Efficiency parametrization of b-tagging classifier using Graph Neural Networks

CMS DP-2022/051

Krunal Gedia,
On behalf of the CMS Collaboration

3rd November 2022
ML4Jets 2022
Rutgers University

- **Introduction and DeepCSV b-tagging classifier**

- Jet classifier efficiency measurement algorithms
  - Using selection cuts
    - Direct Tagging
  - Using efficiency weights
    - Efficiency Map (2D parameterization)
    - GNN-based MVA approach (nD parameterization)

- GNN approach in detail

- Results

- Conclusion

Jets

- Due to QCD confinement, free quarks produced in collision undergo hadronization to create an colorless objects which are reconstructed as jets by jet algorithms.
- This study uses jets reconstructed using the anti-kt algorithm[0].

Tagging of Jet flavour for MC events

Gen jets:
Apply anti-kt algorithm on final gen-state particles (excluding neutrinos).

Reco jets:
Apply anti-kt algorithm on Particle-Flow (PF) candidates.

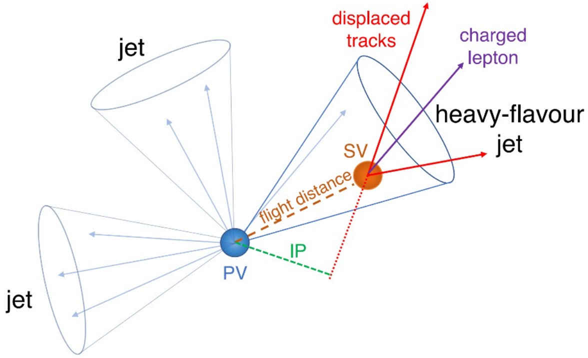
Jet flavour is determined by **matching gen and reco jets** (with $\Delta R < 0.25$)

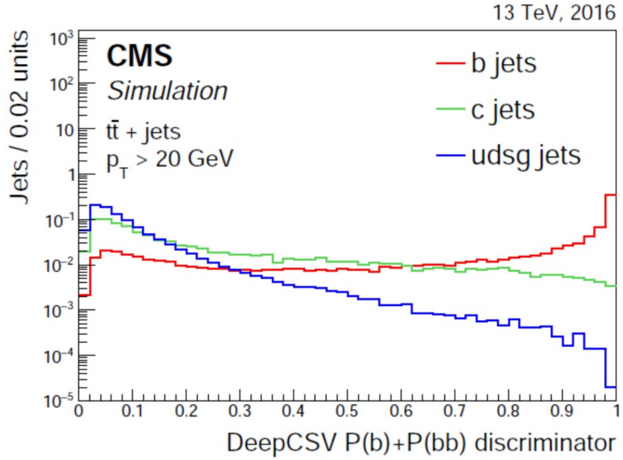
- If gen jet have **b-hadron** → **b-jet**
- If gen jet have **c-hadron** → **c-jet**
- Remaining jets → **l(udsg)-jets**

DeepCSV [1]:
Jet flavour classification algorithm

- Dense layer neural network.
- Training data: QCD and TT MC.
- Input variables: Track, secondary & global vertex features.
- Architecture: 4 hidden layers (100 nodes each)
- Output classes: P(b), P(bb), P(c), P(cc), P(udsg)

- **DeepCSV classifier score used in this talk = <u>P(b) + P(bb)</u>**





[JINST 13 (2018) P05011]

- Different thresholds (working points s_{wp}) on the classifier score (s) are used to quantify the performance of the classifier.

$$\frac{N_{udsg}\left(s > s_{\mathrm{wp}}\right)}{N_{udsg}} \in [0, 1]$$

where N_{udsg} is the number of light jets, s is the classifier score and s_{wp} is the classifier score at working point.

- Three standard working points are used in CMS based on the light-flavour mis-tag rate.

Working point	Light-flavour jet mis-tag rate
Loose	10 %
Medium	1 %
Tight	0.1 %

- Introduction and DeepCSV b-tagging classifier
- **Jet classifier efficiency measurement algorithms**
  - **Using selection cuts**
    - Direct Tagging
  - Using efficiency weights
    - Efficiency Map (2D parameterization)
    - GNN-based MVA approach (nD parameterization)
- GNN approach in detail
- Results
- Conclusion

Selection cut approach to estimate classifier efficiency wrt to a working point

Direct tagging

- Apply a classification cut and select events above the working point threshold.
- The efficiency is the ratio of number of selected events by total number of events.

$$\varepsilon = \frac{N_{\text{selected}}}{N_{\text{total}}} \in [0, 1]$$

- <u>Limited statistical precision</u>
  - Uncertainty of the efficiency measurement by this approach depends on the number of events that can be simulated in a given region of phase space.

- Introduction and DeepCSV b-tagging classifier
- **Jet classifier efficiency measurement algorithms**
  - Using selection cuts
    - Direct Tagging
  - **Using efficiency weights**
    - Efficiency Map (2D parameterization)
    - GNN-based MVA approach (nD parameterization)
- GNN approach in detail
- Results
- Conclusion

Efficiency weight approach to estimate classifier efficiency wrt to a working point

Efficiency weighting

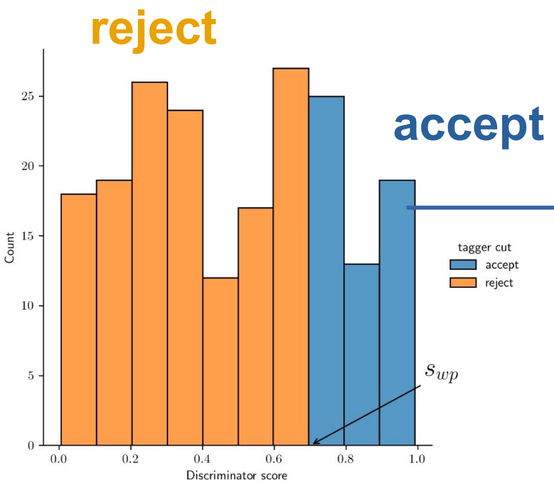
- Apply a weight to each jet instead of just accepting/rejecting cut based-approach of direct tagging.
- The weight corresponds to classifier efficiency for a given working point.

