The Rise of the Tagging Machines

Tilman Plehn

Universität Heidelberg

Debrecen 8/2017

Generation One to Three

From deterministic taggers to deep networks

1994 QCD-algorithm W-tagger for heavy Higgs [Seymour]

1994 QCD-algorithm top tagger for fun [Seymour]

2008 QCD-algorithm BDRS Higgs tagger [Butterworth, Davison, Rubin, Salam]

2008 QCD-algorithm JH/CMS top tagger [Kaplan, Rehermann, Schwartz, Tweedie]

2009 QCD-algorithm HEPTopTagger [TP, Salam, Spannowsky]

. . .

2009 template top tagger [Almeida, Lee, Perez, Sterman, Sung, Virzi]

2011 N-Subjettiness [Thaler, van Tilburg]

2011 Shower Deconstruction [Soper, Spannowsky]

2015 Multi-variate HEPTopTagger [Kasieczka, TP, Schell, Strebler, Salam]

. . .

2014 image recognition W-tagger [Cogan, Kagan, Strass, Schwartzman]

2017 image recognition top tagger [Kasieczka, Plehn, Russell, Schell]

2017 language recognition W-tagger [Louppe, Cho, Becot, Cranmer]

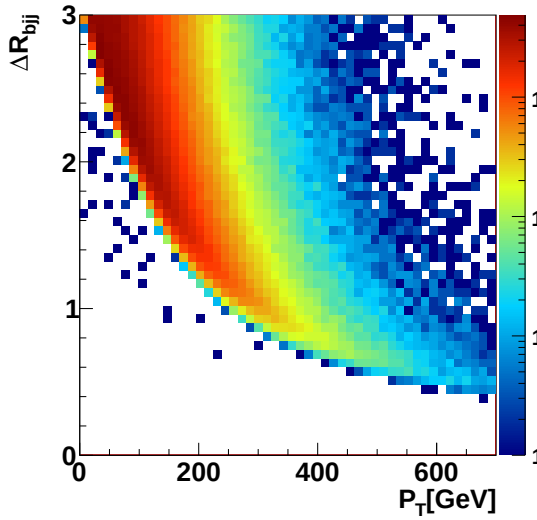
2017 4-vector-based top tagger [Butter, Kasieczka, Plehn, Russel]

Hadronic $t\bar{t}$ resonances

Sub-jet top tagging

- hadronic top identification and reconstruction
- hadronic decays vs QCD splittings
- SM sample: semileptonic $t\bar{t}$ events

$\Rightarrow$ *t*-tagging the new *b*-tagging?

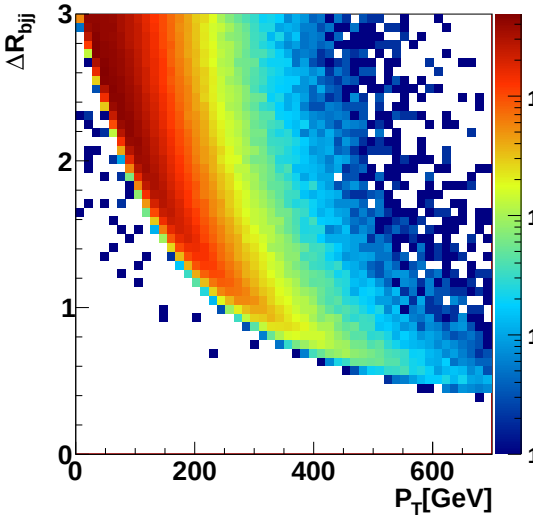


Hadronic $t\bar{t}$ resonances

Sub-jet top tagging

- hadronic top identification and reconstruction
- hadronic decays vs QCD splittings
- SM sample: semileptonic $t\bar{t}$ events

$\Rightarrow$ *t*-tagging the new *b*-tagging?



Mass drop HEPTopTagger [TP, Salam, Spannowsky, Takeuchi]

- algorithm based on QCD-controlled variables

1– C/A fat jet, $R = 1.5$ and $p_T > 200$ GeV [FastJet limitation]

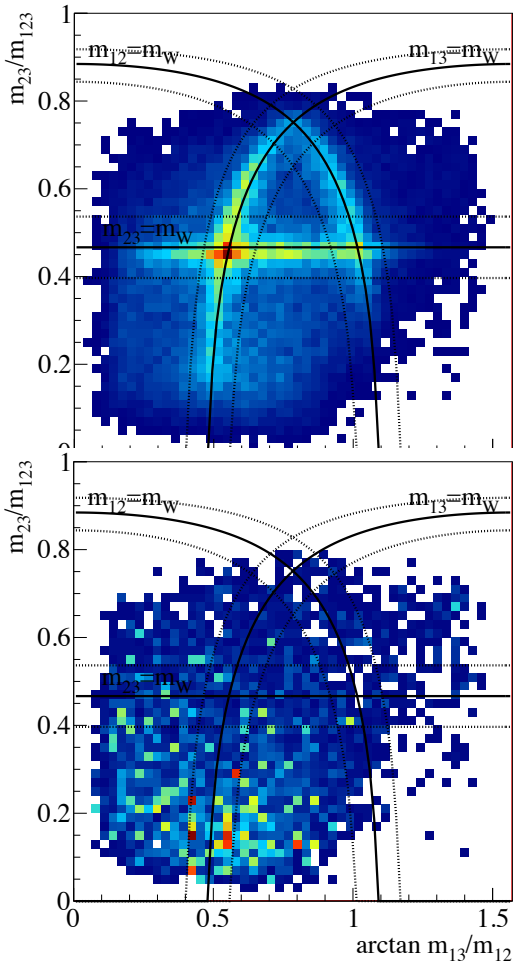
2– mass drop, cutoff $m_{\text{sub}} > 30$ GeV

3– filtering leading to hard substructure triple

4– top mass window $m_{123} = [150, 200]$ GeV

5– A-shaped mass plane cuts as function of m_W/m_t

6– consistency condition $p_T^{(\text{tag})} > 200$ GeV

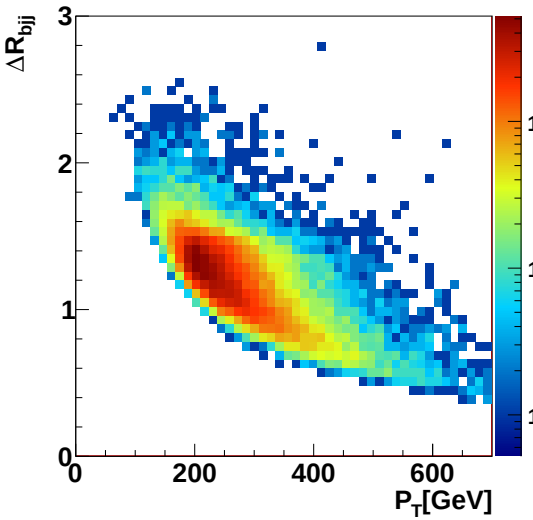


Hadronic $t\bar{t}$ resonances

Sub-jet top tagging

- hadronic top identification and reconstruction
- hadronic decays vs QCD splittings
- SM sample: semileptonic $t\bar{t}$ events

$\Rightarrow$ *t*-tagging the new *b*-tagging?



