Deep learning in the collider physics

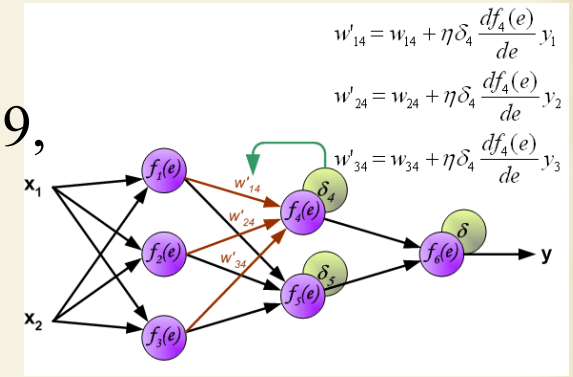
- About the history
- Main applications of DNN in collider experiments
- Measurements and Searches with DNN

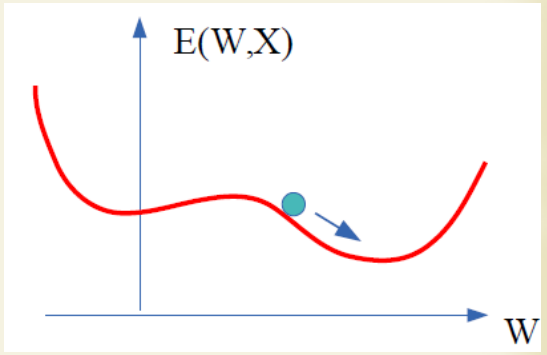
Lev Dudko

SINP MSU

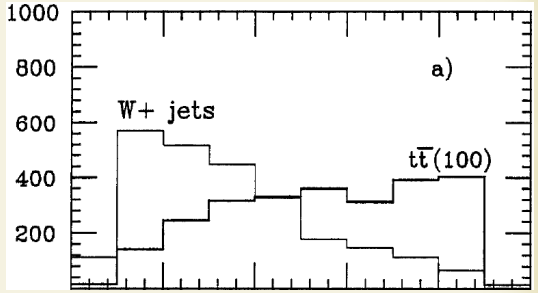
The short history of NN in the collider physics

1) Hornick, Stinchcombe, White. «Multilayer Feedforward Networks are Universal Approximators.» Neural Networks, 1989, v. 2, 5. ;
2) Cybenko. «Approximation by Superpositions of a Sigmoidal Function.» Mathematical Control Signals Systems, 1989, 2.;
3) Funahashi. «On the Approximate Realization of Continuous Mappings by Neural Networks.» Neural Networks, 1989, v. 2, 3.

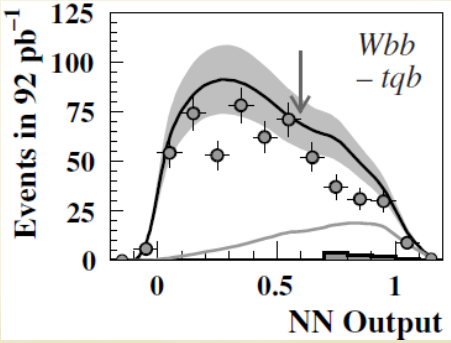




------ AIHEP'1990 (first in series)
- «Status of HEP neural net research in the USA.» Bruce H. Denby, Stephan L. Linn Comp.Phys.Com. 57 (**1989**) 297-300, 5cit.
- «Snagging the top quark with a neural net» Howard Baer, Debra Dzialo Karatas, Gian Giudice, Phys.Rev.D 46 (**1992**) 4901-4906, 6 cit.

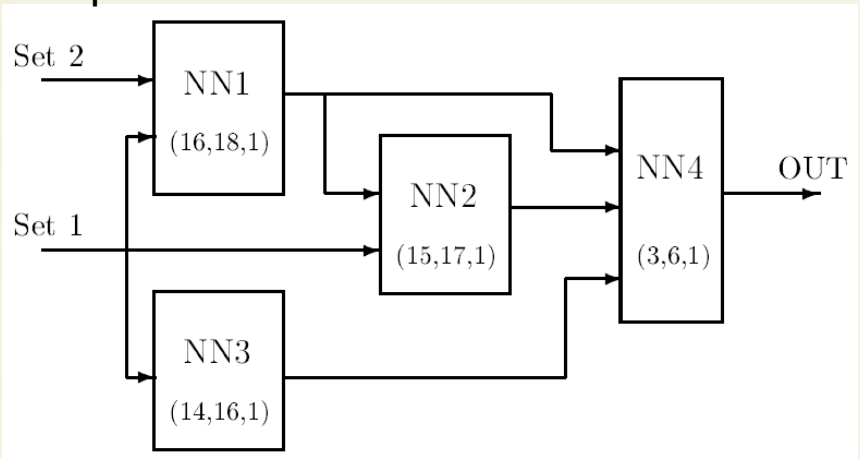


...
- Phys.Rev.D 63 (**2000**) 031101 – D0 collab. Single top search, classical analysis; 86 cit.
Phys.Lett.B 517 (**2001**) 282-294 – D0 collab. Single top search neural network analysis, 107 citations.

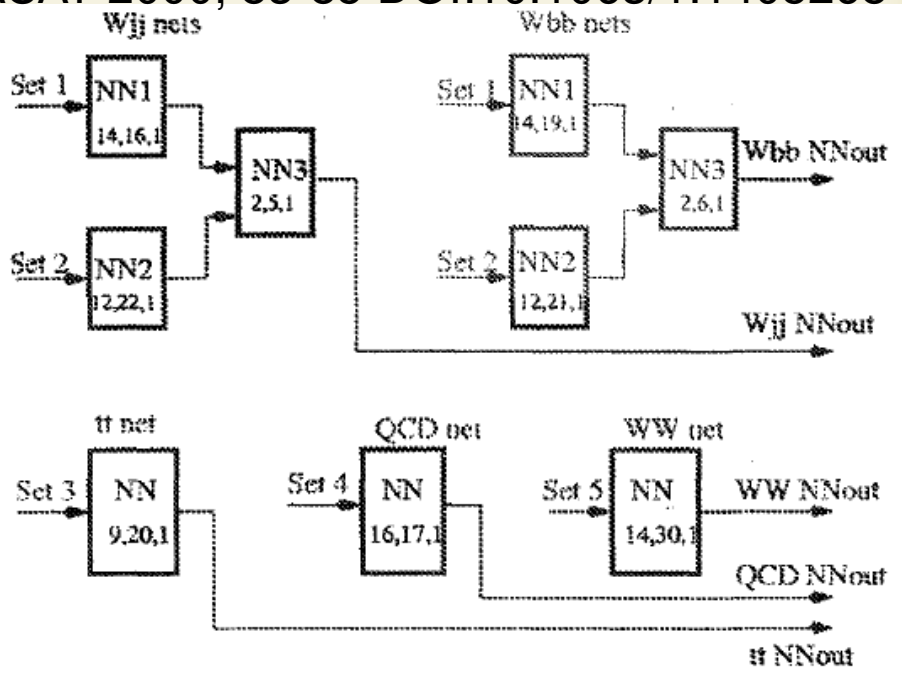


Optimization of NN application for the single top search in D0 experiment.

