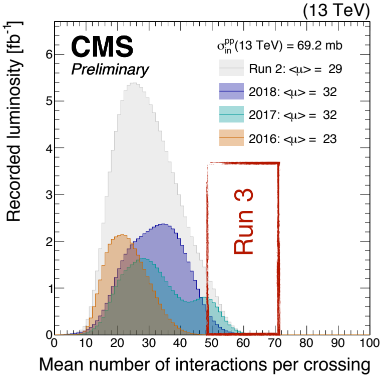
Optimal transport solutions for pileup mitigation at hadron colliders

L. Gouskos [1] **F. Iemmi** [2] S. Liechti [4] B. Maier [1]
V. Mikuni [3] H. Qu [1]

[1]European Organization for Nuclear Research (CERN), Geneva
[2]Institute of High Energy Physics (IHEP), Beijing
[3]National Energy Research Scientific Computing Center (NERSC), Berkeley
[4]University of Zurich (UZH), Zurich

ML4Jets2022, New Brunswick, USA
based on arXiv:2211.02029

PU mitigation at hadron colliders



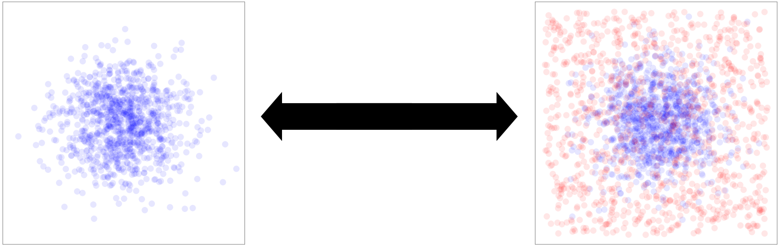
- **Pileup**: additional pp collisions superimposing to main collision
- **PU** has **increased** in Run3 ($\langle nPU \rangle = 50$) and will increase in HL-LHC ($\langle nPU \rangle = 140$)
- Will severely **degrade quality of observables** (jet multiplicity, jet substructure, ...) if not properly treated
- **Easy task for charged** particles: use tracking information to disentangle particles
- **Very challenging for neutral** particles

Overview of PU mitigation techniques

- Currently in use (e.g., CMS): **PUPPI** [1407.6013]
  - **Rule-based** algorithm
  - For each neutral particle, consider the energy of neighboring particles
  - Extract a probability for the particle to be LV or PU
  - Relies on properties of charged particles and **extrapolates to neutrals**
- Nature and complexity of task inspired **machine-learning-based approaches**
  - PUMML: treat jets as images, reconstruct LV neutral radiation [1707.08600]
  - Semi-supervised PUPPI: train on charged, apply on neutrals [2203.15823]
- **Recurring problem**: lack of truth **"labels" for neutrals**
- We developed a **new ML-based approach** to overcome this bottleneck
  - Use Attention-Based Cloud Network (ABCNet, [2001.05311]) combined with optimal transport
  - **TOTAL**: **T**raining **O**ptimal **T**ransport with **A**ttention **L**earning
  - Train model on a Delphes-based simulation of the CMS Phase2 detector

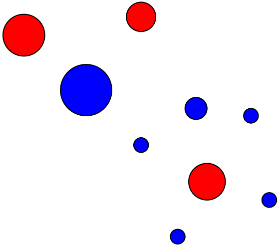
A novel approach to PU mitigation

- Definition of truth labels is highly non trivial in simulations at hadron colliders
- **Our approach**: simulate **identical** proton-proton **collisions in two scenarios**
  1. Only the hard interaction is simulated: **no-PU sample**
  2. Pileup is superimposed to the hard interaction: **PU sample**
- **Do not assign per-particle labels**: rather just assign a "global" label to samples
- Train network to **learn differences between the two samples**



How to learn: OT concepts for a loss function

- We build a custom **loss inspired by optimal transport** ideas (OT)



