서울시립대학교
UNIVERSITY OF SEOUL

Zero-Permutation Jet-Parton Assignment

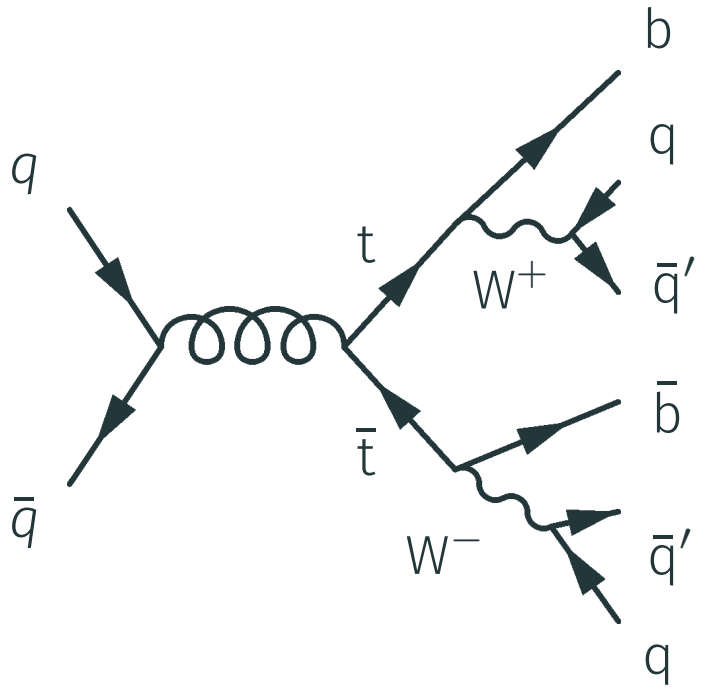
Top quark pair reconstruction using an attention-based neural network

Jason Sang Hun Lee, Inkyu Park, Ian James Watson, and Seungjin Yang
University of Seoul

4th Inter-experiment Machine Learning Workshop
October 21, 2020

Introduction

- The reconstruction of events with intermediate states decaying to jets requires a technique to assign jets to partons.
- The number of jets can be greater than the number of partons because of additional QCD radiation. It makes the assignment harder.
- We introduce **the self-attention for jet assignment (SAJA) network without requiring jet permutations**. We apply SAJA to find jet-parton assignments of fully-hadronic $t\bar{t}$ events to test the performance.



Preceding Research

- χ^2 method [CMS, Eur. Phys. J. C 79 (2019) 313]

$$\chi^2 = \sum_{j \in \text{jets}} \left[\frac{(p_{T,j}^{\text{reco}} - p_{T,j}^{\text{fit}})^2}{\sigma_{p_{T,j}^2}} - \frac{(\eta_j^{\text{reco}} - \eta_j^{\text{fit}})^2}{\sigma_{\eta_j^2}} - \frac{(\phi_j^{\text{reco}} - \phi_j^{\text{fit}})^2}{\sigma_{\eta_j^2}} \right]$$

- Kinematic likelihood fitting [J. Erdmann, Nucl.Instrum.Meth.A 748 (2014) 18-25]

$$\mathcal{L} = B(m_{q_1 q_2 q_3} | m_t, \Gamma_t) \cdot B(m_{q_1 q_2} | m_W, \Gamma_W) \cdot B(m_{q_4 q_5 q_6} | m_t, \Gamma_t) \cdot B(m_{q_4 q_5} | m_W, \Gamma_W) \cdot \prod_{i=1}^{6} W_{\text{jet}}(E_{\text{jet},i}^{\text{meas}} | E_{\text{jet},i})$$

- Machine Learning [M. Erdmann, JINST 12 (2017) P08020], [J. Erdmann, JINST 14 (2019) P11015]

$$\text{correct or wrong} = \text{model}\big(\text{assignment}\big)$$

All of the above methods follow the same steps.

1. Compute all (promising) jet permutations.
2. Evaluate how well each permutation agrees with the underlying event topology. (χ^2, $\mathcal{L}$, the sigmoid output of NN, BDT score)
3. Choose the best permutation.

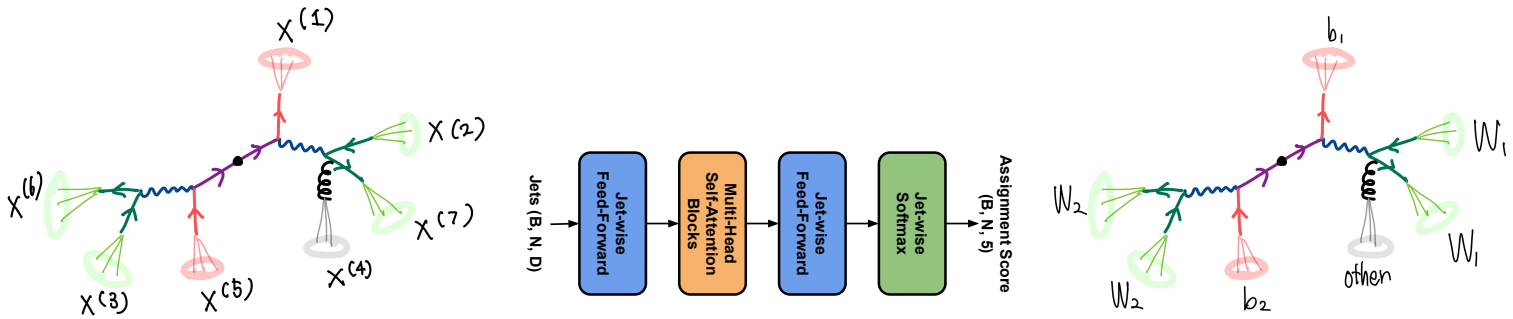
$\Rightarrow$ Combinatorial explosion O(n!)

Modeling

Jet-parton assignment as jet-wise classification.

$$f^{\theta}: \begin{bmatrix} \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(N)} \end{bmatrix} \rightarrow \begin{bmatrix} \hat{y}^{(1)}_{b_1} & \hat{y}^{(1)}_{W_1} & \hat{y}^{(1)}_{b_2} & \hat{y}^{(1)}_{W_2} & \hat{y}^{(1)}_{\text{other}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{y}^{(N)}_{b_1} & \hat{y}^{(N)}_{W_1} & \hat{y}^{(N)}_{b_2} & \hat{y}^{(N)}_{W_2} & \hat{y}^{(N)}_{\text{other}} \end{bmatrix}$$

- Since it is hard to distinguish jets originating from t and jets originating from $\bar{t}$, arbitrary indices 1 and 2 are introduced.
- When the assignment doesn't agree with the underlying topology (e.g. no b_1 or 3 W_2), the assignment is said to be topologically invalid and is not selected.