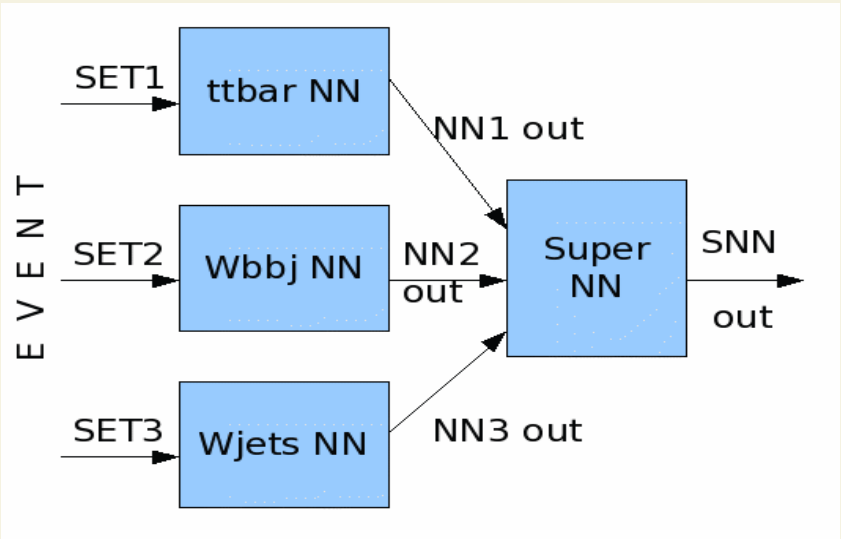
hep-ex/9907041

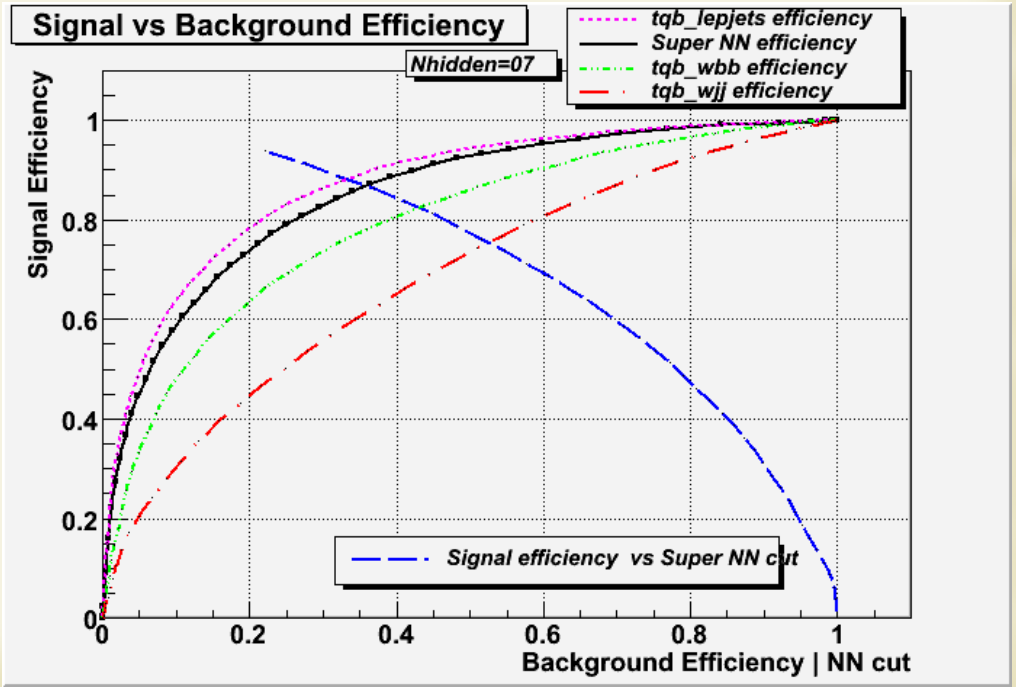


ACAT 2000, 83-85 DOI:10.1063/1.1405268

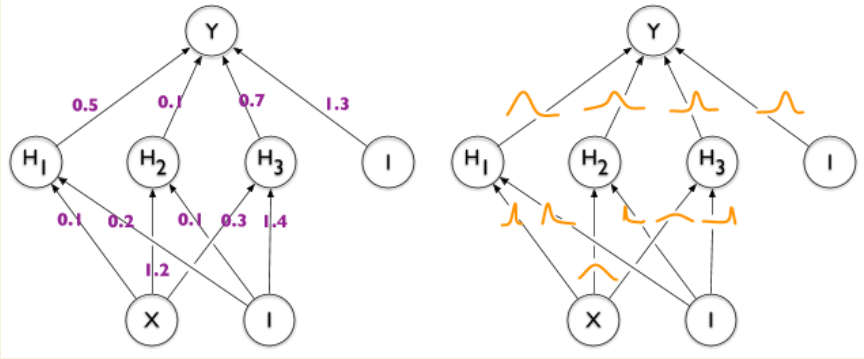


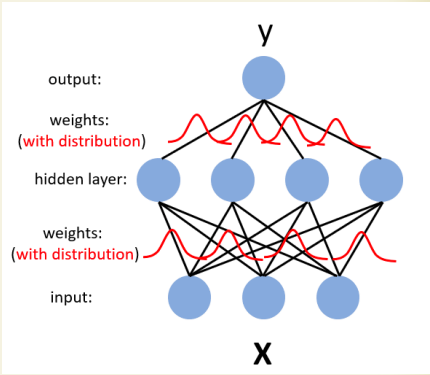
Phys.Lett.B 517 (2001) 282-294





Bayesian Neural Networks (BNN)



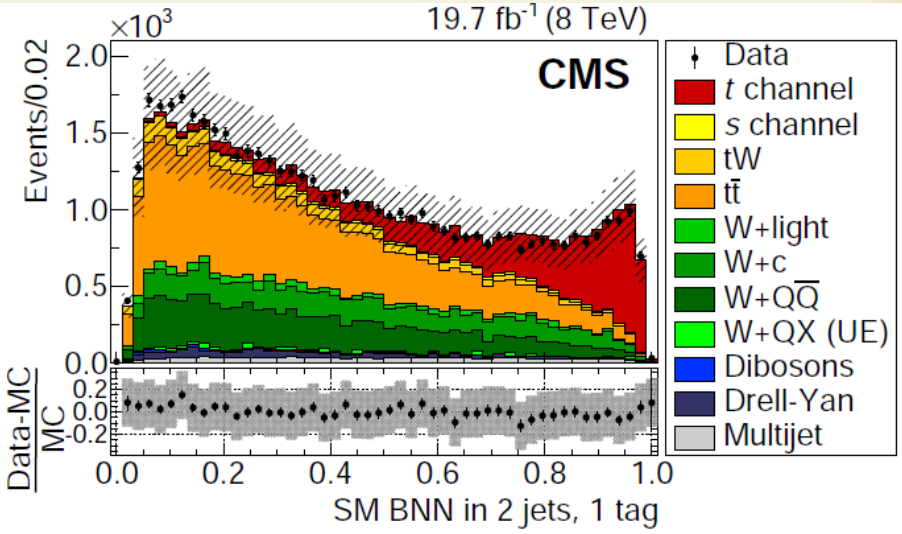


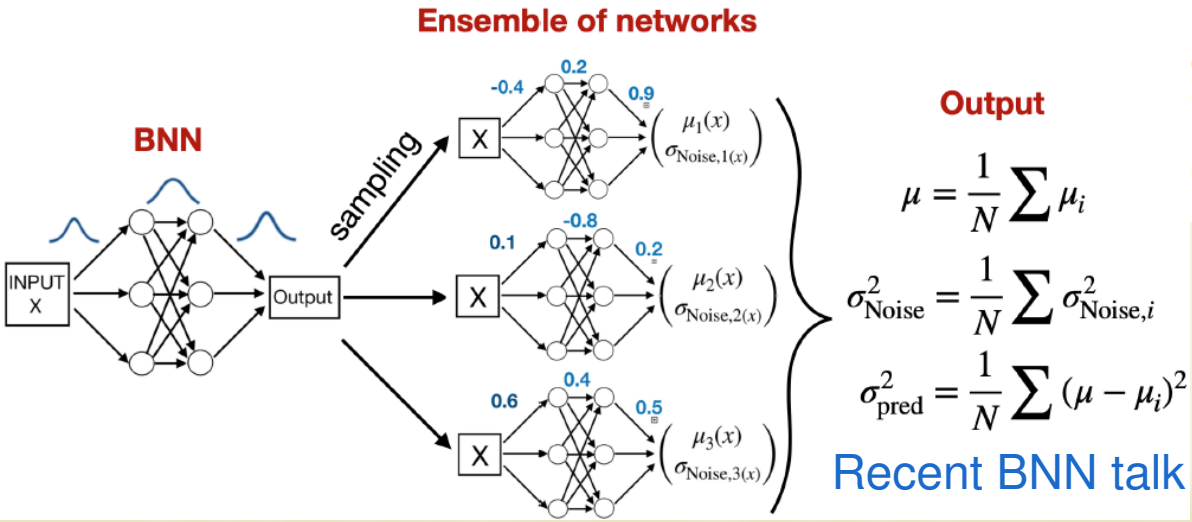
P. C. Bhat and H. B. Prosper, "Bayesian Neural Networks" PHYSTAT 2005;
R. M. Neal, Bayesian Learning of Neural Networks (1996); FBM package;

All of D0 analyses after 2005 use BNN not NN
e.g. D0, Observation of Single Top Quark Production
Phys.Rev.Lett. 103 (2009) 092001

Partial realisation in deep NN: tensorflow_probability, variational dropout, … , with fixed form of distribution.