- OT example: the **Earth Mover's Distance** is the minimum work to move **earth** to fill some **holes**

$$EMD(\vec{x}, \vec{y}) = \min_{f} W(f, \vec{x}, \vec{y})$$

- With OT you can **match distributions** (e.g., earth-holes)
- We want to match the distributions for the no-PU particles and **PU particles weighted by an ABCNet weight** ($\vec{\omega}$)

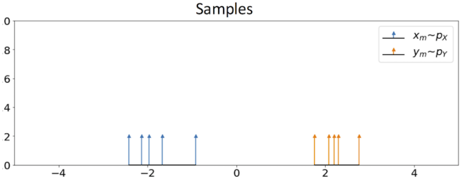
Efficient OT: sliced Wasserstein distance (SWD)

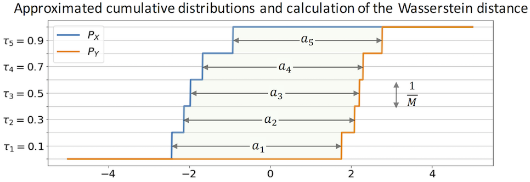
- The optimal transport problem has a **closed form for 1D problems**:

$$W_c(p_X, p_Y) = \int_0^1 c\left(P_X^{-1}(\tau), P_Y^{-1}(\tau)\right) \mathrm{d}\tau$$

where p_X, p_Y are 1D PDFs and $P_X^{-1}(\tau), P_Y^{-1}(\tau)$ are the respective CDFs

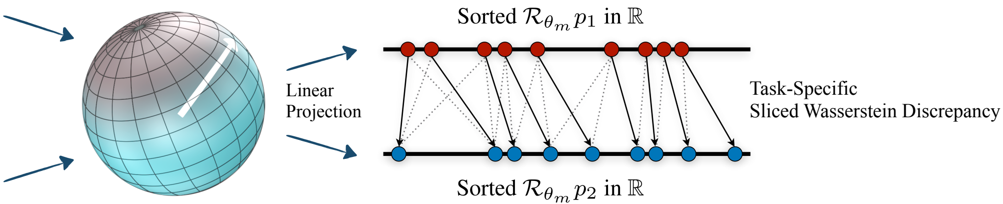
- If we only have samples from the distributions, $x \sim p_X, y \sim p_Y$ the task becomes evens simpler
- The problem is reduced to a sorting problem
- **Fast and easy to solve**





Efficient OT: sliced Wasserstein distance (SWD)

- **SWD**: take our *n*-D feature space and **project (slice)** it **to 1D**
- Project on a vector belonging to S^{n-1}
- For robustness, take **multiple random slices**
- Now can **solve the 1D OT problem for each slice**
- **Sort particles by slice**
- The **average SWD on all slices and particles** becomes the **loss function**



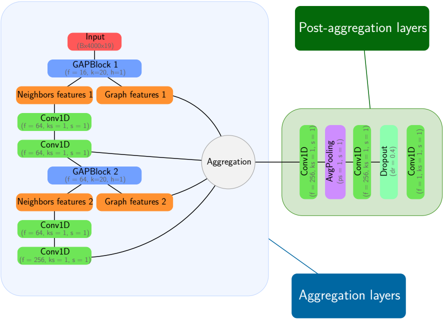
Energy conservation in OT: MET constraint

- SWD focuses on the optimal matching between individual particles in no-PU and PU samples
  - No guarantee that energy is conserved between the two
- Add an **event-level MET constraint** term to the loss
  - Enforce energies in no-PU and PU events to be similar
- Final loss function:

$$\mathcal{OT} = \mathrm{SWD}(\vec{x}_p \cdot \vec{\omega}, \vec{x}_{np}) + \mathrm{MSE}\left(\mathrm{MET}(\vec{x}_p \cdot \vec{\omega}) - \mathrm{MET}(\vec{x}_{np})\right)$$

where $\vec{x}_p$ = PU sample; $\vec{x}_{np}$ = no-PU sample; MSE = mean squared error