Objective function

Since the indices 1 and 2 are arbitrary, the general objective function cannot be used here.
Therefore, a new cross-entropy based objective function is introduced.

$$J(\theta) = \frac{1}{N}\sum_{j=1}^{N}[\min\left(\pi_{12}^{(j)}, \pi_{21}^{(j)}\right) + y_{\text{other}}^{(j)} \log \hat{y}_{\text{other}}^{(j)}],$$

where

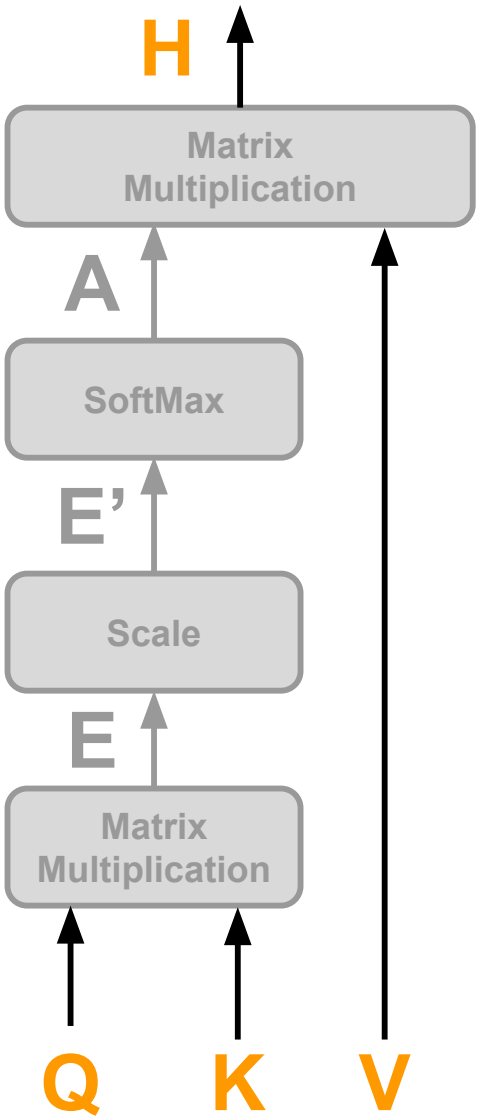
$$\pi_{\alpha\beta}^{(j)} = y_b^{(j)} \log \hat{y}_{b_\alpha}^{(j)} + y_{\bar{b}}^{(j)} \log \hat{y}_{b_\beta}^{(j)} + y_{W^+}^{(j)} \log \hat{y}_{W_\alpha}^{(j)} + y_{W^-}^{(j)} \log \hat{y}_{W_\beta}^{(j)}.$$

$\Rightarrow$ A deep learning model of any architecture can become a **zero-permutation jet-parton assignment model** if it is fit by minimizing the above objective function.

Self-Attention based Model Architecture

- The model implementation is based on TRANSFORMER, which is the neural machine translation model and features **self-attention**. [A. Vaswani, arXiv:1706.03762]
- Self-attention performs a weight sum of the elements (vectors) of the input set, where the weight matrix is also computed from the elements.
  - Sequence = a set of words
  - Image = a set of pixels
  - Event = a set of jets
- Self-attention based model can learn the dependency between elements.

Scaled Dot-Product Attention (0)



- The set is represented as the matrix.

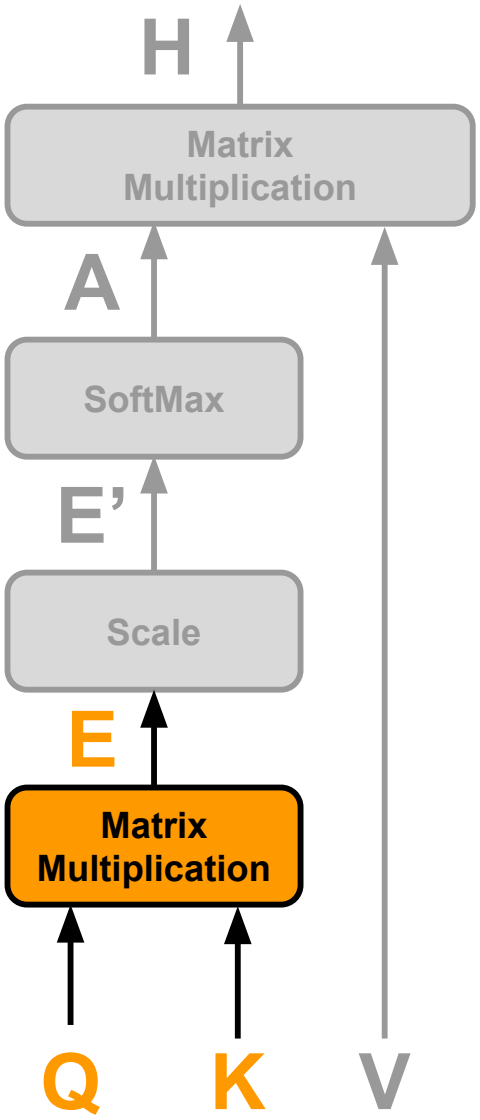
$$X = \begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vec{x}_3 \end{bmatrix}$$

- Attention function takes three sets K, V and Q as input and produces a set. In general, K and V are originated from the same set X.

$$\begin{aligned} H &= Attention(K, V, Q) \\ &= Attention(\phi_K(X), \phi_V(X), Q) \end{aligned}$$

