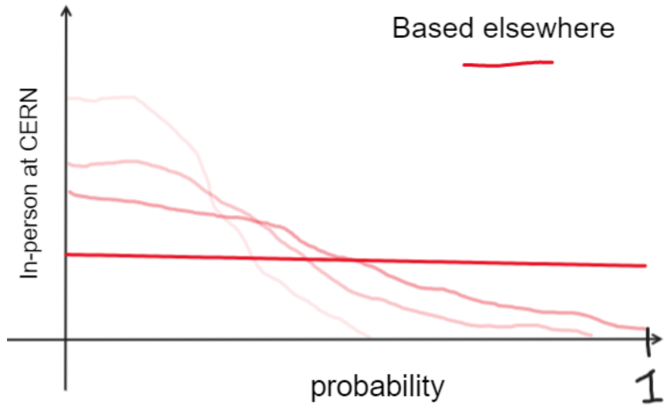
Post training decorrelation

Advantages with post decorrelation:

1. No sampling of conditions required. Train on full stat
2. Unchanged AUC as a function of conditions
3. Improved performance over existing methods



Ex: $\boldsymbol{T(in\ person\ at\ CERN|based\ elsewhere)}$ would decorrelate

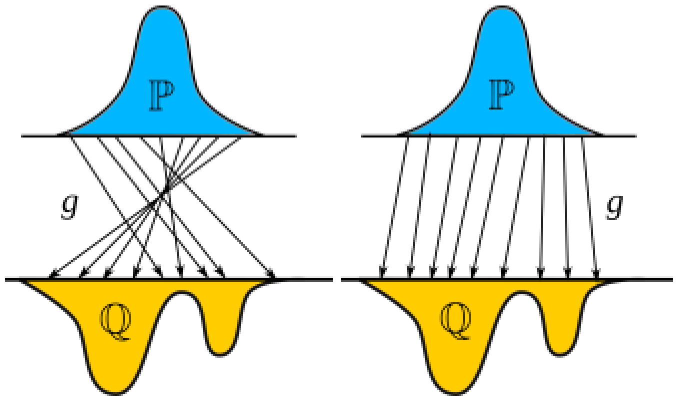


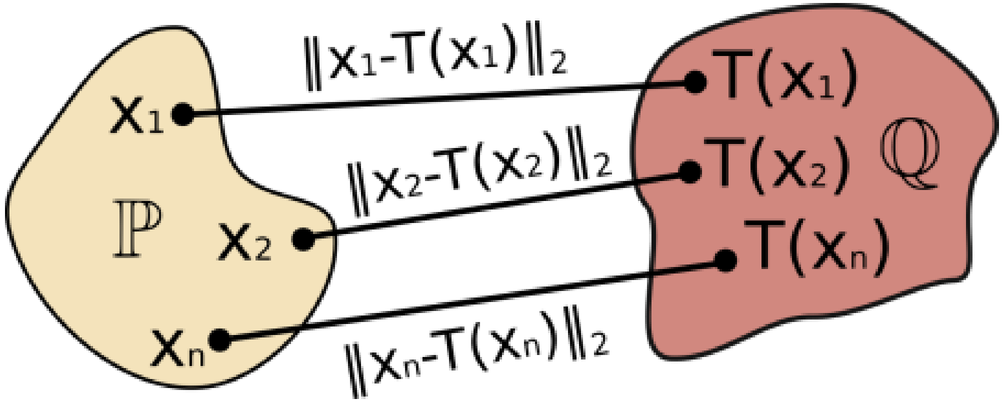
The difficult part is finding a continuous monotonic $T(\cdot\,|c)$

- *1*D: Continuous flows (*cflow*)
- *N*D: **Find optimal transport (OT) with convex neural solver**
    - Using OT we can find a continuous monotonic in *N*D

Optimal Transport

- Estimate the Wasserstein distance (w^2)
- Transformation between arbitrary distributions
- Attempt to minimize transport cost
- Scales to higher dimensions and conditions