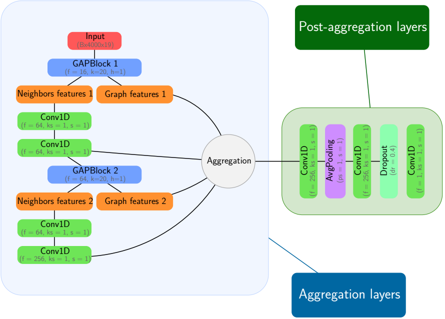
The model



- **9 input features**:
  - (p_T, η, ϕ, E)
  - Charge
  - PDG ID
  - dXY & dZ impact parameters
  - PUPPI weight
- **Loss**: $\mathrm{SWD}(\vec{x}_p \cdot \vec{\omega}, \vec{x}_{np}) + \mathrm{MET}$ constraint
- **Sliced features**: (p_T, η, ϕ, E)
- **Output**: per-particle weight $\vec{\omega}$
- **Optimizer**: Adam

- Train on **300k events**, equally split between QCD multijet, $t\bar{t}$ dileptonic and VBF Higgs(4ν) processes
- Consider **9000 particles per event** (zero-padding included)
- Gather the **20 *k*-nearest neighbors** for each particle when building graph

The model



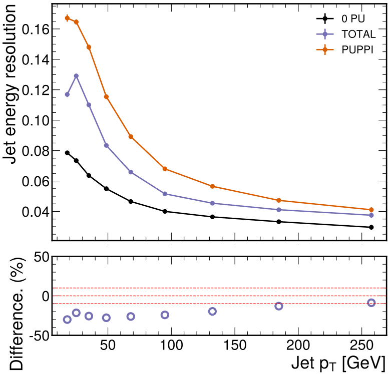
- **Compare** TOTAL **with PUPPI and no-PU** scenario
- **Reweight** each particle's 4-momentum by the network weight
- **Cluster** TOTAL jets and TOTAL MET

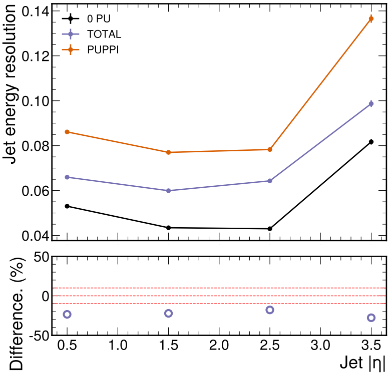
- We define the resolution as:

$$\delta = \frac{q_{75\%} - q_{25\%}}{2}$$

where $q_{X\%}$ is the X-th quantile of the considered response distribution

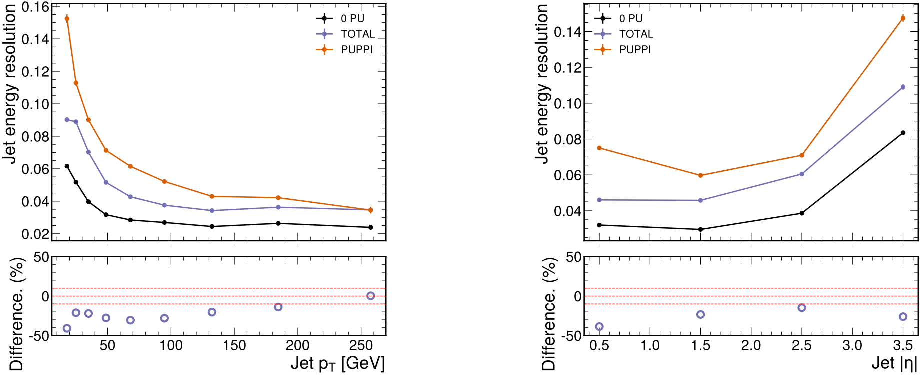
Results: QCD multijet





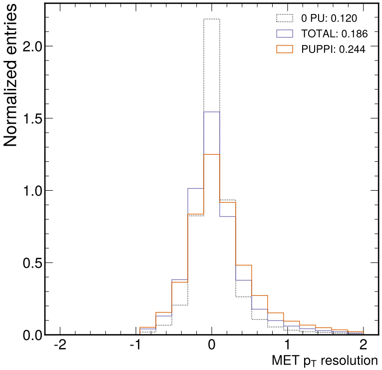
- Jet energy resolution as a function of jet p_T (left) and jet η (right)
- Improvement up to 30% in JER, up to 25% in η resolution