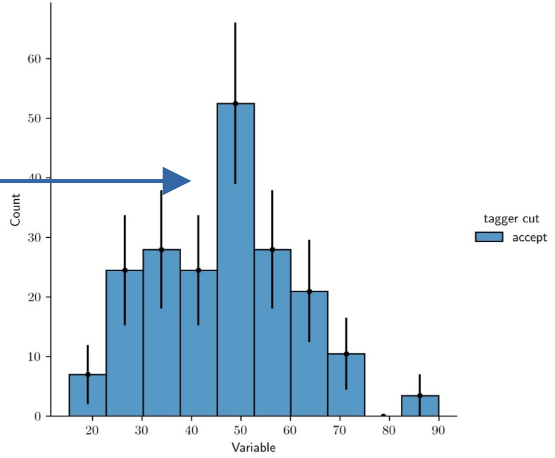
$$\varepsilon = \frac{N\left(s > s_{\mathrm{wp}}\right)}{N} \in [0, 1]$$

- This method is also called Truth Tagging (TT).

Ways to estimate classifier efficiency wrt to a working point

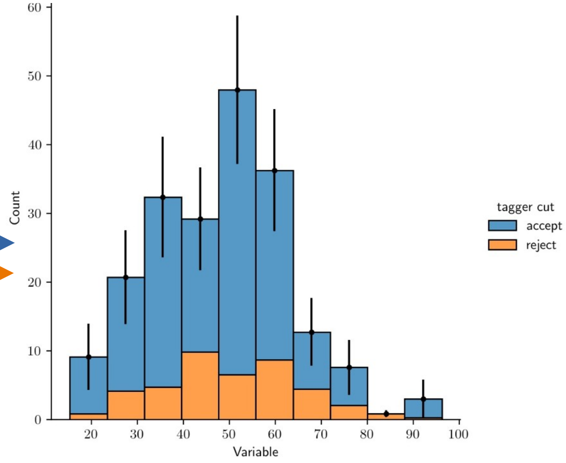
Direct tagging





Efficiency weighing

Efficiency weights for each jet/event wrt s_{WP}



How to calculate efficiency weights?

- In order to achieve parameterization of classifier efficiency weights, a set of low level observables (θ) are chosen which capture the dependency of classifier score (s(x)) where x are generally the high level variables.

$$\varepsilon(\theta) = \frac{N\left(s(x) > s_{\mathrm{wp}} \mid \theta\right)}{N(\theta)}$$

- Example choices for θ are variables like
  - transverse momentum (p_T) which captures dependencies corresponding to secondary vertex reconstruction
  - Pseudo-rapidity (η) which captures dependencies corresponding to the track reconstruction.

- Way to estimate (approximate) efficiency weights:
  - Efficiency maps (2 dimensional parameterization)
  - GNN (multidimensional parameterization)

- Introduction and DeepCSV b-tagging classifier
- **Jet classifier efficiency measurement algorithms**
  - Using selection cuts
    - Direct Tagging
  - **Using efficiency weights**
    - Efficiency Map (2D parameterization)
    - GNN-based MVA approach (nD parameterization)
- GNN approach in detail
- Results
- Conclusion

Efficiency Maps

- Here, the efficiency weights of each flavour of jet (f) are parameterized in bins of p_T and η.
- From the per-jet weights, per-event weights are estimated depending on the number of b-tagged jets required in the analysis

$$\varepsilon_{\mathrm{f},i,j} = \frac{N_{\mathrm{f}}\left(s > s_{\mathrm{wp}} \wedge p_{\mathrm{T}_i} \leq p_{\mathrm{T}} < p_{\mathrm{T}_{i+1}} \wedge \eta_j \leq \eta < \eta_{j+1}\right)}{N_{\mathrm{f}}\left(p_{\mathrm{T}_i} \leq p_{\mathrm{T}} < p_{\mathrm{T}_{i+1}} \wedge \eta_j \leq \eta < \eta_{j+1}\right)}$$

- Main limitation:
  - Finite correlation corresponding to 2D are captured.
  - Can not account for environment effects.
  - Correlations between jets are also neglected.

- Introduction and DeepCSV b-tagging classifier
- **Jet classifier efficiency measurement algorithms**
  - Using selection cuts
    - Direct Tagging
  - **Using efficiency weights**
    - Efficiency Map (2D parameterization)
    - GNN-based MVA approach (nD parameterization)
- GNN approach in detail
- Results
- Conclusion

GNN approach

- First proposed by researchers at ATLAS [2].
- Takes full event as input and provides simultaneous efficiency weights for each jet flavour and for each of the standard working points.
- This approach could also capture higher order correlations, environmental effects, correlations among jets of an event.
- No binning of parameterization variables required as done in efficiency map approach.

- Introduction and DeepCSV b-tagging classifier
- Jet classifier efficiency measurement algorithms
  - Using selection cuts
    - Direct Tagging
  - Using efficiency weights
    - Efficiency Map (2D parameterization)
    - GNN-based MVA approach (nD parameterization)
- **GNN approach in detail**
- Results
- Conclusion

MC dataset used in the study

$t\bar{t}$ sample

- dilepton decay channel
- 3 million training events

QCD sample

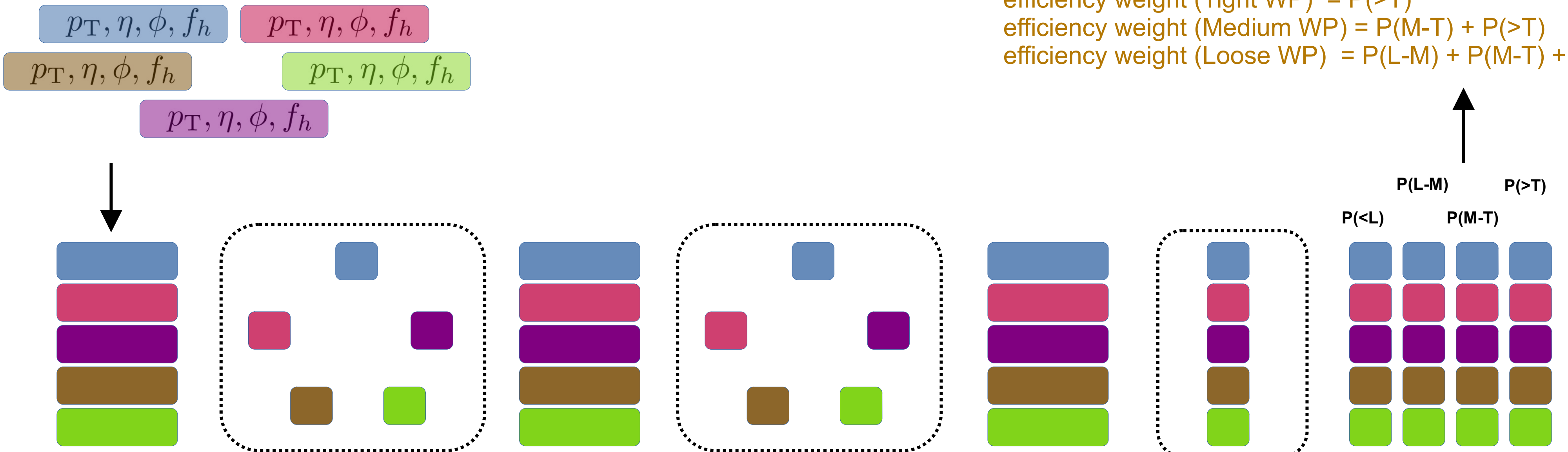
- pp collisions into multi-jet events
- Leading jet p_T slice: 300-470 GeV
- Enriched in muons
- 600k training events