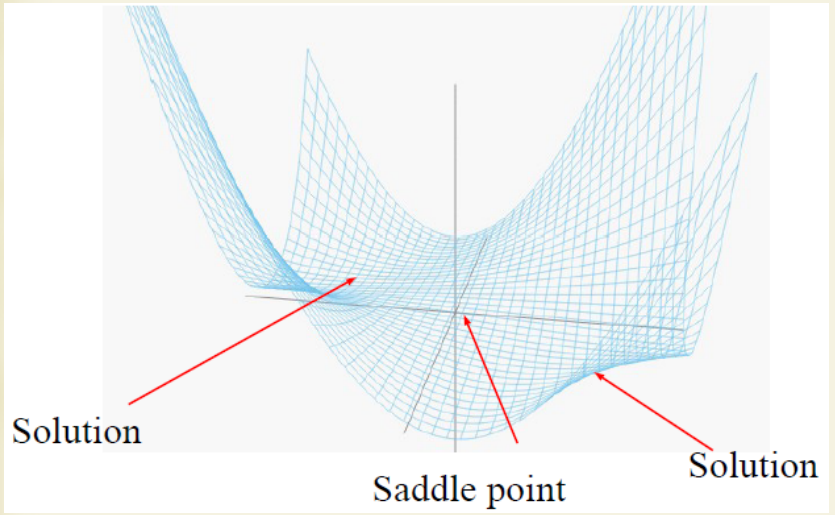
BNN in CMS (LHC) JHEP 02 (2017) 028

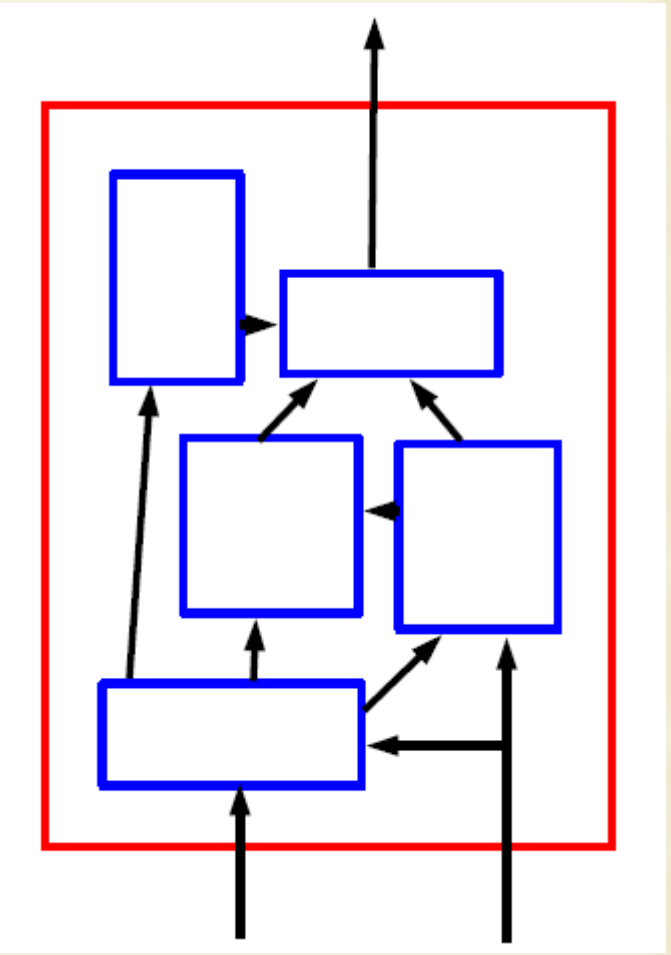




Recent BNN talk T.Plehn (06/22)

Deep Learning Neural Networks, DNN





Hinton, G. E., Osindero, S., & Teh, Y. W. (2006).
A fast learning algorithm for deep belief nets.
Neural computation, 18(7), 1527-1554.

The main advantage of DNN is the ability to analyze raw, not preprocessed data.

Probably, the first application in HEP: Nature Commun. 5 (2014) 4308

	Discovery significance		
Technique	Low-level	High-level	Complete
NN	2.5σ	3.1σ	3.7σ
DN	4.9σ	3.6σ	5.0σ

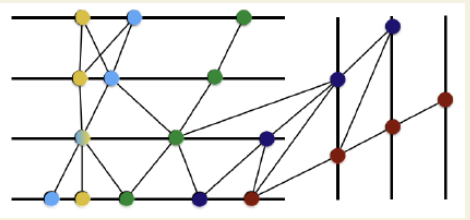
$$gg \to H^0 \to W^{\mp}H^{\pm} \to W^{\mp}W^{\pm}h^0$$

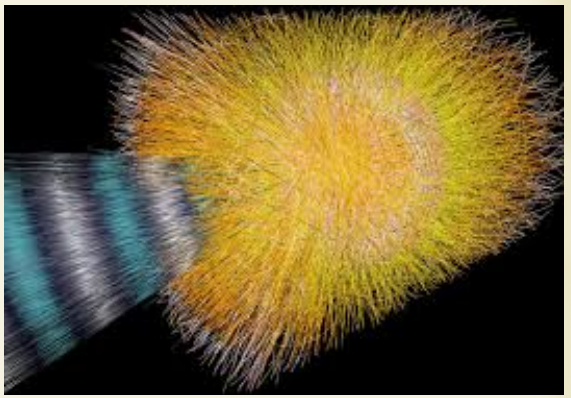
Modern applications of DNN in collider experiments

- **Triggers**, online event selection, hardware and software DNN implementations
- **Reconstruction and identification of the objects**, software implementations of DNN.
- **Classification of events**, distinguishing of some signal process from background processes. Problem of event negative weights.
- **Anomaly detection**, search for some deviations in data, unsupervised training, autoencoders. Low efficiency.
- **Fast simulation**, using GAN to simulate more events, or detector response. Usually does not decrease statistical uncertainty
- Parton density functions **NNPDF** – most used PDF now
- **Unfolding**, back from detector level to parton level
- **Regression tasks** to estimate some model parameter(s)
- **Symbolic regression**, to estimate an analytic function from data
- **Self-driving lab**oratory

Reconstruction and identification of the objects

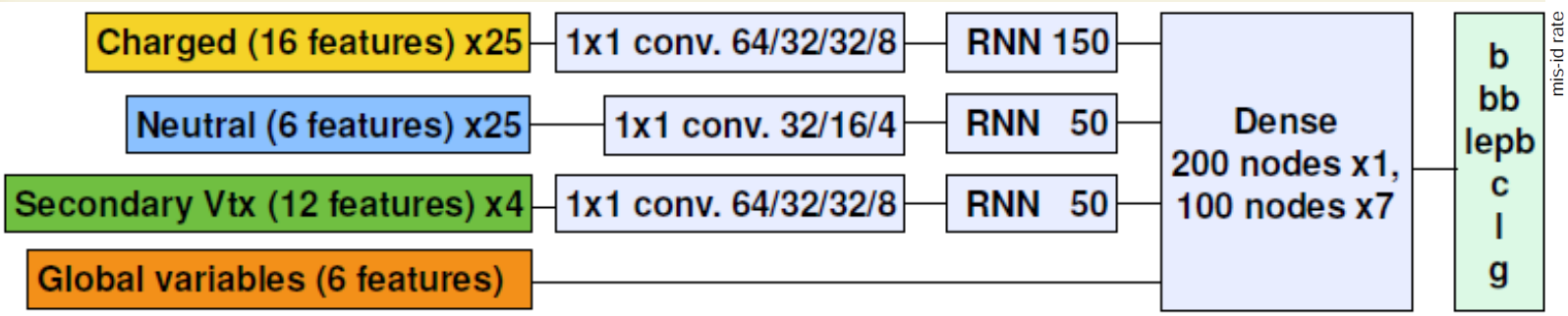
- Track reconstruction (GraphNN, ...)