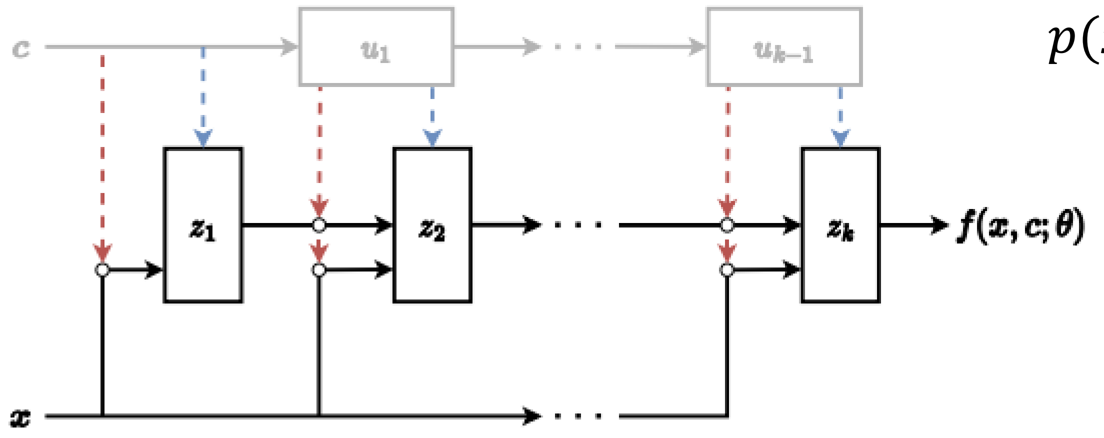


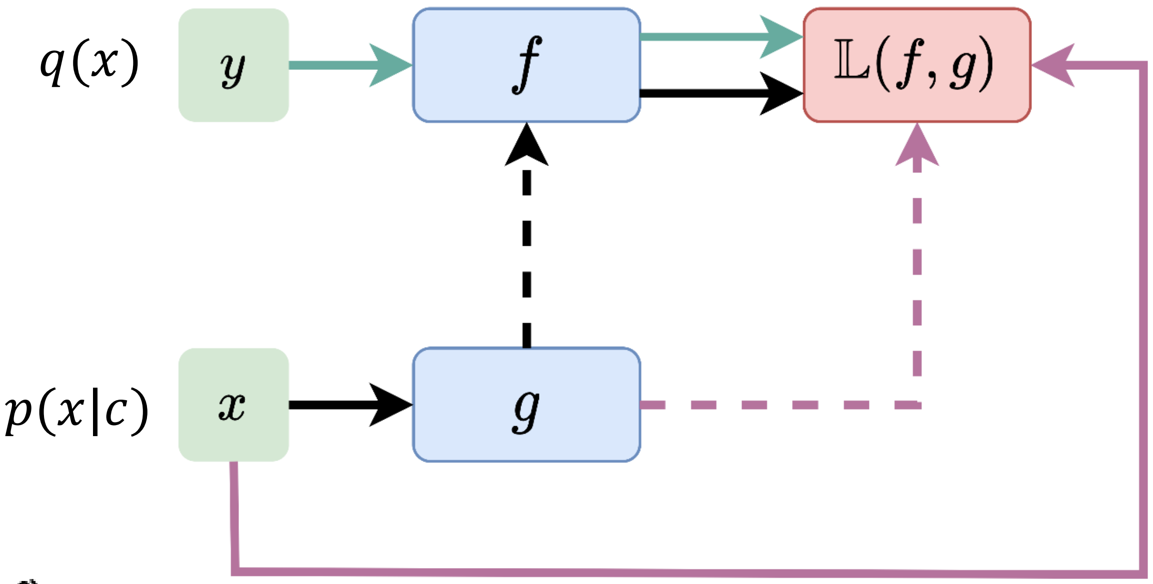
Optimal transport using convex neural solvers

Find the optimal transport between two densities

- Using Partially-input-convex neural networks (PICNN)
- Minimize the 2-Wasserstein distance between a correlated space and a decorrelated one.

PICNN





$$\mathbb{L}(f, g) = \min_f \max_g \sum f_\phi(y; \theta) + x \cdot \nabla_x g_\psi(x; \theta') - f_\phi(\nabla_x g_\psi(x; \theta'), \theta')$$ with the $\hat{T} = \nabla_x g(x; \theta')$

* Colors in $\mathbb{L}(\phi, \psi)$ indicate the flow of information in *Optimal transport architecture*