GNN architecture

Event with 5 jets

Efficiency weight predictions

efficiency weight (Tight WP) = P(>T)
efficiency weight (Medium WP) = P(M-T) + P(>T)
efficiency weight (Loose WP) = P(L-M) + P(M-T) + P(>T)



Njets X jet features

5 blocks of GNN
with skip connection after each block

Njets X hidden features

GATv2* [4]

Njets X hidden features

5 hidden layer
feed forward neural
network

Njets X 4 output nodes

Message passing function : MLP
Aggregation function : Sum
Update function : MLP

* Graph attention network (v2) is used only for QCD multi-jet training/evaluation

GNN architecture

$$
\begin{aligned}
\text{Node Input features } (\mathbf{v}_f) &= p_\mathrm{T}, \eta, \phi \text{ (azimuthal angle)}, f_h \text{ (jet flavour)} \\
\text{Embedding vector dimension for } f_h &= 2 \\
\text{Edge Input features } (\mathbf{e}_f) &= \Delta R \\
\text{Output classes} &= \text{4 DeepCSV WP categories (<Loose, Loose-Medium, Medium-Tight, >Tight)} \\
\text{train : val : test} &= 0.95 * 0.75 : 0.05 * 0.75 : 0.25 \\
\mathrm{GNN} &= \left(\text{five blocks with } d_{\mathbf{e}'_h} = 256 \text{ and } d_{\mathbf{v}'_h} = 512 \right), \\
\mathrm{GATv2} &= (\text{ eight heads with } d_{\mathbf{head}} = 64 \text{ and total output features per node} = 512) , \\
\text{feed forward hidden layers } &= \{512, 256, 128, 50\}. \\
p^{\text{edge}}_{\text{dropout}}, p^{\text{node}}_{\text{dropout}}, p^{\text{GATv2}}_{\text{dropout}}, p^{\text{ffNN}}_{\text{dropout}} &= 30\%, 30\%, 10\%, 30\%
\end{aligned}
$$

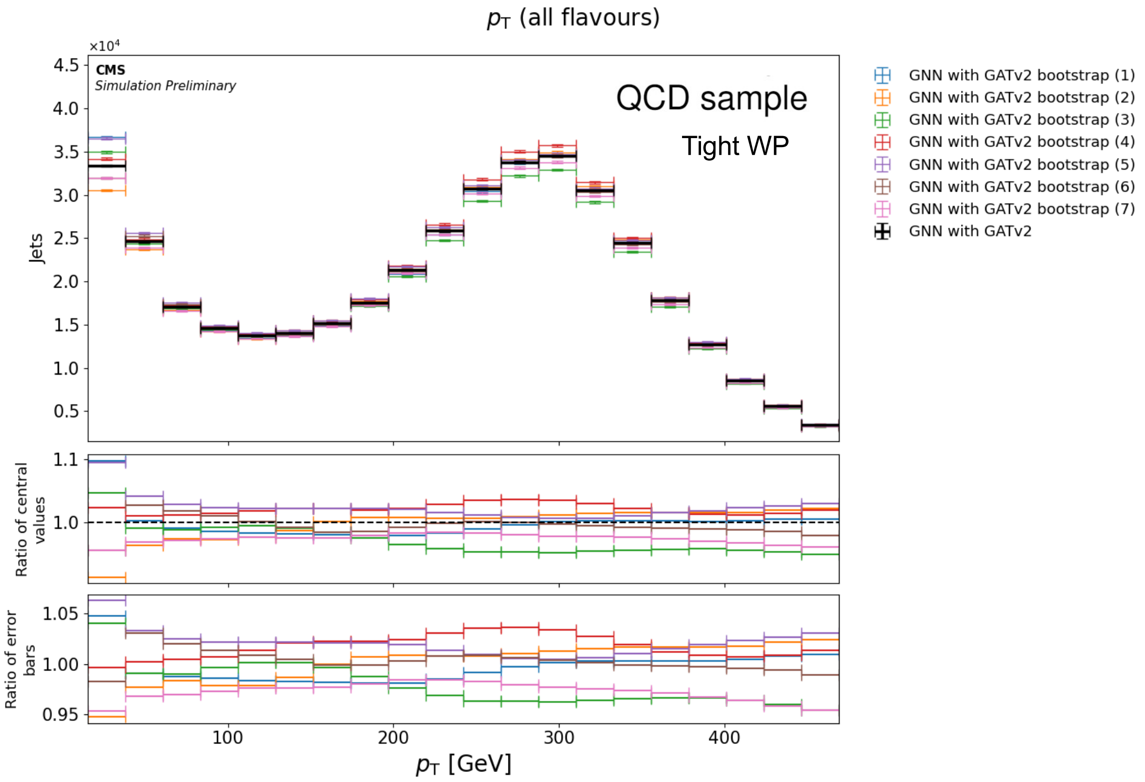
Bagging

- Bootstrap aggregation is performed to get the central value and uncertainty of the final GNN predictions.
- Multiple trainings are performed using a training set sampled with replacement.
- For each bootstrap training, corresponding predictions are evaluated on the common test set.
- The central value of an observable in the final result is the median values of the central value of the histogram bins of all the bootstrap sample.
- The uncertainty on the histogram bin of the final result is the corresponding statistical uncertainty of the bin.

Top pad:
Distribution evaluated from each of the bootstrap trainings and the ensemble using the median efficiency predictions.

Middle pad:
Ratio of the bin values of individual bootstrap trainings with the ensemble.

Bottom pad:
Ratio of the statistical uncertainty of the individual bootstrap trainings with the ensemble.



