Progress in Machine Learning

Tilman Plehn

Universität Heidelberg

CMS DESY 4/2019

Story's starting point: Nothing is ever new

LHC visionaries

– 1991: NN-based quark-gluon tagger [visionary: Lönnblad, Peterson, Rögnvaldsson]

USING NEURAL NETWORKS TO IDENTIFY JETS

Leif LÖNNBLAD*, Carsten PETERSON** and Thorsteinn RÖGNVALDSSON***

Department of Theoretical Physics, University of Lund, Sölvegatan 14A, S-22362 Lund, Sweden

Received 29 June 1990

A neural network method for identifying the ancestor of a hadron jet is presented. The idea is to find an efficient mapping between certain observed hadronic kinematical variables and the quark-gluon identity. This is done with a neuronic expansion in terms of a network of sigmoidal functions using a gradient descent procedure, where the errors are back-propagated through the network. With this method we are able to separate gluon from quark jets originating from Monte Carlo generated e^+e^- events with $\sim 85\%$ approach. The result is independent of the MC model used. This approach for isolating the gluon jet is then used to study the so-called string effect.

Story's starting point: Nothing is ever new

LHC visionaries

- 1991: NN-based quark-gluon tagger [visionary: Lönnblad, Peterson, Rögnvaldsson]
- 1994: jet-algorithm W/top-tagger [Seymour]

USING NEURAL NETWORKS TO IDENTIFY JETS

Leif LÖNNBLAD*, Carsten PETERSON** and Thorsteinn RÖGNVALDSSON***

Department of Theoretical Physics, University of Lund, Sölvegatan 14A, S-22362 Lund, Sweden

Received 29 June 1990

A neural network method fo
is to find an efficient mapping be
quark-gluon identity. This is don
functions using a gradient descen
network. With this method we ar
Carlo generated e^+e^- events w
model used. This approach for i
effect.

Searches for new particles using cone and cluster jet algorithms: a comparative study

Michael H. Seymour

Department of Theoretical Physics, University of Lund, Sölvegatan 14A, S-22362 Lund, Sweden

Received 18 June 1993; in revised form 16 September 1993

Abstract. We discuss the reconstruction of the hadronic decays of heavy particles using jet algorithms. The ability to reconstruct the mass of the decaying particle is compared between a traditional cone-type algorithm and a recently proposed cluster-type algorithm. The specific examples considered are the semileptonic decays of a heavy Higgs boson at $\sqrt{s}=16$ TeV, and of top quark–antiquark pairs at $\sqrt{s}=1.8$ TeV. We find that the cluster algorithm offers considerable advantages in the former case, and a slight advantage in the latter. We briefly discuss the effects of calorimeter energy resolution, and show that a typical resolution dilutes these advantages, but does not remove them entirely.

except that the invariant mass of a pair is replaced by the transverse momentum of the softer particle relative to the other.

More recently, this algorithm was extended to collisions with incoming hadrons [5], and a longitudinally-invariant $k_\perp$-clustering algorithm for hadron-hadron collisions was proposed [6]. This algorithm has been compared with the more commonly used cone algorithm from the viewpoints of a parton-shower Monte Carlo program [6, 7], and a fixed-order matrix lation [8], and advantages of the cluster reported in both cases. This paper is a comparison between the algorithms reconstructing the hadronic decays of which was also studied in a preliminary

The only as-yet unobserved particles Standard Model are the top quark and H search for, and study of, these particles most important goals of current and p hadron collider experiments. In both cas

∼ 1970: *People with visions should see a doctor* [Helmut Schmidt, wrong for once]

Fat jet taggers

Look what makes jets [Pre-LHC, jets were just annoying]

– top jets from $t \to bq\bar{q}'$ vs QCD jets

– top decays well-defined in theory

– labelled sample: semileptonic $t\bar{t}$ events

$\Rightarrow$ Fat jets as LHC physics playground

Simple top tagging [BDRS; TP, Salam, Spannowsky, Takeuchi]