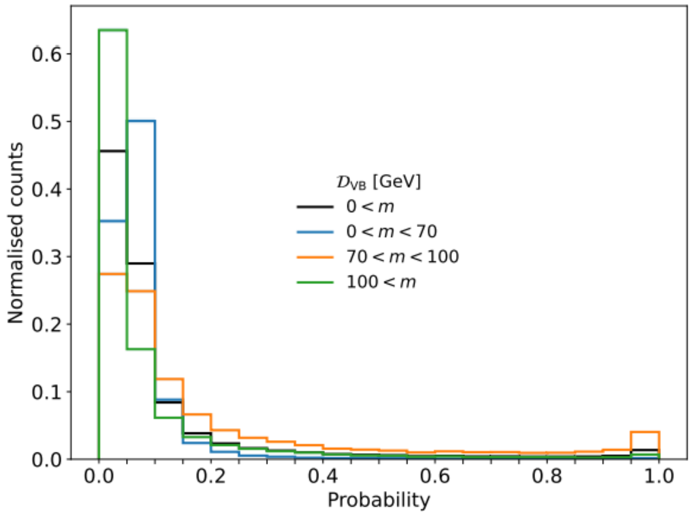
Method/Training

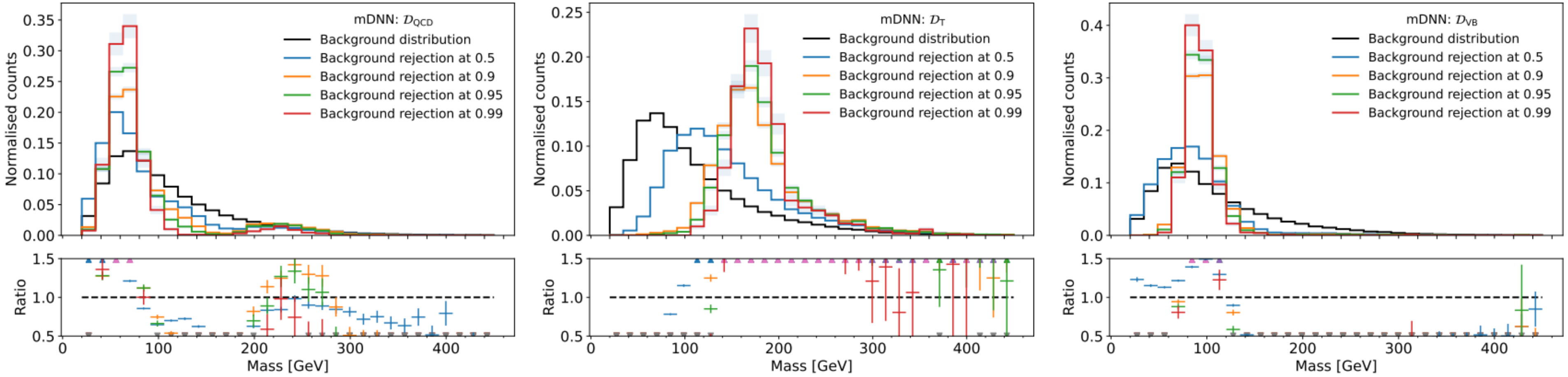
1. Train classifier to get *D*(pQCD,pTOP, pVB)

   - Probability change as a function of mass

   - Mass sculpting after a selection

$$\mathcal{D}_{\rm QCD} = \frac{p_{QCD}}{p_T + p_{VB}}, \quad \mathcal{D}_{\rm T} = \frac{p_T}{p_{QCD} + p_{VB}}, \quad \mathcal{D}_{\rm VB} = \frac{p_{VB}}{p_T + p_{QCD}}.$$