Mass drop HEPTopTagger [TP, Salam, Spannowsky, Takeuchi]

- algorithm based on QCD-controlled variables

1– C/A fat jet, $R = 1.5$ and $p_T > 200$ GeV [FastJet limitation]

2– mass drop, cutoff $m_{\text{sub}} > 30$ GeV

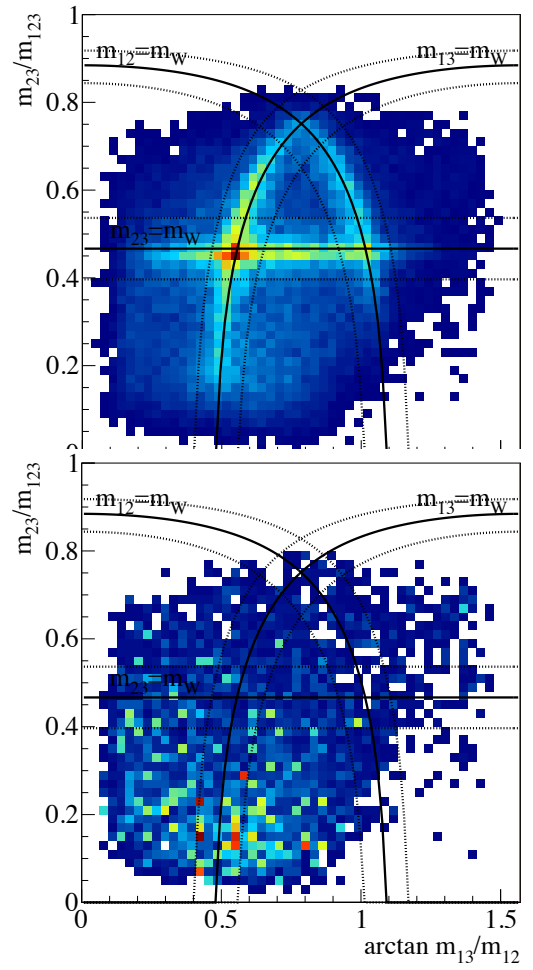
3– filtering leading to hard substructure triple

4– top mass window $m_{123} = [150, 200]$ GeV

5– A-shaped mass plane cuts as function of m_W/m_t

6– consistency condition $p_T^{(\text{tag})} > 200$ GeV

$\Rightarrow$ G1: experimental break-through



Multi-variate top taggers

OptimalR and N-Subjettiness [Kasieczka, TP, Salam, Schell, Strebler]

- multivariate analysis old idea [Lonnblad, Peterson, Rognvaldsson]
  HEPTopTaggerv2 to keep up with shower deconstruction [Soper, Spannowsky]
- optimal fat jet size R_{opt} [large to decay jets, small to avoid combinatorics, compute from kinematics]

  $$|m_{123} - m_{123}^{(R_{\text{max}})}| < 0.2\, m_{123}^{(R_{\text{max}})} \quad \Rightarrow \quad R_{\text{opt}}$$
- add subjet counting [Thaler, van Tilburg]
- $\{m_{123}, f_W, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}, \tau_j, \tau_j^{(\text{filt})}\}$

Multi-variate top taggers

OptimalR and N-Subjettiness [Kasieczka, TP, Salam, Schell, Strebler]

- multivariate analysis old idea [Lonnblad, Peterson, Rognvaldsson]
  HEPTopTaggerv2 to keep up with shower deconstruction [Soper, Spannowsky]
- optimal fat jet size R_{opt} [large to decay jets, small to avoid combinatorics, compute from kinematics]

  $$|m_{123} - m_{123}^{(R_{\text{max}})}| < 0.2\, m_{123}^{(R_{\text{max}})} \quad \Rightarrow \quad R_{\text{opt}}$$
- add subjet counting [Thaler, van Tilburg]
- $\{m_{123}, f_W, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}, \tau_j, \tau_j^{(\text{filt})}\}$

Qjets [Ellis, Hornig, Roy, Krohn, Schwartz]

- more than one clustering history, weighted by

  $$\omega_{ij} = \exp\left[-\alpha \frac{y_{ij} - y_{ij}^{\text{min}}}{y_{ij}^{\text{min}}}\right]$$

  then using distributions like $\langle m^2 \rangle - \langle m \rangle^2$
- $\{\ldots, \{m_{123}^{\text{Qjets}}\}\}$

Multi-variate top taggers

OptimalR and N-Subjettiness [Kasieczka, TP, Salam, Schell, Strebler]

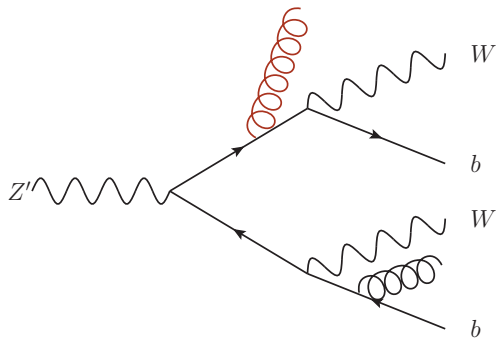
- multivariate analysis old idea [Lonnblad, Peterson, Rognvaldsson]
  HEPTopTaggerv2 to keep up with shower deconstruction [Soper, Spannowsky]
- optimal fat jet size R_{opt} [large to decay jets, small to avoid combinatorics, compute from kinematics]

  $$|m_{123} - m_{123}^{(R_{\text{max}})}| < 0.2\, m_{123}^{(R_{\text{max}})} \quad \Rightarrow \quad R_{\text{opt}}$$
- add subjet counting [Thaler, van Tilburg]
- $\{m_{123}, f_W, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}, \tau_j, \tau_j^{(\text{filt})}\}$

Fat jet and top kinematics

- FSR major problem for Z' search

$\Rightarrow \{\ldots, m_{tt}, p_{T,t}, m_{jj}^{(\text{filt})}, p_{T,j}^{(\text{filt})}\}$



Multi-variate top taggers

OptimalR and N-Subjettiness [Kasieczka, TP, Salam, Schell, Strebler]

