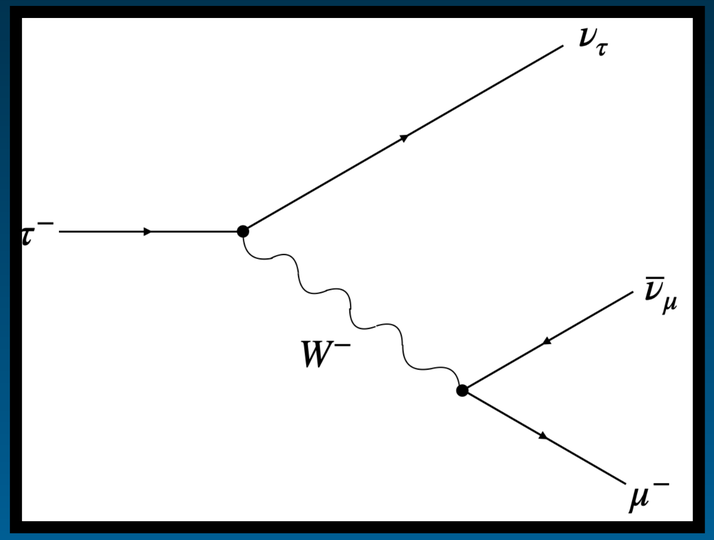
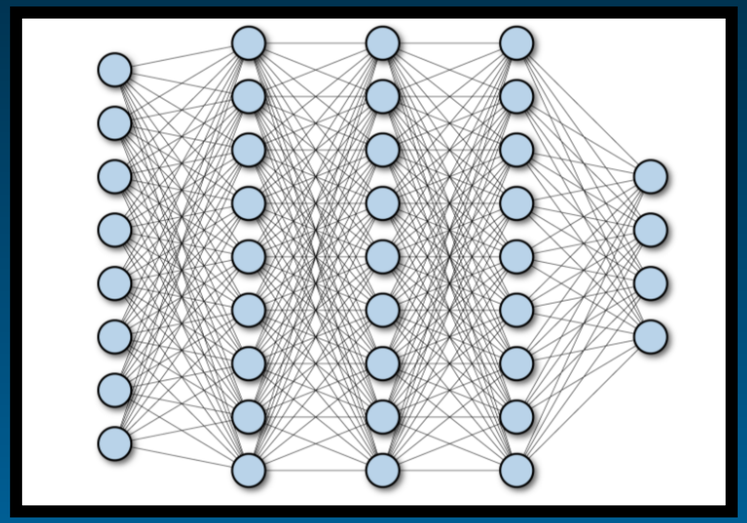
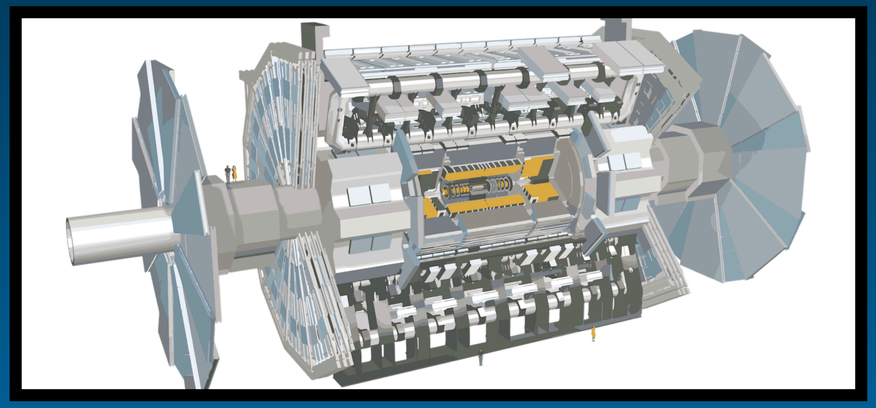
Studies on the Separation of Prompt and Tau Muons with the ATLAS Detector using Machine Learning Methods



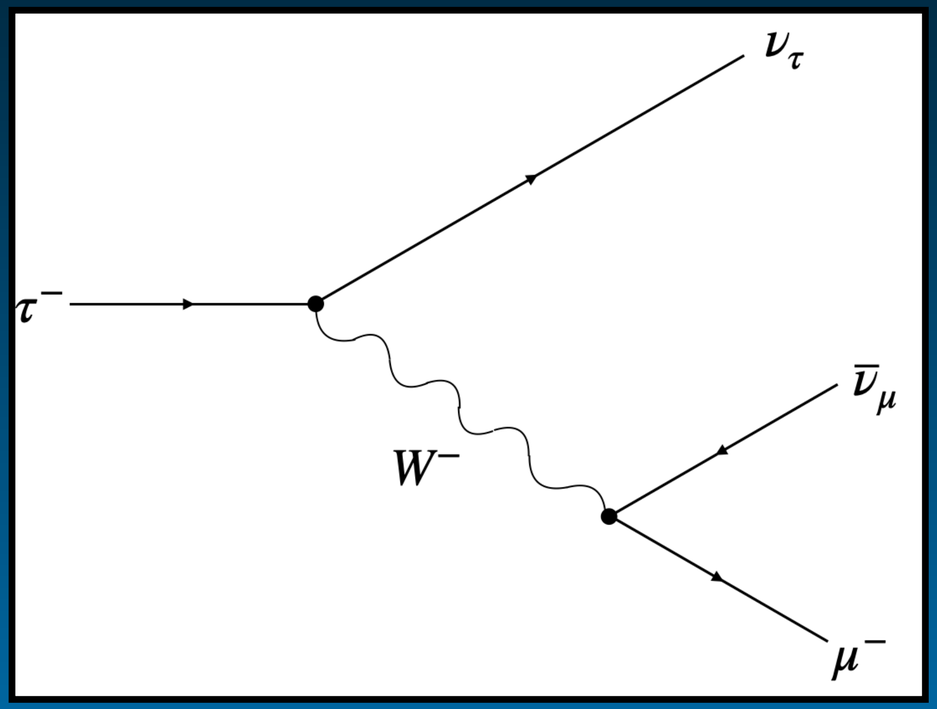


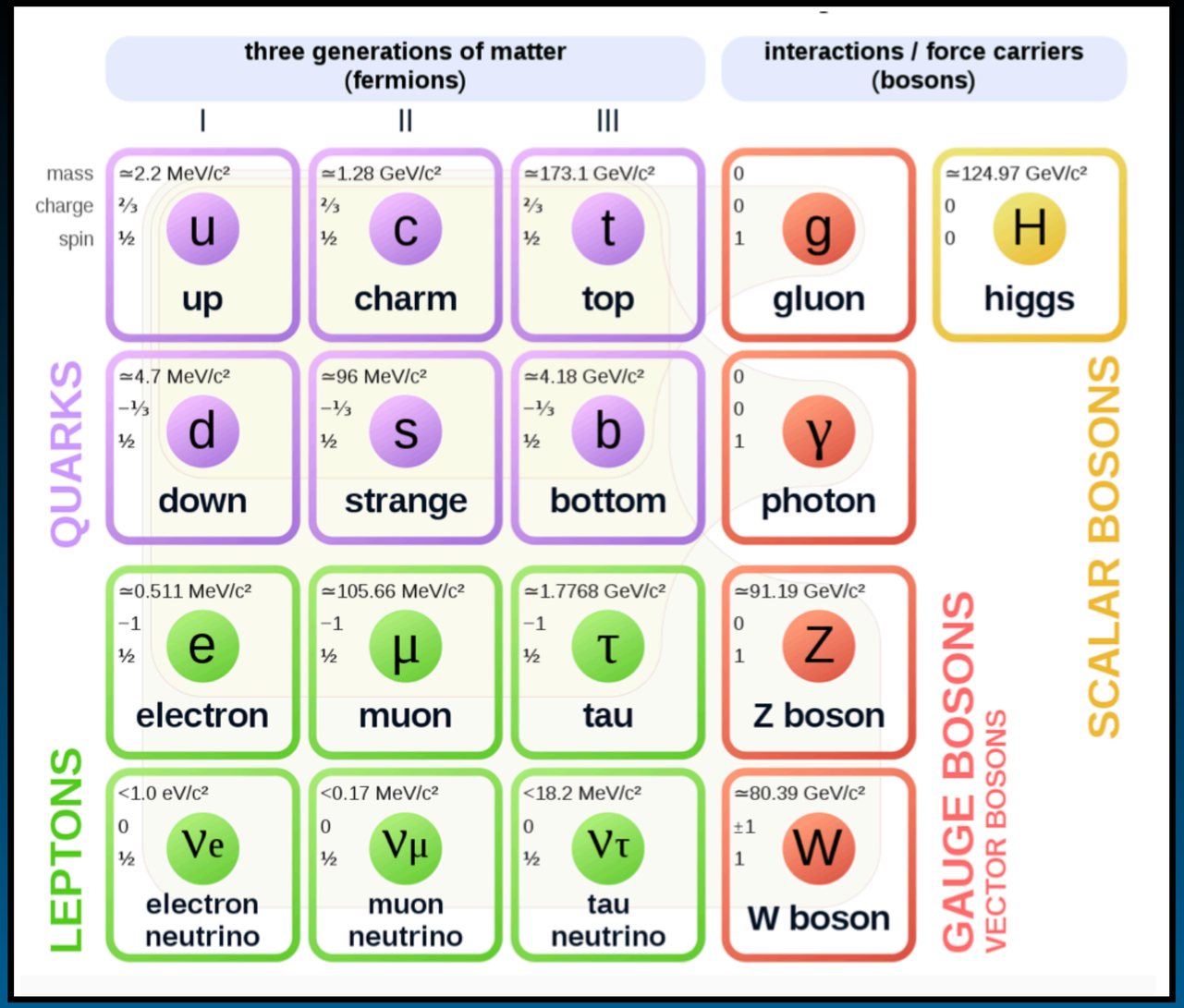


Physics at the LHC Seminar, Daniel Ortmann, 7/25/2022

Outline

1 Motivation





- Process of interest: decay of the tau lepton τ into the muon μ (mediated by the W boson)
- Conservation of lepton family number $\longrightarrow$ neutrinos in the final state

1 Motivation

- Muons originating from tau decays ($W, Z \to \tau + \ldots \to \mu + \ldots$), referred to as *tau muons* - are reconstructed as *prompt muons* ($W, Z \to \mu + \ldots$)
- Tau decay length $l_\tau \approx 0.09\text{mm} \longrightarrow \tau$ is not directly trackable by detector
- How can we distinguish between those ?
- Closest approach of the muon track to the beam spot perpendicular to the beam axis: d_0
- Current analyses use only d_0 to distinguish between tau and prompt muons $\longrightarrow$ use other variables as well
- Potential application in several analyses:
  - Verification of the Lepton-Flavour Universality: $\mathrm{R}(\tau/\mu) = \mathrm{B}(W \to \tau\nu_\tau) \,/\, \mathrm{B}(W \to \mu\nu_\mu) \approx 1$
    - Measure $\mathrm{B}(W \to \tau\nu_\tau \to \mu\nu_\mu\nu_\tau\nu_\tau)$ and $\mathrm{B}(W \to \mu\nu_\mu)$ using a dedicated classifier
    - Extrapolate $\mathrm{B}(W \to \tau\nu_\tau \to \mu\nu_\mu\nu_\tau\nu_\tau)$ to $\mathrm{B}(W \to \tau\nu_\tau)$ using the branching ratio of taus to muons $\approx 1/6$
  - Application in the search for BSM physics (supersymmetry in particular)

2 Approach and Data

- Approach: Supervised Learning with deep neural networks (DNNs)
- Datasets: Muons satisfying the *Medium* WP and originating from different Monte-Carlo simulated processes (with POWHEG-BOX interfaced with PYTHIA 8 as Event-Generator):
  - $t\bar{t} \rightarrow bb + qq' + \ell\nu_\ell$ and $bb + \ell\nu_\ell + \ell'\nu_{\ell'}$ $\quad (\ell = \tau, \mu)$
  - $Z \rightarrow \ell\ell$
  - $W \rightarrow \ell\nu_\ell$
  - $W^* \rightarrow \tau\nu_\tau$ $\longrightarrow$ in order to generate a large amount of high-p_T tau muons
- No isolation WP is used to suppress fake muons, since they are optimized for prompt but not tau muons
- Signal: *tau muon* μ_τ
- Background: *prompt muons* μ and *fake muons* (objects misclassificd as muons or non-prompt muons, e.g. hadron decay in jets)
- No distinction between fake and prompt muons is made (tertiary classification was tested and no improvement was observed)

3 Feature Variables and Classifier Architecture

- Eight features were selected for their distinctiveness between tau muons and background

Impact parameter related	Transverse momentum related	Others
$\lvert d_0 \rvert$	p_{T}	$\lvert \eta \rvert$
Δz_0	$\frac{\Delta p_{\mathrm{T}}}{p_{\mathrm{T}}}$	Calorimeter Isolation
$\lvert z_0 \sin\theta \rvert$	ρ'	

$$\Delta z_0 = |z_{\mathrm{PV}} - z_0|$$

$$\rho' = \frac{p_{\mathrm{T}}^{\mathrm{ID}} - p_{\mathrm{T}}^{\mathrm{MS}}}{p_{\mathrm{T}}^{\mathrm{CB}}}$$

$$\eta = -\ln\tan(\frac{\theta}{2})$$

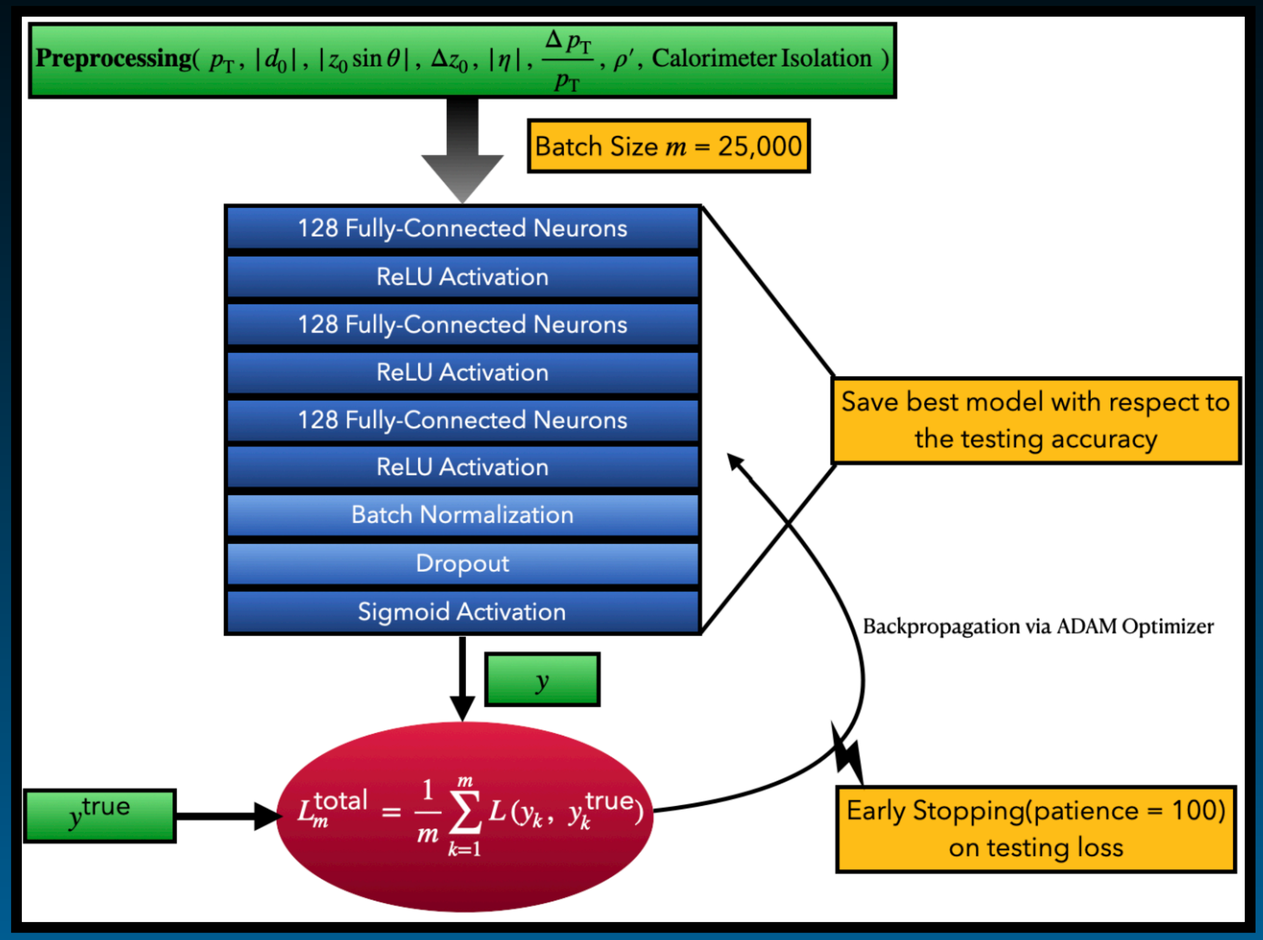
3 Feature Variables and Classifier Architecture