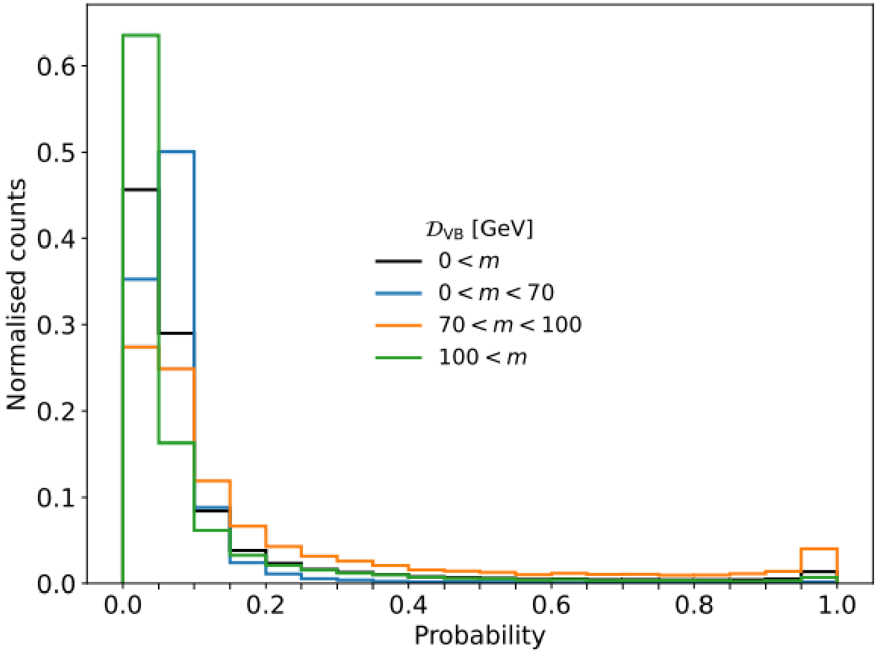


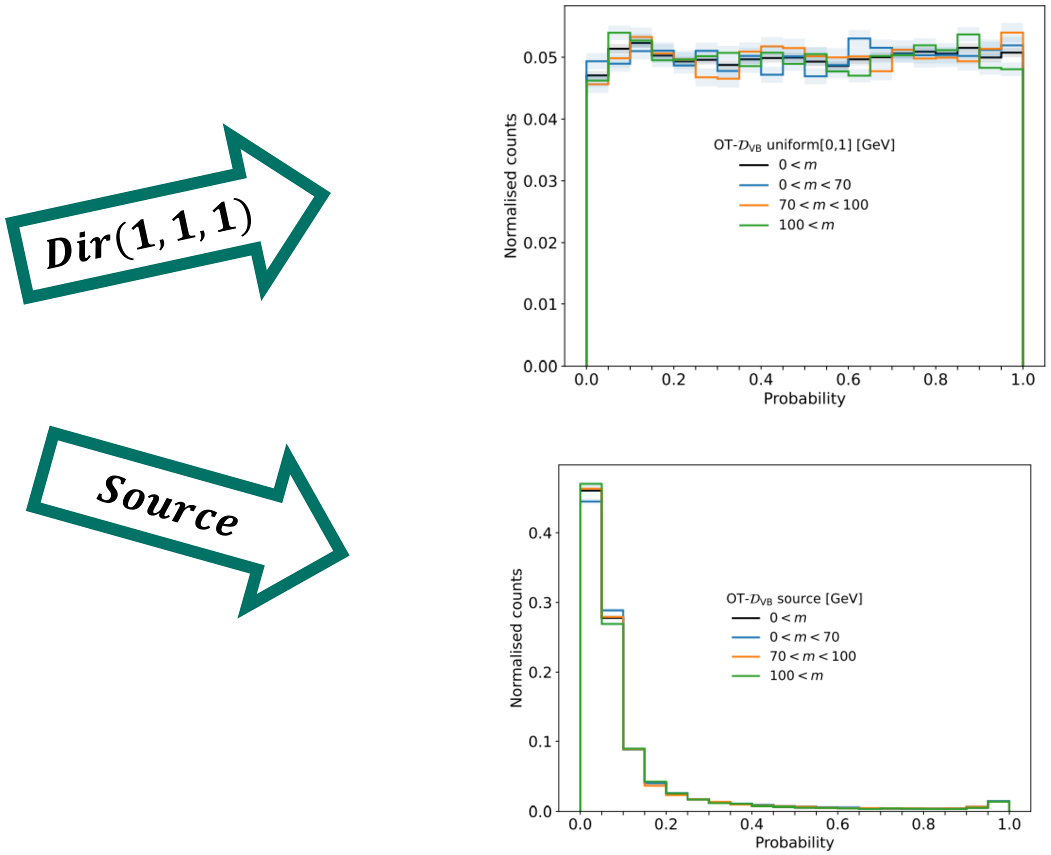
(a) $\mathcal{D}_{\rm QCD}$ projection (b) $\mathcal{D}_{\rm Top}$ projection (c) $\mathcal{D}_{\rm VB}$ projection