Evaluation metrics

Evaluation metrics (chi-squared distance (χ2)):

- Quantify closure of efficiency weight based methods (efficiency map and GNN approach) wrt to direct tagging.
- We use χ2 distance metric to quantify closure of histograms of an observable of efficiency weight based methods (H_{ew}) (efficiency map and GNN approach) wrt to direct tagging (H_{dt}).
- For the GNN approach, we apply a bagging procedure as described previously to obtain the central value and uncertainty of an observable.

$$\mathrm{D}_{\chi^2}\,(H_{\mathrm{ew}}, H_{\mathrm{dt}}) = \sum_{i=1}^{N_{\text{bins}}} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{N_{\text{bins}}} \frac{(b_{\mathrm{ew},i} - b_{\mathrm{dt},i})^2}{b_{\mathrm{dt},i}}$$

- Introduction and DeepCSV b-tagging classifier
- Jet classifier efficiency measurement algorithms
  - Using selection cuts
    - Direct Tagging
  - Using efficiency weights
    - Efficiency Map (2D parameterization)
    - GNN-based MVA approach (nD parameterization)
- GNN approach in detail
- **Results**
- Conclusion

Results

Efficiency predictions obtained for Tight WP

$t\bar{t}$ sample

χ^2		
	Efficiency map	GNN
p_T(j)	203.86	113.78
η(j)	350.01	103.53
ϕ(j)	145.34	61.81
m(j)	232.11	186.12
area(j)	105.30	85.79
m(jj)	22.16	11.71
ΔR(jj)	25.85	11.00

QCD sample

χ^2		
	Efficiency map	GNN with GATv2
p_T(j)	20.02	14.66
η(j)	50.11	13.31
ϕ(j)	24.67	18.12
m(j)	25.84	14.96
area(j)	15.67	7.15
m(jj)	22.55	6.81
ΔR(jj)	23.83	8.18

Efficiency predictions obtained for Medium WP

$t\bar{t}$ sample

	χ^2	
	Efficiency map	GNN
p_T(j)	388.09	303.01
η(j)	5997.29	2441.21
ϕ(j)	192.82	153.69
m(j)	358.32	314.00
area(j)	199.80	174.52
m(jj)	48.52	26.17
ΔR(jj)	48.02	26.89

QCD sample

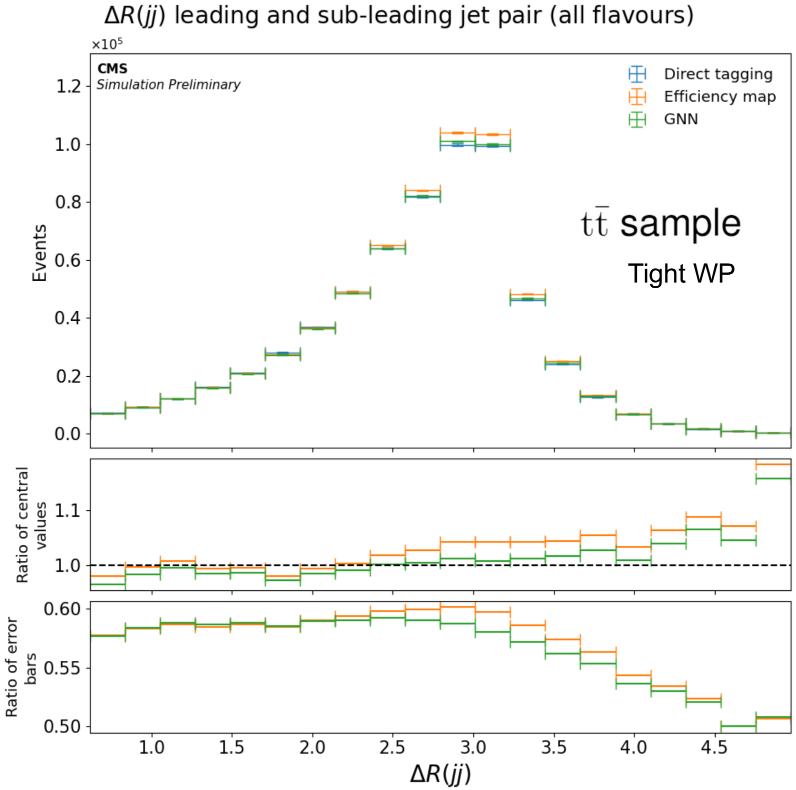
	χ^2	
	Efficiency map	GNN with GATv2
p_T(j)	71.18	37.04
η(j)	731.08	67.73
ϕ(j)	34.24	20.29
m(j)	81.17	44.84
area(j)	36.52	17.12
m(jj)	24.72	6.56
ΔR(jj)	24.81	7.33

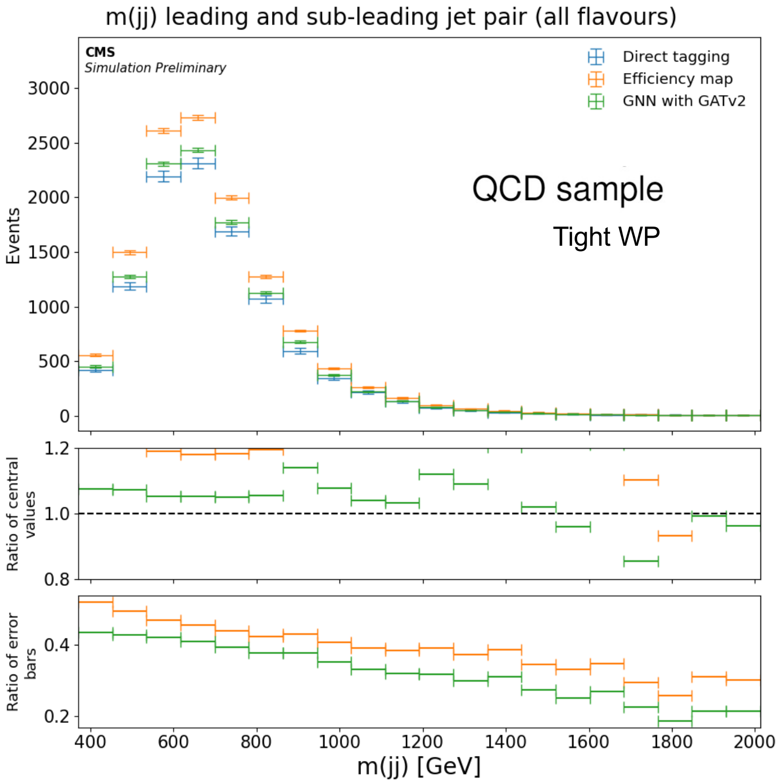
Based on the χ^2 metric, we thus see that the GNN is able to capture higher order correlations and environment effects and outperforms the efficiency map approach for both Tight and Medium working point predictions.

Top pad:
Histograms resulting from direct tagging, the efficiency map, and the GNN model.

Middle pad:
Ratio of the central values as predicted by the efficiency map and GNN to direct tagging.

Bottom pad:
Ratio of the statistical uncertainty of the bin values resulting from efficiency map and GNN to direct tagging.