- multivariate analysis old idea [Lonnblad, Peterson, Rognvaldsson]
  HEPTopTaggerv2 to keep up with shower deconstruction [Soper, Spannowsky]
- optimal fat jet size R_{opt} [large to decay jets, small to avoid combinatorics, compute from kinematics]

$$|m_{123} - m_{123}^{(R_{\text{max}})}| < 0.2\, m_{123}^{(R_{\text{max}})} \quad \Rightarrow \quad R_{\text{opt}}$$

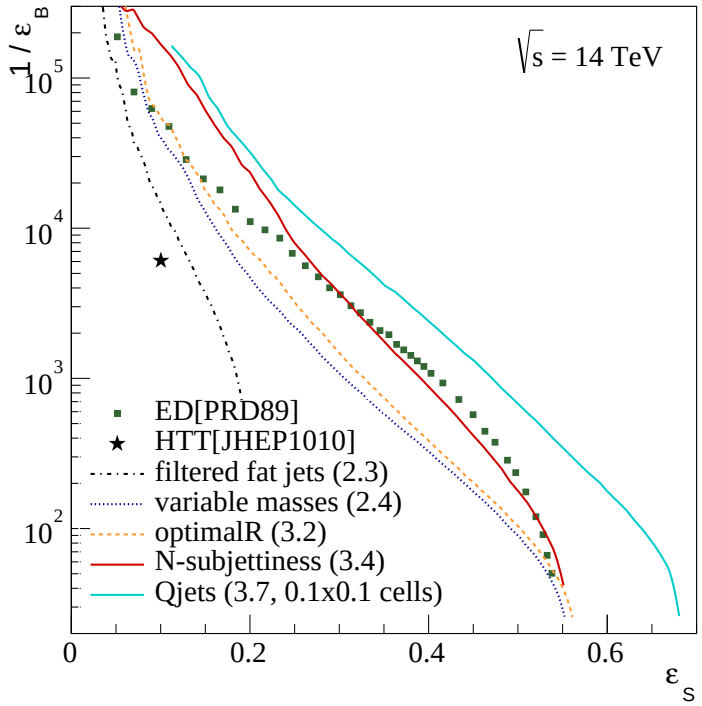
- add subjet counting [Thaler, van Tilburg]
- $\{m_{123}, f_W, R_{\text{opt}} - R_{\text{opt}}^{(\text{calc})}, \tau_j, \tau_j^{(\text{filt})}\}$

Fat jet and top kinematics

- FSR major problem for Z' search

$\Rightarrow$ $\{\ldots, m_{tt}, p_{T,t}, m_{jj}^{(\text{filt})}, p_{T,j}^{(\text{filt})}\}$

$\Rightarrow$ G2: deterministic taggers terminated!



Jet images

Image regognition for jets

- wavelet transformation [Rentala, Shepherd, Tait; Monk]
- W-tagging with image recognition [Cogan etal, Oliveira etal, Baldi etal]
- top-tagging attempt [Almeida, Backovic, Cliche, Lee, Perelstein]
- QCD and shower study [Barnard etal]
- quark-gluon discrimination including tracks [Komiske etal]

$\Rightarrow$ G3: new avenue in jet physics

Jet images

Image regognition for jets

- wavelet transformation [Rentala, Shepherd, Tait; Monk]
- W-tagging with image recognition [Cogan etal, Oliveira etal, Baldi etal]
- top-tagging attempt [Almeida, Backovic, Cliche, Lee, Perelstein]
- QCD and shower study [Barnard etal]
- quark-gluon discrimination including tracks [Komiske etal]

$\Rightarrow$ G3: new avenue in jet physics

Experimental questions

- does it work?
- what is the training sample? [Metodiev, Nachman, Thaler]
- how do we get it past the jets people?

Jet images

Image regognition for jets

- wavelet transformation [Rentala, Shepherd, Tait; Monk]
- W-tagging with image recognition [Cogan etal, Oliveira etal, Baldi etal]
- top-tagging attempt [Almeida, Backovic, Cliche, Lee, Perelstein]
- QCD and shower study [Barnard etal]
- quark-gluon discrimination including tracks [Komiske etal]

$\Rightarrow$ G3: new avenue in jet physics

Experimental questions

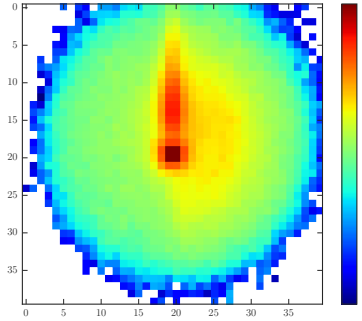
- does it work?
- what is the training sample? [Metodiev, Nachman, Thaler]
- how do we get it past the jets people?

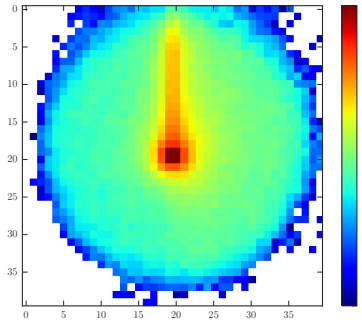
Theoretical questions

- what does the neutal network learn?
- how much of it is hard QCD?
- how can be improve the setup? [the future has not been written]

$\Rightarrow$ Established benchmark crucial

Jet images





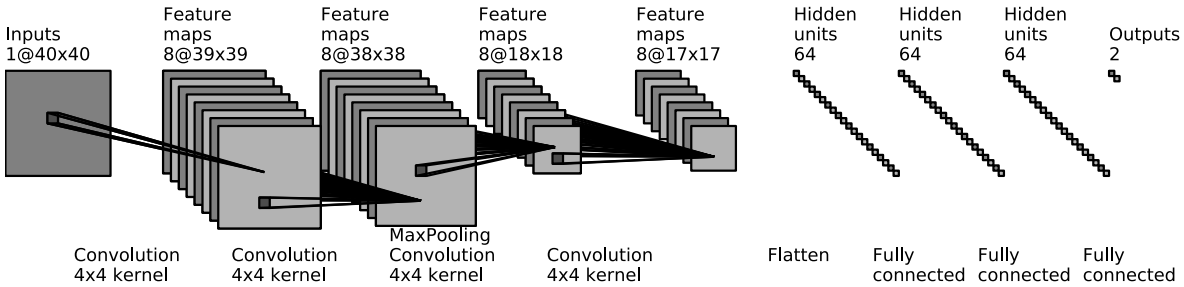
Start with anti-k_T fat jet [$p_T = 350 \ldots 450$ GeV, $R = 1.5$]