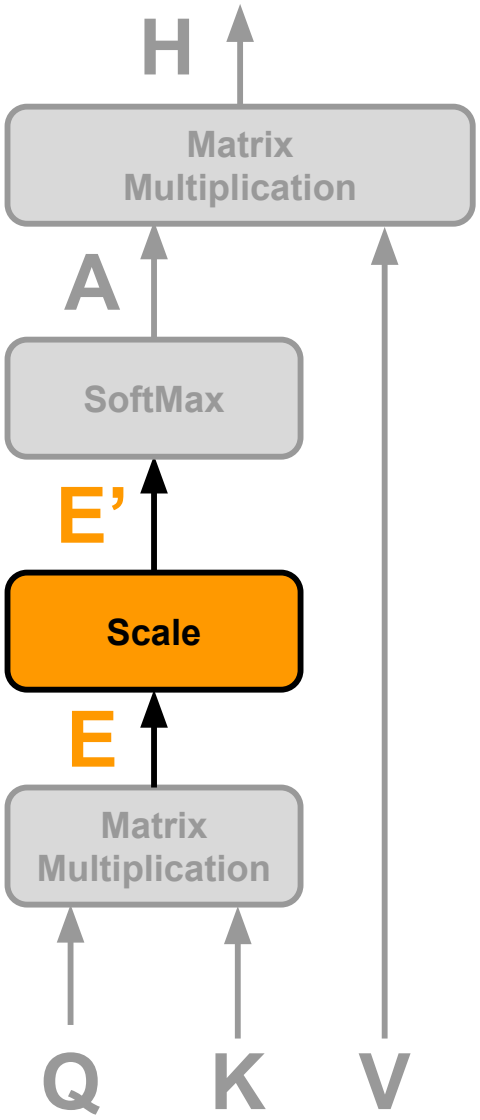
- Self-Attention is a special case of attention, where Q is also came from X.

Scaled Dot-Product Attention (1)



$$
\begin{aligned}
E &= QK^T \\
&= \begin{bmatrix} \vec{q}_1 \\ \vec{q}_2 \\ \vec{q}_3 \end{bmatrix} \begin{bmatrix} \vec{k}_1 & \vec{k}_2 & \vec{k}_3 \end{bmatrix} \\
&= \begin{bmatrix} \vec{q}_1 \cdot \vec{k}_1 & \vec{q}_1 \cdot \vec{k}_2 & \vec{q}_1 \cdot \vec{k}_3 \\ \vec{q}_2 \cdot \vec{k}_1 & \vec{q}_2 \cdot \vec{k}_2 & \vec{q}_2 \cdot \vec{k}_3 \\ \vec{q}_3 \cdot \vec{k}_1 & \vec{q}_3 \cdot \vec{k}_2 & \vec{q}_3 \cdot \vec{k}_3 \end{bmatrix}.
\end{aligned}
$$

Scaled Dot-Product Attention (2)



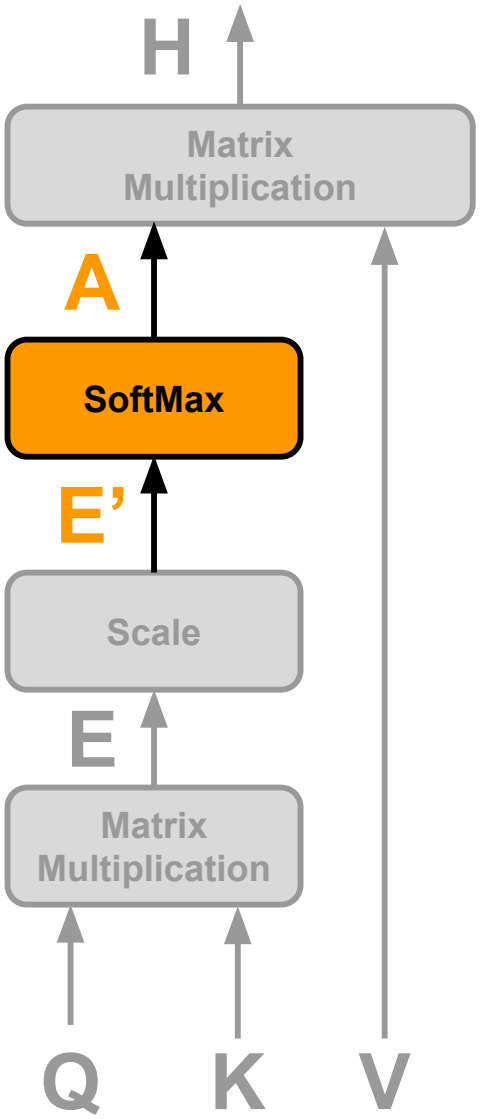
$$E = QK^T$$

$$= \begin{bmatrix} \vec{q}_1 \\ \vec{q}_2 \\ \vec{q}_3 \end{bmatrix} \begin{bmatrix} \vec{k}_1 & \vec{k}_2 & \vec{k}_3 \end{bmatrix}$$

$$= \begin{bmatrix} \vec{q}_1 \cdot \vec{k}_1 & \vec{q}_1 \cdot \vec{k}_2 & \vec{q}_1 \cdot \vec{k}_3 \\ \vec{q}_2 \cdot \vec{k}_1 & \vec{q}_2 \cdot \vec{k}_2 & \vec{q}_2 \cdot \vec{k}_3 \\ \vec{q}_3 \cdot \vec{k}_1 & \vec{q}_3 \cdot \vec{k}_2 & \vec{q}_3 \cdot \vec{k}_3 \end{bmatrix}.$$

$$E' = \frac{1}{\sqrt{dim(\vec{k})}} E.$$

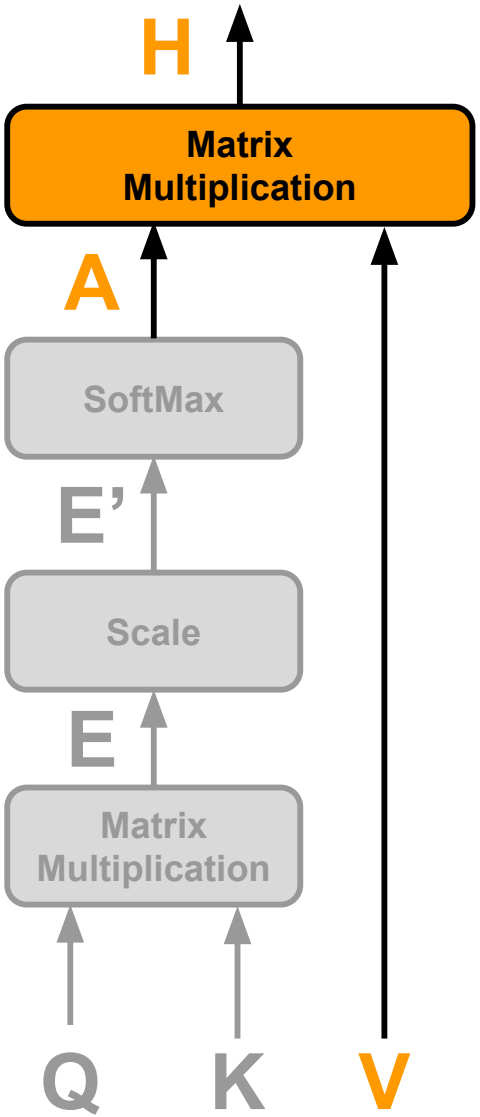
Scaled Dot-Product Attention (3)



$$A = \begin{bmatrix} \frac{e^{E'_{11}}}{Z_1} & \frac{e^{E'_{12}}}{Z_1} & \frac{e^{E'_{13}}}{Z_1} \\ \frac{e^{E'_{21}}}{Z_2} & \frac{e^{E'_{22}}}{Z_2} & \frac{e^{E'_{23}}}{Z_2} \\ \frac{e^{E'_{31}}}{Z_3} & \frac{e^{E'_{32}}}{Z_3} & \frac{e^{E'_{33}}}{Z_3} \end{bmatrix},$$

where $Z_j = \sum_j e^{E'_{ij}}$.