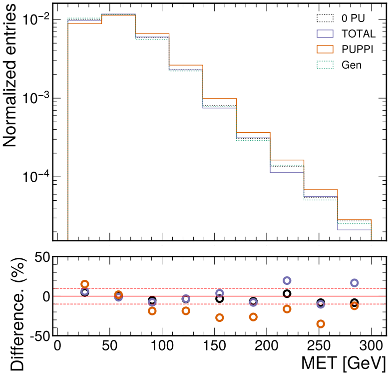
Results: dileptonic $t\bar{t}$



- Jet energy resolution as a function of jet p_T (left) and jet η (right)
- Improvement up to 40% in JER, up to 40% in η resolution

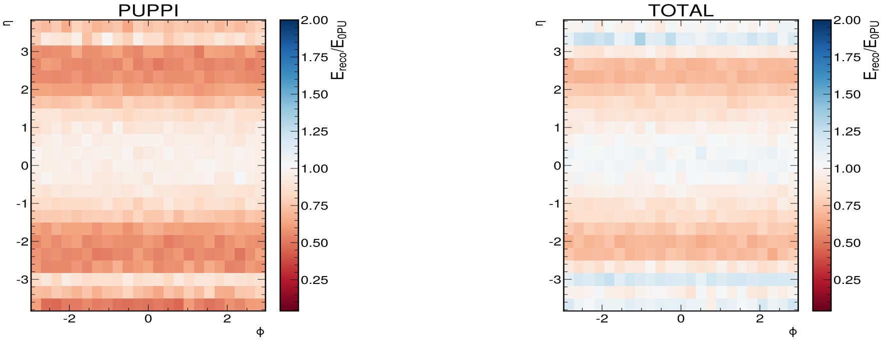
Results: dileptonic $t\bar{t}$





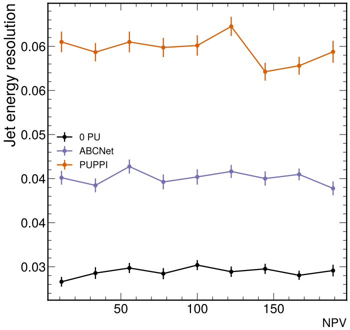
- MET p_T resolution (left); MET p_T distribution (right)
  - MET resolution is reduced by 24%

Inspecting particles weights



- Ratio $\frac{p_T \times \omega}{p_{T,\text{noPU}}}(\eta, \phi)$ for PUPPI and TOTAL (right) in QCD multijet events
- Smoother behavior for TOTAL in central and forward region
- Still room for improvement in transition region

Robustness



- Evaluate resolution on **processes and PU scenarios unseen during training**
- Network is trained on QCD+$t\bar{t}$+VBF with $\langle \mathrm{NPV} \rangle = 140$
- Evaluate on W+jets production, flat NPV between 0 and 200

Conclusions

- We presented **novel algorithm to reject PU particles** at high-intensity hadron colliders
  - Trained and tested on Delphes simulation of Phase2 CMS detector
- We are Training Optimal Transport with Attention Learning: **TOTAL**
- **We do not rely on explicit, per-particle labeling**
- Such an algorithm will be **crucial at the High-Luminosity LHC**, where much harsher data-taking conditions are expected
- Our **approach can be generalized** to a wide range of denoising problems
  - Only needed input is a reliable simulation of signal and noise
- Can this method be generalized to statistically independent samples?
  - Simulation-free pileup rejection?