- Introduction and DeepCSV b-tagging classifier
- Jet classifier efficiency measurement algorithms
  - Using selection cuts
    - Direct Tagging
  - Using efficiency weights
    - Efficiency Map (2D parameterization)
    - GNN-based MVA approach (nD parameterization)
- GNN approach in detail
- Results
- **Conclusion**

Conclusion

- Estimating the efficiency of a process using selection cuts is statistically limited by the number of events that could be simulated in a given phase space.
- Efficiency weighting approaches could help to mitigate this issue.
- Traditional approach of efficiency map parameterized in p_T & η suffers from curse of dimensionality. It also can not account for higher order correlations, environment effects or correlations among jets in an event.
- GNN based approach presented here helps to solve these issues and as shown, outperforms the traditional efficiency map approach for both Tight and Medium working point predictions.
- Also, as expected, the efficiency weights approach leads to significant gains in statistical uncertainty with respect to the direct tagging approach.

References

[0]: M. Cacciari, G. P. Salam and G. Soyez, “The anti-kt jet clustering algorithm,” JHEP 0804 (2008) 063.

[1]: CMS Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”, JINST 13 (2018) no.05, P05011, doi:10.1088/1748-0221/13/05/P05011

[2]: Di Bello, Francesco Armando, et al. "Efficiency parameterization with neural networks." Computing and software for big science 5.1 (2021): 1-12.

[3]: Brody, Shaked, Uri Alon, and Eran Yahav. "How attentive are graph attention networks?." arXiv preprint arXiv:2105.14491 (2021).

Back - up

Gain in statistical uncertainty due to efficiency weighting technique over direct tagging

If b is the number of events in a bin, σ_b is statistical error of the bin and w_i is the weight of the ith event in the bin,

$$b = \sum_{i=1}^{N} w_i \quad \sigma_b = \sqrt{\sum_{i=1}^{N} w_i^2}$$

Since number of events by theory are same in a bin for efficiency weighted histogram (ew) and direct tagging histogram (dt)

$$b_{\text{ew}} = N_{\text{total}} \cdot \varepsilon = N_{\text{selected}} = b_{\mathrm{dt}}$$

where ε is the efficiency weight of an event in the efficiency weighted histogram.

Assuming weight of 1 for direct tagging histogram,

$$\sigma_{b_{\mathrm{dt}}} = \sqrt{b_{\mathrm{dt}}},$$

Assuming ε as the weight of each event in bin of efficiency weighted histogram,

$$\sigma_{b_{\text{ew}}} = \sqrt{N_{\text{total}} \cdot \varepsilon^2} = \sqrt{\varepsilon} \cdot \sqrt{b_{\mathrm{ew}}},$$

$$\sigma_{b_{\mathrm{ew}}} = \sqrt{\varepsilon} \cdot \sigma_{b_{\mathrm{dt}}}$$

Thus, the statistical uncertainty is reduced by a factor of $\sqrt{\varepsilon}$.

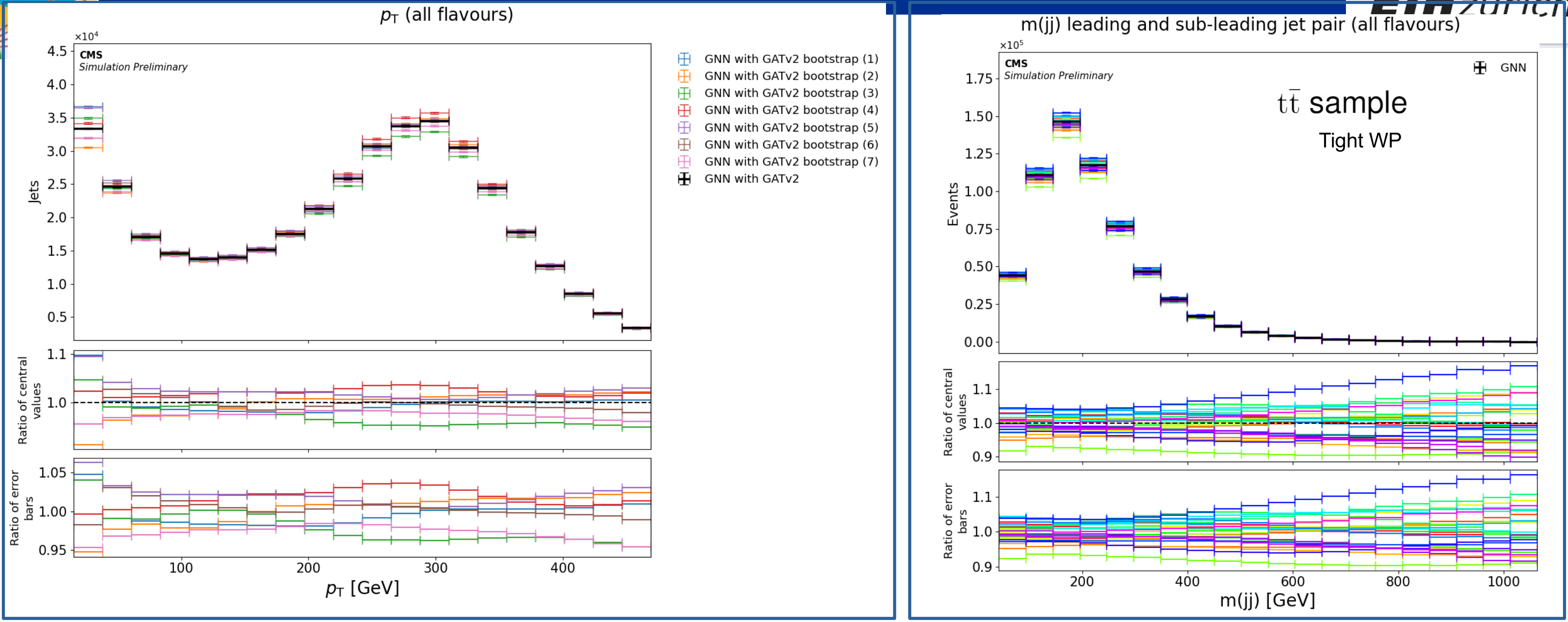


Figure 1: The left plot shows the distribution of transverse momentum of the jet (p_T(j)) for QCD sample and the right plot shows the distribution of dijet invariant mass (m(jj)) of leading and subleading jet in $t\bar{t}$ sample. The top pad shows the predictions for the bootstrap sample (colorful markers) and the ensemble (black marker) obtained using the median of the central values of the bootstrap samples. For the left plot, all 7 individual bootstrap samples are listed in the legend. The right plot contains 25 bootstrap samples. The middle pad shows the ratio of the central values of each of the bootstrap samples with the ensemble. The predictions of individual bootstrap samples are around 5% of their median value. The bottom pad shows the ratio of statistical uncertainty of individual bootstrap training with the ensemble.