- DNN is implemented using *Keras* a library on top of *TensorFlow*
- Preprocessing: transform each input variable x_j on same scale (mean $\bar{x}_j$ and standard deviation σ_j over all samples)

$$x_j \longmapsto x'_j = \frac{x_j - \bar{x}_j}{\sigma_j}$$

- 3 layers containing 128 (fully-connected) neurons
- Introduce non-linearities: ReLU(x) = max(0, x)
- Batch normalization after and Dropout on last neuron layer
- Sigmoid function $S(z) = \frac{1}{1+e^{-z}} \longrightarrow$ squeeze values into [0, 1]
- Binary crossentropy loss

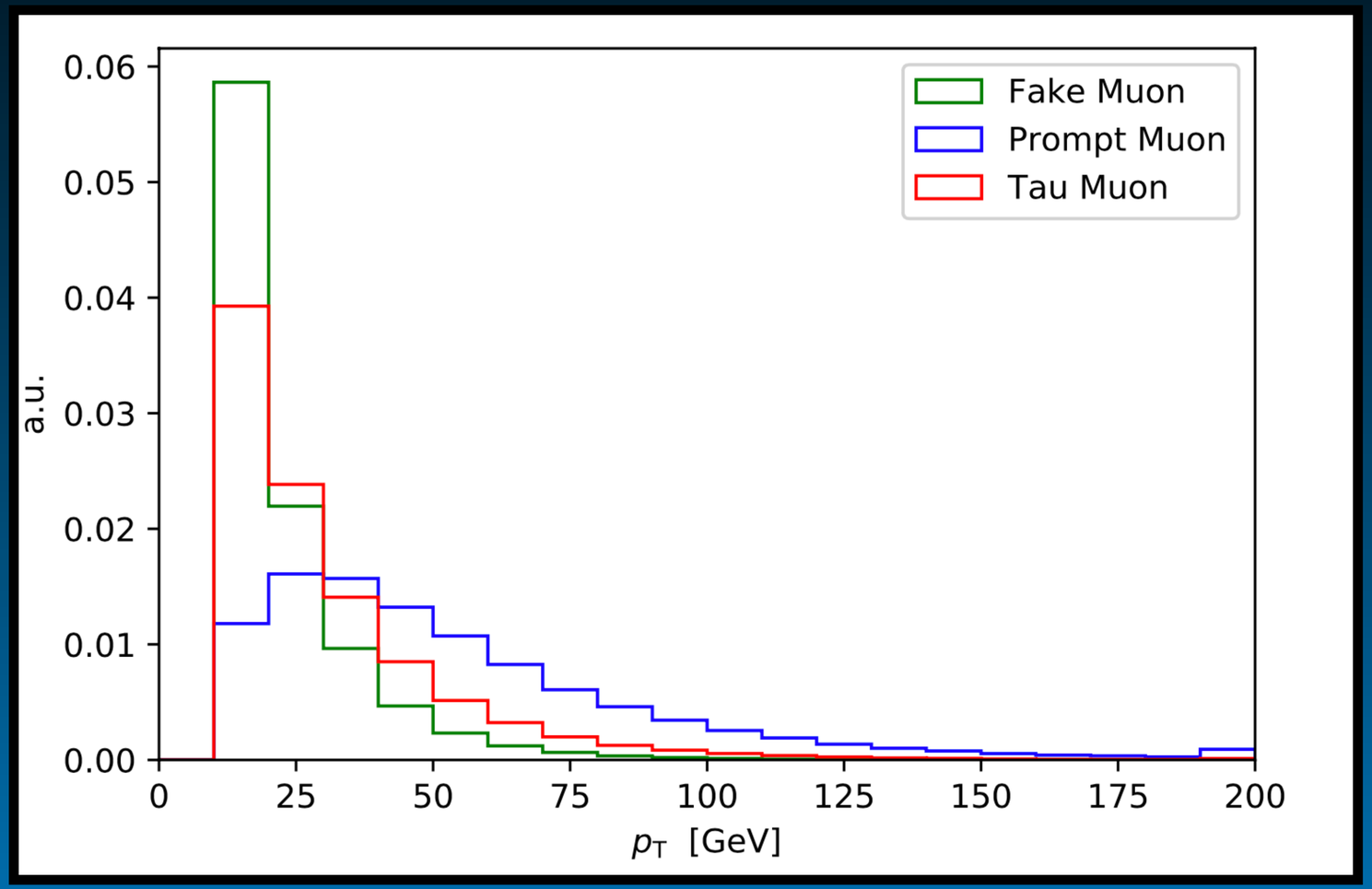
$$L(y, y^{\text{true}}) = -\, y^{\text{true}} \log(y) - (1 - y^{\text{true}}) \log(1 - y)$$



- $t\bar{t}$ dataset is split up into
  - Train set: used for training
  - Test set: evaluated during training to monitor possible overfitting
  - Evaluation set: evaluated after training for analysing the performance of the trained classifier

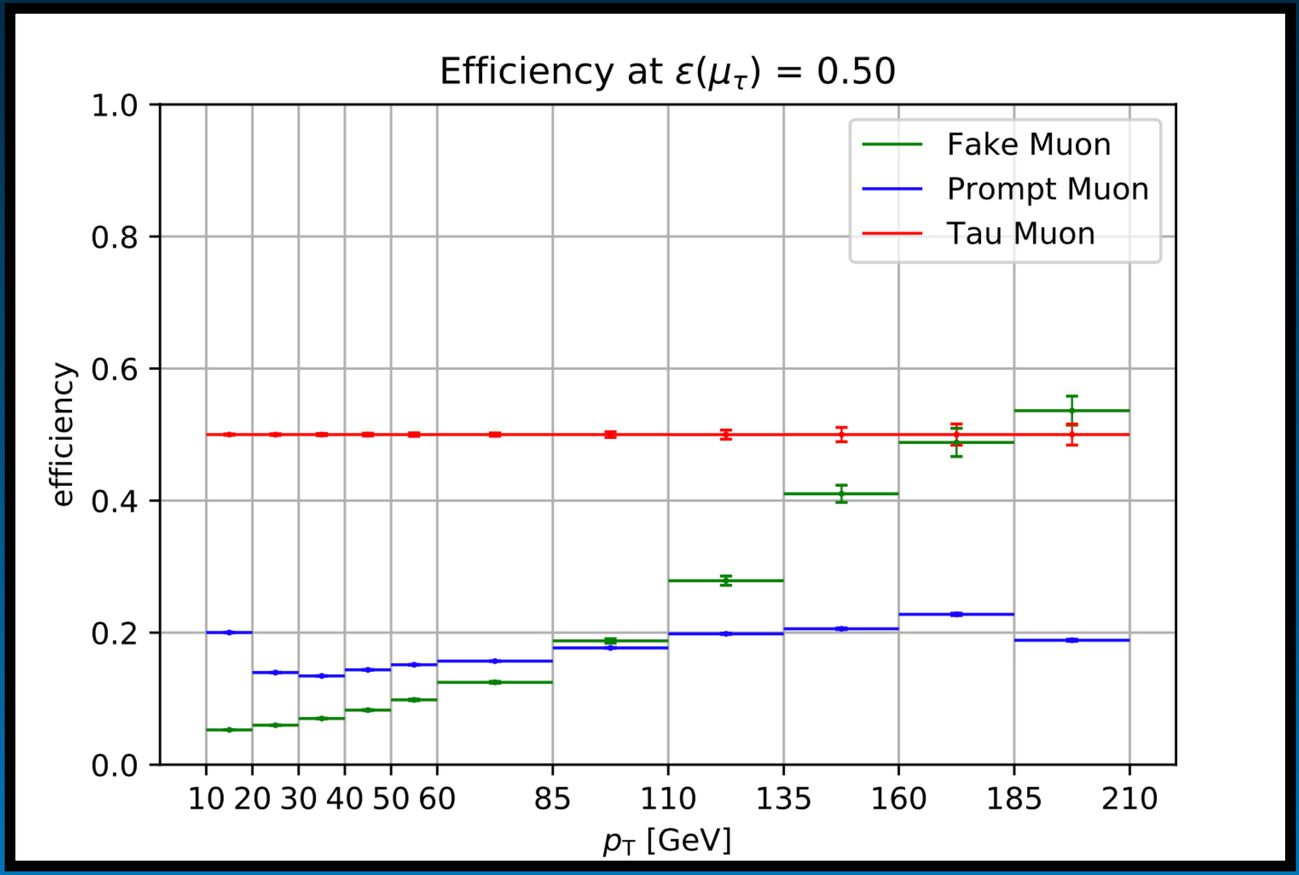
4 Training Set Bias in p_T

- Most tau muons are at low p_T and the high-p_T regime is dominated by prompt muons ⟶ high-p_T muons are with high probabilty classified as prompt muons



4 Training Set Bias in p_T

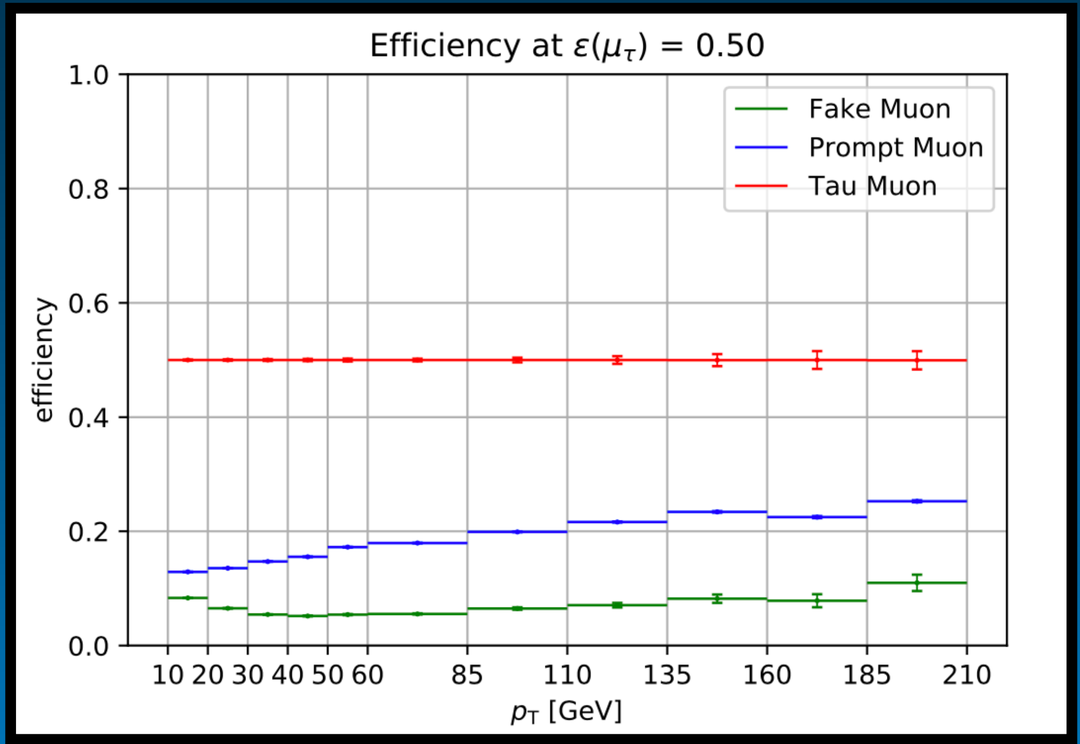
- As an illustration, train a classifier without a dedicated handling of this bias
- Sample 250,000 tau muons and 125,000 prompt and fake muons from $t\bar{t}$, then split into train (75 %) and test dataset (25 %)
- Remaining $t\bar{t}$ samples are defined to be evaluation set
- Use efficiency as performance measurement, i.e. the fraction of samples per class surviving an output score threshold
- Use fixed signal efficiency of 50 % for better comparison



- Here the bias, i.e. signal efficiency drop at high p_T, is equivalent of an increasing high-p_T prompt muon efficiency
- Small amount of high-p_T fake muons ⟶ classifier ignores them resulting in a high fake muon efficiency
- We need a training strategy that copes with that!
- Three strategies were used:
  - Flat sampling during batch calls
  - Sample weights
  - Distance correlation

5 Flat Sampling during Batch Calls

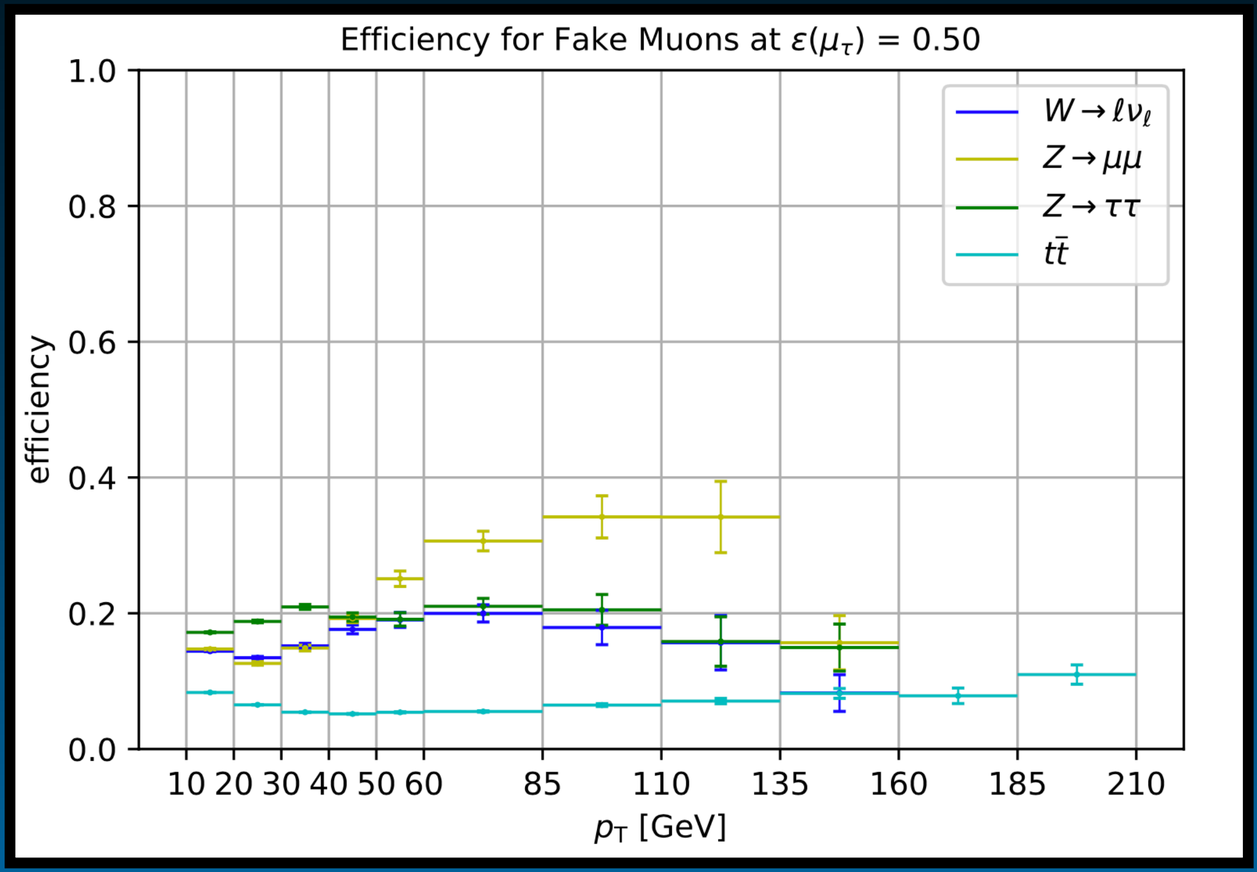
- Approach: Use a distribution flat in the entire p_T regime for the training set, to give high-p_T tau muons equally importance compared to background muons
- Flat sampling during batch calls: full availabe $t\bar{t}$ dataset is split into train and test set, where the latter is defined to be also the evaluation set
- During train time and at each epoch: a different subset is sampled from the train set that is flat in p_T and on which the DNN is trained on $\longrightarrow$ improve generalization power
- 10,000 tau and 5,000 prompt as well as fake muons are sampled from 5 GeV p_T-bins at each epoch, last bin is an overflow bin [195,∞]
- If a subset does not contain enough muons of a particular category, muons may be selected multiple times