Method/Training

1. Train classifier to get D(pQCD,pTOP, pVB)

2. Train decorrelation methods on only QCD events for $\boldsymbol{T}$(**pQCD,pTOP, pVB| mass**)



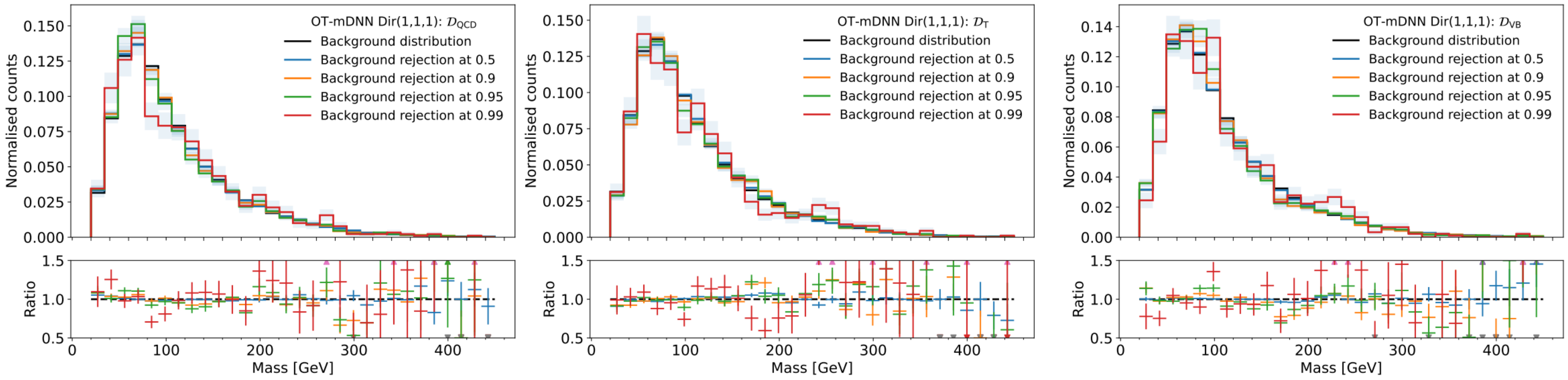


Method/Training

1. Train classifier to get *D*(pQCD,pTOP, pVB)

2. Train decorrelation methods on only QCD events for $\boldsymbol{T}$(**pQCD,pTOP, pVB| mass**)

   - Mass sculpting around resonance has been removed



(a) $\mathcal{D}_{\text{QCD}}$ projection (b) $\mathcal{D}_{\text{Top}}$ projection (c) $\mathcal{D}_{\text{VB}}$ projection

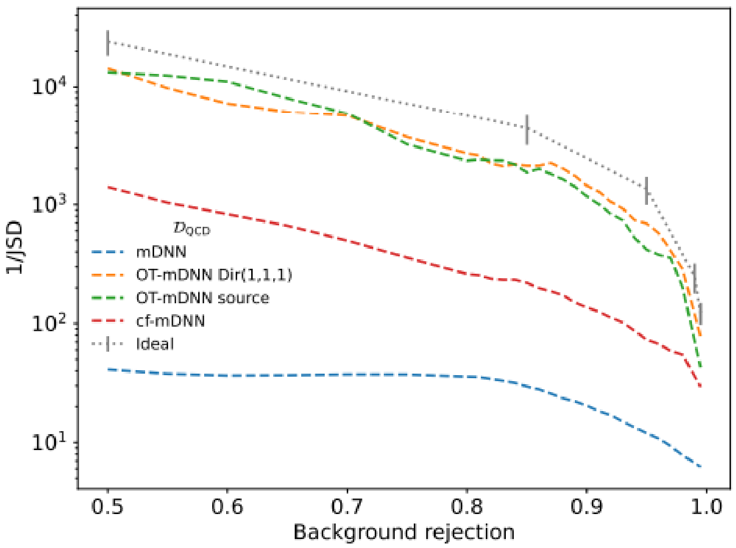
Method/Training

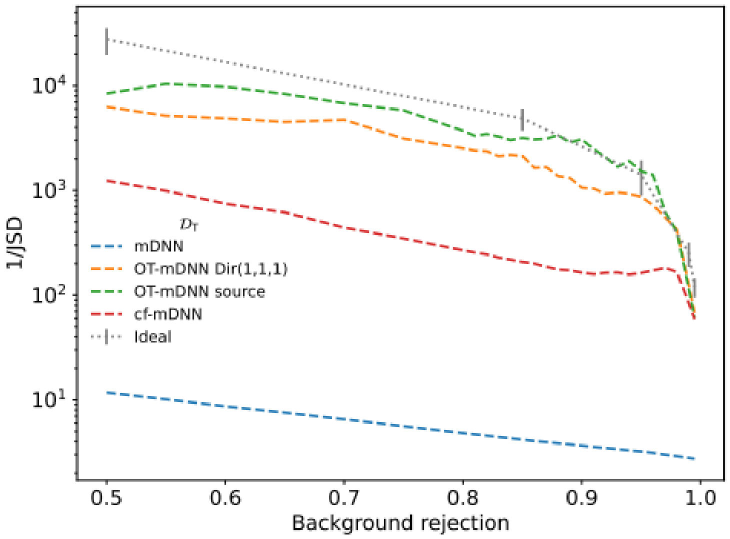
1. Train classifier to get *D*(pQCD,pTOP, pVB)