$t\bar{t}$ sample - Tight working point

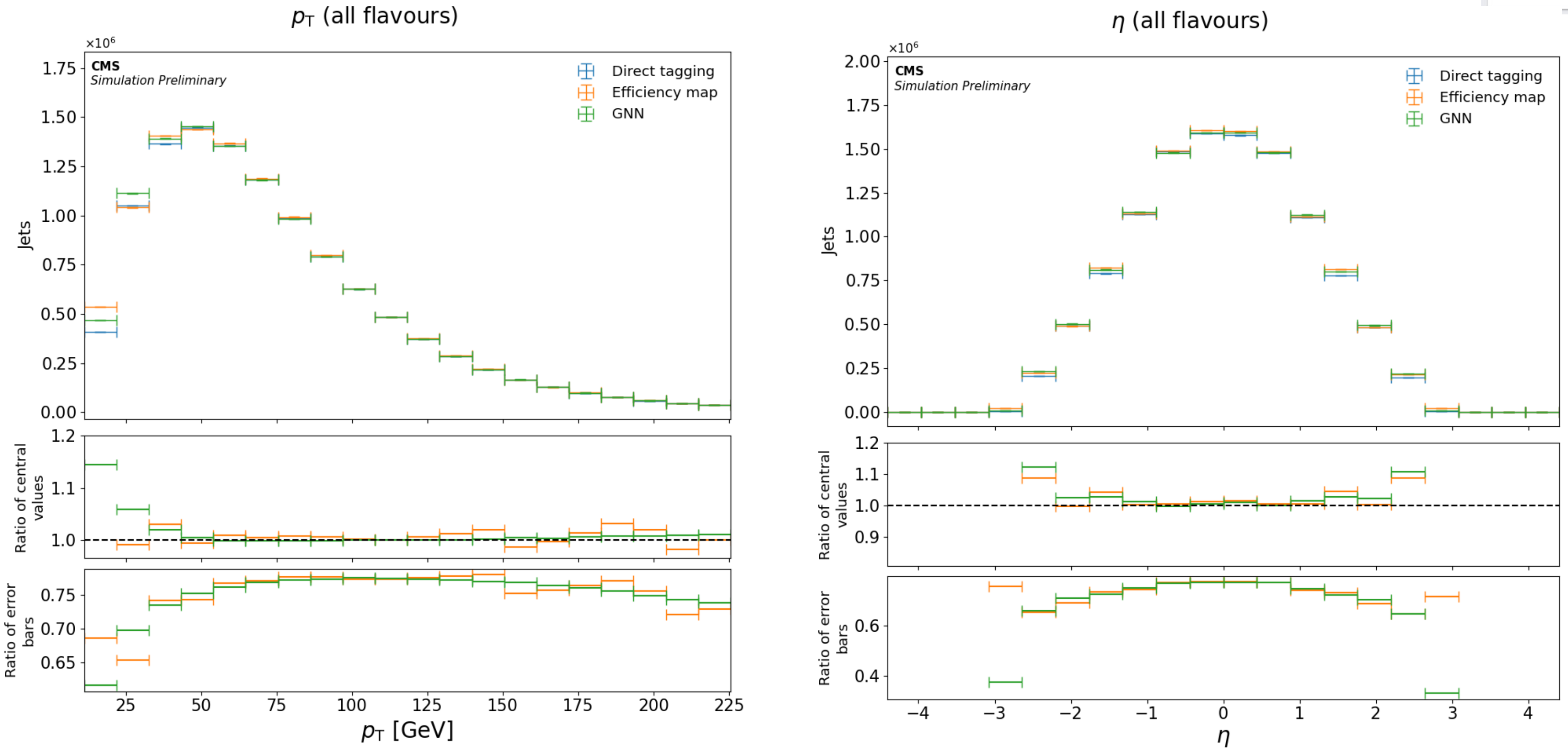


Figure 2: The left plot shows the transverse momentum distribution of the jet while the right plot shows the pseudo-rapidity distribution of the jet of the $t\bar{t}$ sample for the Tight working point. The top pad shows the histograms resulting from the direct tagging, efficiency map, and the GNN approach. The middle pad shows the ratio of the central value of the efficiency map and the GNN approach to direct tagging. The bottom pad shows the ratio of the statistical uncertainty of the bin values resulting from the efficiency map and the GNN approach to direct tagging.