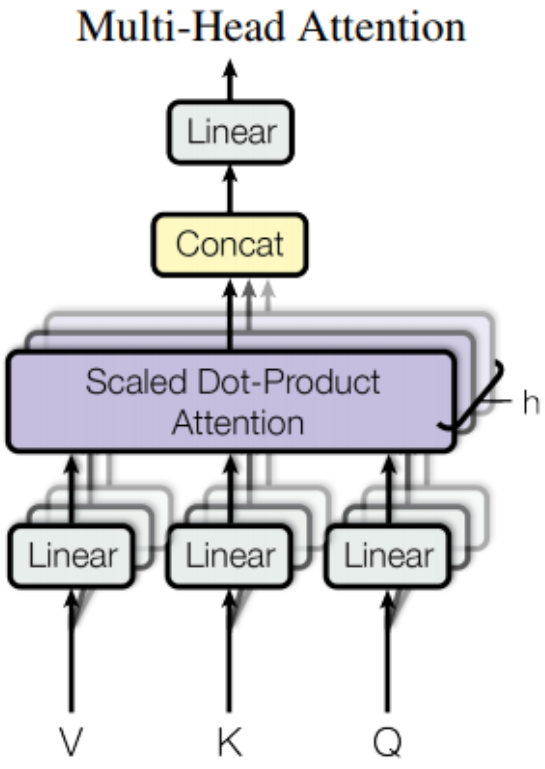
Scaled Dot-Product Attention (4)



$$H = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} \vec{V}_1 \\ \vec{V}_2 \\ \vec{V}_3 \end{bmatrix}$$

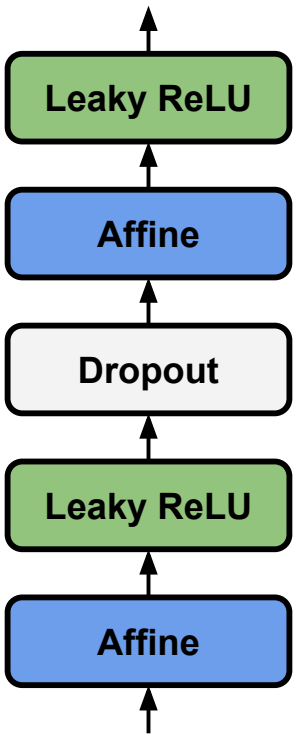
$$= \begin{bmatrix} \sum_i A_{1i}\vec{V}_i \\ \sum_i A_{2i}\vec{V}_i \\ \sum_i A_{3i}\vec{V}_i \end{bmatrix}.$$

Multi-head Attention

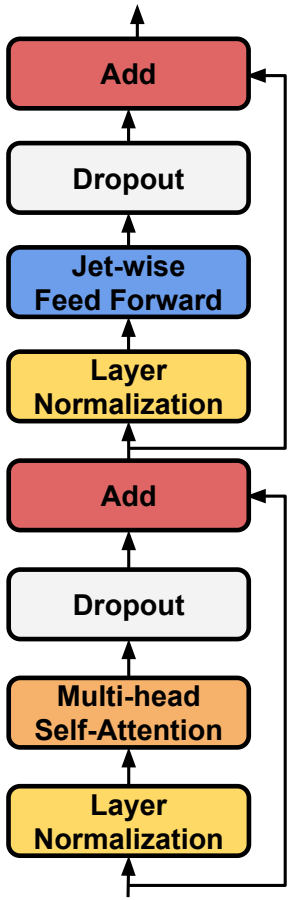


Multi-Head Attention consists of several attention layers running in parallel. Reprinted from "Attention Is All You Need" [arXiv:1706.03762]

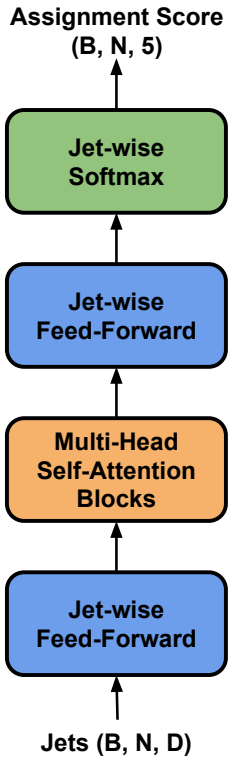
Self-Attention based Model Architecture (SAJA)



(a) Jet-wise feed-forward network



(b) Multi-head self-attention block



(c) SAJA, Jet-parton assignment model

SAJA is invariant to the order of jets and this permutation invariant property can be used to peek inside the black box of SAJA with Monte Carlo information.

Simulation

Generation

- pp collision with $\sqrt{s} = 13\,\text{TeV}$ using MadGraph5_aMC@NLO and Pythia8
- Fully hadronic $t\bar{t}$ with up to two additional jets at NLO, $m_t = 172.5\,\text{GeV}$
- QCD multijet at LO.

Detector response

- Delphes3
- CMS-like detector

Jet Finding

- FastJet3
- anti-k_T algorithm with R=0.4

Selection

Selection follows the trigger selection used in the CMS $t\bar{t}$ all-jets analysis. [CMS, Eur. Phys. J. C 79 (2019) 313].

Jet

- $p_T > 30$ GeV
- $|\eta| < 2.4$
- b-tag: true/fake based on the efficiency for b quark and misidentification rates for gluon, light quark jets and c-jets.