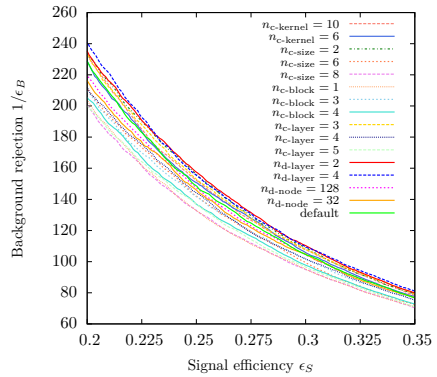
- shift move image to center the global maximum
- rotation rotate the second maximum to 12 o'clock
- flip ensure third maximum is in the right half-plane
- crop crop the image to 40×40 pixels
- decide on E vs E_T for rapidities $\eta \gtrsim 2$

$\Rightarrow$ pre-processing only for illustration

Set up network [Kasieczka, TP, Russell, Schell]

- run on 2-D jet images [$p_T = 350, \ldots, 450$ GeV]
- binning through calorimeter resolution [$\Delta\eta = 0.1$ vs $\Delta\phi = 5^\circ$]
- 150k events for training
- analyze geometric patterns [convolutional network]

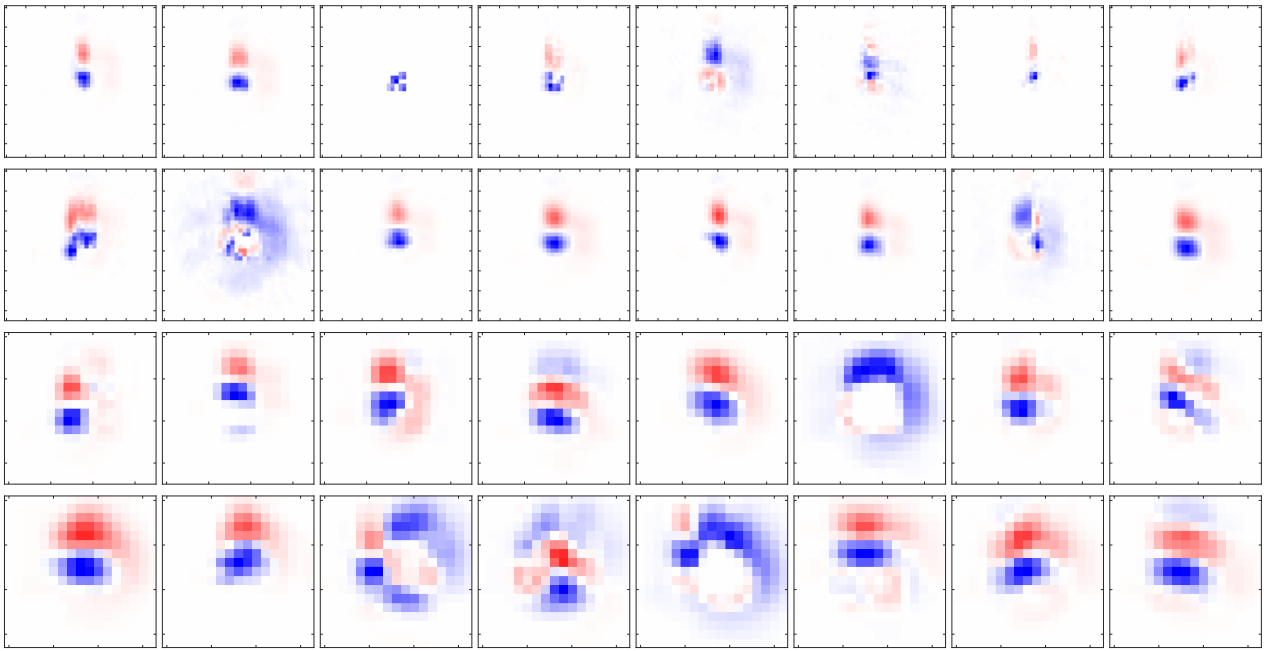




DeepTop tagger

Benchmarking image-based top tagger [Kasieczka, TP, Russell, Schell]

– 4 convolutional layers probing 2D structure with kernel matrix

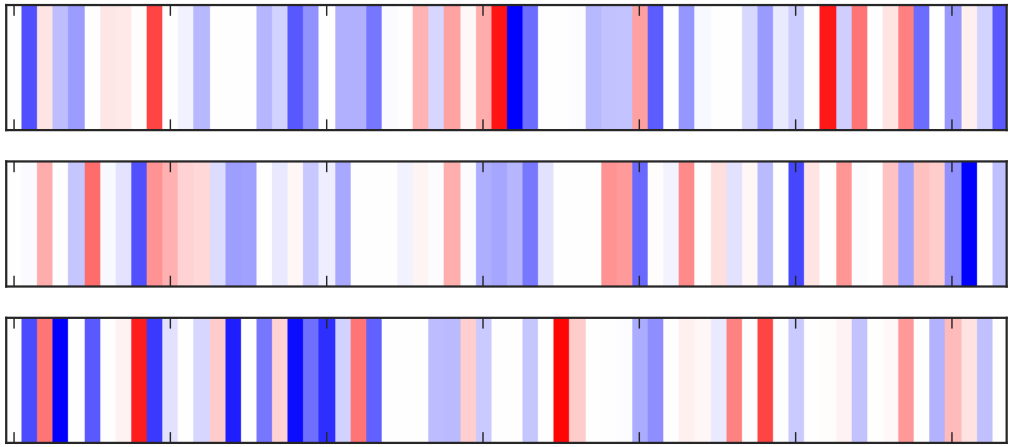


DeepTop tagger

Benchmarking image-based top tagger [Kasieczka, TP, Russell, Schell]

- 4 convolutional layers probing 2D structure with kernel matrix
- 3 fully connected layers weight function linking input and output

$$y_i = \max\left(0, \sum_{j=1}^{n^2} W_{ij} x_j + b_i\right)$$



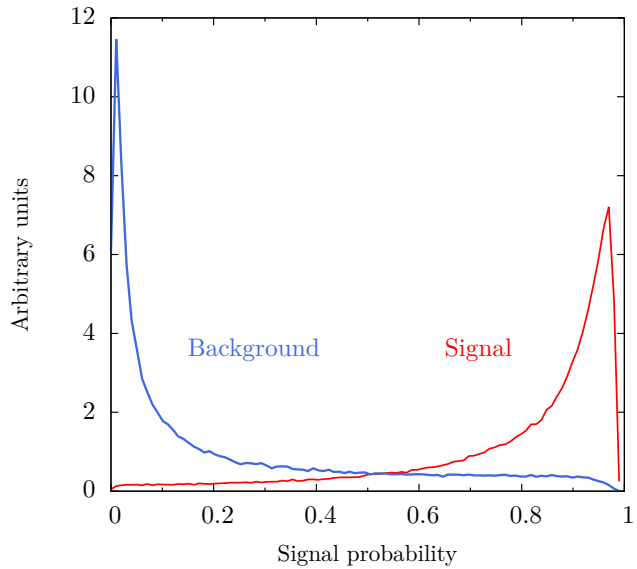
DeepTop tagger

Benchmarking image-based top tagger [Kasieczka, TP, Russell, Schell]

- 4 convolutional layers probing 2D structure with kernel matrix
- 3 fully connected layers weight function linking input and output

$$y_i = \max\left(0, \sum_{j=1}^{n^2} W_{ij} x_j + b_i\right)$$

- signal-ness vs background-ness output [probability?]