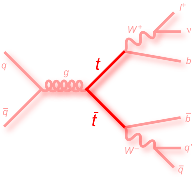
1– fat jet with $p_T > 200$ GeV

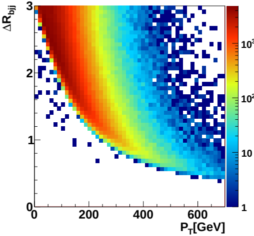
2– filtering defining 3-5 decay jets

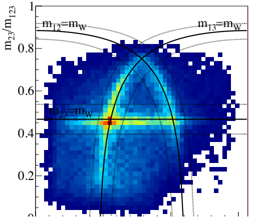
3– top mass window $m_{123} = [150, 200]$ GeV

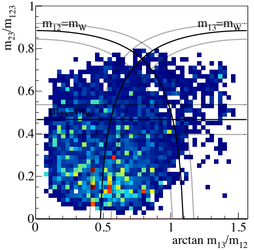
4– mass plane cuts extracting $m_{ij} \approx m_W$

$\Rightarrow$ Not rocket science, but crucial to build trust









Multi-variate taggers

Developing the benchmark

- multivariate analysis generally old news
  multivariate tagger 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]

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

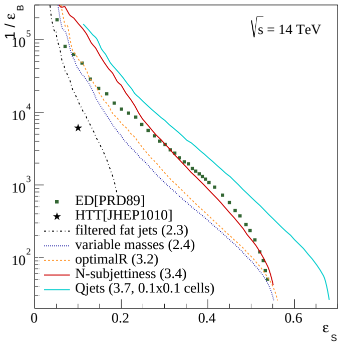
- add N-subjettiness [Thaler, van Tilburg]
- $\{ m_{123}, f_W, R_\text{opt} - R_\text{opt}^\text{(calc)}, \tau_j, \tau_j^\text{(filt)} \}$

$\Rightarrow$ Theory all but precision

Fat jet and top kinematics

- jet radiation major problem for Z' search
- tag and reconstruction in each other's way
- $\{ ..., m_{tt}, p_{T,t}, m_{jj}^\text{(filt)}, p_{T,j}^\text{(filt)} \}$

$\Rightarrow$ Is this all we can do?

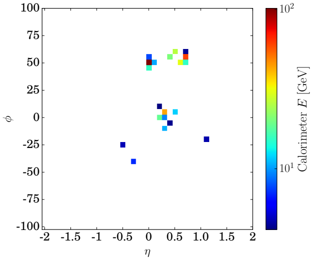


Jet image machines

The natural next step [Cogan etal, Oliveira, Nachman etal, Baldi, Whiteson etal (2014/15)]

– why intermediate high-level variables?
– learn theory through more NN layers
– calorimeter output as image
– as data-based as possible

⇒ Deep learning = modern networks on low-level observables

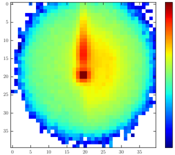


Jet image machines

The natural next step [Cogan etal, Oliveira, Nachman etal, Baldi, Whiteson etal (2014/15)]

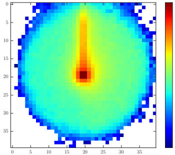
- why intermediate high-level variables?
- learn theory through more NN layers
- calorimeter output as image
- as data-based as possible

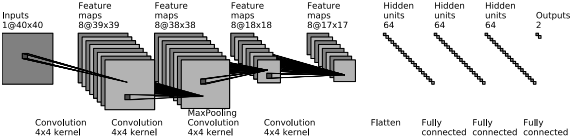
$\Rightarrow$ Deep learning = modern networks on low-level observables



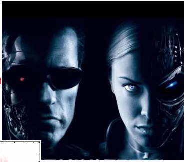
Convolutional network [Kasieczka, TP, Russell, Schell; Macaluso, Shih]

- image recognition standard ML task
- rapidity vs azimuthal angle, colored by energy deposition
- top tagging on 2D jet images
- 40 $\times$ 40 bins through calorimeter resolution