Event

- $N_{\text{jet}} \geq 6$
- At least one b-tagged jet of the six most energetic jets
- $p_T(\text{jet}_6) > 40$ GeV
- $H_T \equiv \sum_{\text{jet}} p_T > 450$ GeV

Jet-Parton Matching

After the event selection, only about 20% of $t\bar{t}$ events satisfy the following jet-parton matching condition and are called **matched**.

$$\Delta R(\text{jet}, \text{parton}) = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$$

Only the matched $t\bar{t}$ events are used to train the DL model.

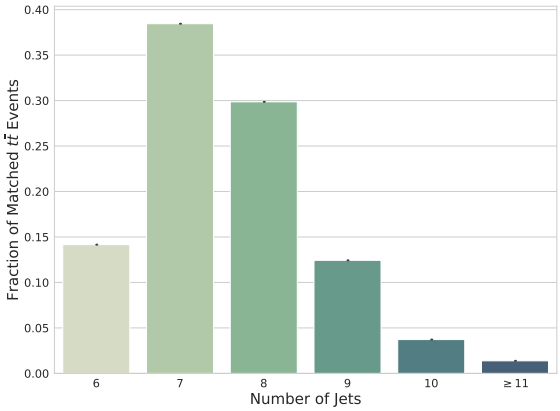


Figure 2: The distribution of the number of jets in the matched $t\bar{t}$ events

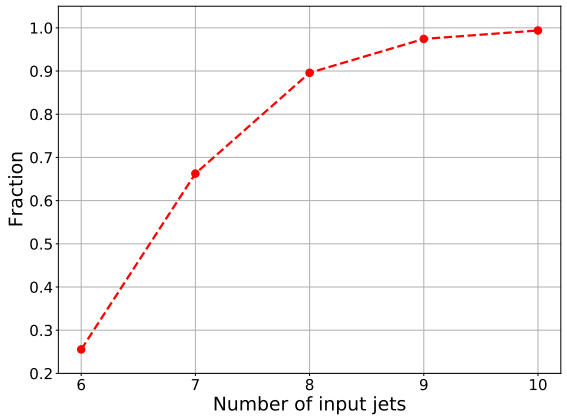


Figure 3: The fraction of matched $t\bar{t}$ events, where all partons can be matched with the most energetic N jets.

Training Dataset

Event

All jets in the event are used as input to the model.

For simplicity, the jet is high level reconstructed variables.

$$(p_T, \eta, \phi, \frac{p_T}{H_T}, \text{b-tag})$$

Jet Shape

Gluon-initiated jets should always be assigned to 'other'. So the effect of the following jet shape variables is studied. [CMS, arXiv:1409.3072]

- $p_T D = \frac{\sum_i p_{T,i}^2}{\sum_i p_{T,i}}$
- Major and minor axes of the jet axis in $\eta - \phi$ space
- Charged hadron, neutral hadron, electrons, muon, and photon multiplicities

Pre-processing

All features except b-tag is scaled to [0, 1] through min-max scaling $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$

Predictive Entropy

NB

In this slides, the uncertainty is ML-side terminology.

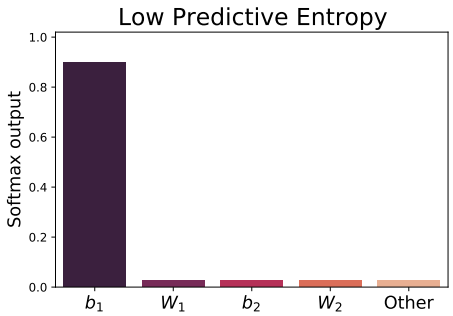
- To suppress poor jet-parton assignments, the DL model uncertainty is studied.

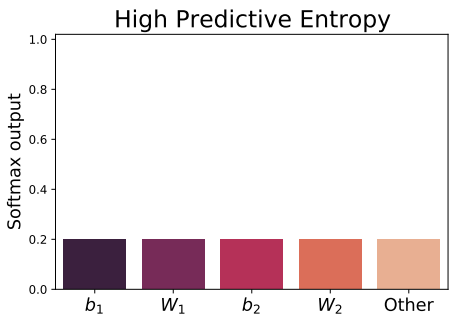
- The **predictive entropy** quantifies the uncertainty in the prediction of the classifier.

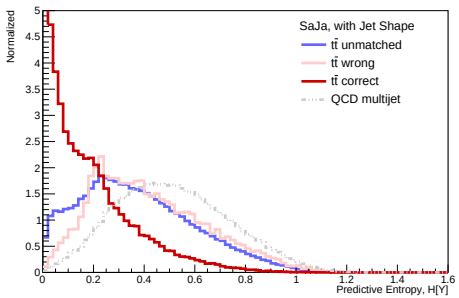
$$H[\hat{Y}] = \frac{1}{N}\sum_{j=1}^{N}[-\sum_{c\in \mathrm{partons}} \hat{y}_c^{(j)} \log \hat{y}_c^{(j)}]$$

- When the jet-parton assignment with the predictive entropy higher than the threshold, the event is not selected.

+ The uncertainty is also used to detect out-of-distribution test data (corresponding to QCD in this study).







Benchmark: Kinematic Likelihood Fitting

• Kinematic likelihood fitting is performed as a baseline study, the KLFitter library is used. [J. Erdmann, Nucl.Instrum.Meth.A 748 (2014) 18-25]

$\mathcal{L} = B(m_{q_1q_2q_3}|m_t, \Gamma_t) \cdot B(m_{q_1q_2}|m_W, \Gamma_W) \cdot B(m_{q_4q_5q_6}|m_t, \Gamma_t) \cdot B(m_{q_4q_5}|m_W, \Gamma_W) \cdot \prod_{i=1}^{6} W_{\text{jet}}(E_{\text{jet},i}^{\text{meas}}|E_{\text{jet},i})$,
where B indicates the Breit-Wigner distribution and W does the the transfer functions

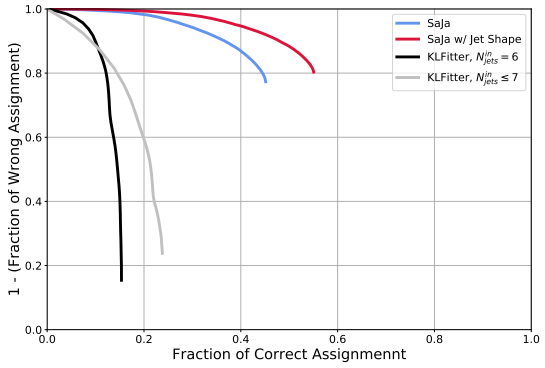
• As more jets are used, the number of permutations increases explosively. Therefore, two case are studied.