DeepTop tagger

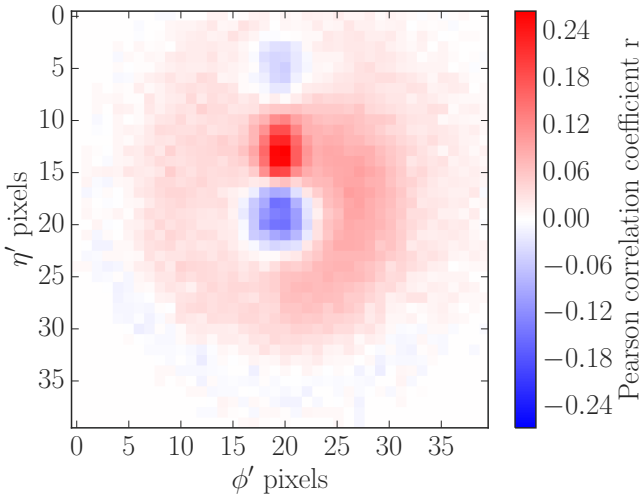
Benchmarking image-based top tagger [Kasieczka, TP, Russell, Schell]

- 4 convolutional layers probing 2D structure with kernel matrix
- 3 fully connected layers weight function linking input and output

$$y_i = \max\left(0, \sum_{j=1}^{n^2} W_{ij} x_j + b_i\right)$$

- signal-ness vs background-ness output [probability?]
- Pearson input-output correlation [pixel x vs label y]

$$r_{ij} \approx \sum_{\text{images}} \left(x_{ij} - \bar{x}_{ij}\right)\left(y - \bar{y}\right)$$



DeepTop tagger

Benchmarking image-based top tagger [Kasieczka, TP, Russell, Schell]

- 4 convolutional layers probing 2D structure with kernel matrix
- 3 fully connected layers weight function linking input and output

$$y_i = \max\left(0, \sum_{j=1}^{n^2} W_{ij} x_j + b_i\right)$$

- signal-ness vs background-ness output [probability?]
- Pearson input-output correlation [pixel x vs label y]

$$r_{ij} \approx \sum_{\text{images}} \left(x_{ij} - \bar{x}_{ij}\right)\left(y - \bar{y}\right)$$

- comparison to G2 MotherOfTaggers

$$\{m_{\text{sd}}, m_{\text{fat}}, m_{\text{rec}}, f_{\text{rec}}, \Delta R_{\text{opt}}, \tau_2, \tau_3, \tau_2^{\text{sd}}, \tau_3^{\text{sd}}\}$$

DeepTop tagger

Benchmarking image-based top tagger [Kasieczka, TP, Russell, Schell]

- 4 convolutional layers probing 2D structure with kernel matrix
- 3 fully connected layers weight function linking input and output

$$y_i = \max\left(0, \sum_{j=1}^{n^2} W_{ij} x_j + b_i\right)$$

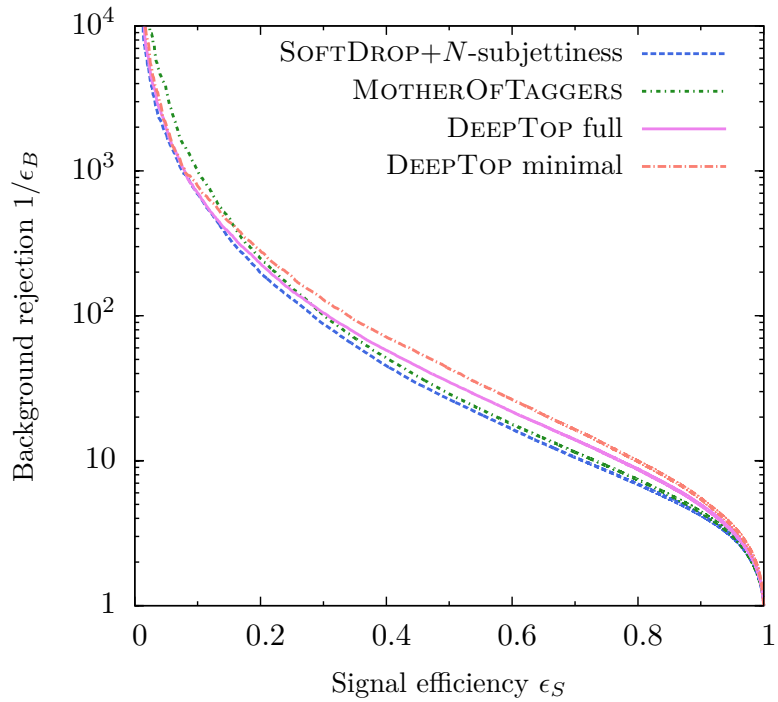
- signal-ness vs background-ness output [probability?]
- Pearson input-output correlation [pixel x vs label y]

$$r_{ij} \approx \sum_{\text{images}} \left(x_{ij} - \bar{x}_{ij}\right)\left(y - \bar{y}\right)$$

- comparison to G2 MotherOfTaggers

$$\{m_{\text{sd}}, m_{\text{fat}}, m_{\text{rec}}, f_{\text{rec}}, \Delta R_{\text{opt}}, \tau_2, \tau_3, \tau_2^{\text{sd}}, \tau_3^{\text{sd}}\}$$