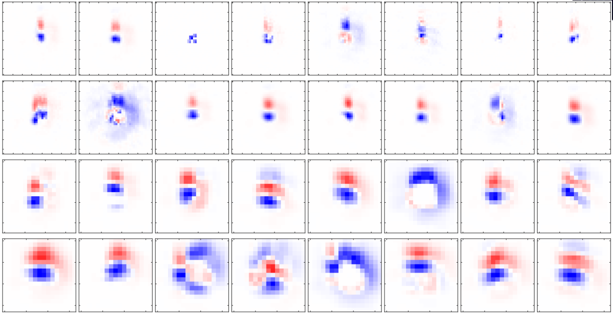


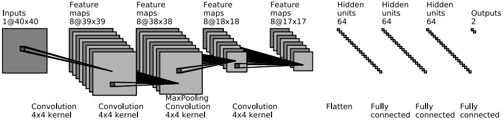
Inside DeepTop



Particle physicists as weird users [Kasieczka, TP, Russell, Schell; Macaluso & Shih]

– 2+2 convolutional layers





Inside DeepTop

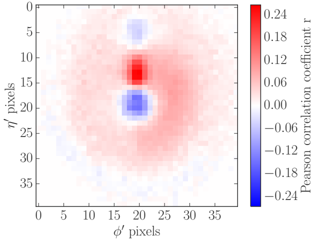
Particle physicists as weird users [Kasieczka, TP, Russell, Schell; Macaluso & Shih]

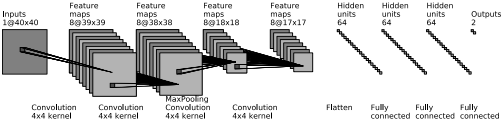
– 2+2 convolutional layers

– 3 fully connected layers

– Pearson input-output correlation [pixel x vs label y]

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





Inside DeepTop

Particle physicists as weird users [Kasieczka, TP, Russell, Schell; Macaluso & Shih]

– 2+2 convolutional layers

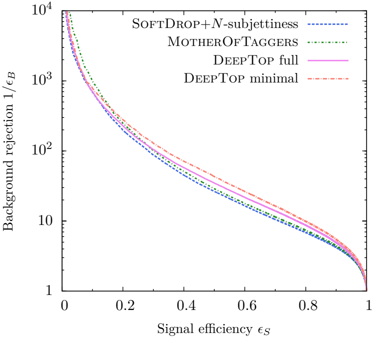
– 3 fully connected layers

– 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 MotherOfTaggers BDT

⇒ Understandable performance gain



Inside DeepTop

Particle physicists as weird users [Kasieczka, TP, Russell, Schell; Macaluso & Shih]

– 2+2 convolutional layers

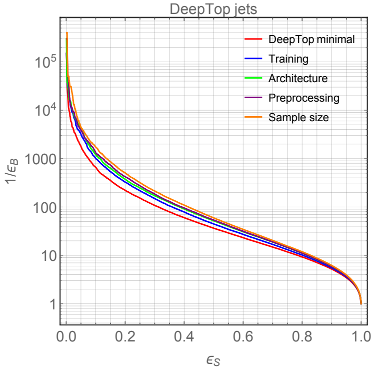
– 3 fully connected layers

– 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 MotherOfTaggers BDT

⇒ Understandable performance gain



Inside DeepTop

Particle physicists as weird users [Kasieczka, TP, Russell, Schell; Macaluso & Shih]

- 2+2 convolutional layers
- 3 fully connected layers
- Pearson input-output correlation [pixel x vs label y]

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

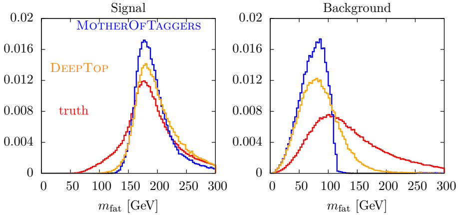
- comparison to MotherOfTaggers BDT

⇒ Understandable performance gain

Typical reaction: ‘F*** you, you f***ing machine’