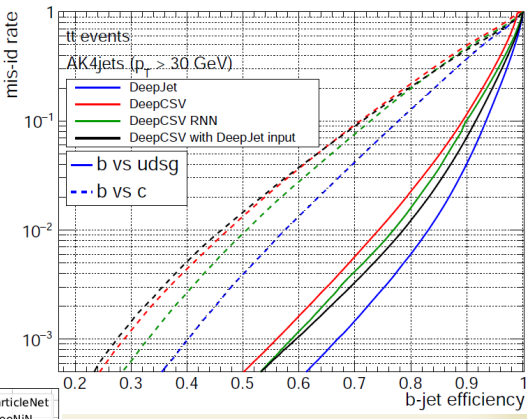




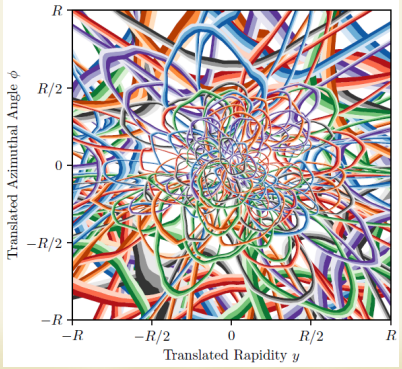
- Identification of the objects (jet b-tagging, top-tagging, …)

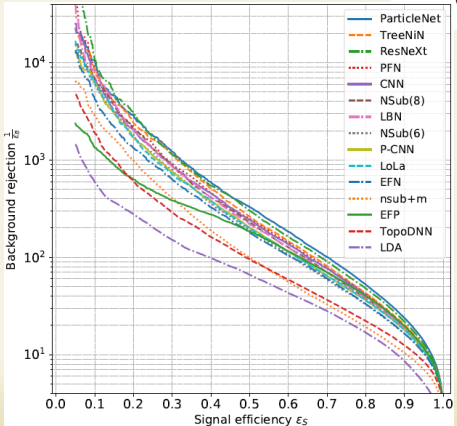
DeepJet CMS, JINST 15 (2020) 12, P12012





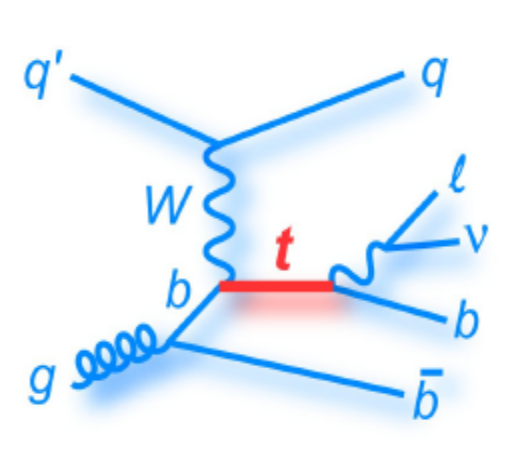
SciPost Phys. 7 (2019) 014 – landscape of top-taggers





Classification of events

Signal process signature



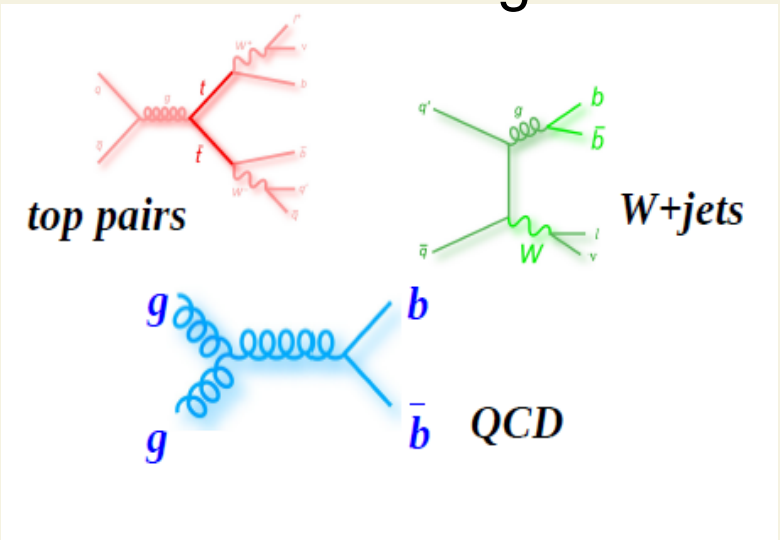
Light flavor jet

Lepton
Missing Et

High Pt b-jet

Low Pt b-jet

Irreducable and reducable backgrounds

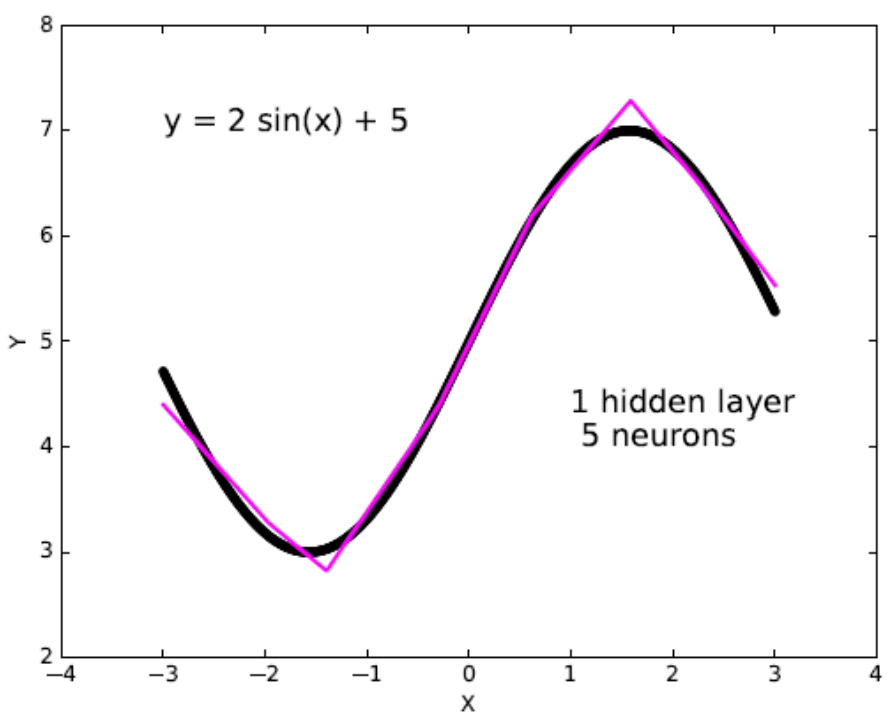


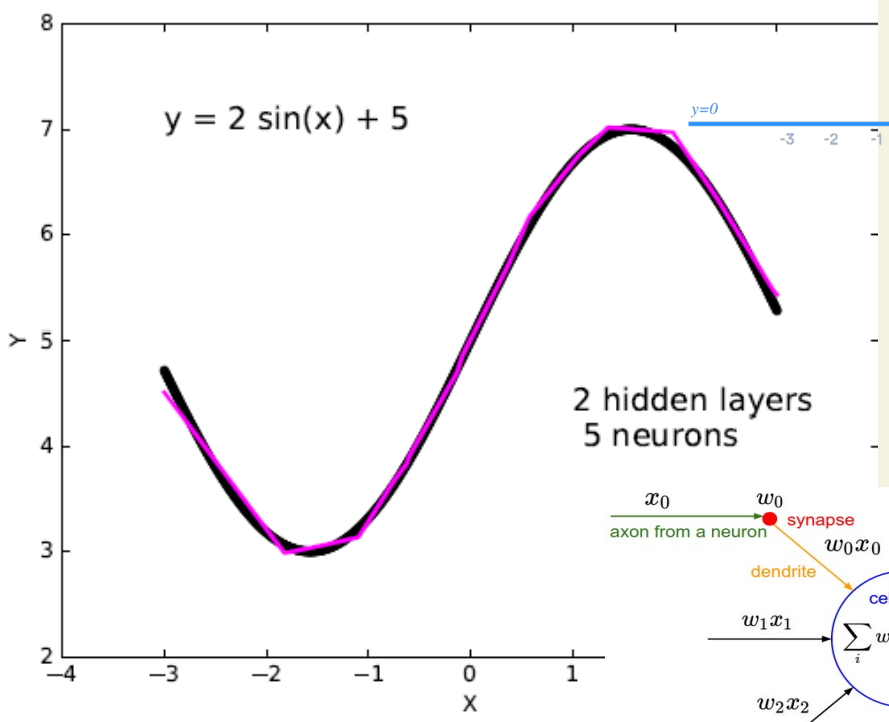
Total and differential cross sections $d\sigma \sim M^2(p_i \cdot p_f, s, t, u)$ are the functions of scalar products of four-momenta and Mandelstam variables.
Example of squared matrix element for u,d→t,b process:

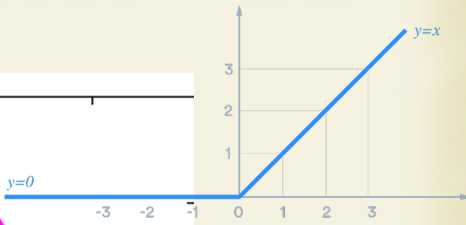
$$|M|^2 = V_{tb}^2 V_{ud}^2 (g_W)^4 \frac{(p_u p_b)(p_d p_t)}{\left(\hat{s} - m_W^2\right)^2 + \Gamma_W^2 m_W^2},$$

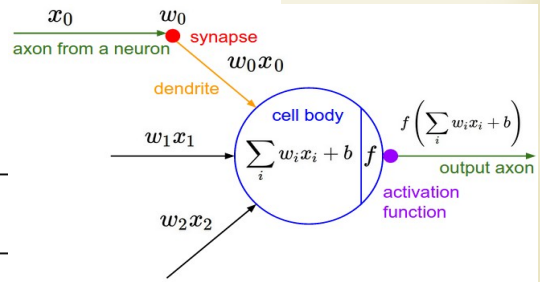
$$|M|^2 = V_{tb}^2 V_{ud}^2 (g_W)^4 \frac{\hat{t}\left(\hat{t} - M_t^2\right)}{\left(\hat{s} - m_W^2\right)^2 + \Gamma_W^2 m_W^2}.$$