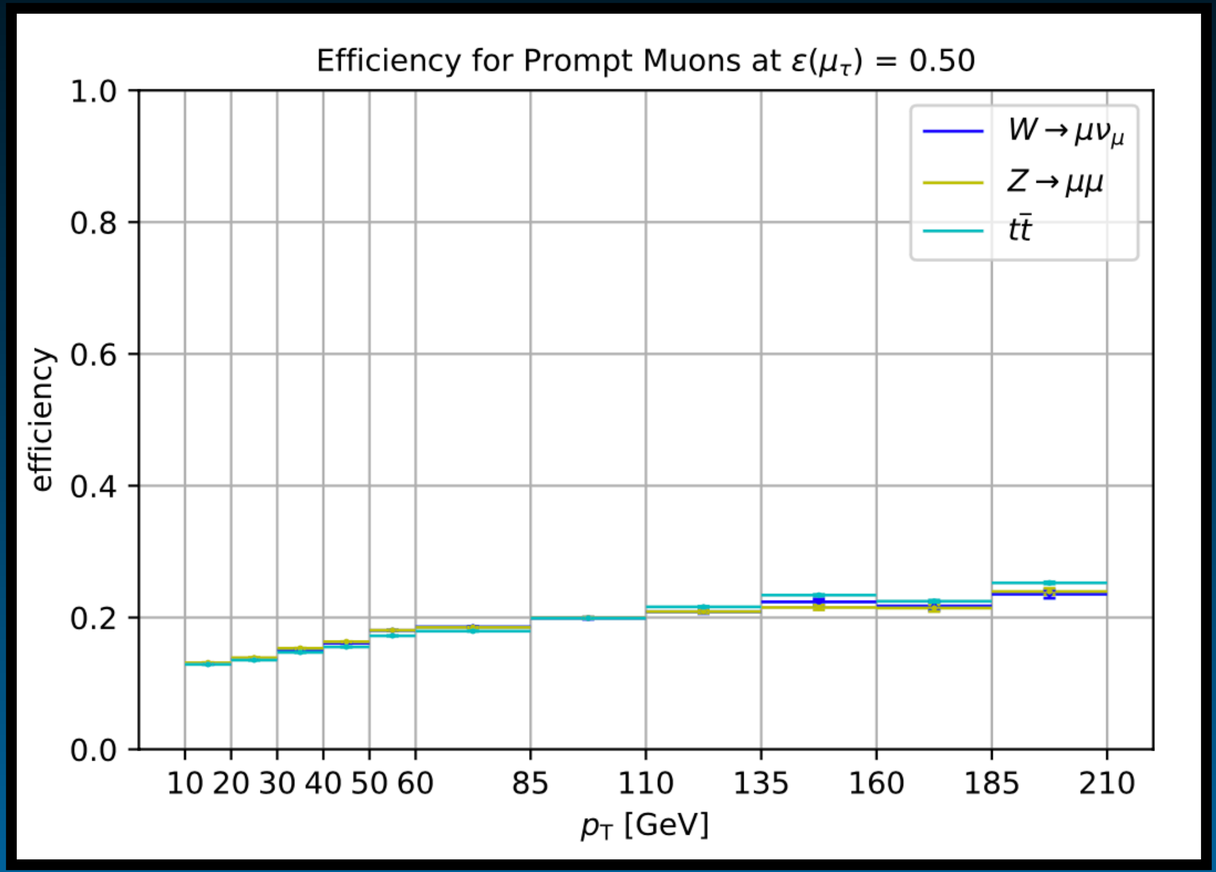


- Overall background is reduced:
  - At low p_T prompt muons rejection increased, fake muon rejection decreases only slightly
  - At high p_T fake muons are highly rejected at the expense of a slight increasement of the prompt muon efficiency

6 Generalization to Other Processes

- Output score thresholds are reused

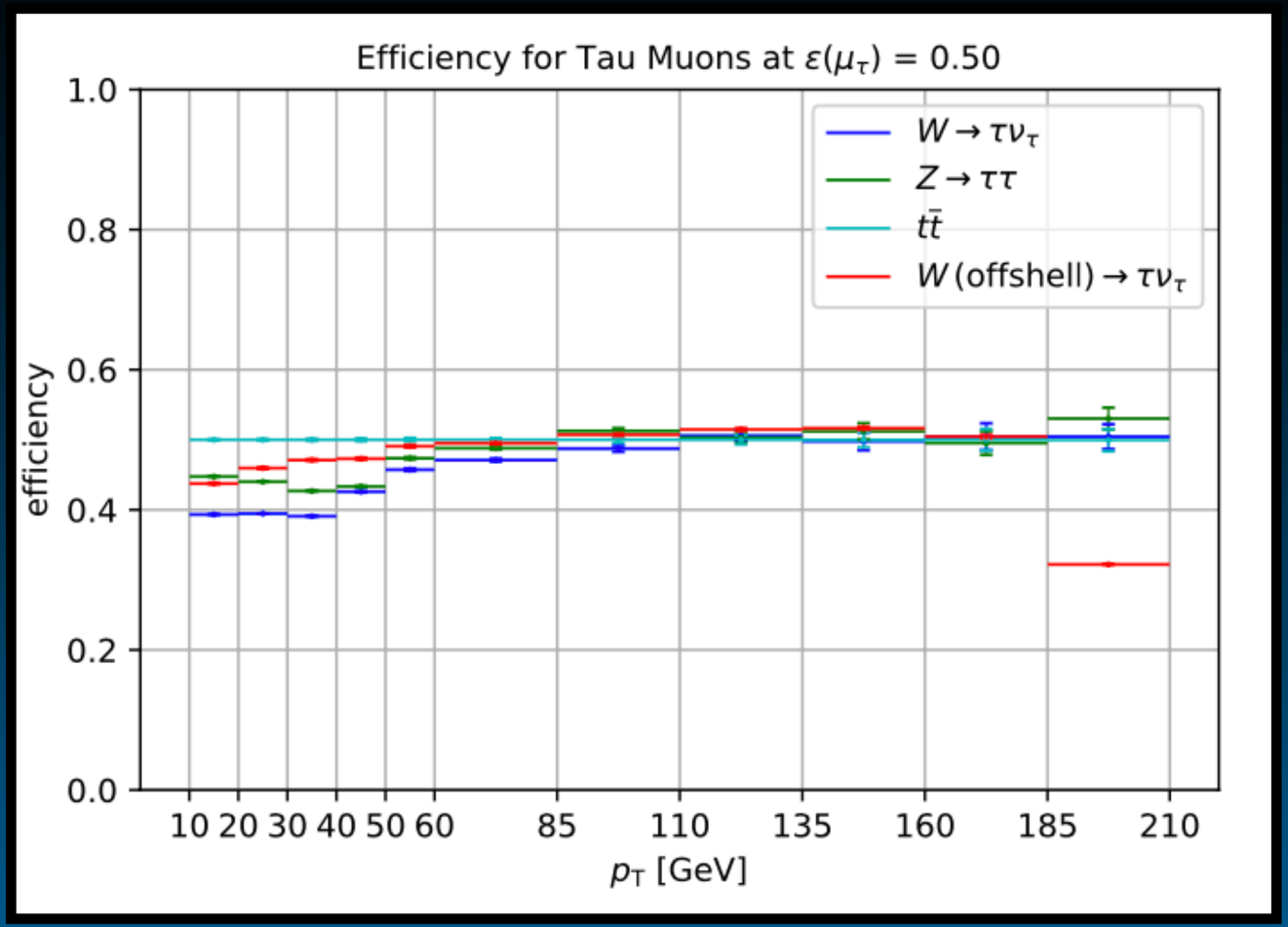




- Significant less rejection power of fake muons from W/Z events than in $t\bar{t}$ ⟶ likely originates in different fake compositions

- Low process dependency ⟶ has seen full availabe $t\bar{t}$ dataset during training

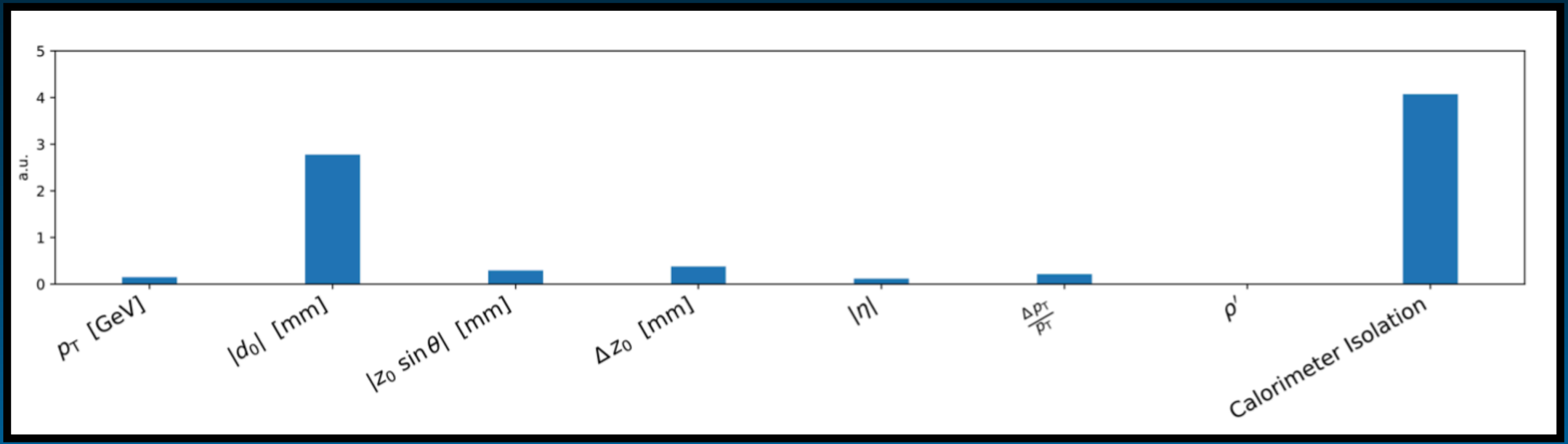
6 Generalization to Other Processes



- $t\bar{t}$ tau muon efficiency is 0.5 (per definition)
- Signal efficiency drop up to 0.1 in the regime 10 GeV $\leq p_T \leq$ 60 GeV. Small process dependency for larger p_T
- Strong W^* efficiency drop for $p_T > 185$ GeV $\longrightarrow$ W^* sample is dominated by high-p_T muons and the regime $p_T > 185$ GeV was not covered by enough muons during training which results in a strong efficiency drop $\longrightarrow$ train dedicated classifier for high-p_T regime

7 Feature Importance

- Shuffle the j-th feature variable across all samples and keep the others fixed ⟶ break its correlation with the other variables
- If model performance does not decrease substantially w.r.t. a certain metric ⟶ shuffled feature variable provides only slight importance to the model
- Metric: area under ROC curve (AUC)
- Measurement: $\mathrm{AUC}^{\text{unshuffeld}} - \mathrm{AUC}_j^{\text{shuffled}}$ (normalized by the mean of all eight feature importances)

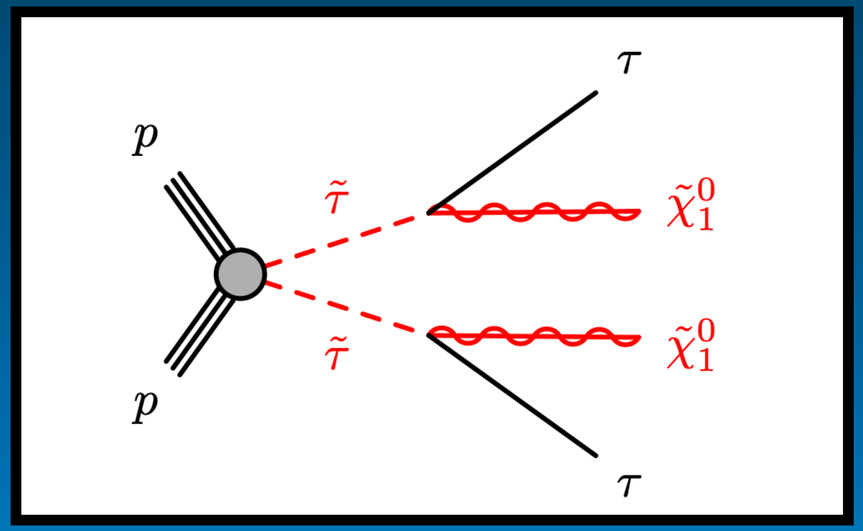


- Fake muons are much less isolated than tau muons, hence the classifier relies heavily on the calorimeter isolation variable
- Except of the calorimeter isolation, the other variables do not contribute much to the discrimination power compared to $|d_0|$
- ρ' could be dropped from the training without performance loss

8 Application on BSM Physics Signal

- At least an extension of the SM is necessary to address big open questions in elementary particle physics
  $\longrightarrow$ supersymmetry (SUSY)
  - Could solve the hierarchy problem and would offer an Dark Matter candidate
  - Postulates the existence of partner particle for each particle in the SM (*superpartner*), which differs in spin by half a unit
  - Consider the Minimal Supersymmetric Standard Model (MSSM)
    - Some superpartners:
      - Tau lepton $\tau \longrightarrow$ *stau* $\tilde{\tau}$
      - Neutral gauge bosons W^0, $B^0 \longrightarrow$ neutral *gauginos* $\tilde{W}^0$, $\tilde{B}^0$
      - Neutral components of the Higgs field (H_u^0 and H_d^0) $\longrightarrow$ neutral higgsinos ($\tilde{H}_u^0$ and $\tilde{H}_d^0$)
    - Electroweak symmetry breaking $\longrightarrow$ $\tilde{H}_u^0$, $\tilde{H}_d^0$ and $\tilde{W}^0$, $\tilde{B}^0$ mix to four neutral mass eigenstates: *neutralinos* ($\tilde{\chi}_j^0$, j= 1, 2, 3, 4)
    - $\tilde{\chi}_1^0$ would be the Dark Matter candidate (LSP and R-parity)

- Hypothesized process of interest:
  - Stau pair production: $\tilde{\tau}\tilde{\tau} \rightarrow \tau\tau + \tilde{\chi}_1^0\tilde{\chi}_1^0 \rightarrow \mu\nu_\mu\nu_\tau + \tau_{\text{had}} + \tilde{\chi}_1^0\tilde{\chi}_1^0$
- Application of classifier for the search in the semileptonic decay channel:
  - Suppress W + jets $\rightarrow \mu\nu_\mu$ + jets (if jet is misidentified as τ_{had}

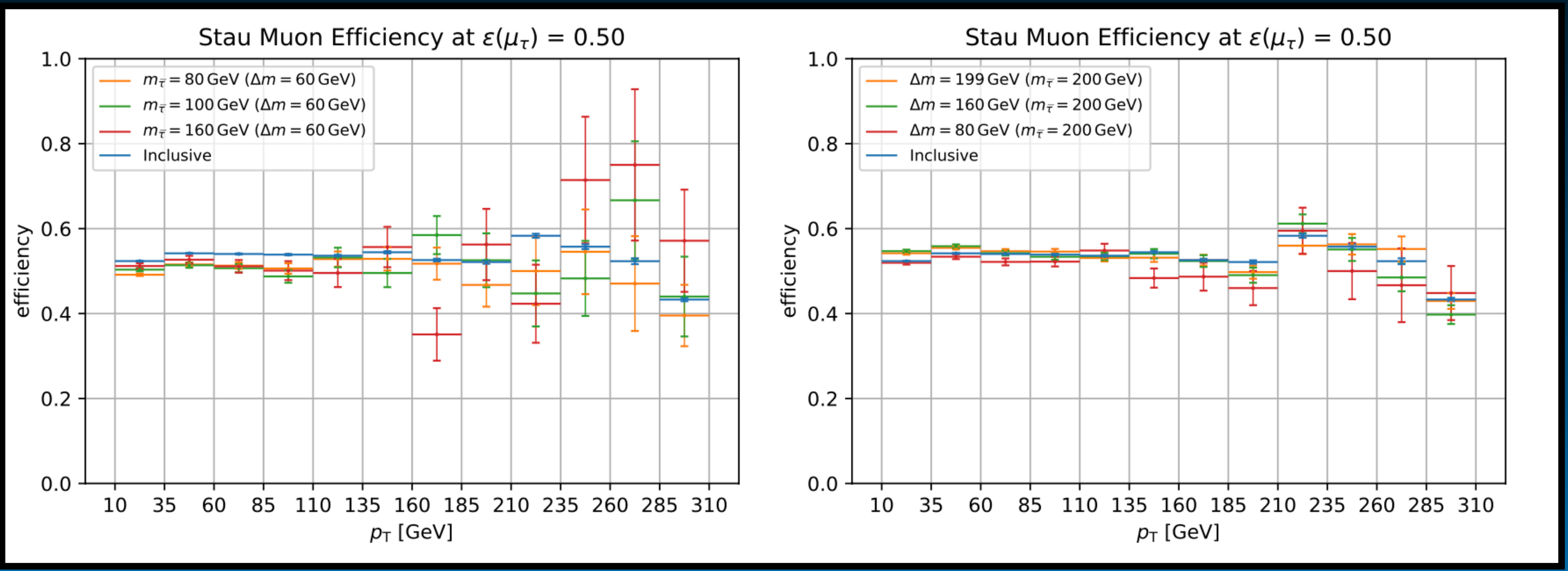


8 Application on BSM Physics Signal

- Evaluate on muons originating from stau-pair production (*stau muons*) in dependency of stau mass $m_{\tilde{\tau}}$ and mass splitting $\Delta m := m_{\tilde{\tau}} - m_{\tilde{\chi}_1^0}$
- In addition, evaluated on stau muons corresponding to a large variety of mass parameters ($m_{\tilde{\tau}} \in [80,440]$ GeV and $m_{\tilde{\chi}_1^0} \in [1,200]$ GeV)
- Reuse output score thresholds (i.e. such that tagging efficiency of $t\bar{t}$ tau muons is 0.5)