$t\bar{t}$ sample - Tight working point

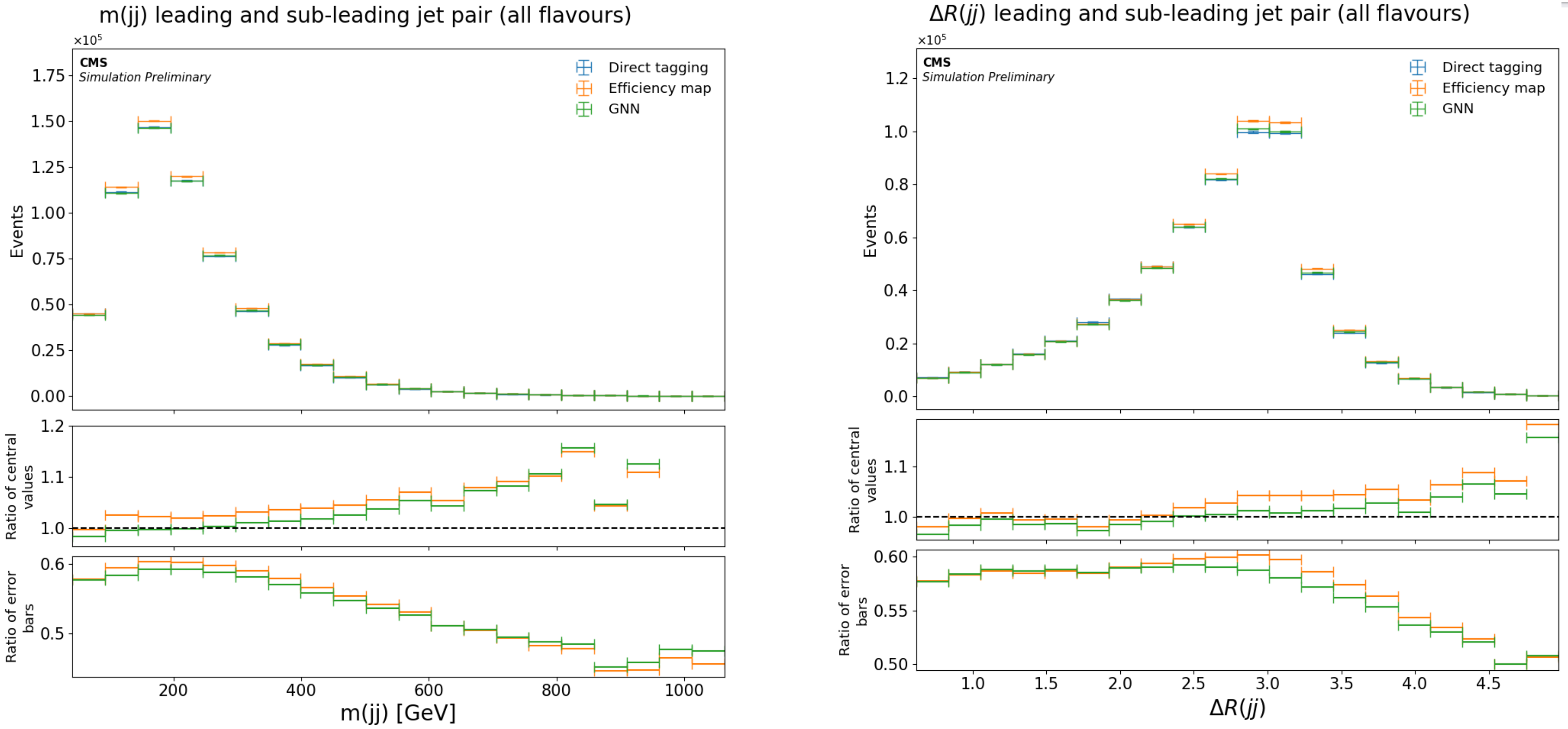


Figure 3: The left plot shows the dijet invariant mass distribution of the leading and sub-leading jet pair while the right plot shows the ΔR distribution of the leading and sub-leading jet pair for the Tight working point. The top pad shows the histograms resulting from the direct tagging, efficiency map, and the GNN approach. The middle pad shows the ratio of the central value of the efficiency map and the GNN approach to direct tagging. The bottom pad shows the ratio of the statistical uncertainty of the bin values resulting from the efficiency map and the GNN approach to direct tagging.

QCD sample - Tight working point

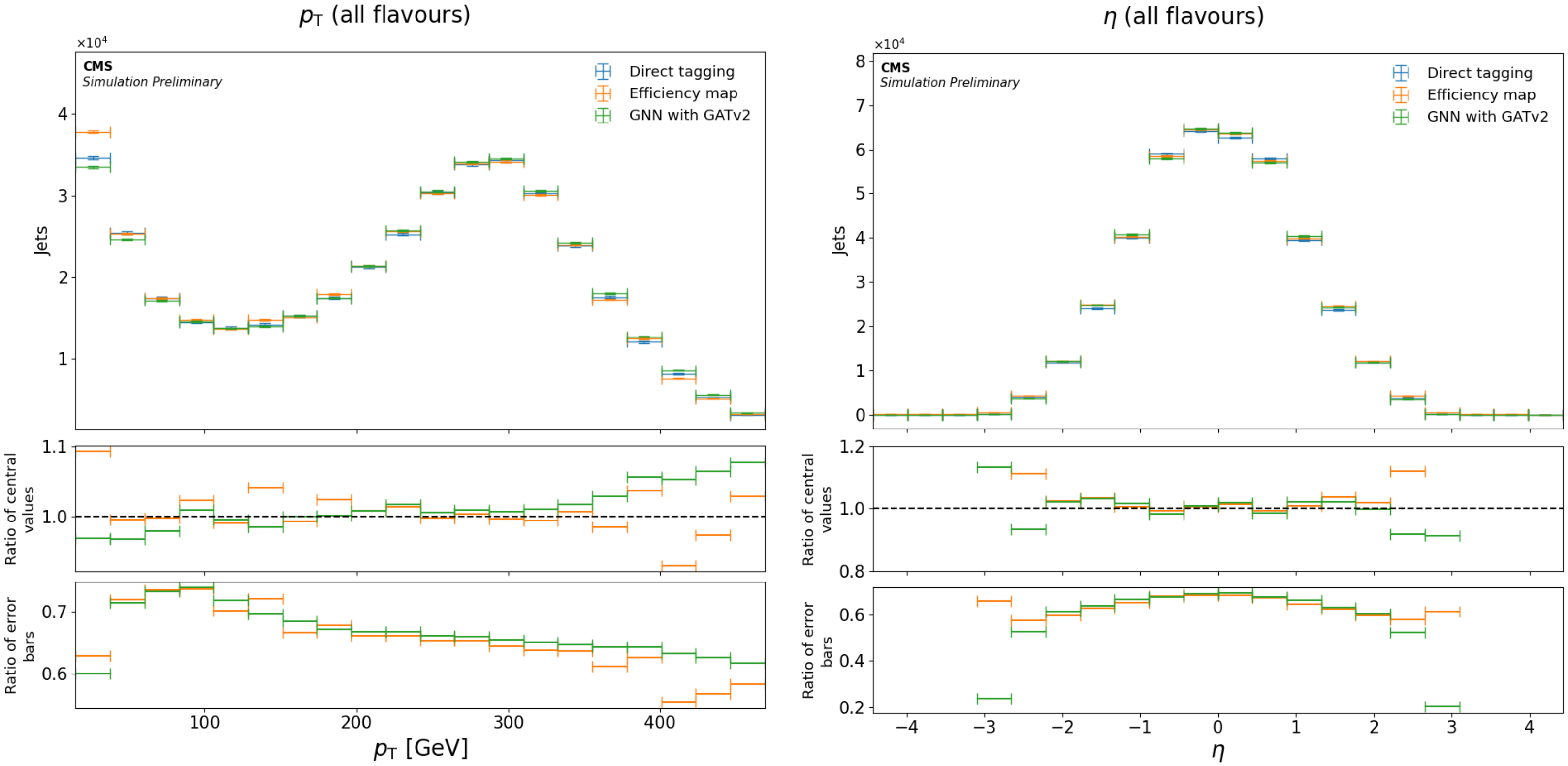


Figure 4: The left plot shows the transverse momentum distribution of the jet while the right plot shows the pseudo-rapidity distribution of the jet of the QCD sample for the Tight working point. The top pad shows the histograms resulting from the direct tagging, efficiency map, and the GNN approach. The middle pad shows the ratio of the central value of the efficiency map and the GNN approach to direct tagging. The bottom pad shows the ratio of the statistical uncertainty of the bin values resulting from the efficiency map and the GNN approach to direct tagging.