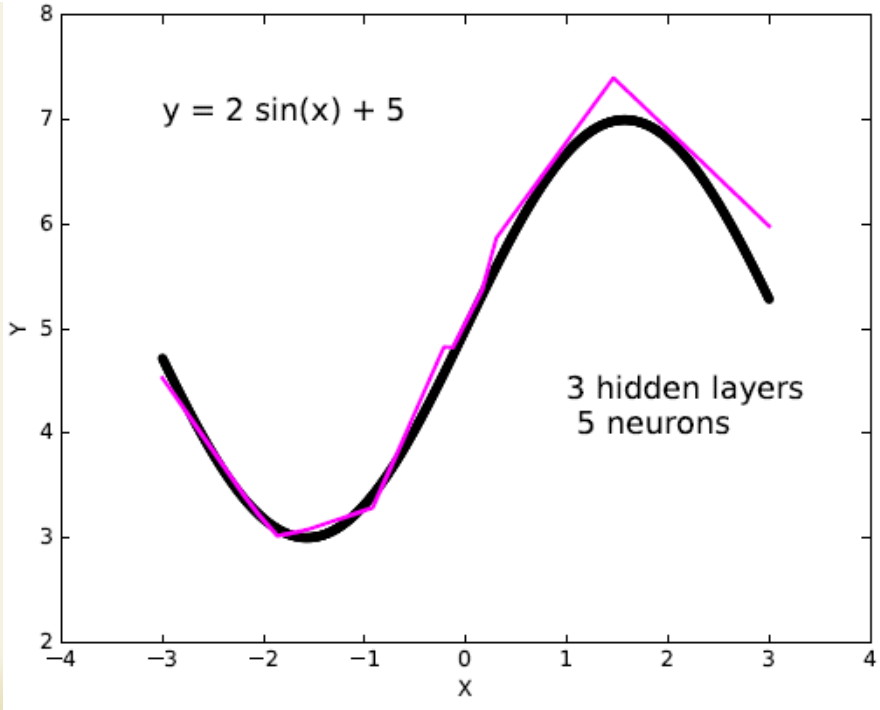
Simple example of NN

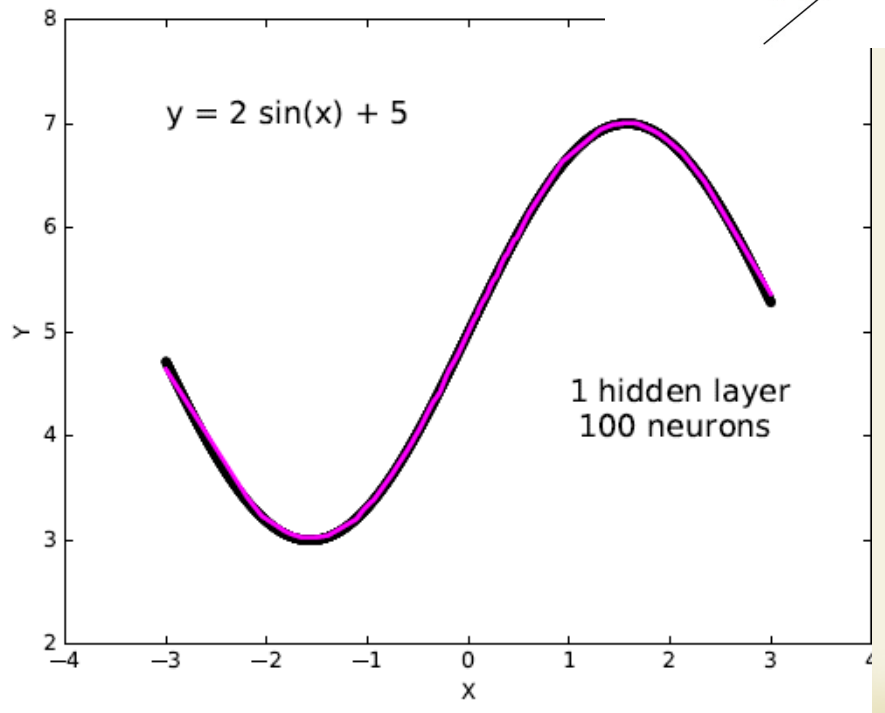












Some general remarks, based on experience

1) Increasing the complexity of DNN (number of nodes, layers) leads to complicate training and usually decrease efficiency of DNN.

2) Input information (input vector) should contains complete set of important observables, without overabundant information which complicates training.

3) Decrease the order of nonlinearity in the task for DNN
e.g. $F(x)=x^2 \rightarrow NN(x^2)$ not $NN(x)$

3) Preprocessing of input data improves training. Understand your data.

4) Use minimally sufficient size of DNN (number of nodes, layers).

5) Do control and minimize overfitting (dropout, regularisation, test samples) to keep stability of the result.

5) Do control on propagation of uncertanties to DNN output (precision means low uncertanty, not only efficient classification).

Method of high level “optimal observables”

- **Provides general recipe how to choose most sensitive high-level variables to separate signal and background**
  - It is based on the analysis of Feynman diagrams (FD) contributing to signal and background processes
  - Distinguish **three classes** of sensitive variables for the signal and each of kinematically different backgrounds: **Singular** variables (denominators of FD), **Angular** variables (numerators of FD) and **Threshold** variables (Energy thresholds of the processes)
  - Set of variables can be extended with other type of information, like detector relative variables (jet width, b-tagging discriminant)

  Described in different examples for the top and Higgs searches
  - E.Boos, L.Dudko, T.Ohl Eur.Phys.J. C11 (1999) 473-484
  - E.Boos, L.Dudko Nucl.Instrum.Meth. A502 (2003) 486-488
  - E.Boos, V.Bunichev, L.Dudko, A.Markina, M.Perfilov Phys.Atom.Nucl. 71 (2008) 388-393
- **Applied in different experimental analysis in D0 and CMS**
  - Phys.Lett. B517 (2001) 282-294 and other D0 publications
  - JHEP02(2017)028 , ...

Three Classes of Variables

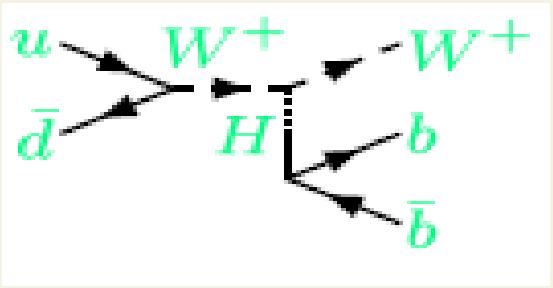
1) "Singular" Sensitive Variables

(denominator of Feynman diagrams)

Most of the rates of signal and background processes come from the integration over the phase space region close to the singularities. If some of the singular variables are different or the positions of the singularities are different the corresponding distributions will differ most strongly

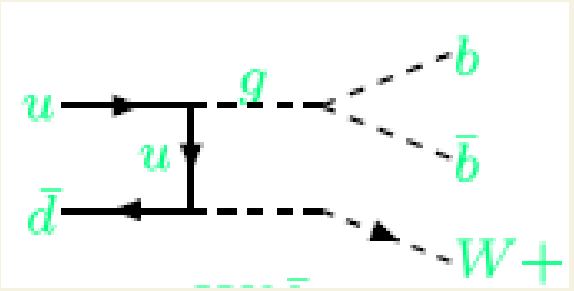
s-channel singularities

$$M^2_{f1,f2} = (p_{f1} + p_{f2})^2$$



t-channel singularities

$$\hat{t}_{i,f} = (p_f - p_i)^2 = -\sqrt{\hat{s}}e^{Y}\, p_T^f e^{-|y_f|}$$



Three Classes of Variables

2) "Angular" variables, Spin effects

(numerator of Feynman diagrams)

e.g. $$\frac{1}{\Gamma_T}\frac{d\Gamma}{d(\cos\chi_\ell^W)} = \frac{3}{4}\frac{m_t^2\sin^2\chi_\ell^W + 2m_W^2\frac{1}{2}(1-\cos\chi_\ell^W)^2}{m_t^2+2m_W^2}$$

G. Mahlon, S. Parke Phys.Rev. D55 (1997) 7249-7254

3) "Threshold" variables

e.g. s_hat and Ht variables relate to the fact that various signal and background processes may have very different energy thresholds