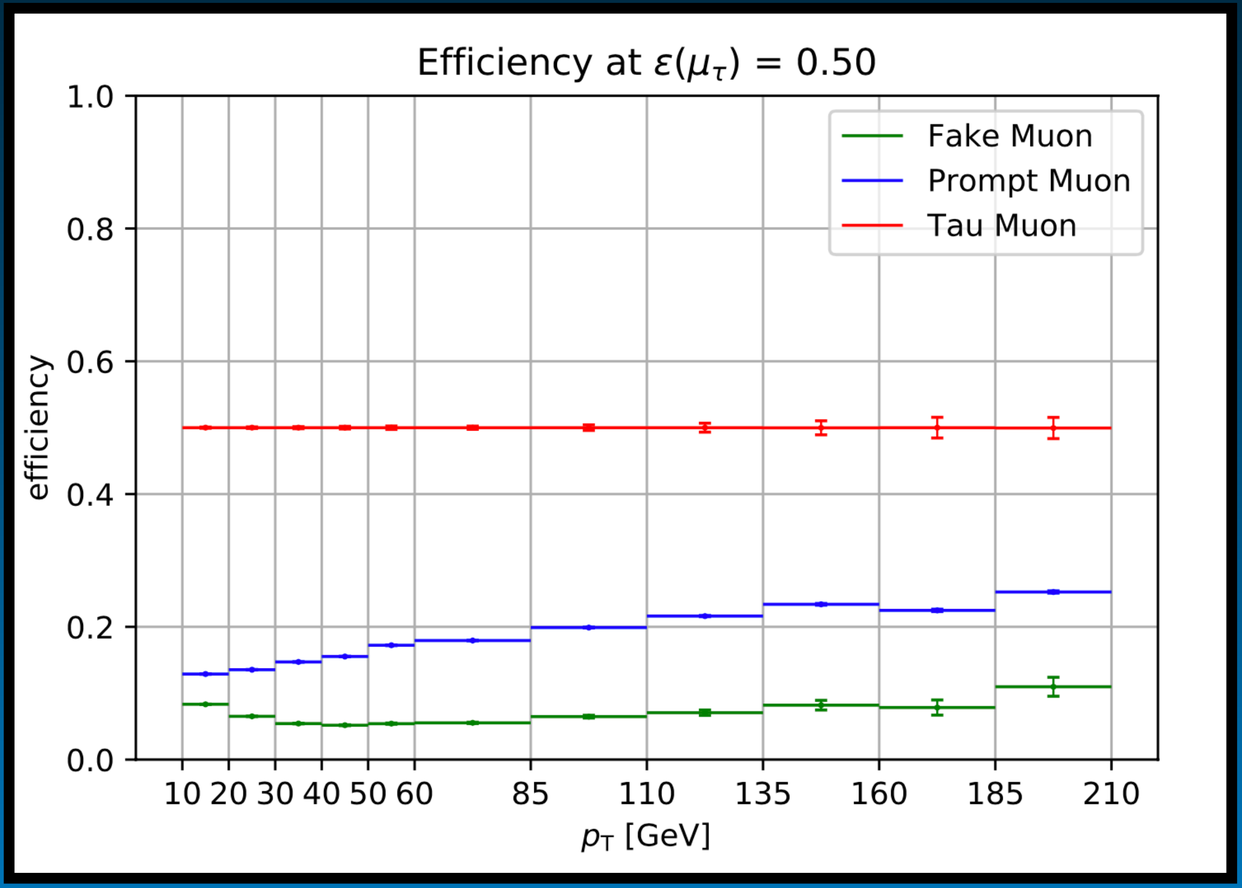
8 Application on BSM Physics Signal

- BSM signal efficiency around 50% and independent of the $m_{\tilde{\tau}}$ and $m_{\tilde{\chi}_1^0}$ up to the first order



9 Conclusion and Outlook

- Tagging of muons originating from tau decays interesting for SM measurements as well as BSM searches
- Presented development of ML-based tagger trained on e.g. $t\bar{t}$ events
- Flat sampling during batch calls classifier shows best performance:
  - At a p_{T}-flat signal efficiency of 50 %: prompt and fake muon efficiency of 12.5 % to 22.5 % and 5 % and 10 %, respectively
  - Evaluating the classifier for processes other than $t\bar{t}$: signal efficiency reduces up to 10 % ⟶ include other processes in the training as well

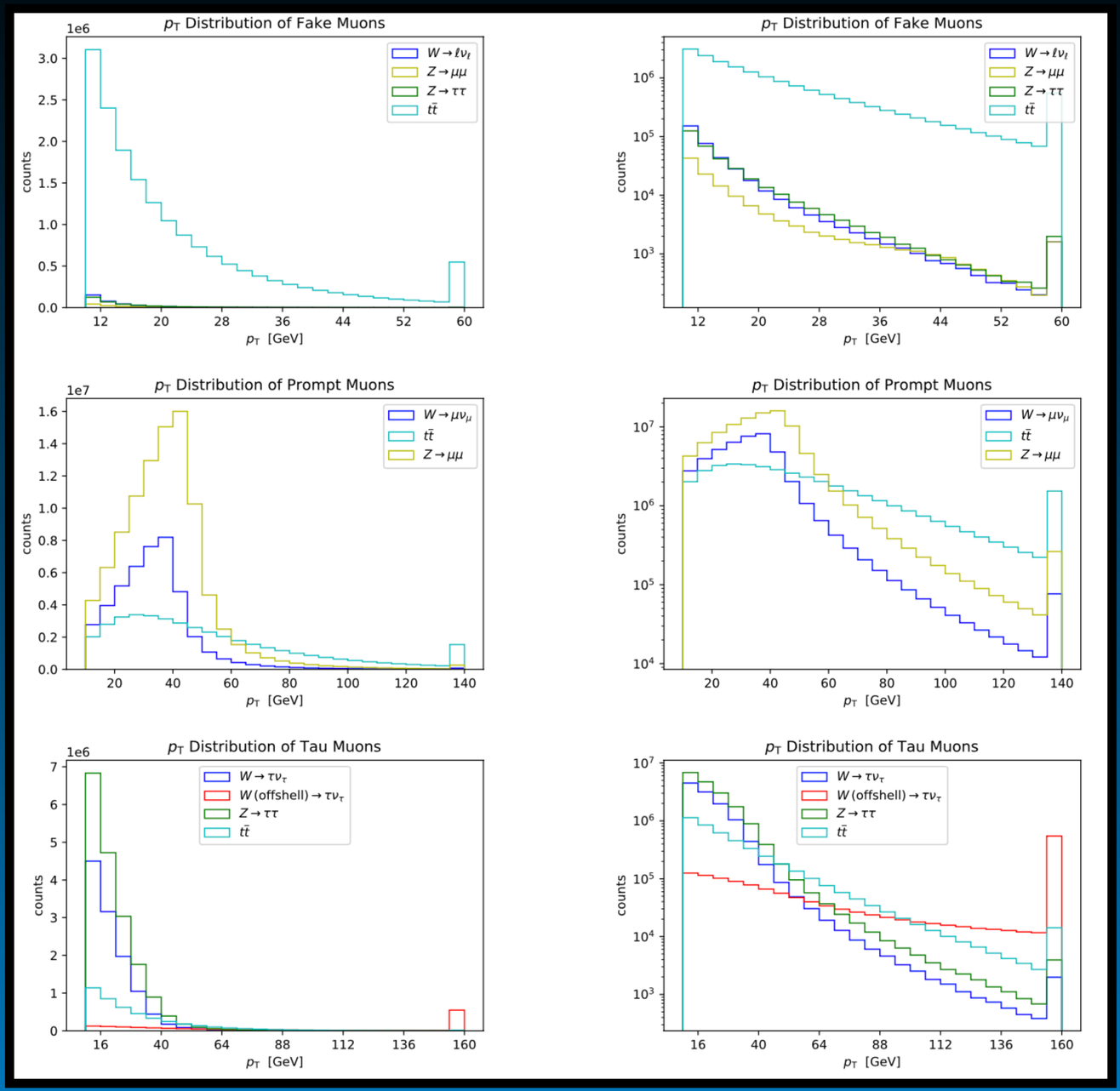


9 Conclusion and Outlook

- Before the application in an actual analysis could take place, several additional steps would be required first:
  - ATLAS event simulation not perfect ⟶ calibrate taggers, i.e. match the efficiency in simulation with the one measured in data
  - Isolation WPs on leptons are typically used in ATLAS analyses ⟶ study interplay between WPs and classifiers (WPs are optimized for prompt muons not tau muons)
  - Decays from taus to muons and electrons occur at the same rate ⟶ develop classifier dedicated to electrons and apply same techniques in the electron channel ⟶ increase the sensitivity reach of a search for new physics

Backup

p_T Distribution of the Simulated Processes

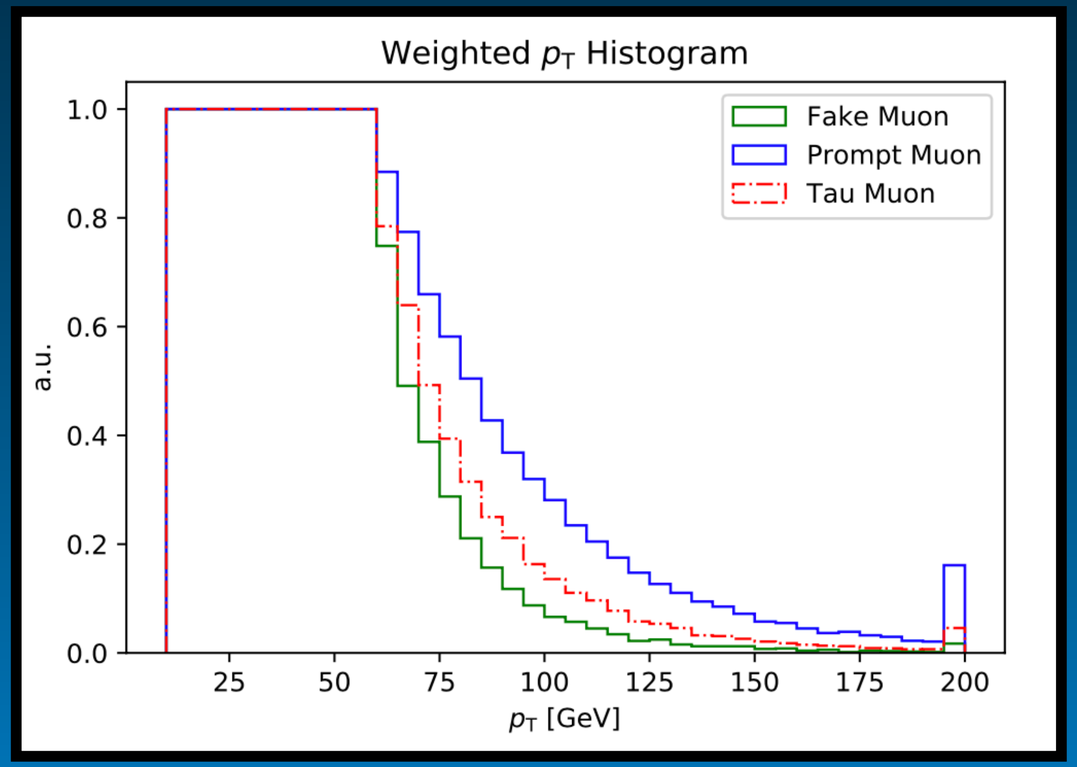


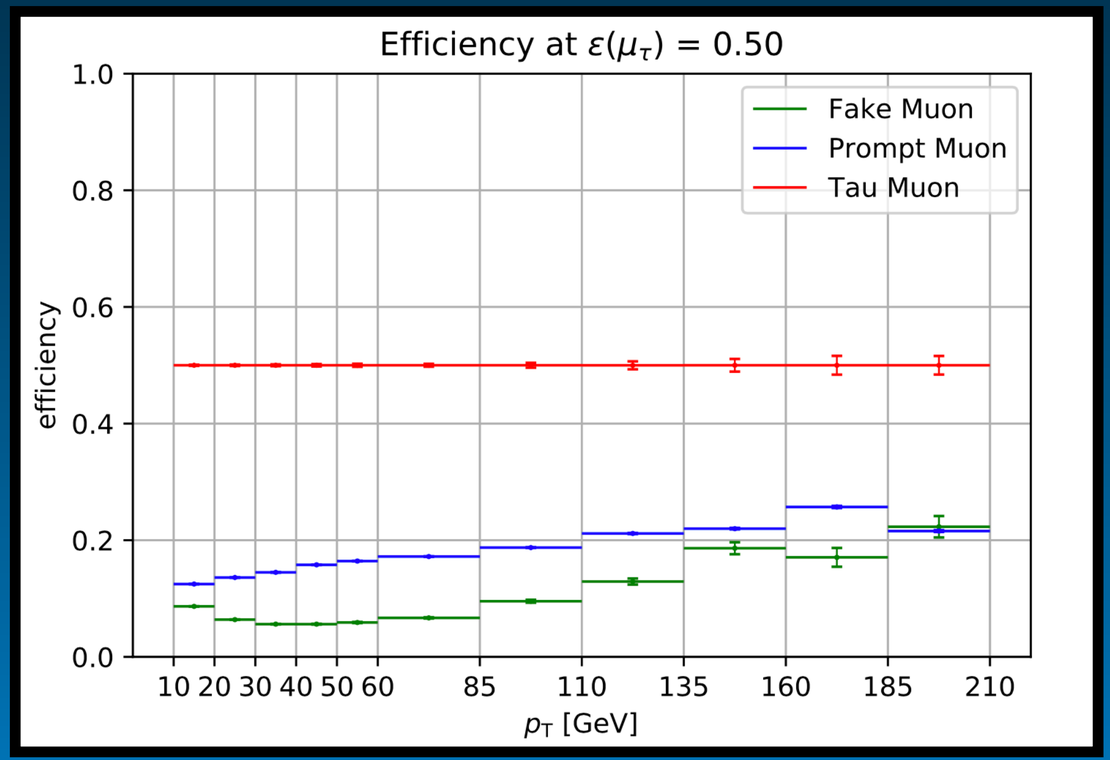
Sample Weights

- Same dataset as in the naive approach with the biased training set
- Determine sample weights in 5 GeV steps up to 60 GeV such that the distribution is flat in p_T

$$L_m^{\text{total}} \longmapsto L_m'^{\text{total}} = \frac{1}{m} \sum_{k=1}^{m} w_k \cdot L(y_k, y_k^{\text{true}})$$

- Sample weights corresponding to the interval [55,60) GeV are applied on every sample with $p_T > 60$ GeV, to prevent large weights and thus fluctuations in the loss surface (may spoil its convergence)