- full control for supervised learning
  easy checks for correctly identified signal/background
- MC truth vs MotherOfTaggers vs DeepTop
  fat jet mass
  N-subjettiness
  transverse momenta

⇒ The box is not black



Grand theory ideas

Networks with 4-vector input [Butter, Kasieczka, TP, Russell; many more by now]

– sparsely filled picture: graph CNN

– physics objects from calorimeter and tracker

– distance measure known from e&m [alternatively: Erdmann, Rath, Rieger]

Inspired by jet algorithm — combination layer

– input 4-vectors

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

– combining them

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

Grand theory ideas

Networks with 4-vector input [Butter, Kasieczka, TP, Russell; many more by now]

– sparsely filled picture: graph CNN
– physics objects from calorimeter and tracker
– distance measure known from e&m [alternatively: Erdmann, Rath, Rieger]

Inspired by jet algorithm — combination layer

– input 4-vectors $(k_{\mu,i})$
– combining them $k_{\mu,i} \xrightarrow{\text{CoLa}} \widetilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij}$

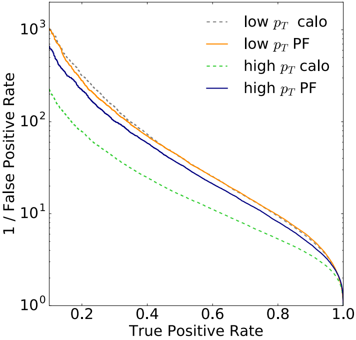
Inspired by Jackson — Lorentz layer

– DNN on Lorentz scalars

$$\tilde{k}_j \xrightarrow{\text{LoLa}} \hat{k}_j = \begin{pmatrix} m^2(\tilde{k}_j) \\ p_T(\tilde{k}_j) \\ \vdots \end{pmatrix}$$

$\Rightarrow$ Learn Minkowski metric

$g = $diag$(0.99 \pm 0.02, -1.01 \pm 0.01, -1.01 \pm 0.02, -0.99 \pm 0.02)$



Meet the professionals

A brief history of achievement

– 2014/15: first jet image papers

– 2017: first (working) ML top tagger

– ML4Jets 2017: What architecture works best?

– ML4Jets 2018: Lots of architectures work [1902.09914]

⇒ Jet classification understood

SciPost Physics Submission

The Machine Learning Landscape of Top Taggers

G. Kasieczka (ed)[1], T. Plehn (ed)[2], A. Butter[2], K. Cranmer[3], D. Debnath[4], M. Fairbairn[5], W. Fedorko[6], C. Gay[6], L. Gouskos[7], P. T. Komiske[8], S. Leiss[1], A. Lister[6], S. Macaluso[3,4], E. M. Metodiev[8], L. Moore[9], B. Nachman,[10,11], K. Nordström[12,13], J. Pearkes[6], H. Qu[7], Y. Rath[14], M. Rieger[14], D. Shih[4], J. M. Thompson[2], and S. Varma[5]

1 Institut für Experimentalphysik, Universität Hamburg, Germany
2 Institut für Theoretische Physik, Universität Heidelberg, Germany
3 Center for Cosmology and Particle Physics and Center for Data Science, NYU, USA
4 NHECT, Dept. of Physics and Astronomy, Rutgers, The State University of NJ, USA
5 Theoretical Particle Physics and Cosmology, King's College London, United Kingdom
6 Department of Physics and Astronomy, The University of British Columbia, Canada
7 Department of Physics, University of California, Santa Barbara, USA
8 Center for Theoretical Physics, MIT, Cambridge, USA
9 CP3, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
10 Physics Division, Lawrence Berkeley National Laboratory, Berkeley, USA
11 Simons Inst. for the Theory of Computing, University of California, Berkeley, USA
12 National Institute for Subatomic Physics (NIKHEF), Amsterdam, Netherlands
13 LPTHE, CNRS & Sorbonne Université, Paris, France
14 III. Physics Institute A, RWTH Aachen University, Germany

gregor.kasieczka@uni-hamburg.de
plehn@uni-heidelberg.de

April 12, 2019

Abstract

Based on the established task of identifying boosted, hadronically decaying top quarks, we compare a wide range of modern machine learning approaches. We find that they are extremely powerful and great fun.