General method of low level “optimal observables”

The main advantage of Deep NNs (many layers, neurons) is the possibility to analyze raw, not preprocessed, information.
$2 \to n$ particles hard process has (3n-4) independent variables
What are the general low level observables?
[Int.J.Mod.Phys.A 35 (2020) 21, 2050119]

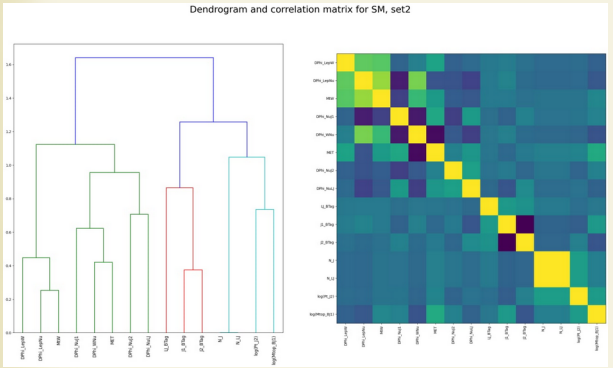
The proposed recipe is simple, need to use the following classes:

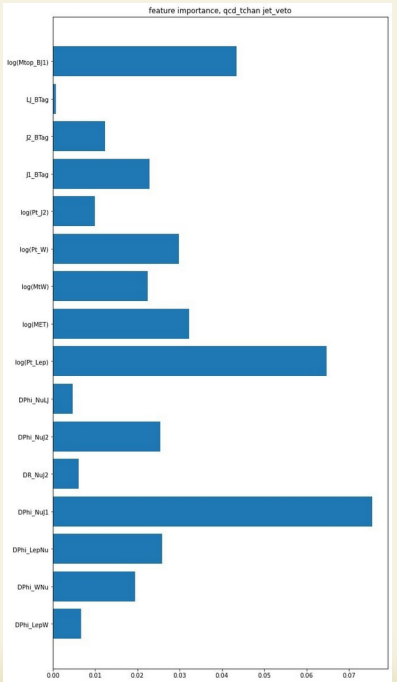
- scalar products of 4-momenta of the final particles,
- Mandelstam variables (only **s** are available for pp; **t**,**u** for lepton colliders),
- transverse momenta and pseudorapidity of the final particles (to approximate t-channel Mandelstam variables which depends on initial particles momenta).

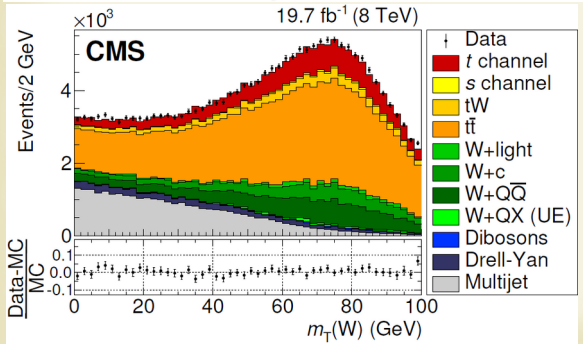
The proposed set of raw observables covers the kinematic differences in hard processes. In additional, it is possible to add some other type of information (e.g. b-tagging discriminant, charge of lepton, …)

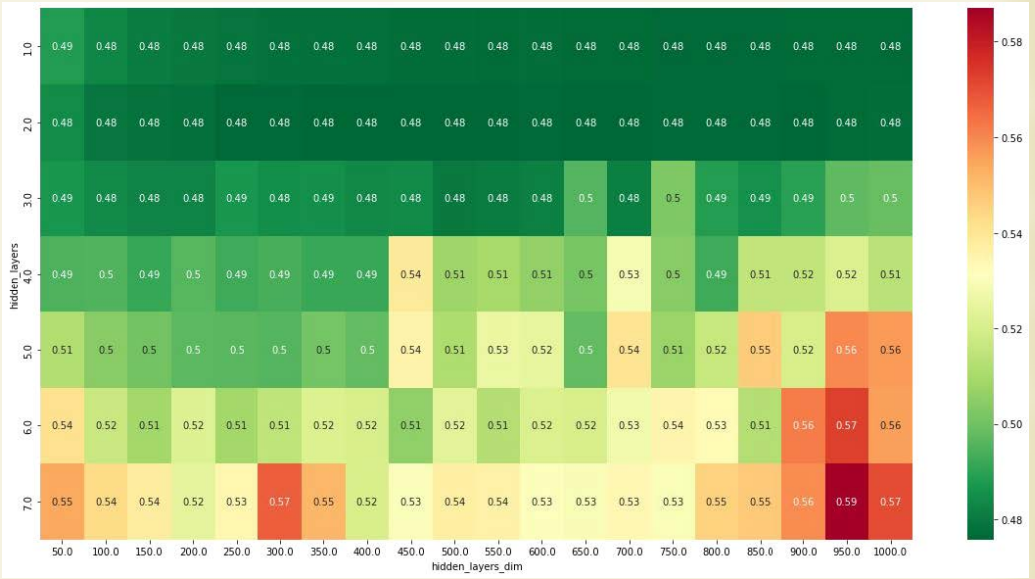
Optimization of input variables, architecture and training parameters of DNN

- Exclude variables with high linear correlations
- Check simulation of each variable (compare with data, or different simulations)
- Can check an importance of each variable
- Transform variables to the same dynamic scale [$(x-\bar{x})/\sigma$, log()]
- Tune architecture of DNN (number of nodes and hidden layers, dropout) and training parameters (e.g. use Keras tuner), based on ROC/AUC, Score or other metrics.