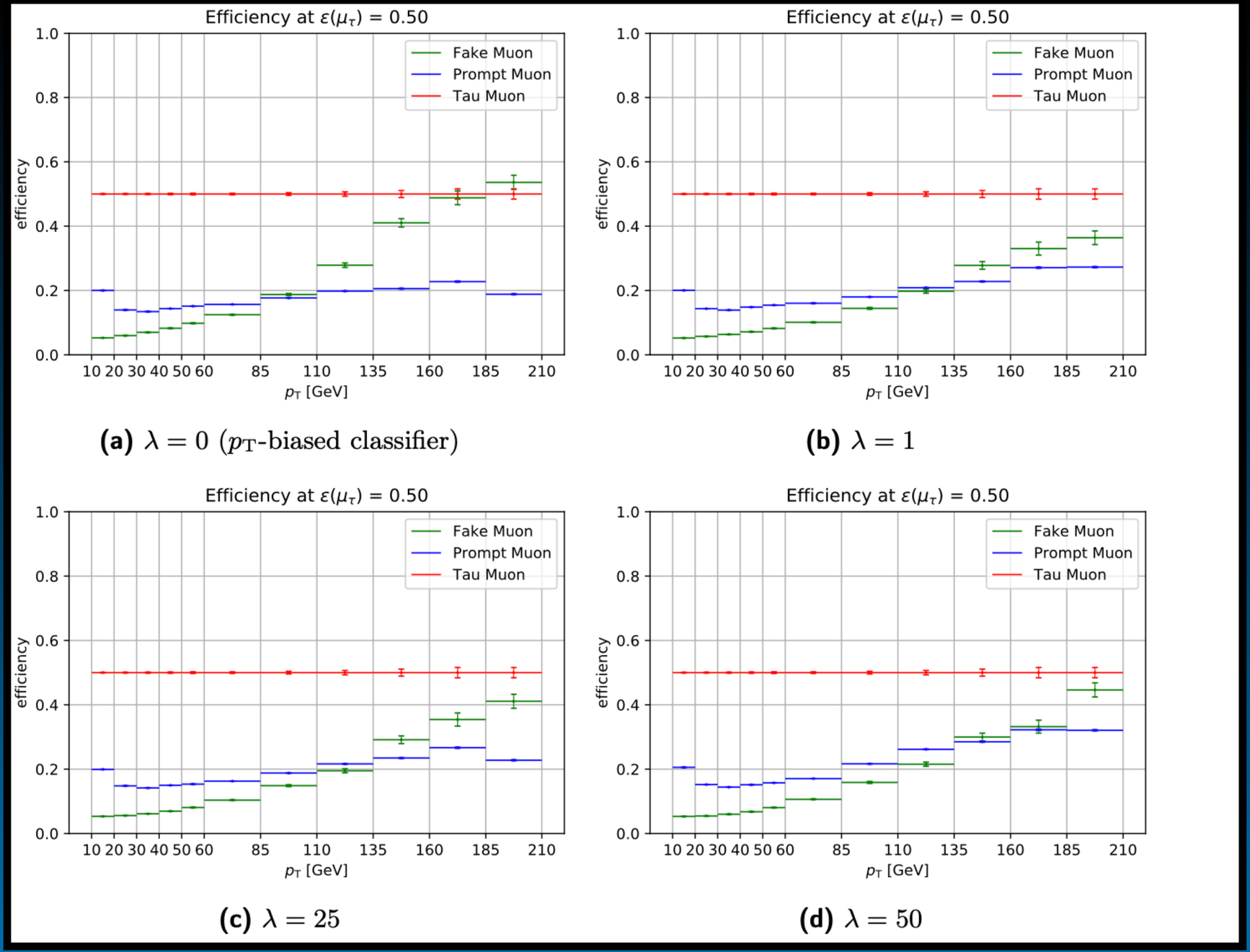


Distance Correlation

- Same datasets used as before. no modification of the dataset is made, but rather on the loss function itself
- if output score and p_{T} of muon are not correlated, this may also ensure a similar tagging performance over the whole p_{T} regime $\longrightarrow$ Add new term to the loss function
- Distance correlation $\mathrm{dCorr}_n^2(X, Y)$ as measurement of correlation between two variables
- Modify loss function: $L_m^{\mathrm{total}} \longmapsto L_m^{'\mathrm{total}} = L_m^{\mathrm{total}}(y, y^{\mathrm{true}}) + \lambda \cdot \mathrm{dCorr}_m^2(y, p_{\mathrm{T}})$, with tunable hyperparameter λ

Distance Correlation

- Fake muon rejection is improved but on the price of a lower prompt muon rejection
- Training set is still dominated by low-p_{T} muons and hence the decorrelation occurs mostly at low p_{T} as well



Distance Correlation

Observed n paired samples $(X, Y) = (x_i, y_i)_{i=1}^n$ with $a_{ij} = |x_i - x_j|$ and $b_{ij} = |y_i - y_j|$ define

$$A_{ij} = a_{ij} - \frac{1}{n}\sum_{i=1}^{n} a_{ij} - \frac{1}{n}\sum_{j=1}^{n} a_{ij} + \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}$$

$$B_{ij} = b_{ij} - \frac{1}{n}\sum_{i=1}^{n} b_{ij} - \frac{1}{n}\sum_{j=1}^{n} b_{ij} + \frac{1}{n^2}\sum_{i,j=1}^{n} b_{ij}$$

The empirical distance covariance $\mathrm{dCov}_n(X, Y)$ is then defined by

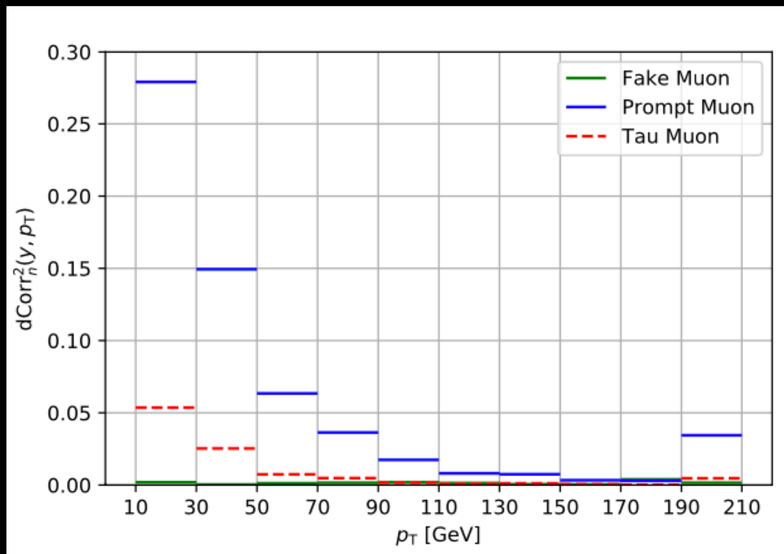
$$\mathrm{dCov}_n^2(X, Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}$$

Finally, the discrete distance correlation $\mathrm{dCorr}_n(X, Y)$ is given by

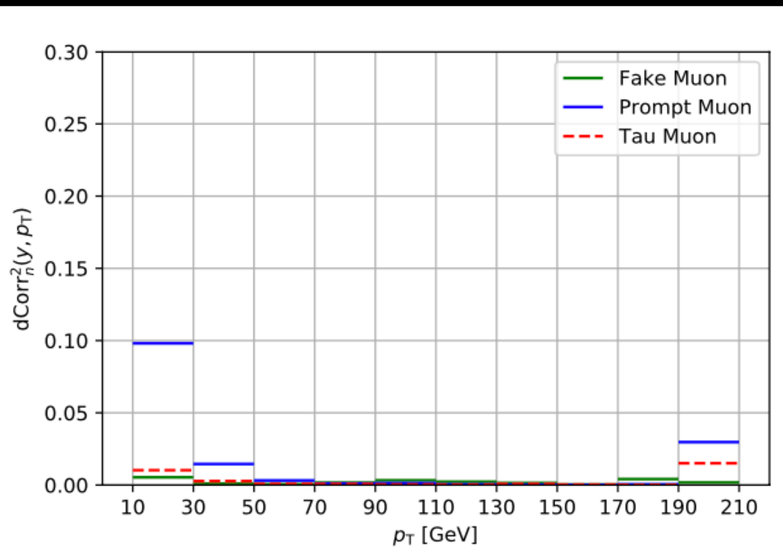
$$\mathrm{dCorr}_n^2(X, Y) = \frac{\mathrm{dCov}_n^2(X, Y)}{\sqrt{\mathrm{dCov}_n^2(X, X)\mathrm{dCov}_n^2(Y, Y)}}$$

Distance Correlation

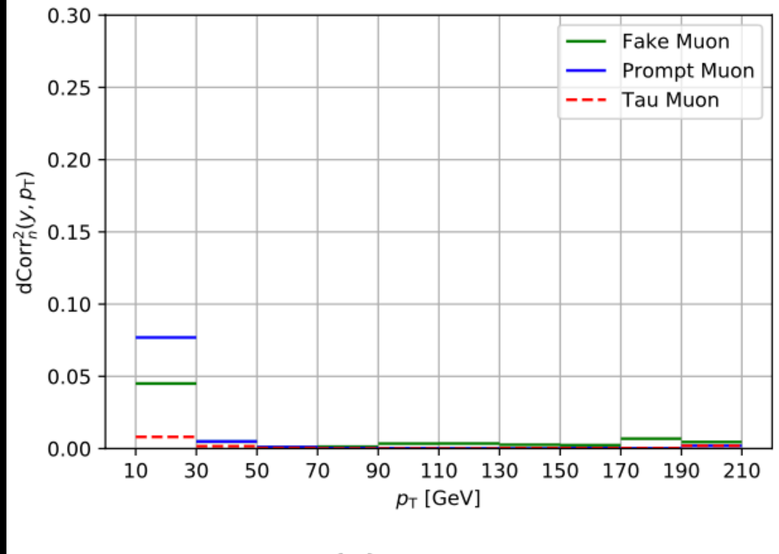
- $\mathrm{dCorr}_n^2(y, p_\mathrm{T})$ with respect to p_T



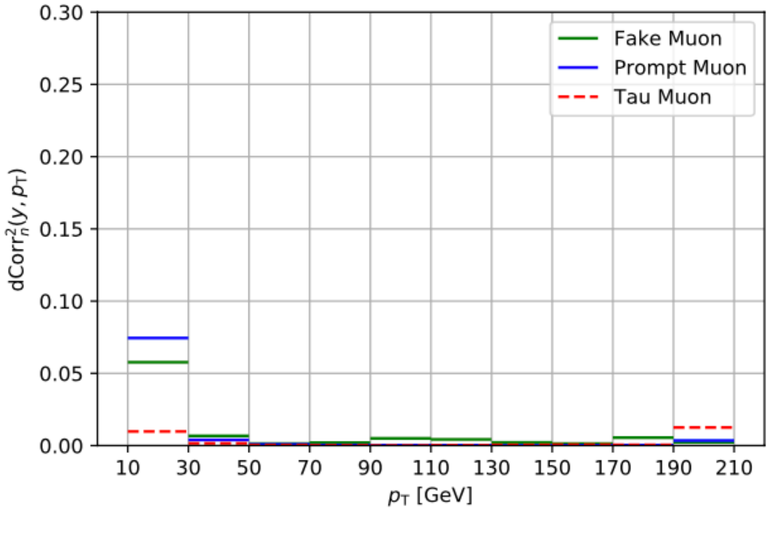
(a) $\lambda = 0$ (p_T-biased classifier)



(b) $\lambda = 1$



(c) $\lambda = 25$

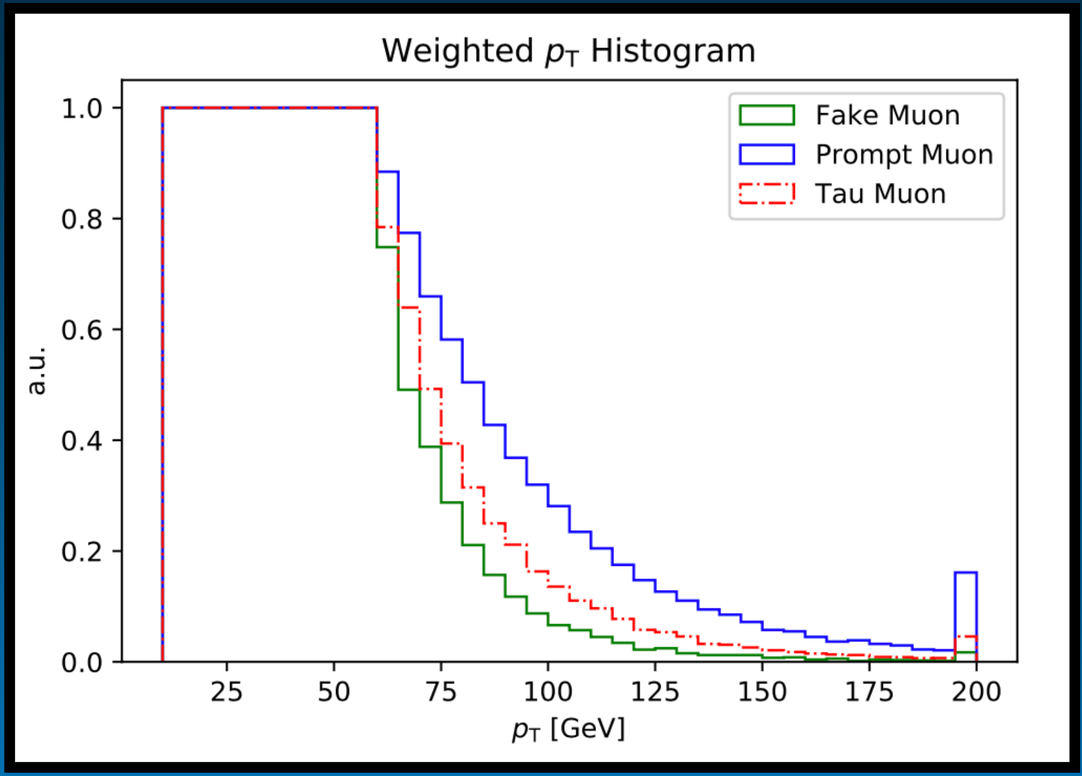


(d) $\lambda = 50$

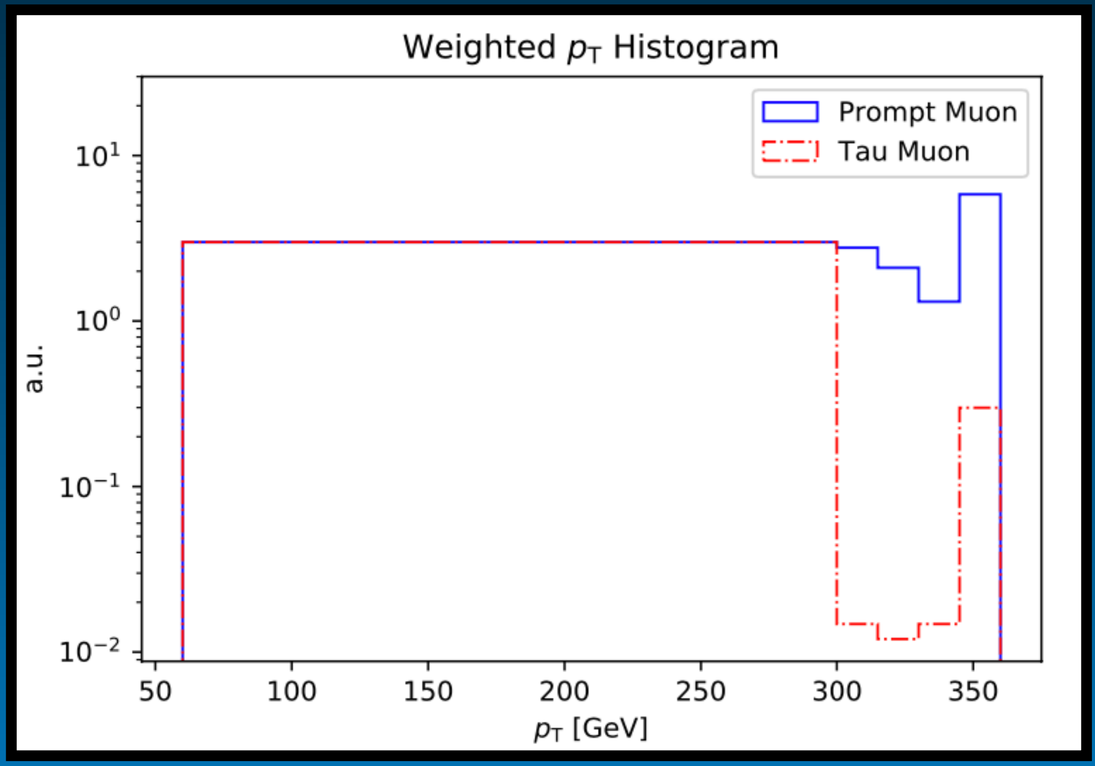
High-p_T Muon Taggers

- Train two high-p_T tau muon tagger exclusively for $p_T \geq 60\,\text{GeV}$ $\longrightarrow$ increase importance for high-p_T muons