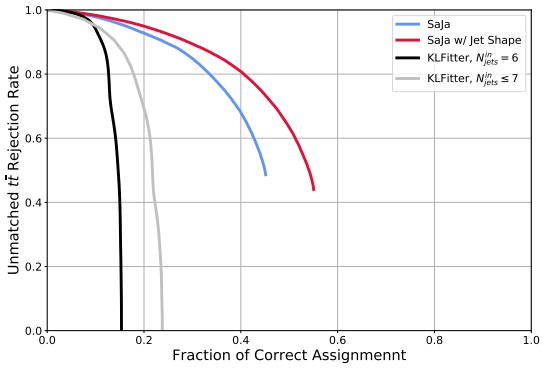
- Most energetic 6 jets, $N_{jets}^{in} = 6 \rightarrow$ Avg. 18 permutations
- Up to 7 most energetic jets $N_{jets}^{in} \leq 7 \rightarrow$ Avg. 126 permutations

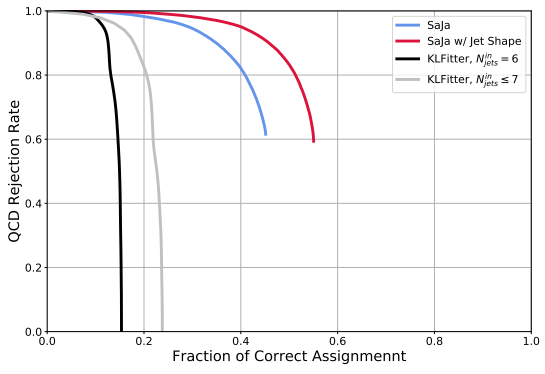
• When the best permutation has a likelihood lower than the threshold, the event is not selected.

Performance

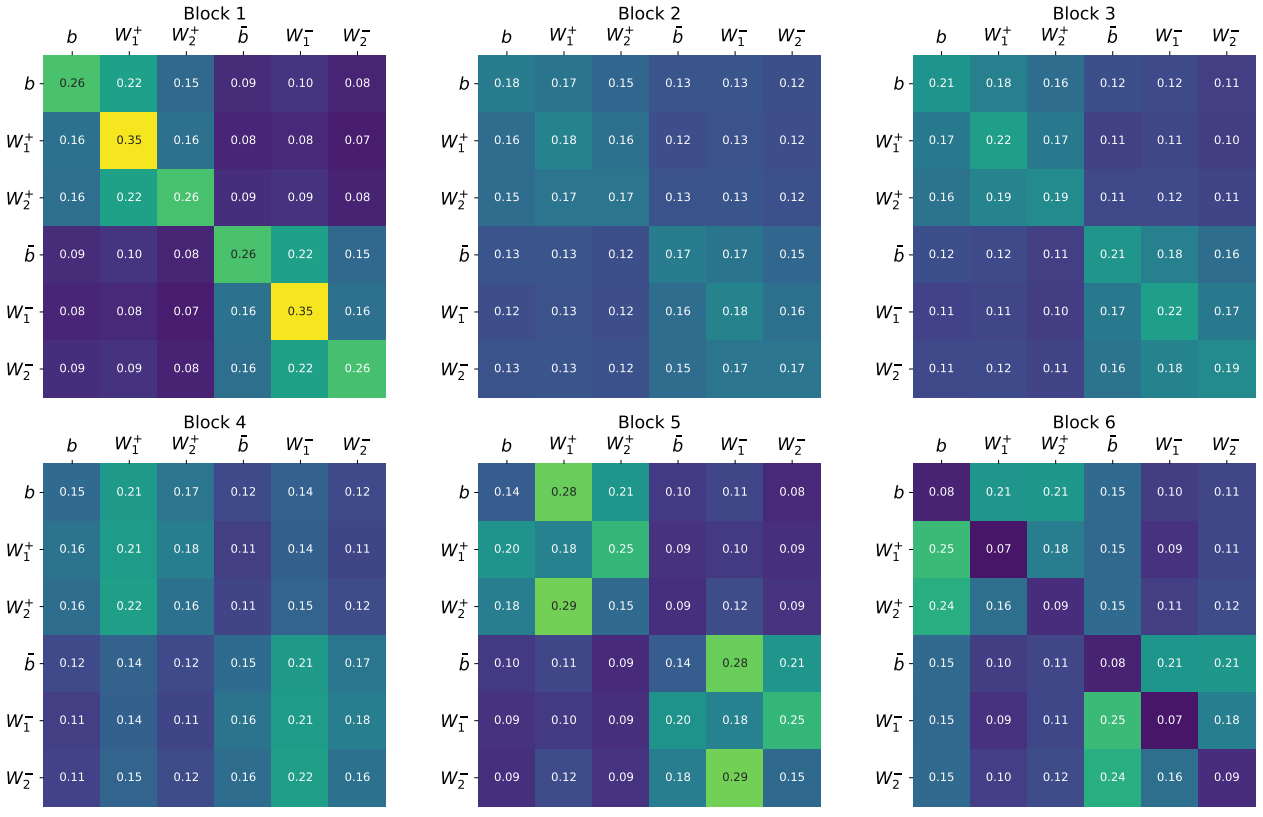
- The assignment performance is visualized as ROC-like curves drawn by varying the threshold value for the predictive entropy of SAJA or the likelihood of KLFITTER.
- SAJA shows more powerful performance than KLFITTER.
- Predictive entropy not only reduces poor jet-parton assignments but also helps reduce unmatched $t\bar{t}$ and QCD multijet events without additional training process.
- Jet shape increases the fraction of correct assignment.