$\Rightarrow$ slight performance gain for CNN

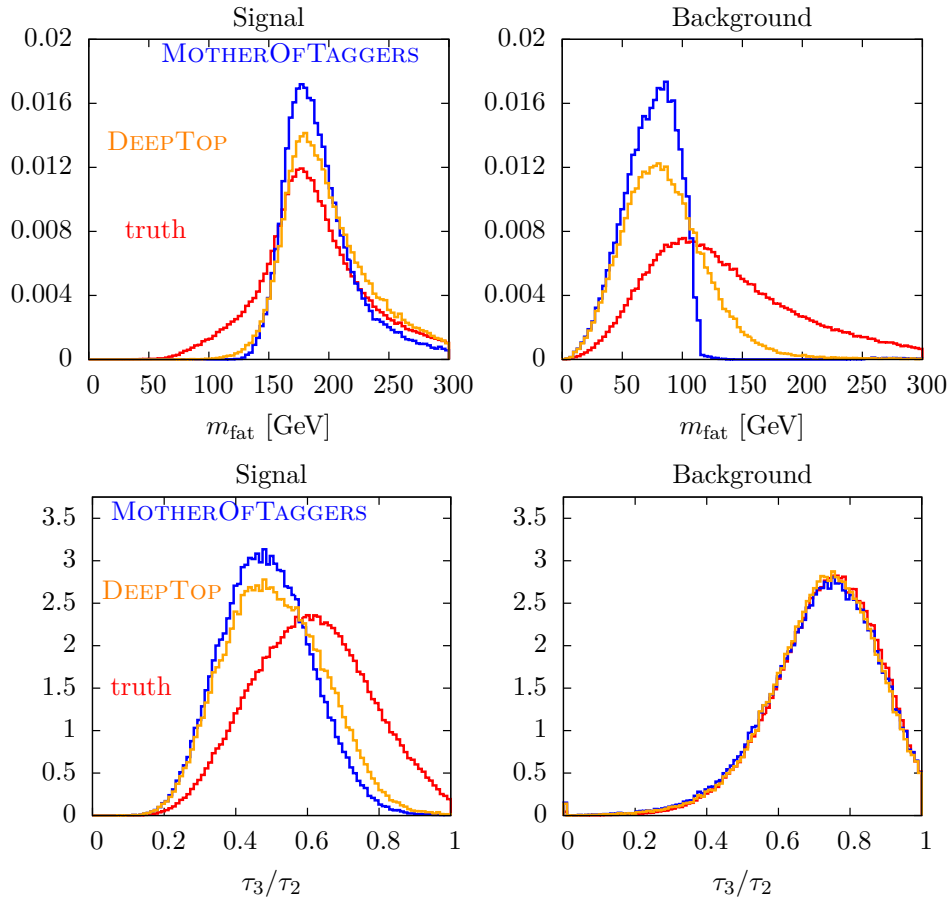


Checking physics

Typical reaction: 'fuck you, you fucking machine'

- in principle, full control for fully supervised learning
- lots of events in the grey zone
  but checks possible for correctly identified signal/background events
- compare truth vs MotherOfTaggers vs DeepTop

1- fat jet mass and N-subjettiness



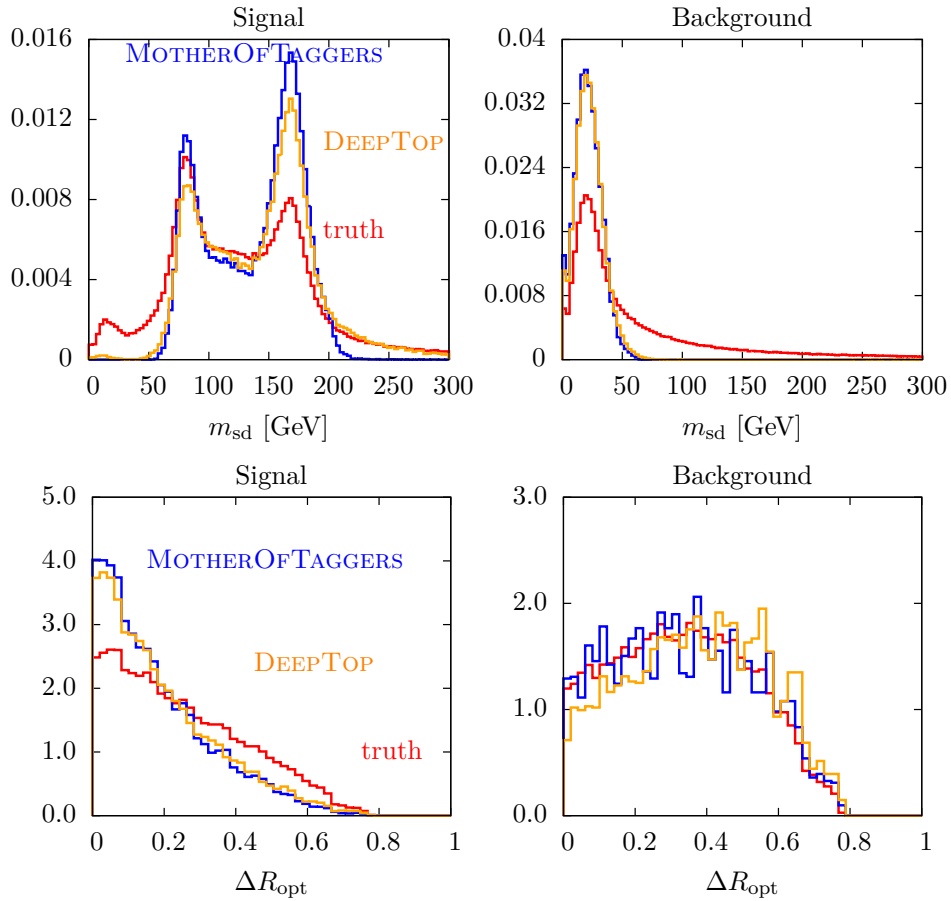
Checking physics

Typical reaction: 'fuck you, you fucking machine'

- in principle, full control for fully supervised learning
- lots of events in the grey zone
  but checks possible for correctly identified signal/background events
- compare truth vs MotherOfTaggers vs DeepTop