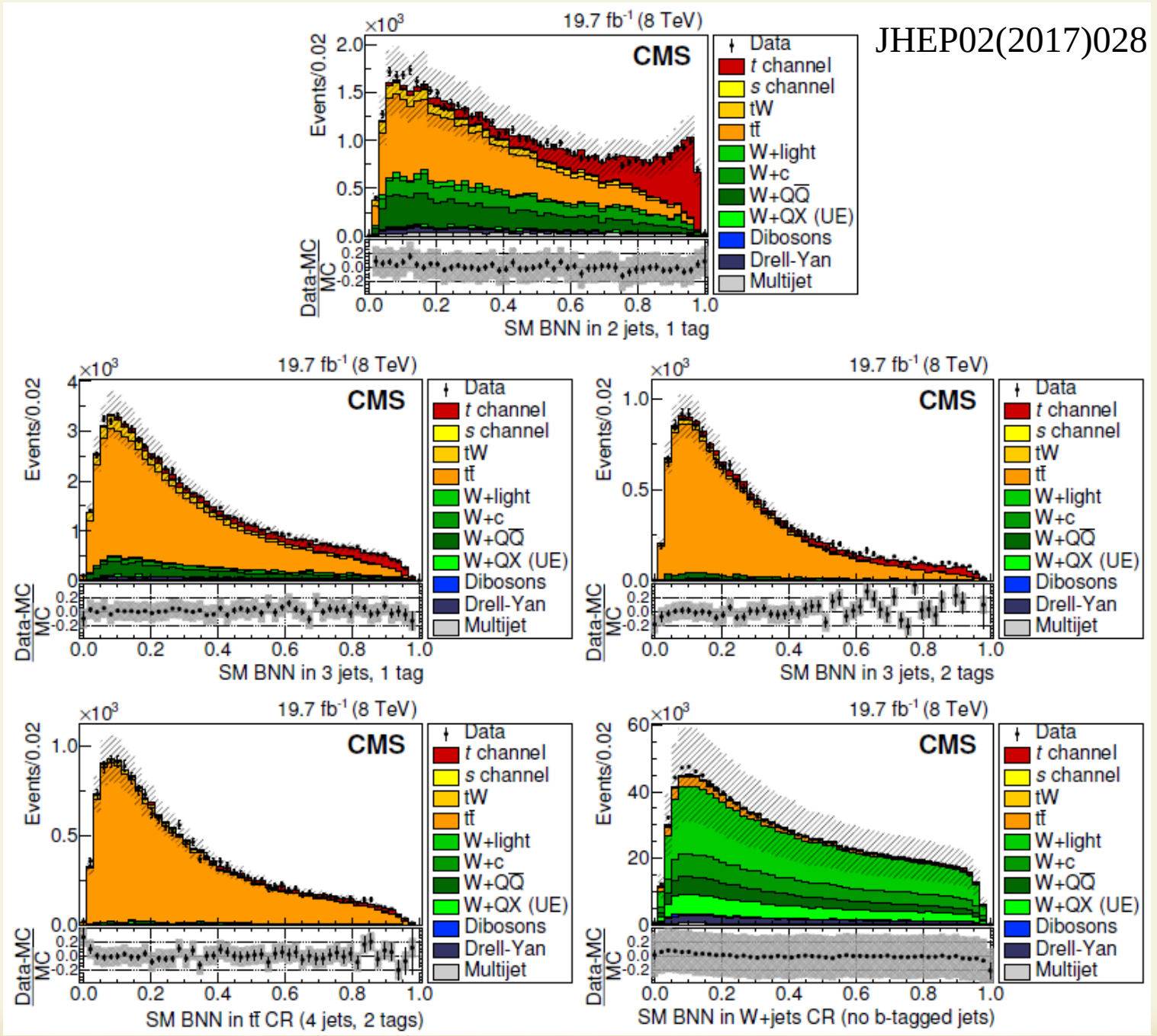




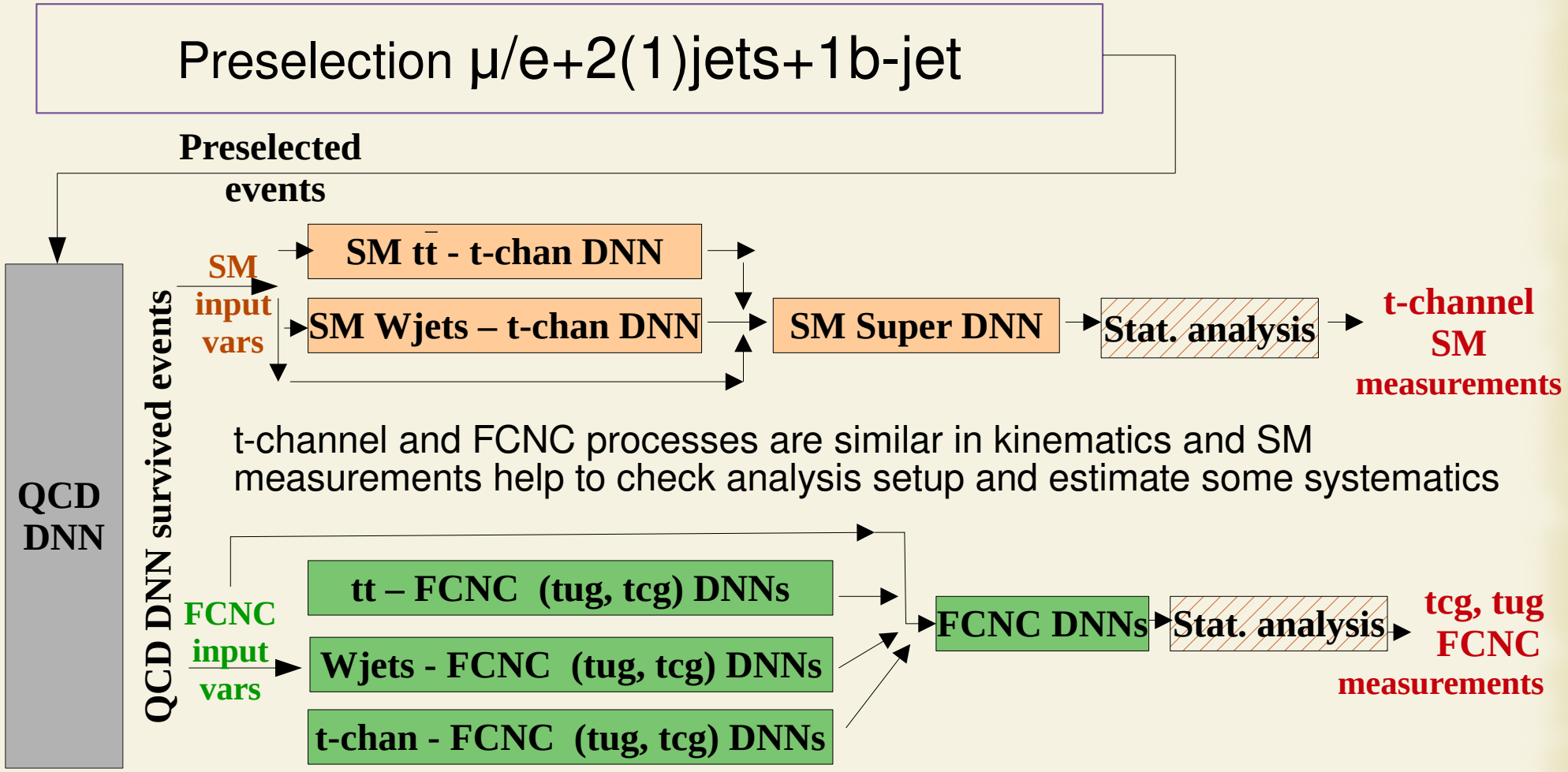


Check network for stability in different control regions, for different backgrounds

JHEP02(2017)028



Current realisations of DNN in our CMS analysis (LHC)



Run II.

Published results:

1) 7&8 TeV JHEP02(2017)028 (FCNC tqg & aWtb)

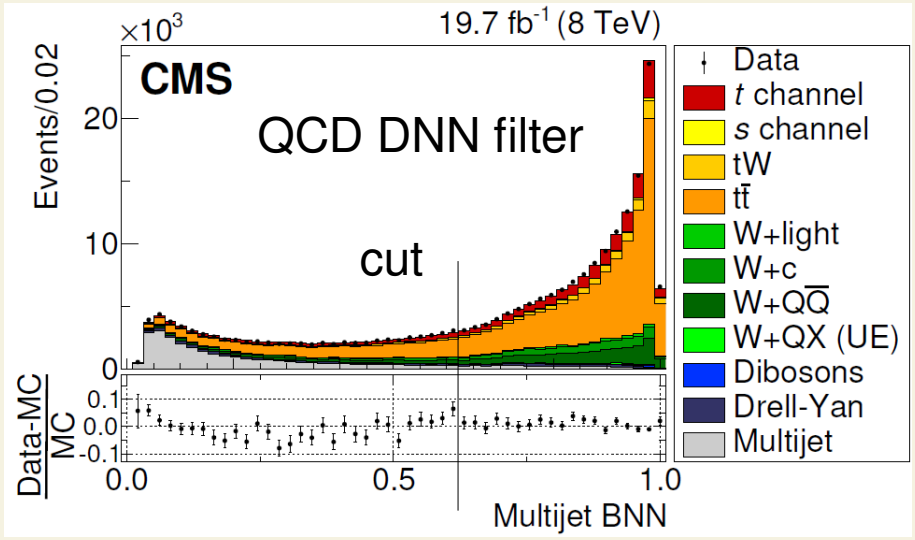
2) 14 TeV HL-LHC YR, PAS-FTR-18-004; extrapolation to HE-LHC (FCNC tqg)

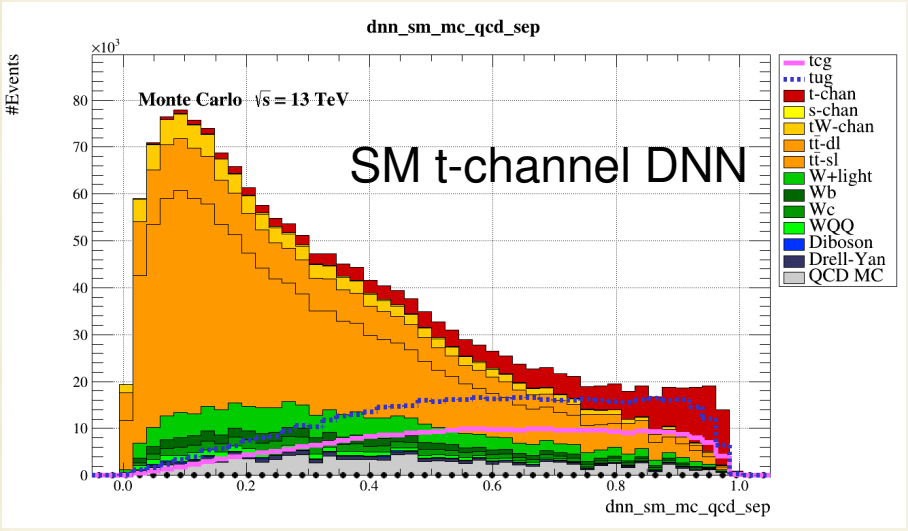
3) FCC 100 TeV, CDR vol.1 (FCNC tqg)