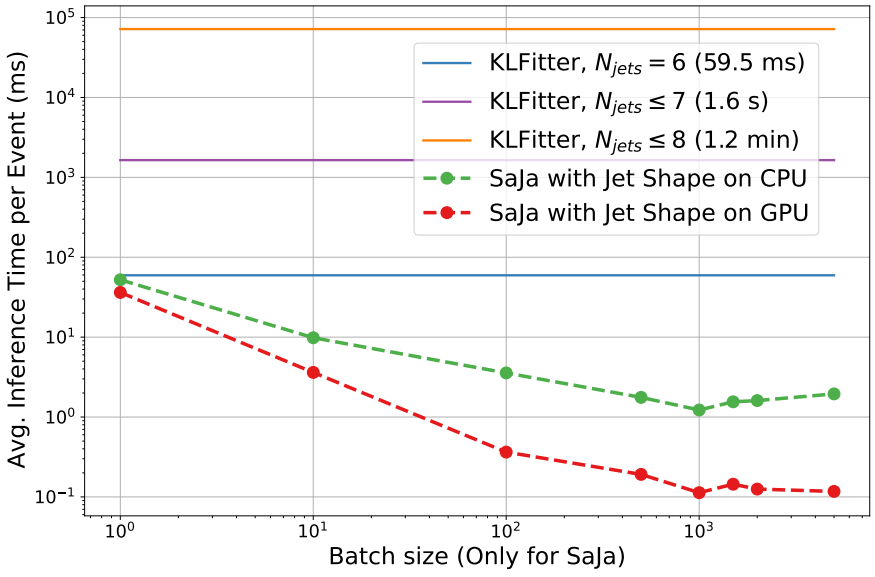


Model Interpretability



- The average of attention matrices for the correct assignments in the matched $t\bar{t}$ events.
- Since SAJA is invariant under the permutation of the jets, it is possible to sort the jets using the jet-parton matching information w.l.o.g., for display purposes.
- W_1^+ indicates higher p_T one of the two quarks, the decay products of W^+ boson.

Inference Latency

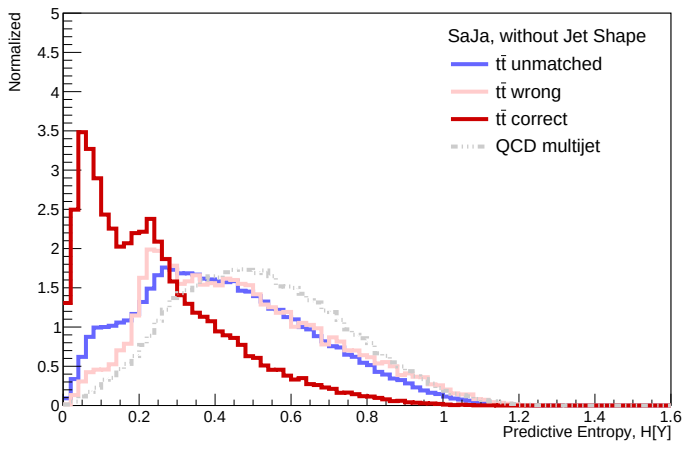


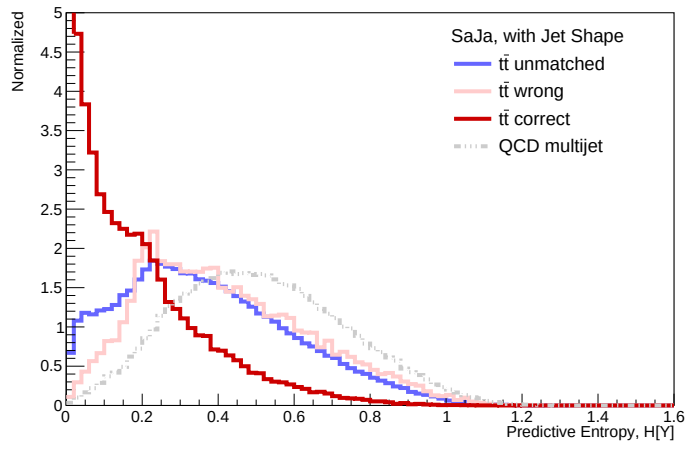
- The graph shows that even for a batch size of one, the zero-permutation SAJA network is outperforming KLFITTER, while a two-order-of-magnitude speed up in inference time is possible by fitting in large batches.

Summary

- We introduce SAJA network for jet-parton assignment without jet permutations.
- SAJA shows better jet-parton assignment and background rejection performance compared to KLFITTER.
- Predictive entropy makes it possible to improve performance without additional training.
- SAJA is fast even without GPU acceleration.