Content

Meet the professionals

A brief history of achievement

- 2014/15: first jet image papers
- 2017: first (working) ML top tagger
- ML4Jets 2017: What architecture works best?
- ML4Jets 2018: Lots of architectures work [1902.09914]

⇒ Jet classification understood

SciPost Physics Submission

The Machine Learning Landscape of Top Taggers

G. Kasieczka (ed)[1], T. Plehn (ed)[2], A. Butter[2], K. Cranmer[3], D. Debnath[4], M. Fairbairn[5], W. Fedorko[6], C. Gay[6], L. Gouskos[7], P. T. Komiske[8], S. Leiss[1], A. Lister[6], S. Macaluso[3,4], E. M. Metodiev[8], L. Moore[9], B. Nachman,[10,11], K. Nordström[12,13], J. Pearkes[6], H. Qu[7], Y. Rath[14], M. Rieger[14], D. Shih[4], J. M. Thompson[2], and S. Varma[5]

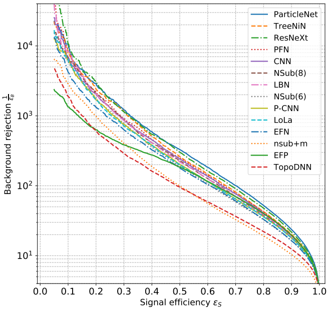
1 Institut für Experimentalphysik, Universität Hamburg, Germany
2 Institut für Theoretische Physik, Universität Heidelberg, Germany
3 Center for Cosmology and Particle Physics and Center for Data Science, NYU, USA
4 NHECT, Dept. of Physics and Astronomy, Rutgers, The State University of NJ, USA
5 Theoretical Particle Physics and Cosmology, King's College London, United Kingdom
6 Department of Physics and Astronomy, The University of British Columbia, Canada
7 Department of Physics, University of California, Santa Barbara, USA
8 Center for Theoretical Physics, MIT, Cambridge, USA
9 CP3, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
10 Physics Division, Lawrence Berkeley National Laboratory, Berkeley, USA
11 Simons Inst. for the Theory of Computing, University of California, Berkeley, USA
12 National Institute for Subatomic Physics (NIKHEF), Amsterdam, Netherlands
13 LPTHE, CNRS & Sorbonne Université, Paris, France
14 III. Physics Institute A, RWTH Aachen University, Germany

gregor.kasieczka@uni-hamburg.de
plehn@uni-heidelberg.de

April 12, 2019

Abstract

Based on the established task of identifying boosted, hadronically decaying top quarks, we compare a wide range of modern machine learning approaches. We find that they are extremely powerful and great fun.



Meet the professionals

A brief history of achievement

– 2014/15: first jet image papers

– 2017: first (working) ML top tagger

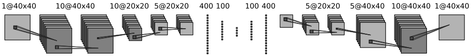
– ML4Jets 2017: What architecture works best?

– ML4Jets 2018: Lots of architectures work [1902.09914]

⇒ Jet classification understood

⇒ **What's new and cool?**

Learning background only



Fully supervised classification boring [Heimel, Kasieczka, TP, Thompson; Farina, Macari, Shih]

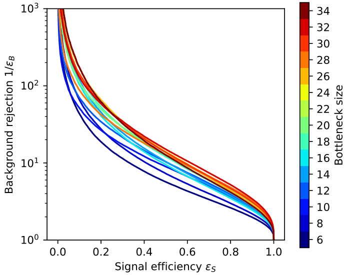
– anomaly searches, only training on ‘background’