- Trained on same weighted dataset as the sample weight classifier, but exclusively for the regime $p_T > 60\,\text{GeV}$



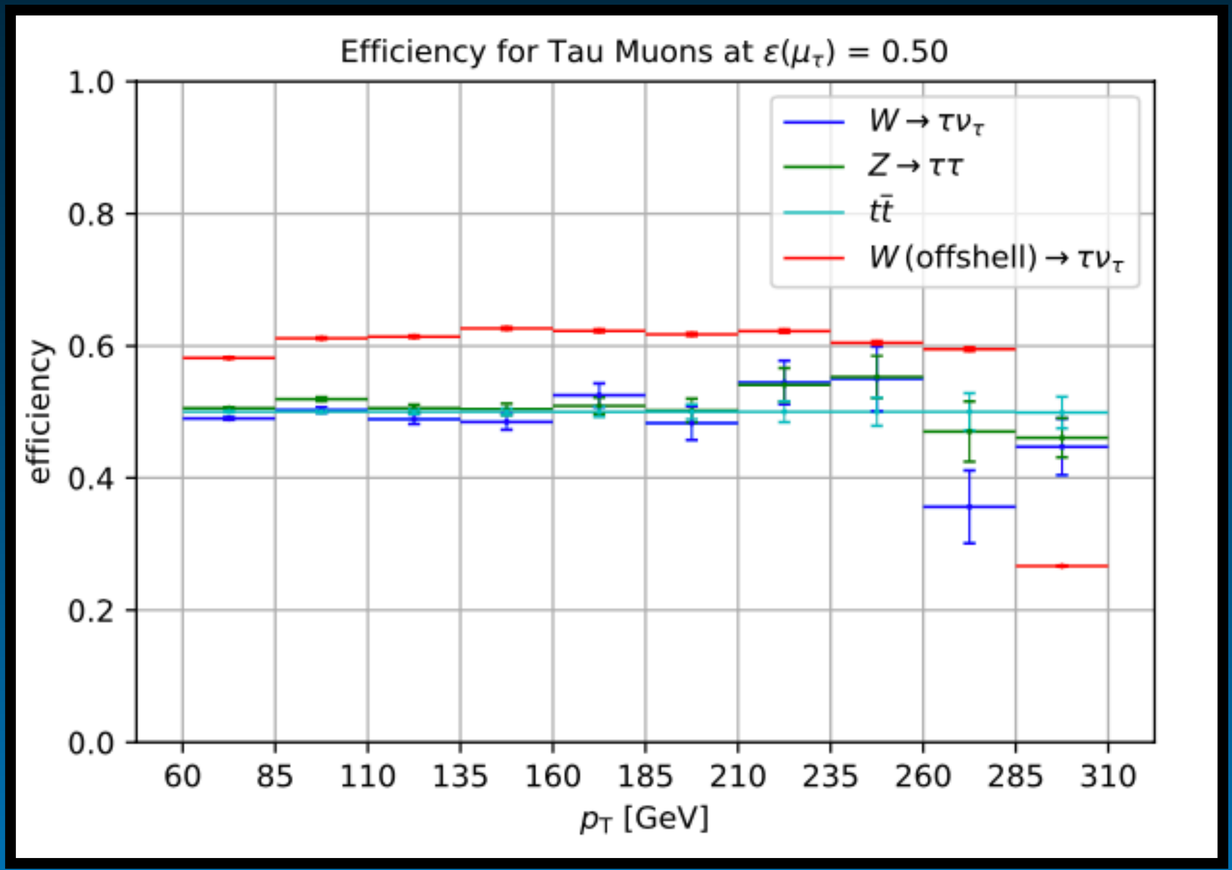
- Trained on tau muons from W^* and on prompt muons from $t\bar{t}$, fake muons are not considered



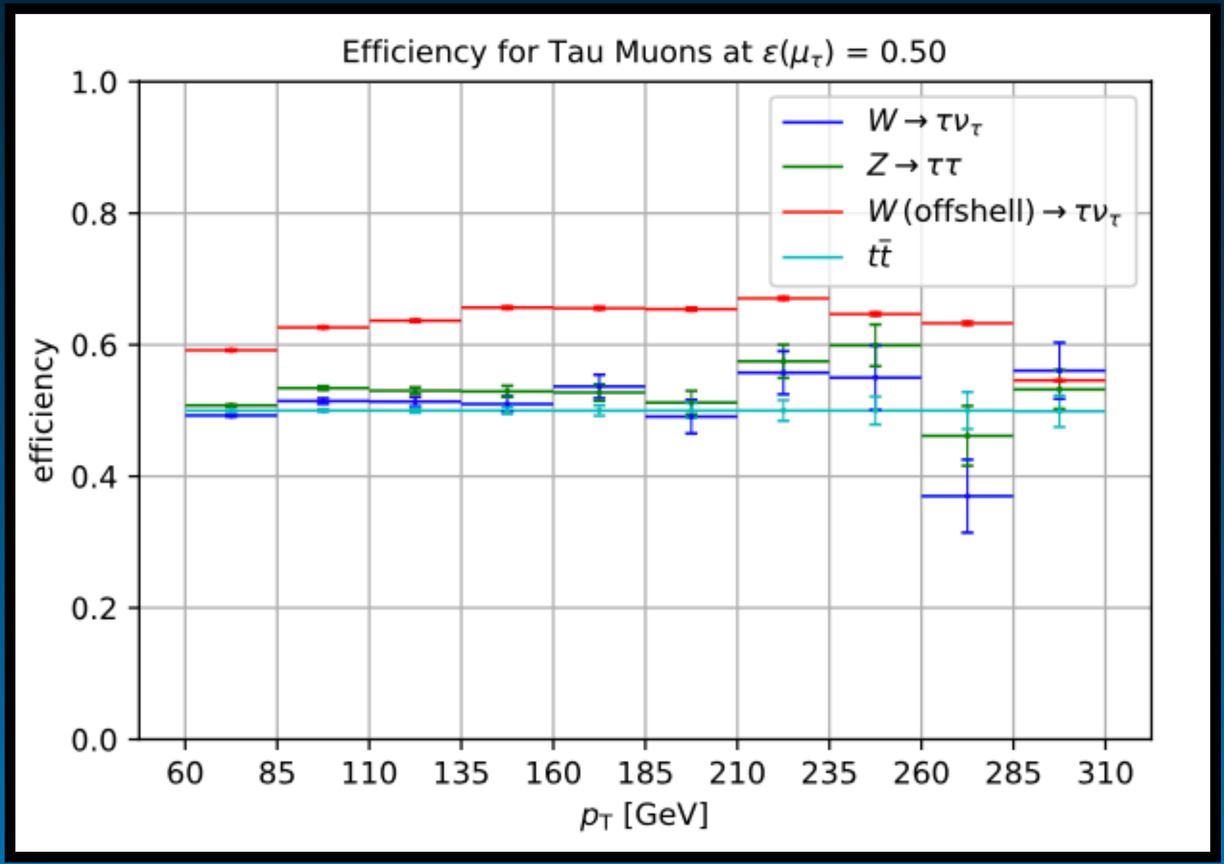
High-p_T Muon Taggers

- High-p_T muon tagger trained on tau muons from $t\bar{t}$ events improves performance only w.r.t. tau muons with $p_T > 285$ GeV

- sample weights classifier

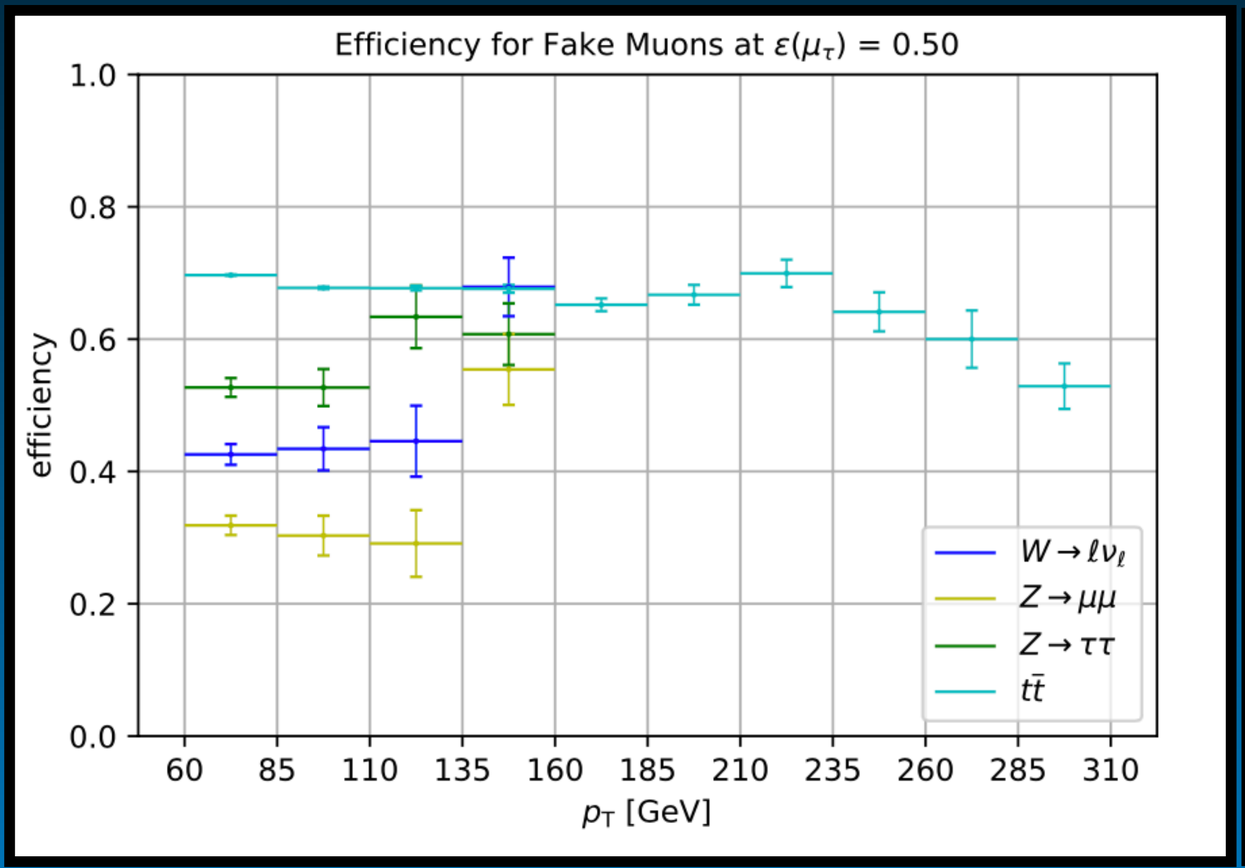


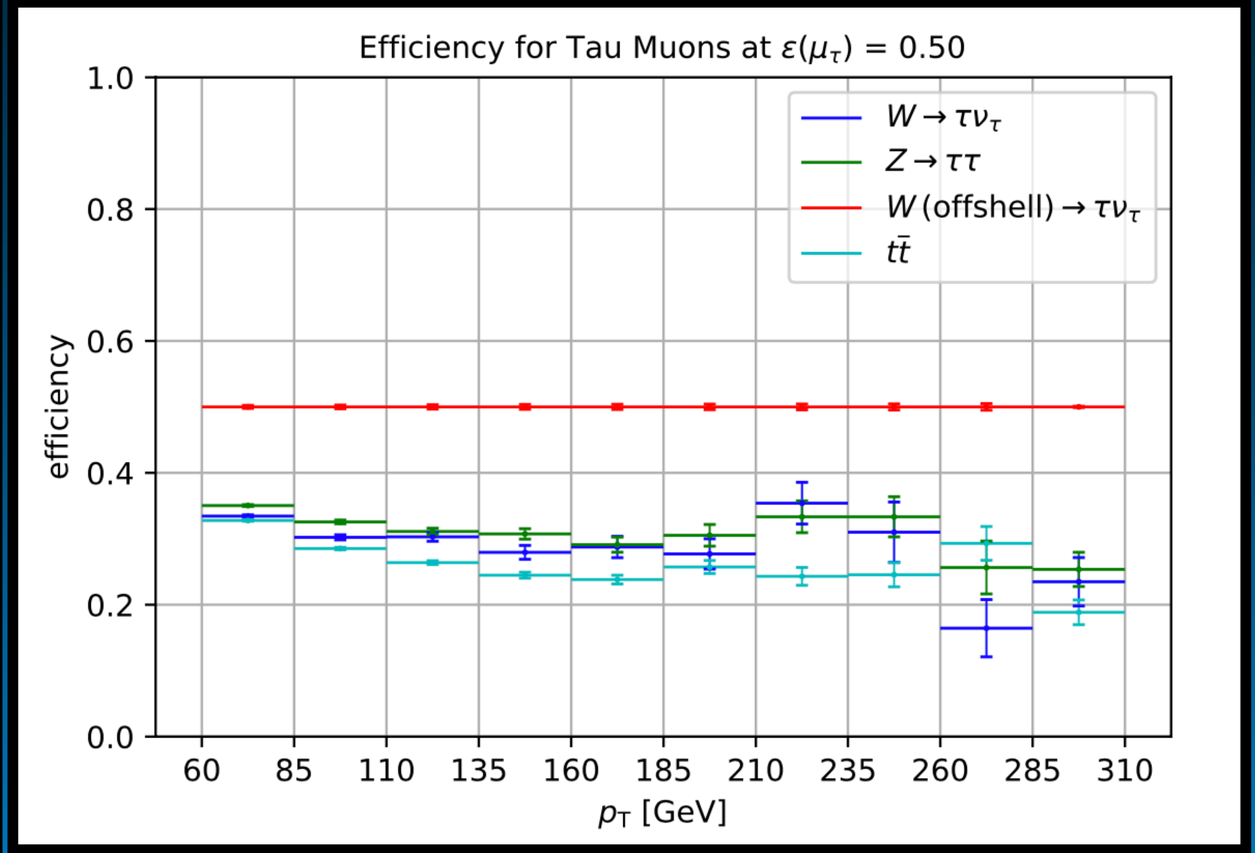
- High-p_T muon tagger



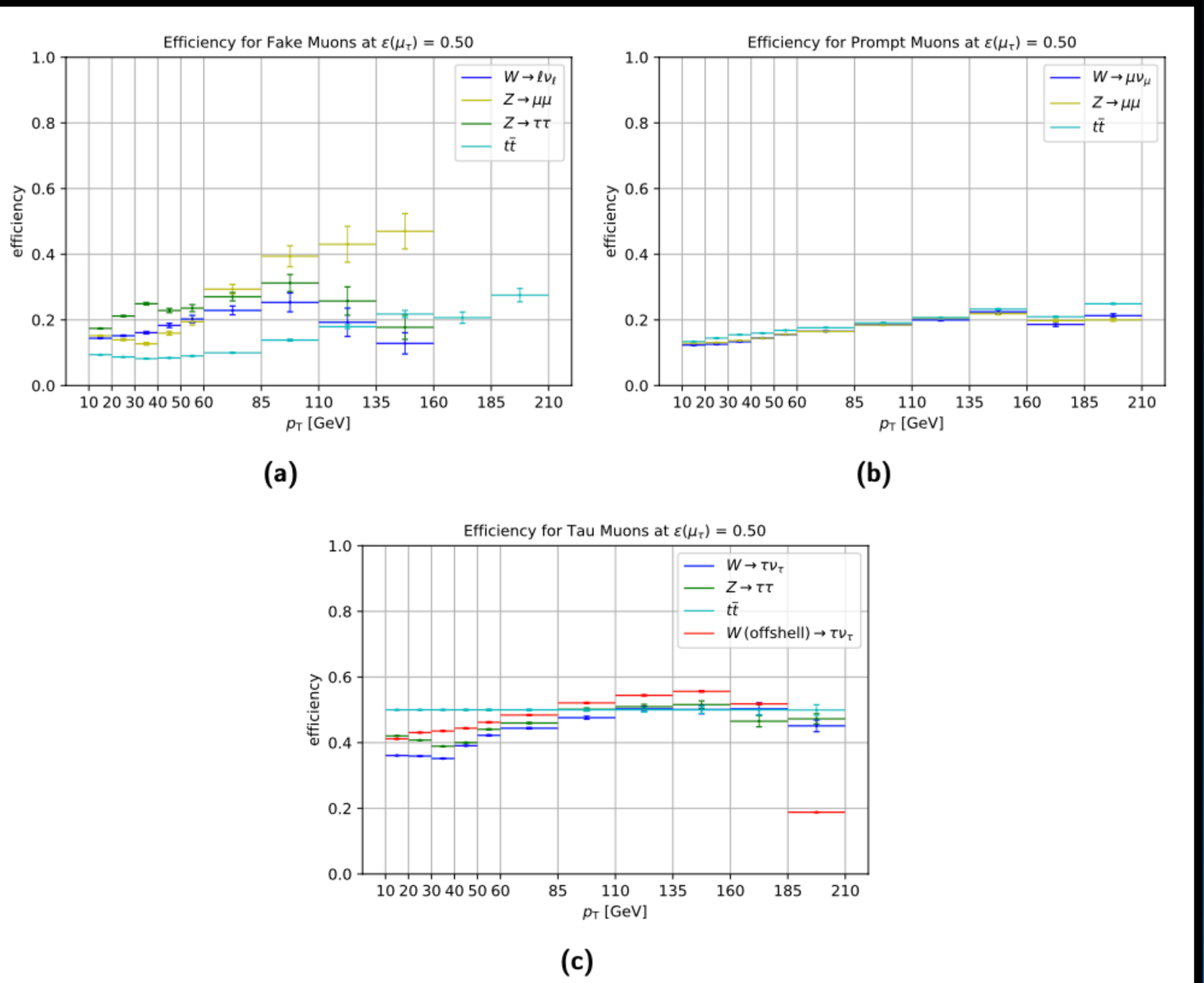
High-p_T Muon Taggers

- High-p_T muon tagger trained on tau muons from W^* events
- Inappropriate for fake muon rejection $\longrightarrow$ not trained on fake muons, i.e. could not learn distinctiveness between fake and tau muons provided by calorimeter isolation
- Bad generalization to tau muons from other processes