Predictive entropy distribution





Effect of Jet Shape

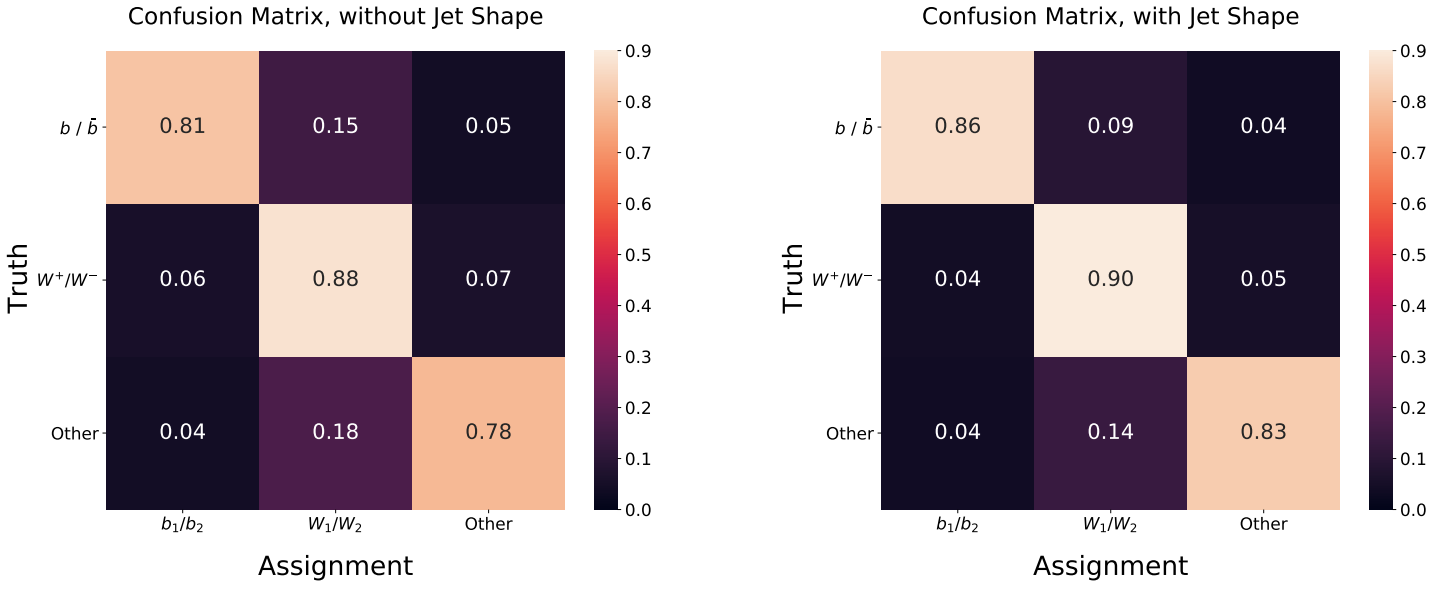
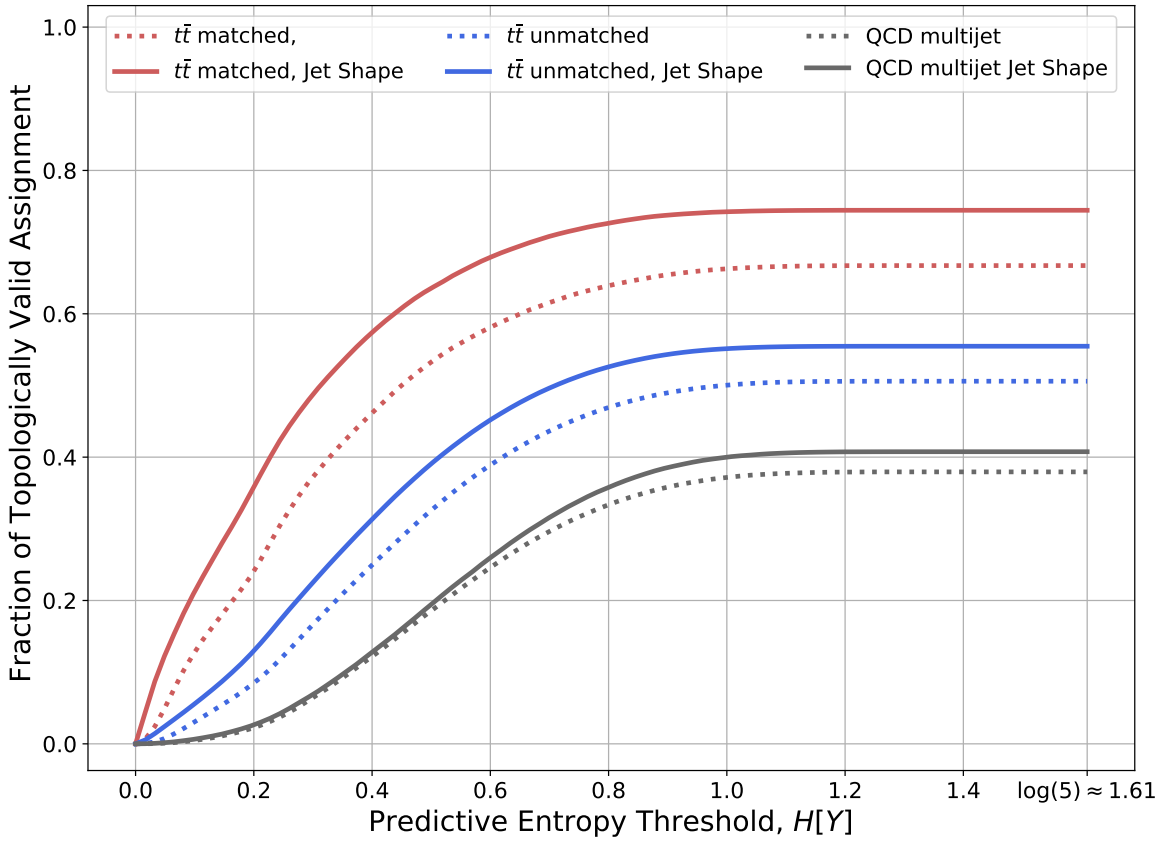


Figure 4: Confusion matrices at the jet-level with an without jet shape.

Jet shape reduces the case where b-initiated jets and additional jets are assigned to W boson.

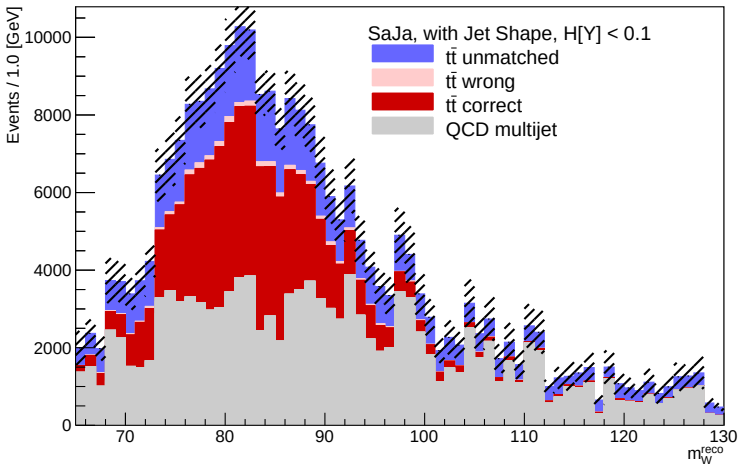
Fraction of Topologically Valid Assignments

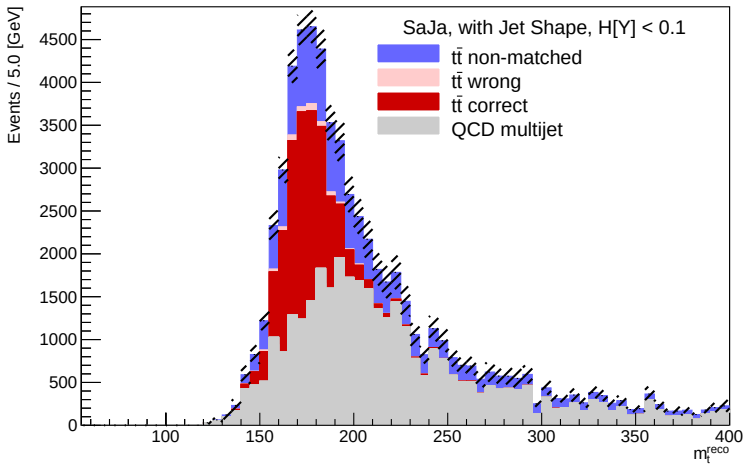


Training in detail

- Training, validation, test set: 311k, 80k, 103k events
- Adam Optimization algorithm
- Learning rate schedule, which reduces the learning rate when the metric on the validation set has stopped improving.
- PyTorch v1.3

Reconstructed W Boson and Top Quark Mass Distributions





(Left) Reconstructed W boson mass distribution and (Right) reconstructed top quark mass distribution.

The Effect of Monte Carlo Simulation