– established ML concept: autoencoder

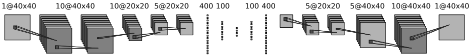
– reconstruct typical QCD jet image from many QCD jets
reduce weights in central layer, compress information to ‘typical’

– search for outliers hard to describe

⇒ Making an okay tagger



Learning background only



Fully supervised classification boring [Heimel, Kasieczka, TP, Thompson; Farina, Macari, Shih]

– anomaly searches, only training on ‘background’

– established ML concept: autoencoder

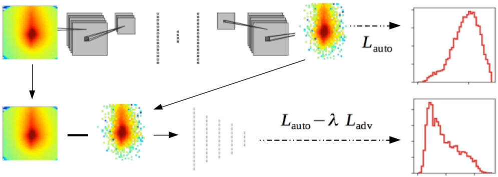
– reconstruct typical QCD jet image from many QCD jets
reduce weights in central layer, compress information to ‘typical’

– search for outliers hard to describe

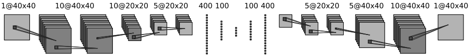
⇒ Making an okay tagger

De-correlate background shaping

– established concept: adversary [Shimmin,...]



Learning background only



Fully supervised classification boring [Heimel, Kasieczka, TP, Thompson; Farina, Macari, Shih]

– anomaly searches, only training on ‘background’

– established ML concept: autoencoder

– reconstruct typical QCD jet image from many QCD jets
reduce weights in central layer, compress information to ‘typical’

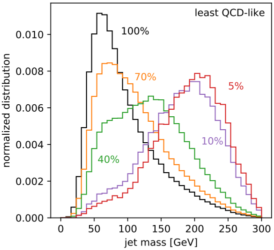
– search for outliers hard to describe

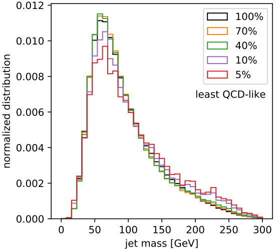
$\Rightarrow$ Making an okay tagger

De-correlate background shaping

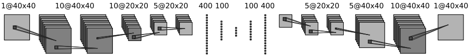
– established concept: adversary [Shimmin,...]

– atypical QCD jets typially with large jet mass
remove jet mass from network training





Learning background only



Fully supervised classification boring [Heimel, Kasieczka, TP, Thompson; Farina, Macari, Shih]

– anomaly searches, only training on ‘background’

– established ML concept: autoencoder

– reconstruct typical QCD jet image from many QCD jets
  reduce weights in central layer, compress information to ‘typical’

– search for outliers hard to describe

$\Rightarrow$ Making an okay tagger

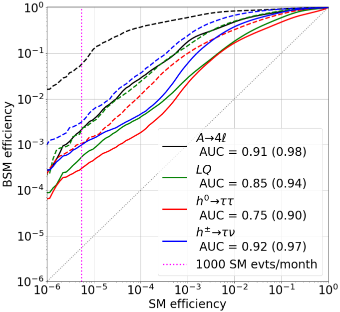
The whole thing on anomalous LHC events [Cerri, Nguyen, Pierini, Spiropulu, Vlimant]

– same thing on full events

– training data a problem

– variational autoencoder more powerful

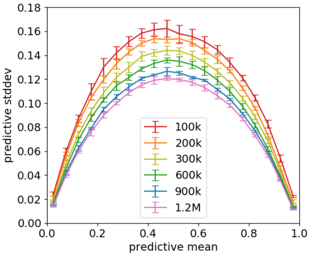
$\Rightarrow$ Proof of concept...