1- fat jet mass and N-subjettiness

2- soft drop mass and ΔR_{opt}



Checking physics

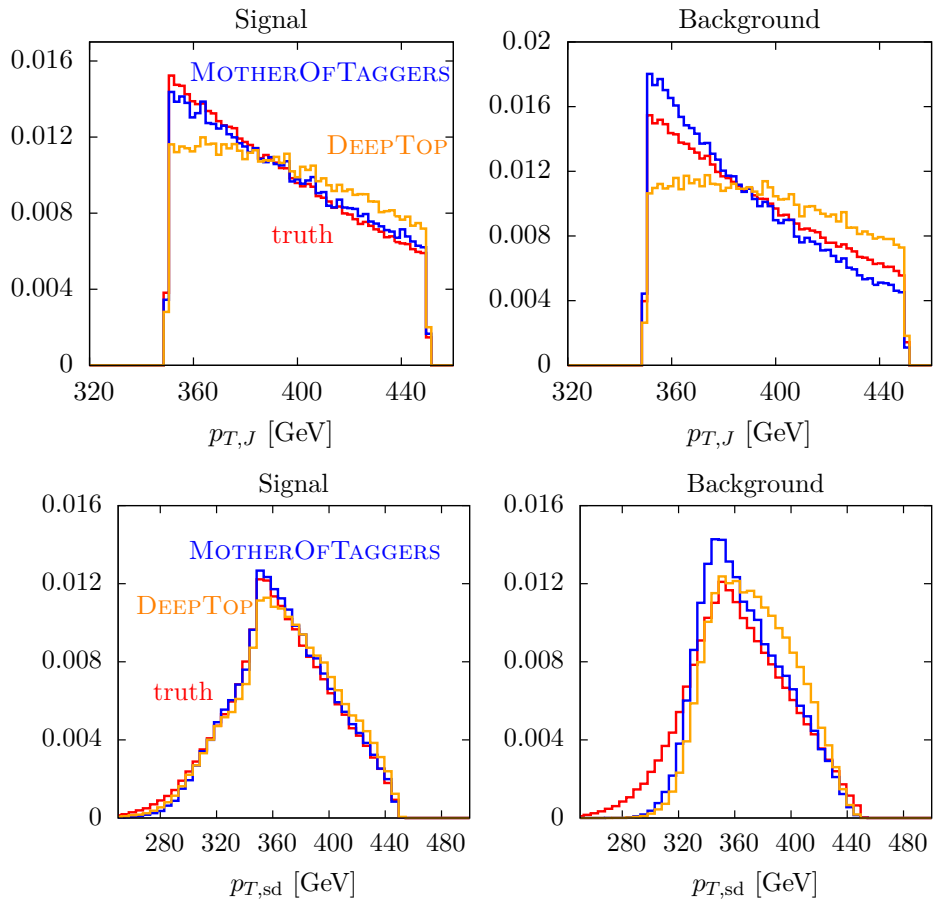
Typical reaction: 'fuck you, you fucking machine'

- in principle, full control for fully supervised learning
- lots of events in the grey zone
  but checks possible for correctly identified signal/background events
- compare truth vs MotherOfTaggers vs DeepTop

1- fat jet mass and N-subjettiness

2- soft drop mass and ΔR_{opt}

3- transverse momenta



Checking physics

Typical reaction: 'fuck you, you fucking machine'

- in principle, full control for fully supervised learning
- lots of events in the grey zone
  but checks possible for correctly identified signal/background events
- compare truth vs MotherOfTaggers vs DeepTop

1- fat jet mass and N-subjettiness

2- soft drop mass and ΔR_{opt}

3- transverse momenta

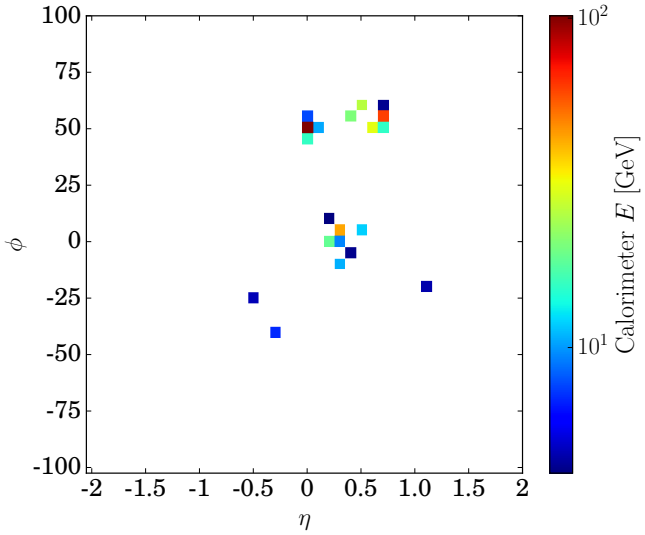
4- what else do you want checked?

$\Rightarrow$ Machine learning works and we know why

DeepTop using Lorentz Layer

Back to 4-vectors? [Butter, Kasieczka, TP, Russell; see also Louppe etal, Pearkes etal]

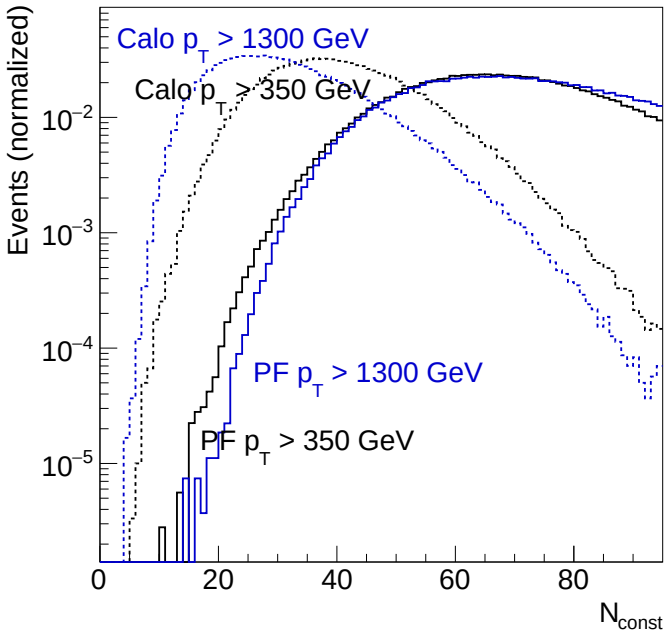
1. appropriate physics objects known (?)
2. tracker/particle flow precision?
3. why rely on image/language tools?
4. link to G1 and G2 taggers? [Larkoski etal]



DeepTop using Lorentz Layer

Back to 4-vectors? [Butter, Kasieczka, TP, Russell; see also Louppe etal, Pearkes etal]

1. appropriate physics objects known (?)
2. tracker/particle flow precision?
3. why rely on image/language tools?
4. link to G1 and G2 taggers? [Larkoski etal]



DeepTop using Lorentz Layer

Back to 4-vectors? [Butter, Kasieczka, TP, Russell; see also Louppe etal, Pearkes etal]

1. appropriate physics objects known (?)
2. tracker/particle flow precision?
3. why rely on image/language tools?
4. link to G1 and G2 taggers? [Larkoski etal]

Avoiding brute force — combination layer

– input 4-vectors

$$(k_{\mu,i}) = \begin{pmatrix} k_{0,1} & k_{0,2} & \cdots & k_{0,N} \\ k_{1,1} & k_{1,2} & \cdots & k_{1,N} \\ k_{2,1} & k_{2,2} & \cdots & k_{2,N} \\ k_{3,1} & k_{3,2} & \cdots & k_{3,N} \end{pmatrix}$$

DeepTop using Lorentz Layer

Back to 4-vectors? [Butter, Kasieczka, TP, Russell; see also Louppe etal, Pearkes etal]

1. appropriate physics objects known (?)
2. tracker/particle flow precision?
3. why rely on image/language tools?
4. link to G1 and G2 taggers? [Larkoski etal]