Comparison to Tertiary Classification



- Investigate whether a distinction between prompt and fake muons offers any advantages for separating them from tau muons

- Loss function: categorical crossentropy

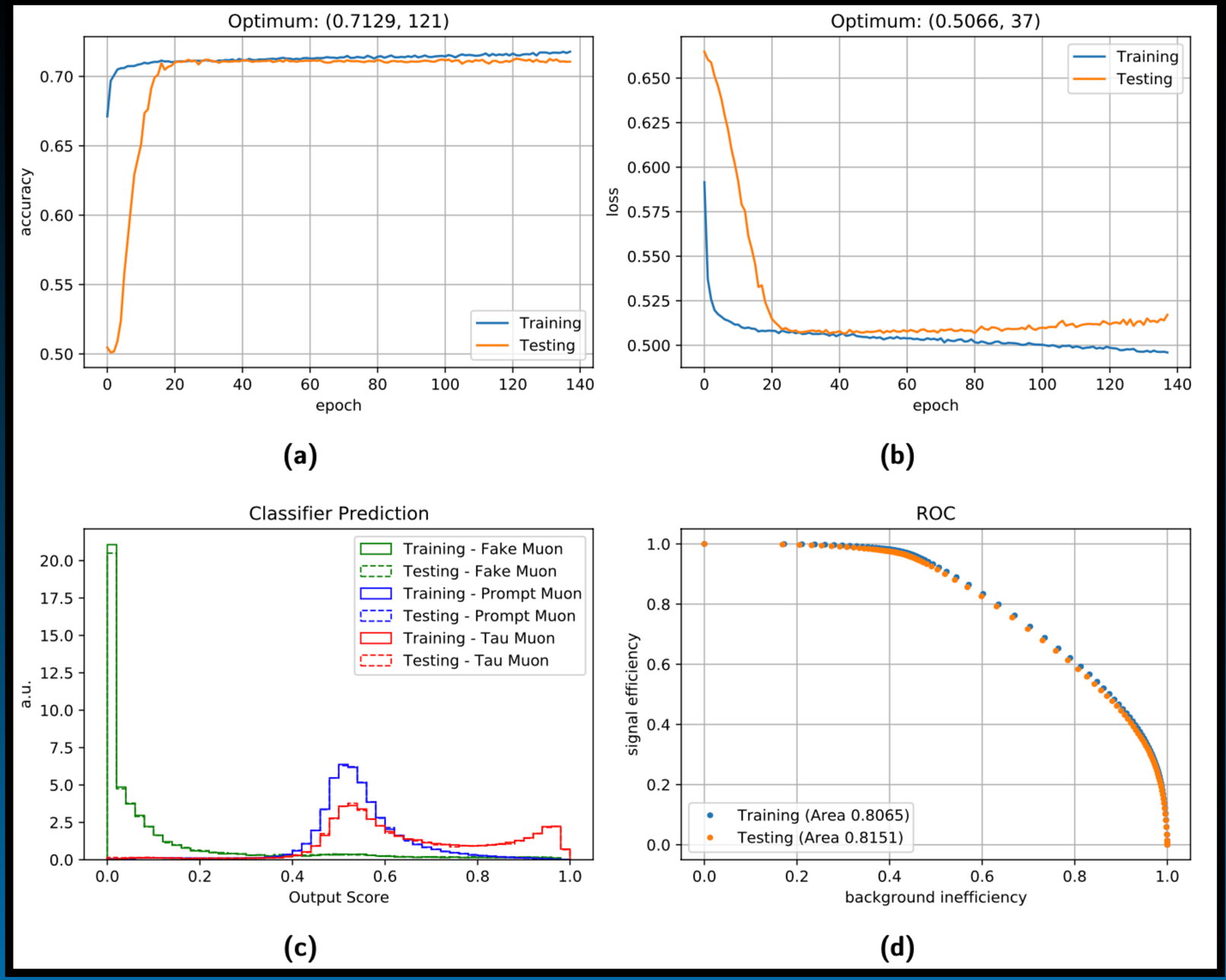
$$L(y, y^{\text{true}}) = -\sum_{i=0}^{2} y_i^{\text{true}} \log(y_i)$$

- Train classifier with flat sampling during batch calls approach, but sample 10,000 prompt as well as fake muons from $t\bar{t}$ during training, so all three classes appear with the same rate
- Consider output score associated with tau muons for all three classes

$\longrightarrow$ Tertiary classification does not improve the performance compared to binary classification

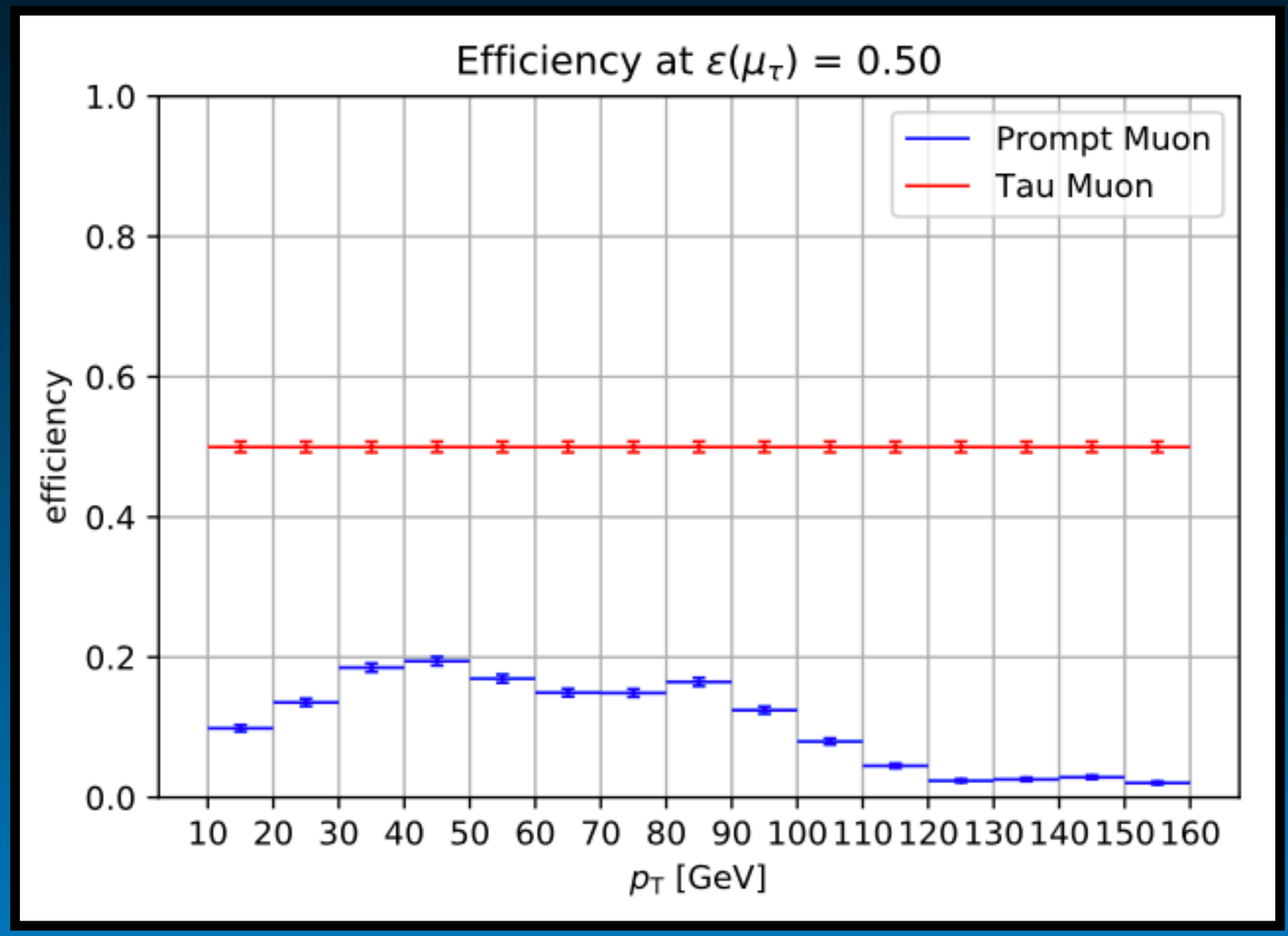
Performance Metrics

- Metrics monitored during training: loss and accuracy
- Normalized output score distributions of all three muon classes from the train and test set
- ROC curve: each point represents a 2D-tupel of signal efficiency and background inefficiency for a certain output score threshold; area under curve (AUC) quantifies discrimination power



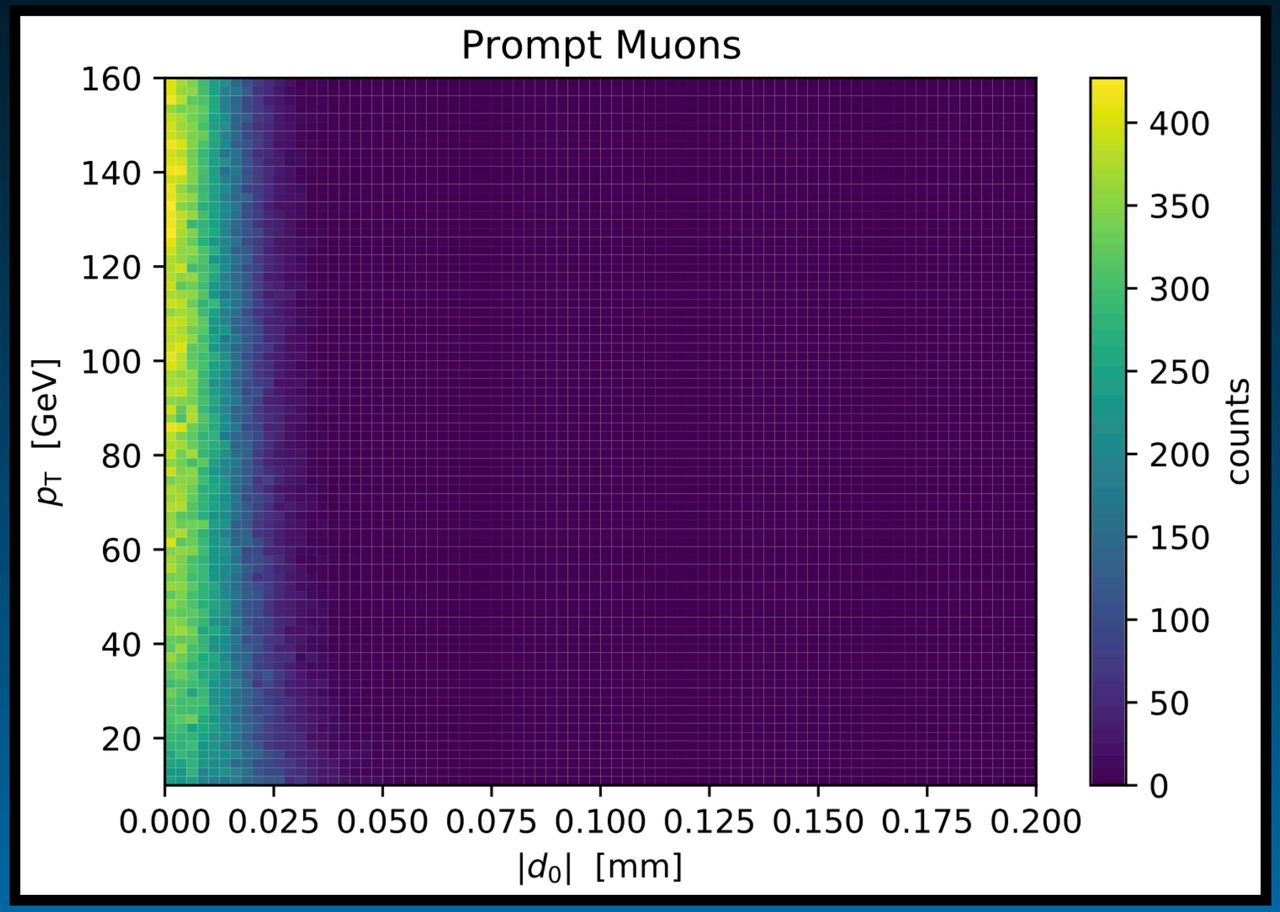
Flat sampling with mixed set

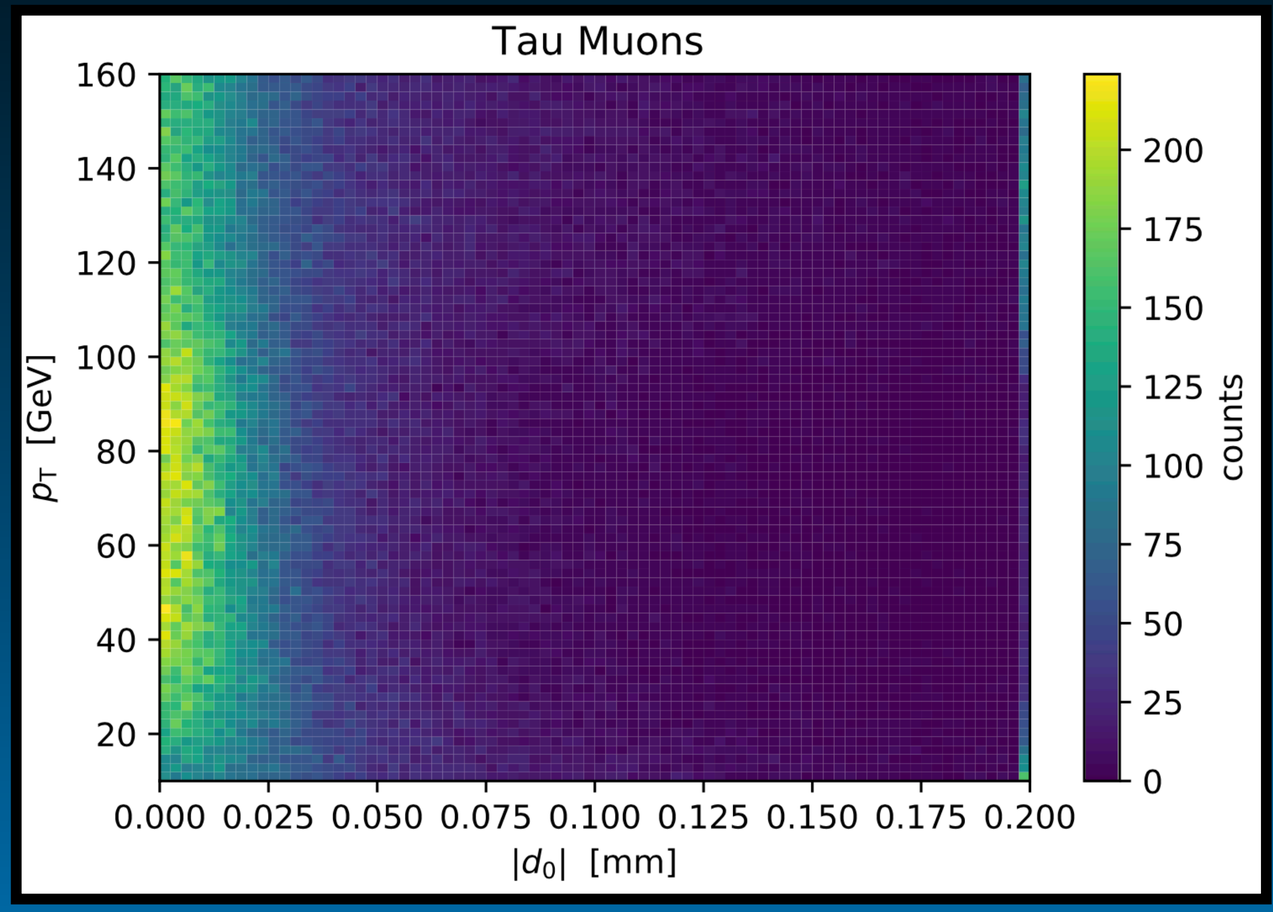
- Approach: Instead of sampling muons multiple time, use not only muons from $t\bar{t}$, but from all generated processes
- 250,000 prompt and tau muons are sampled flat in p_T up to 160 GeV. Fakes are not considered here. This dataset is then split up into train (75 %) and test (25 %) set, where the latter is defined to be also the evaluation set