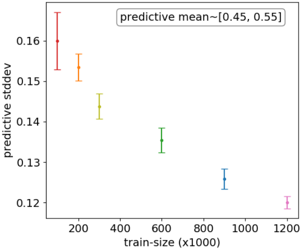


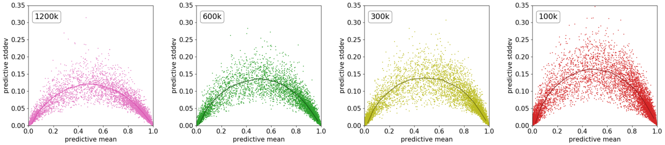
B****ian networks

Simply better networks [Bollweg, Haussmann, Kasieczka, Luchmann, TP, Thompson (soon)]

- learn classification output and uncertainty [(60 ± 30)% top different from (60 ± 1)% top]
- error bars: limited training statistics







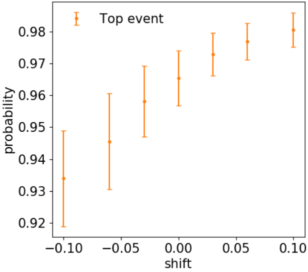
B****ian networks

Simply better networks [Bollweg, Haussmann, Kasieczka, Luchmann, TP, Thompson (soon)]

– learn classification output and uncertainty [

– error bars: limited training statistics

– error bars: jet energy scale (correlated)



B****ian networks

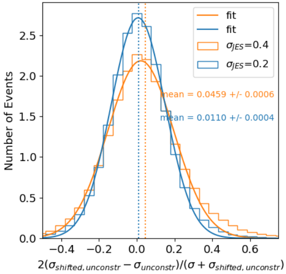
Simply better networks [Bollweg, Haussmann, Kasieczka, Luchmann, TP, Thompson (soon)]

– learn classification output and uncertainty [(60 ± 30)% top different from (60 ± 1)% top]

– error bars: limited training statistics

– error bars: jet energy scale (correlated)

– error bars: jet energy scale (uncorrelated)



B****ian networks

Simply better networks [Bollweg, Haussmann, Kasieczka, Luchmann, TP, Thompson (soon)]

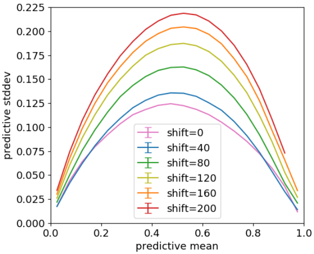
– learn classification output and uncertainty [

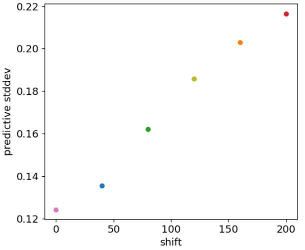
– error bars: limited training statistics

– error bars: jet energy scale (correlated)

– error bars: jet energy scale (uncorrelated)

– stability detection: pile-up

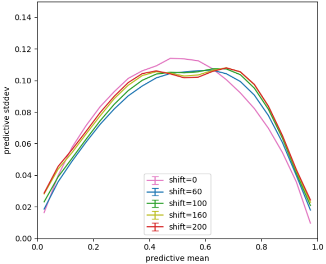




B****ian networks

Simply better networks [Bollweg, Haussmann, Kasieczka, Luchmann, TP, Thompson (soon)]

– learn classification output and uncertainty [(60 ± 30)% top different from (60 ± 1)% top]
– error bars: limited training statistics
– error bars: jet energy scale (correlated)
– error bars: jet energy scale (uncorrelated)
– stability detection: pile-up



B****ian networks

Simply better networks [Bollweg, Haussmann, Kasieczka, Luchmann, TP, Thompson (soon)]

– learn classification output and uncertainty [(60 $\pm$ 30)% top different from (60 $\pm$ 1)% top]

– error bars: limited training statistics

– error bars: jet energy scale (correlated)

– error bars: jet energy scale (uncorrelated)

– stability detection: pile-up

– tagger calibration part of the training

– systematic approach to regularization and drop-out

– performance just like usual taggers

....

– Lots of conceptual and practical advantages at no cost

The future

Machine learning is an amazing tool box...

...LHC physics really is big data

...imagine recognition is a starting point

...deep learning is not just classification

...jets are not the only interesting objects at LHC

...Bayesian networks are extremely likable

Let's find even cooler applications!