Avoiding brute force — combination layer

- input 4-vectors
- crucial on-shell conditions

$$\tilde{k}^2_{\mu,1} = (k_{\mu,1} + k_{\mu,2} + k_{\mu,3})^2 \stackrel{!}{=} m_t^2$$

$$\tilde{k}^2_{\mu,2} = (k_{\mu,1} + k_{\mu,2})^2 \stackrel{!}{=} m_W^2$$

DeepTop using Lorentz Layer

Back to 4-vectors? [Butter, Kasieczka, TP, Russell; see also Louppe etal, Pearkes etal]

1. appropriate physics objects known (?)
2. tracker/particle flow precision?
3. why rely on image/language tools?
4. link to G1 and G2 taggers? [Larkoski etal]

Avoiding brute force — combination layer

- input 4-vectors
- crucial on-shell conditions
- combined 4-vectors

$$k_{\mu,i} \xrightarrow{\text{CoLa}} \widetilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij} \qquad C = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 & C_{1,N+2} & \cdots & C_{1,M} \\ 1 & 0 & 1 & & \vdots & C_{2,N+2} & \cdots & C_{2,M} \\ \vdots & \vdots & \vdots & \ddots & 0 & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 & C_{N,N+2} & \cdots & C_{N,M} \end{pmatrix}$$

- after combination of input 4-vectors

  sum of all momenta, fat jet momentum
  original momenta k_i
  $M - (N+1)$ trainable linear combinations

DeepTop using Lorentz Layer

Avoiding brute force — combination layer

- input 4-vectors $k_{\mu,i}$
- combined 4-vectors $\widetilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij}$

Remember e&m — Lorentz layer

- DNN on Lorentz scalars

$$\tilde{k}_j \xrightarrow{\text{LoLa}} \hat{k}_j = \begin{pmatrix} m^2(\tilde{k}_j) \\ p_T(\tilde{k}_j) \\ w_{jm}^{(E)}\, E(\tilde{k}_m) \\ w_{jm}^{(d)}\, d_{jm}^2 \end{pmatrix}$$

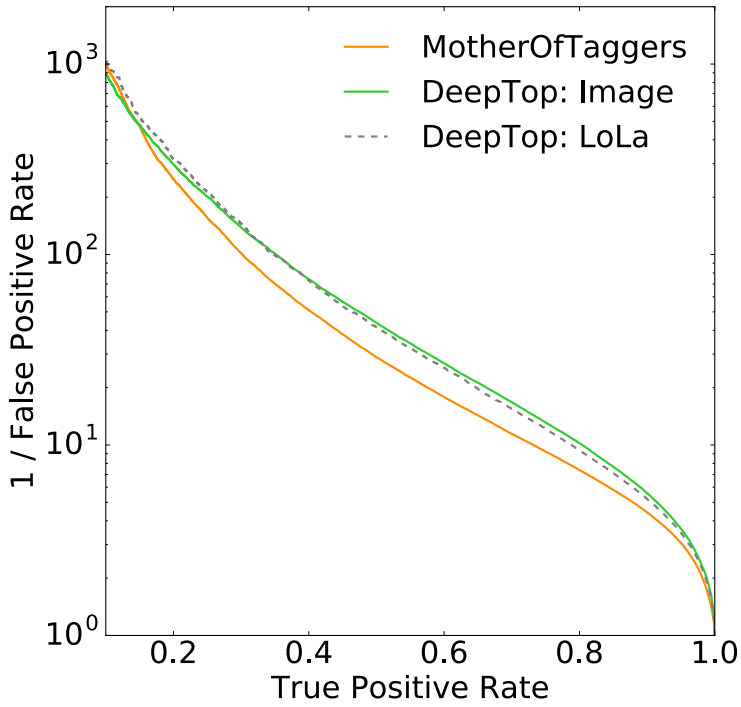
DeepTop using Lorentz Layer

Avoiding brute force — combination layer

- input 4-vectors $k_{\mu,i}$
- combined 4-vectors $\widetilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij}$

Remember e&m — Lorentz layer

- DNN on Lorentz scalars $\hat{k}_j$
- comparison to state of art



DeepTop using Lorentz Layer

Avoiding brute force — combination layer

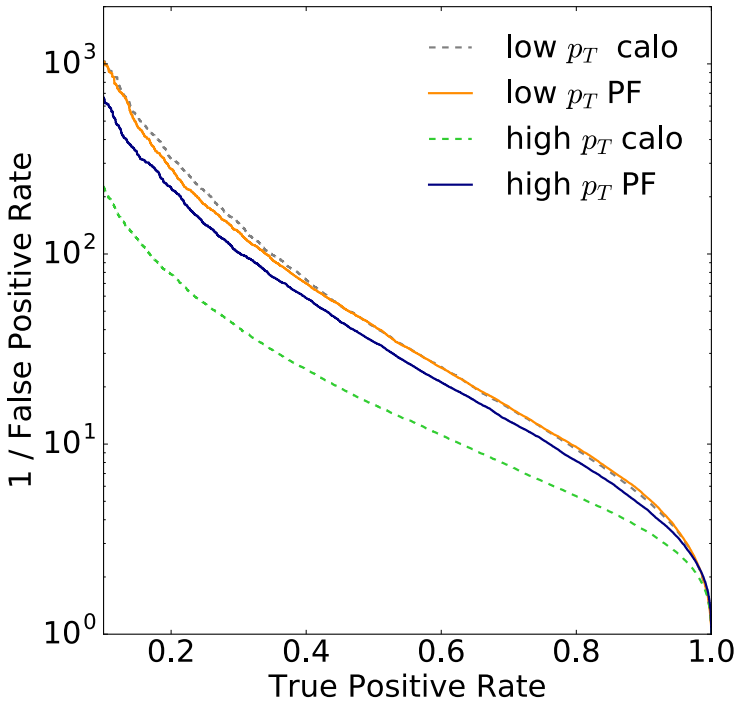
- input 4-vectors $k_{\mu,i}$
- combined 4-vectors $\widetilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij}$

Remember e&m — Lorentz layer

- DNN on Lorentz scalars $\hat{k}_j$
- comparison to state of art
- measuring Minkowski metric

  $g = \text{diag}(0.99 \pm 0.02,$
  $-1.01 \pm 0.01, -1.01 \pm 0.02, -0.99 \pm 0.02)$
- running on particle flow

  [$p_T = 350, \ldots, 450$ GeV and $1300, \ldots, 1400$ GeV]
- 180k training events, 15 GPU minutes
- G3 with nothing but Lorentz invariance



The future

Times are moving fast...

...deterministic taggers estalished/old/boring

...information beyond clustering history helps

...imagine recognition for subjets as starting point

...DeepTop is not a black box

...back to QCD with DeepTopLoLa

ML taggers ready for data!