QCD sample - Tight working point

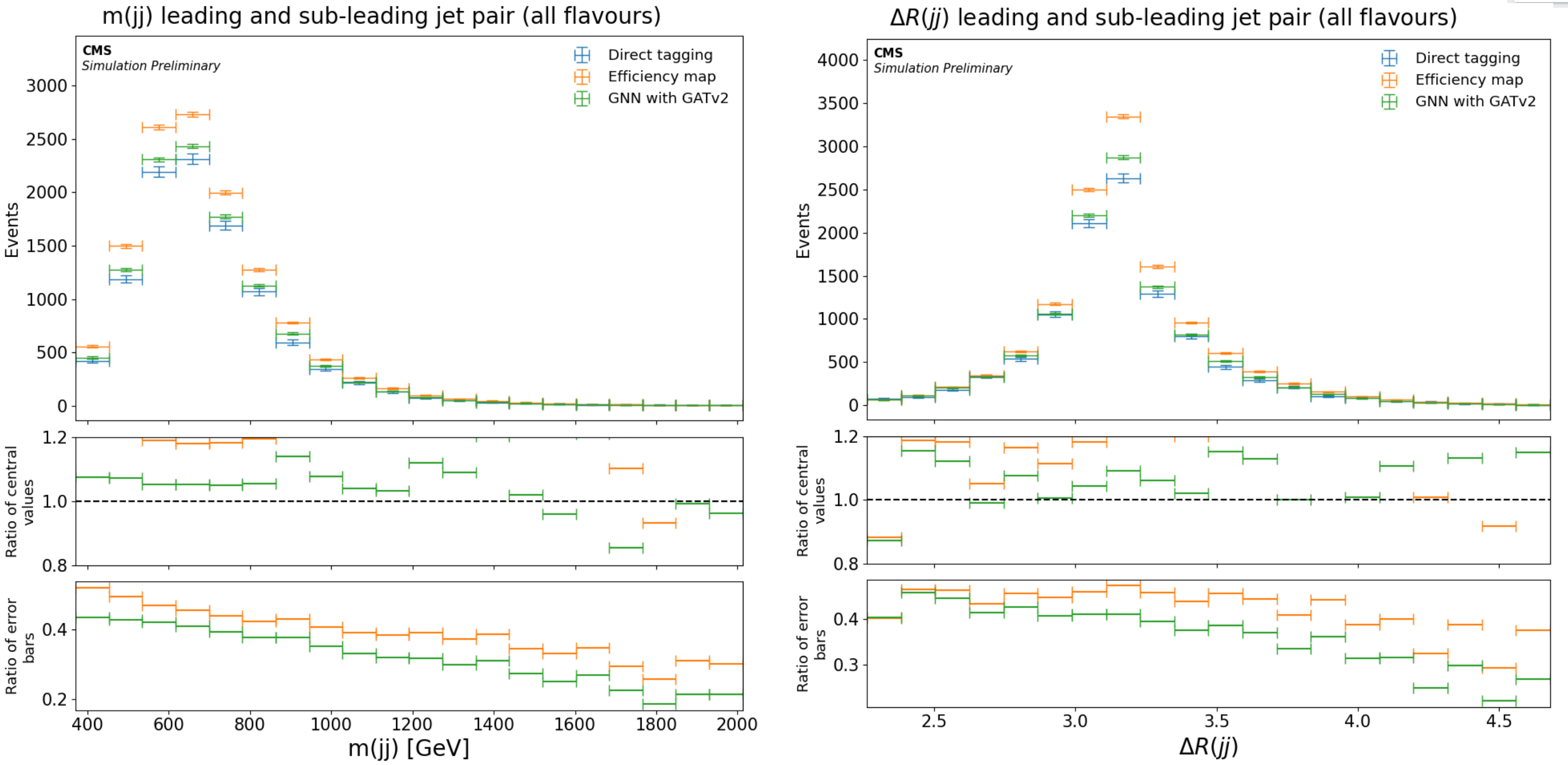


Figure 5: The left plot shows the dijet invariant mass distribution of the leading and sub-leading jet pair while the right plot shows the ΔR distribution of the leading and sub-leading jet pair for the Tight working point. The top pad shows the histograms resulting from the direct tagging, efficiency map, and the GNN approach. The middle pad shows the ratio of the central value of the efficiency map and the GNN approach to direct tagging. The bottom pad shows the ratio of the statistical uncertainty of the bin values resulting from the efficiency map and the GNN approach to direct tagging.

GNN block

Vertices $V = \{\mathbf{v}_i\}$ for $i \in [N^v]$ where $\mathbf{v}_i = (\mathbf{v}_{f,i}, \mathbf{v}_{h,i})$

Edges $E = \{(\mathbf{e}_j, d_j, s_j)\}$ for $j \in [N^e]$ where $\mathbf{e}_j = (\mathbf{e}_{f,j}, \mathbf{e}_{h,j})$

Algorithm 1 Pseudo-code of the computations inside one GN block.

function GRAPHNETWORKBLOCK($G = (V, E)$)
  for $j \in \{1, \ldots, N^e\}$ **do**
    $\mathbf{e}'_{h,j} \leftarrow \phi^e\left(\mathbf{v}_{f,d}, \mathbf{v}_{h,d}, \mathbf{v}_{f,s}, \mathbf{v}_{h,s}, \mathbf{e}_{f,j}\right)$ ▷ Update edge hidden state
    $\mathbf{e}'_{f,j} \leftarrow \mathbf{e}_{f,j}$ ▷ Leave edge features unchanged
  end for
  for $i \in \{1, \ldots, N^v\}$ **do**
    $\mathbf{a} = \phi^v_a(\mathbf{v}_{f,i}, \mathbf{v}_{h,i})$
    let $E'_i = \{(\mathbf{e}'_j, d_j, s_j) \in E \mid d_j = i\}$
    $\bar{\mathbf{e}}'_i \leftarrow \sum_{E'_i} \mathbf{e}'_{h,j}$ ▷ Sum edge hidden states of incoming edges
    $\mathbf{b} = \phi^v_b(\bar{\mathbf{e}}'_i)$
    $\mathbf{v}'_{h,i} \leftarrow$ Normalise(Concatenate($\mathbf{a}, \mathbf{b}$)) ▷ Update vertex hidden state
    $\mathbf{v}'_{f,i} \leftarrow \mathbf{v}_{f,i}$ ▷ Leave vertex features unchanged
  end for
  let $V' = \{\mathbf{v}'_i\}$ for $i \in [N^v]$
  let $E' = \{\mathbf{e}'_j\}$ for $j \in [N^e]$
  return $G' = (V', E')$
end function