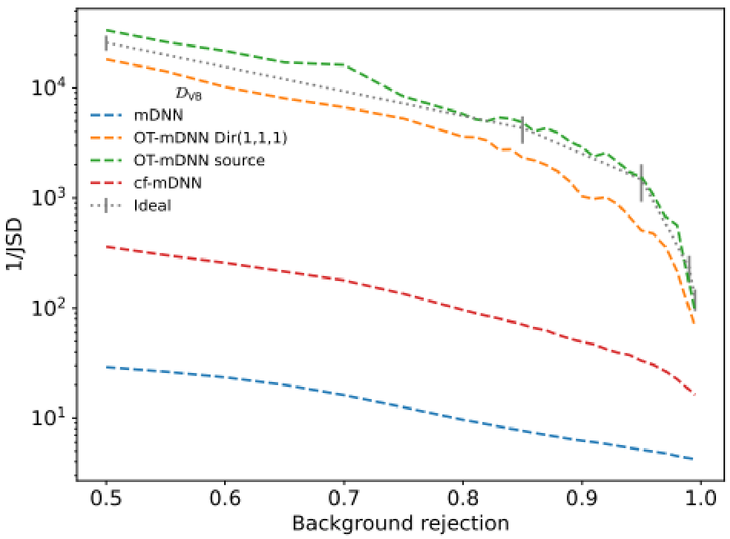
2. Train decorrelation methods on only QCD events for $\boldsymbol{T}$(**pQCD,pTOP, pVB| mass**)

   - Mass sculpting around resonance has been removed

   - Use 1/JSD to measure sculpting (*Higher the better*)







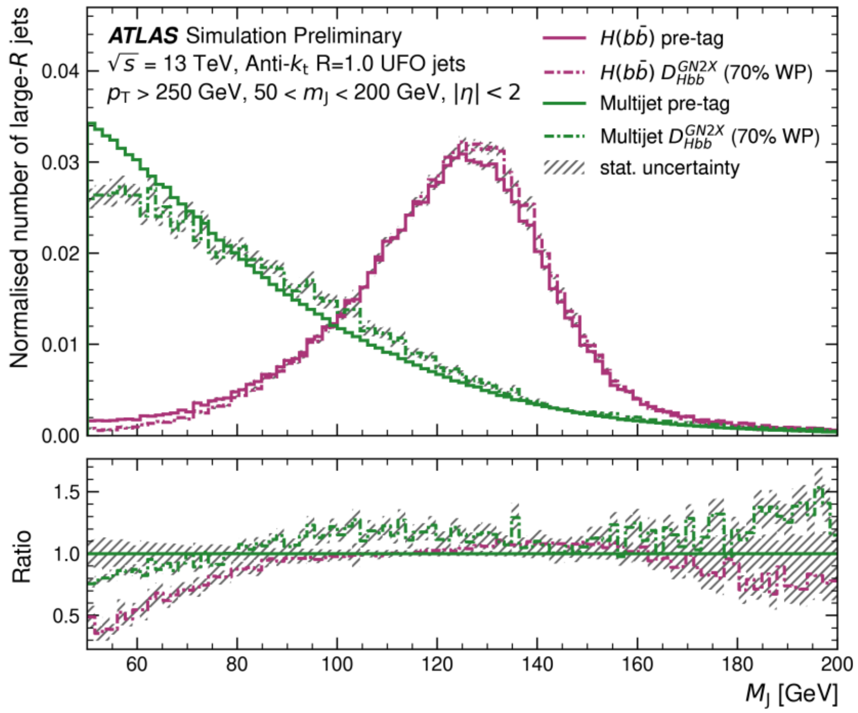
Outlook

Summary:

- OT can find a conditional monotone transformation between two spaces
- This transformation can be used to decorrelate a continuous feature space
- It shows good decorrelation performance and removed mass sculpting
- Works as a post-processing on the correlated probabilities

For more details on the method, see here.

Nail



From ATL-PHYS-PUB-2023-021

Hbb tagger output: (**phbb, phcc, ptop, pqcd**)

Backup

Transportation Theory

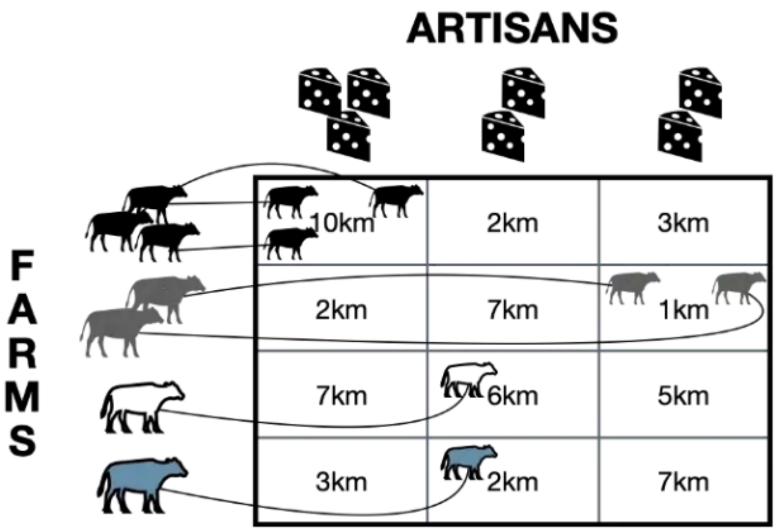
- Transportation theory suggested by Monge in 1781 (no mass splitting)
- Kantorovich relaxation (allowed mass splitting)
- Brenier's theory ($c(x, y) = (x - y)^2$ & $\chi = y = \mathbb{R}^d$)