Flat sampling with mixed set

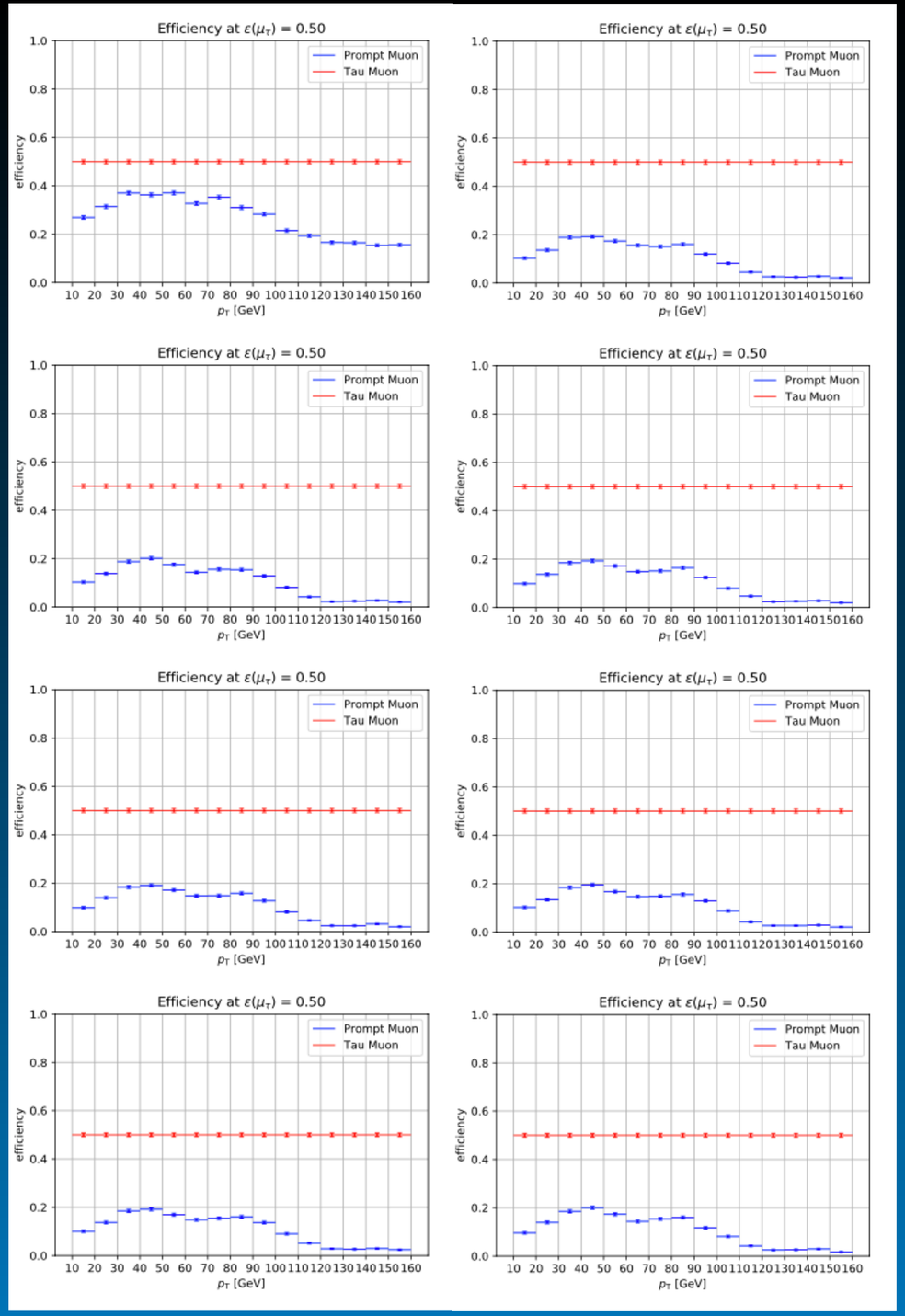
- Cause of the intermediate drop can likely be adressed to the correlation of p_T with the impact parameter related input features





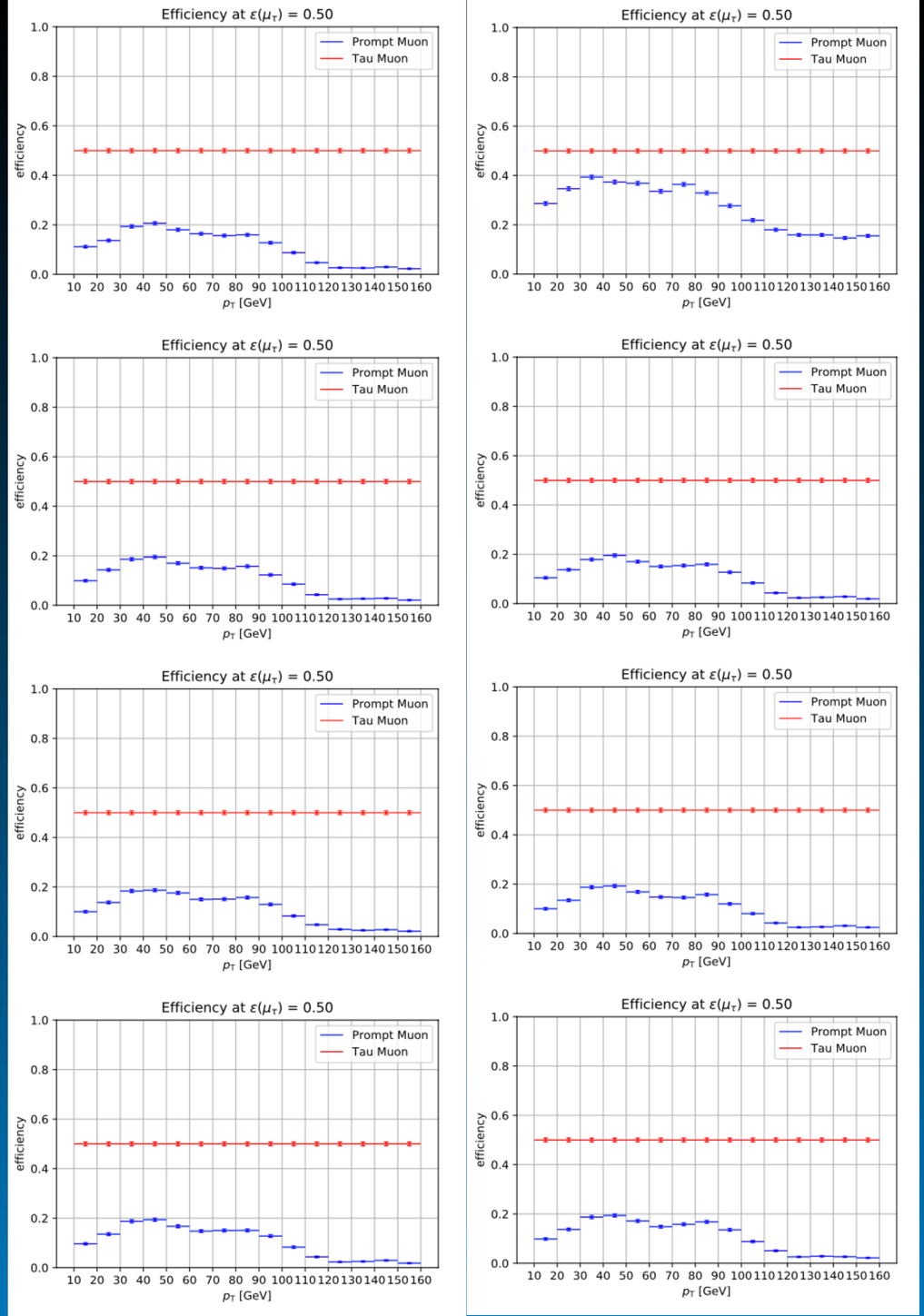
Flat sampling with mixed set

- „N-1“ - feature training: remove one feature during training (in order to resolve strange correlations with p_T)
- Left: p_{T}, $|z_0 \sin\theta|$, $|\eta|$, ρ' were excluded referring to the figures from top to bottom
- Right: $|d_0|$, Δz_0, $\frac{\Delta p_{\mathrm{T}}}{p_{\mathrm{T}}}$, calorimeter isolation were excluded referring to the figures from top to bottom



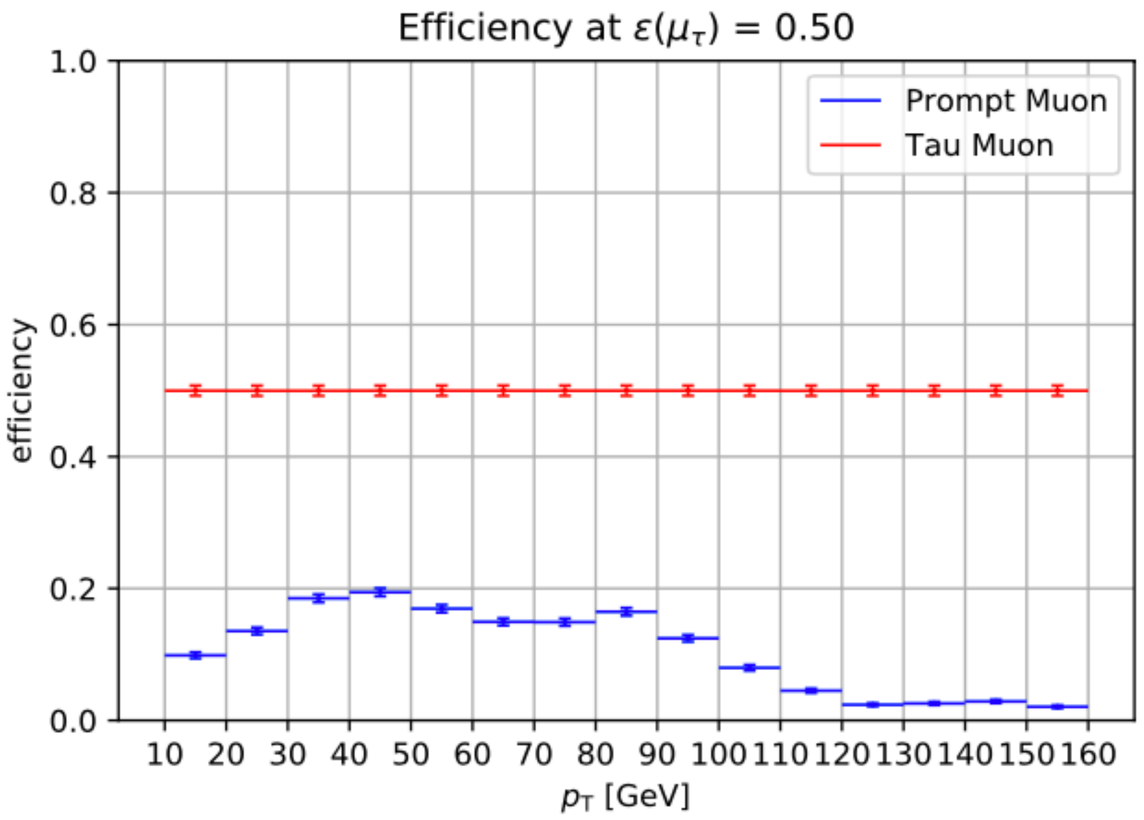
Flat sampling with mixed set

- „Shuffle Training“: shuffle one feature during training (in order to resolve strange correlations with p_T
- Left: p_{T}, $|z_0 \sin\theta|$, $|\eta|$, ρ' were shuffled referring to the figures from top to bottom
- Right: $|d_0|$, Δz_0, $\dfrac{\Delta p_{\mathrm{T}}}{p_{\mathrm{T}}}$, calorimeter isolation were shuffled referring to the figures from top to bottom

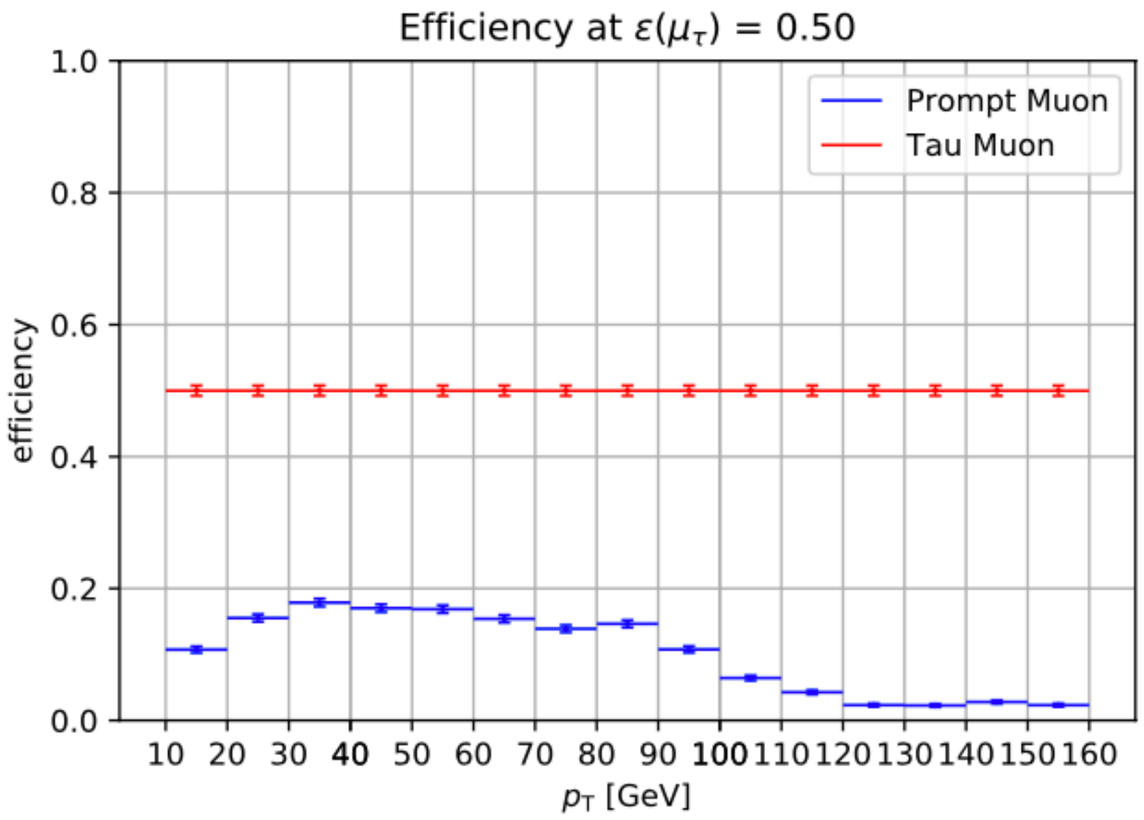


Flat sampling with mixed set

- Bin-wise training: trained a DNN for each section (10,40), (40,90), (90,160) GeV individually



(a) "Inclusive" training



(b) Bin-wise training