The optimal transport is the gradient of a convex function

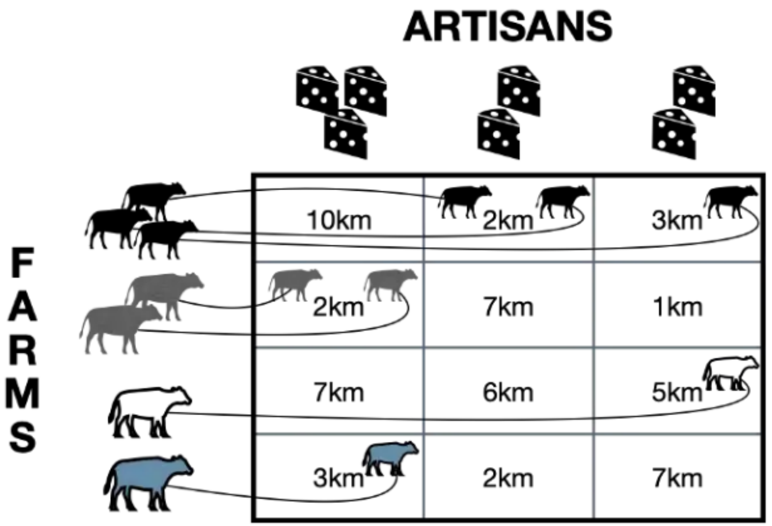
Monge formulation

$$\hat{T}(x) = \operatorname*{arginf}_{T(x):p'(x')\equiv q(x)} \int dx\, p(x)\, c[x, T(x)]$$

Monge Transport theory



Kantorovich Transport theory



Optimal transport

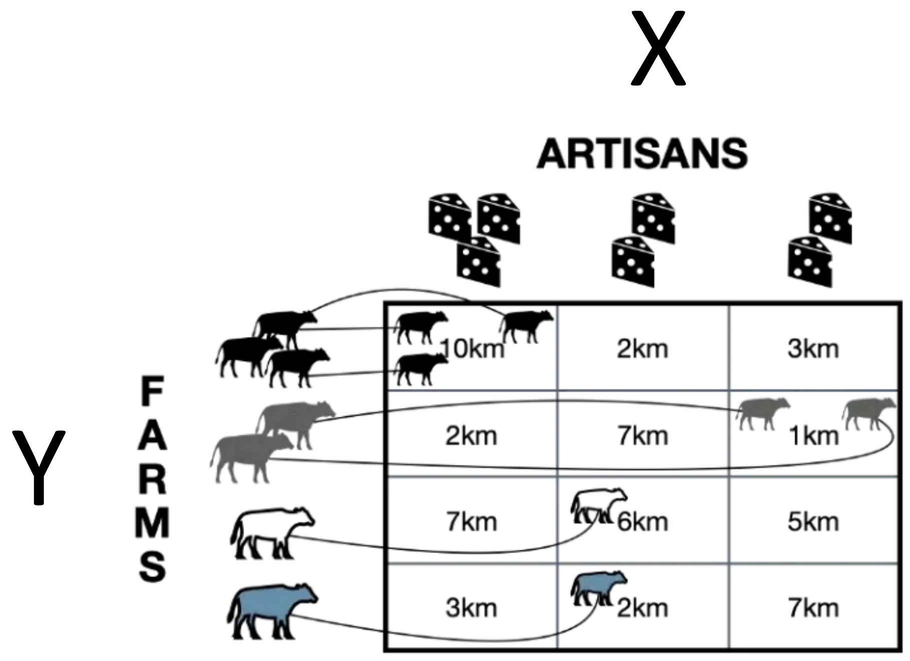
- Monge primal problem

$$\inf_{T(x):p'(x')\equiv q(x)} \int dx\, p(x)\, c[x, T(x)]$$

- Dual problem with constraint $\phi(x) - \psi(y) \leq c(x, y)$ (The outsourcing problem)

$$\sup\left(\int_X \phi(x)d\mu(x) - \int_Y \psi(y)dv(y)\right)$$

- Due to $\phi(x) - \psi(y) \leq c(x, y)$, we can substitute with convex conjugate and with some magic we end up at:



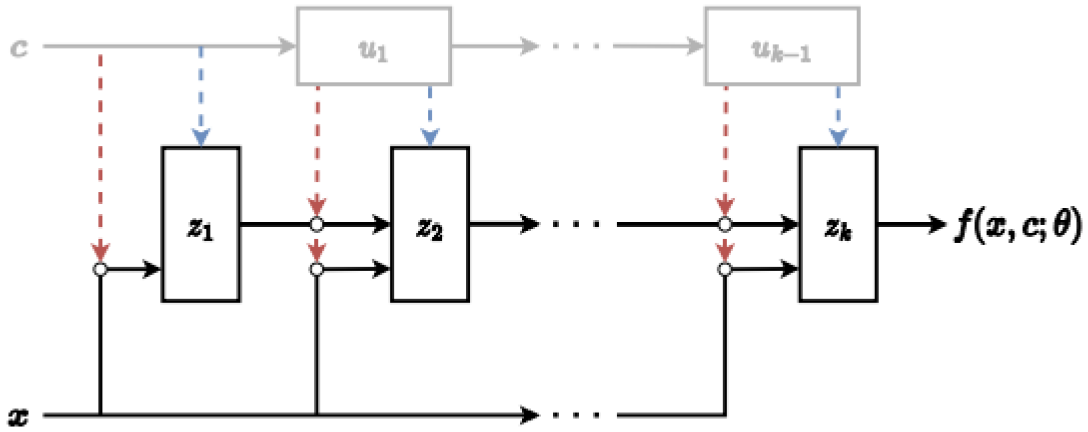
$$\mathbb{L}(\phi, \psi) = \min_f \max_g \sum f_\phi(y;\theta) + x \cdot \nabla_x g_\psi(x;\theta') - f_\phi(\nabla_x g_\psi(x;\theta'), \theta')$$ with the $\hat{T} = \nabla_x g(x;\theta')$

- MinMax problem over two convex neural networks

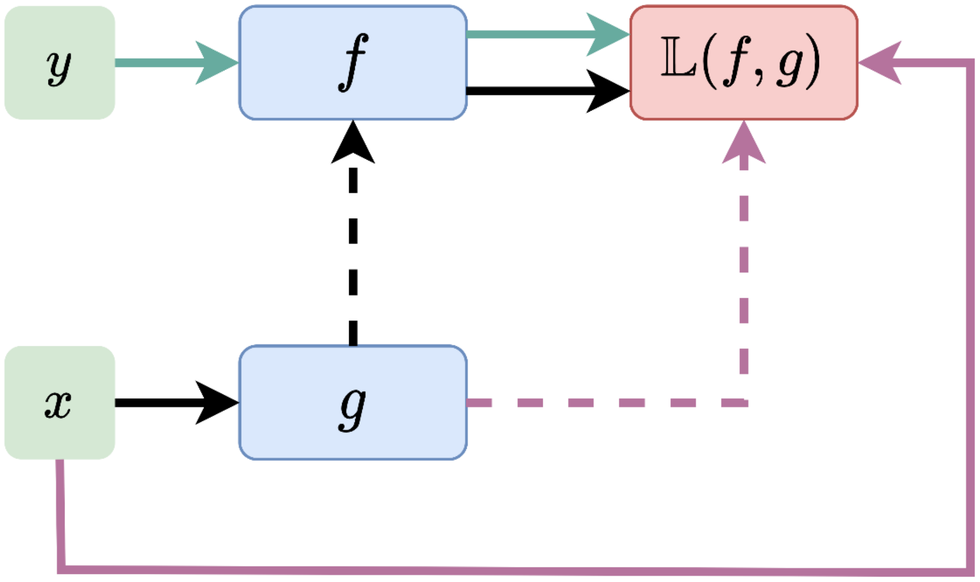
Optimal transport & ML

Overview of Optimal transport:

- f & g has to be convex (see figure below)
- The conditional distributions θ & θ' are required to have the same PDF



Optimal transport architecture



x/y can also contain conditions θ/θ'

$$\mathbb{L}(\phi, \psi) = \min_f \max_g \sum f_\phi(y; \theta) + x \cdot \nabla_x g_\psi(x; \theta') - f_\phi(\nabla_x g_\psi(x; \theta'), \theta')$$ with the $\hat{T} = \nabla_x g(x; \theta')$

* Colors in $\mathbb{L}(\phi, \psi)$ indicate the flow of information in *Optimal transport architecture*

Mass correlation for boosted Higgs tagging

From ATL-PHYS-PUB-2023-021

- Mass sculpting on the Hbb tagger (see figure from paper =>)
- Hbb tagger output: (**phbb, phcc, ptop, pqcd**)
- We can fix this with our decorrelation method:
  - Decorrelate: $\boldsymbol{T}$(**phbb, phcc, ptop, pqcd | mass**)