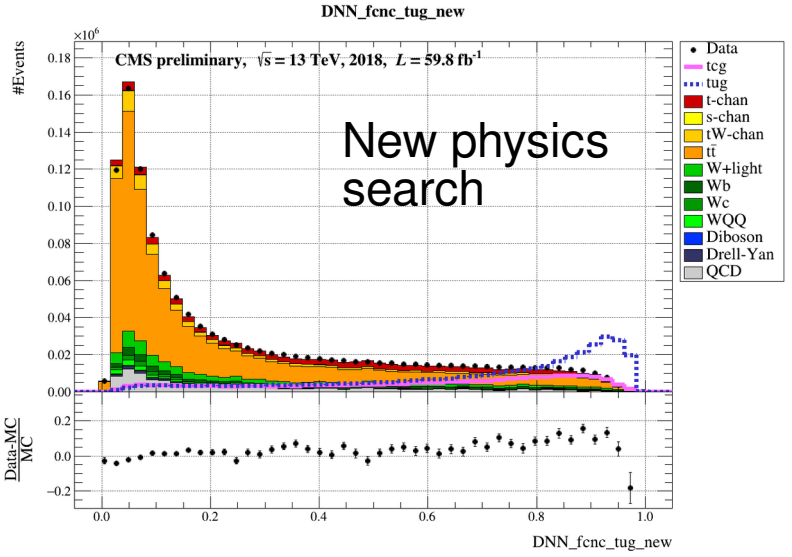
Possible SM measurements: t-channel total cross section, fiducial cross section, differential cross sections, $\mathbf{V_{tb}}$, R(t/t̄) measurements

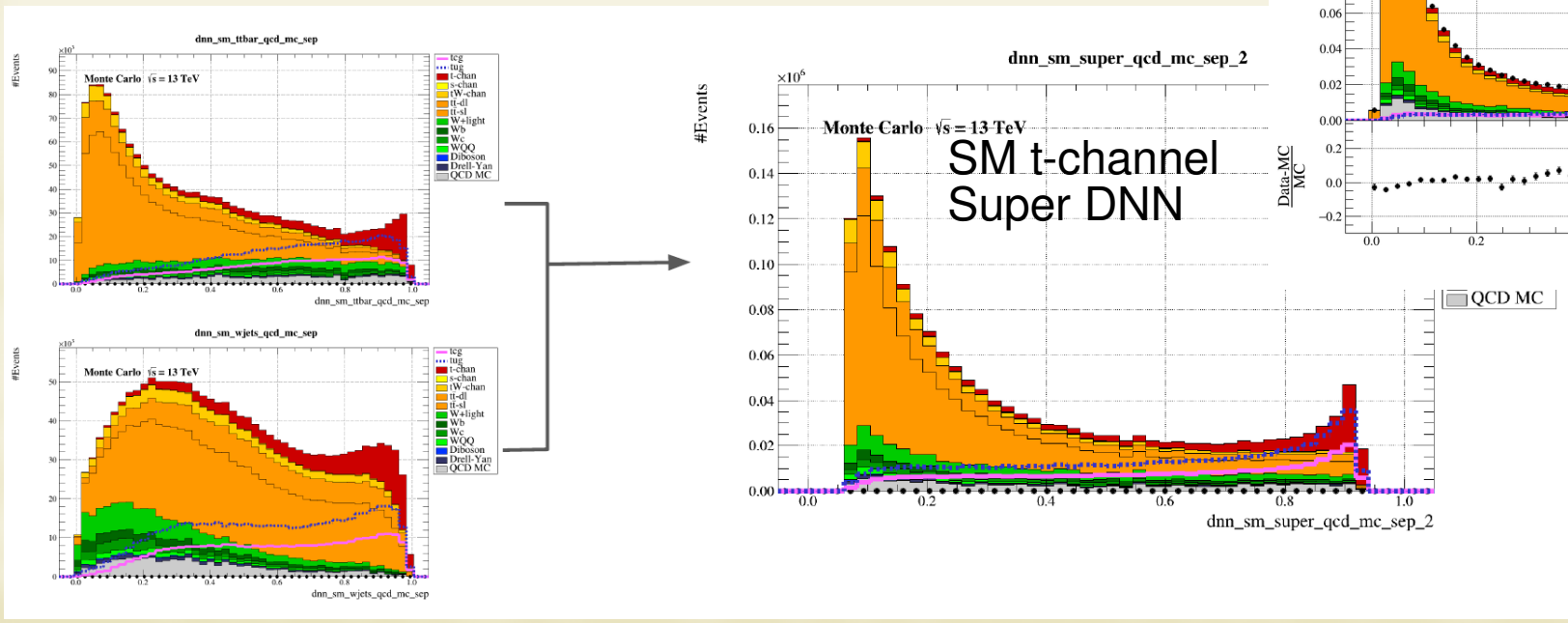
BSM measurements: FCNC tqg couplings, EFT Wilson coeff., ...

Real DNN examples in CMS





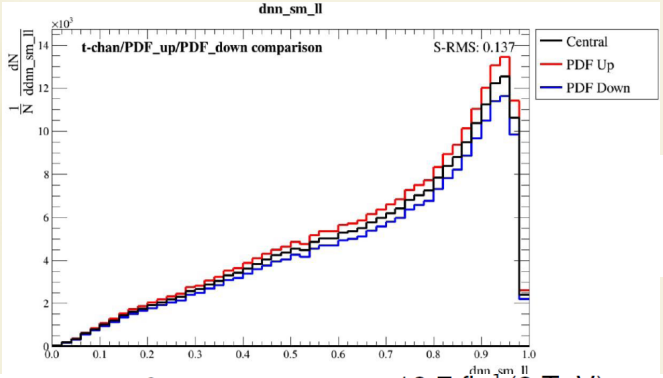




Uncertainties

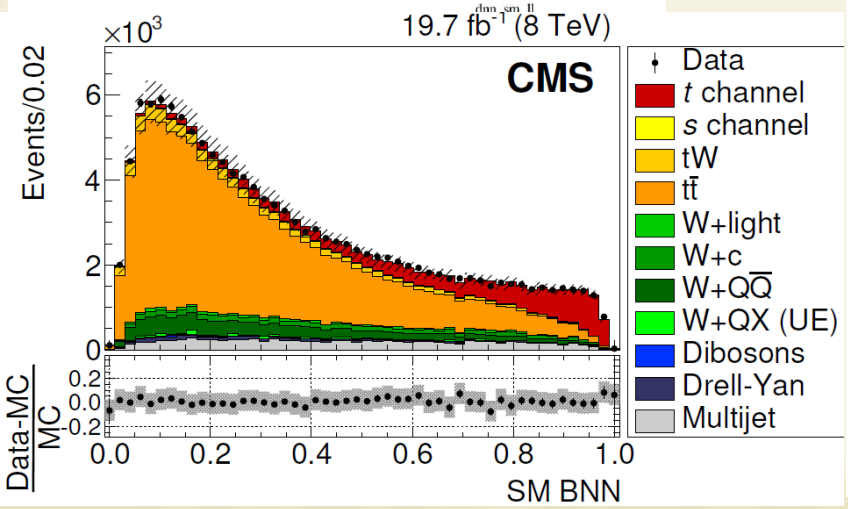
Aleatoric (statistical noise), Epistemic (systematical shift) uncertainties.

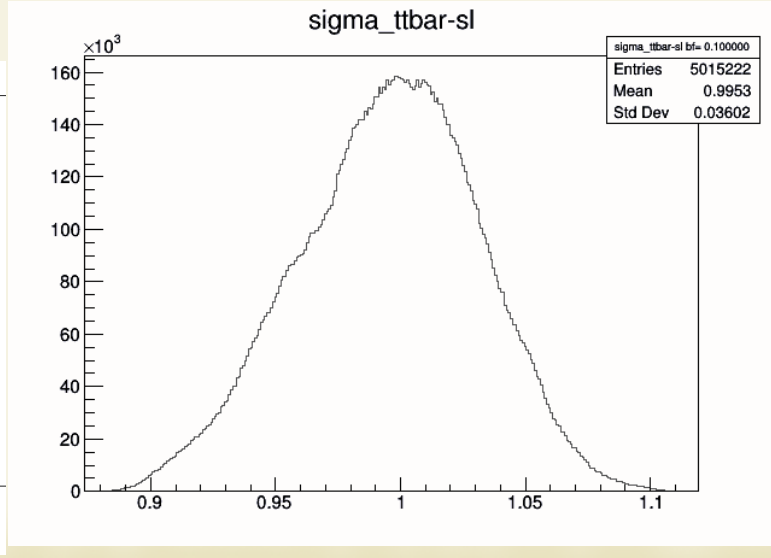
1. statistical uncertainty (data)
2. normalisation systematic uncertainties (cross sections, luminosity)
3. shape systematic uncertainties, correlated shift in all histogram bins (identification,corrections, …)
4. shape systematic uncertainties, uncorrelated shift between bins (scale, PDF, theory uncertainties)
5. finit statistics in Monte-Carlo generated event samples (Barlow-Beaston method)

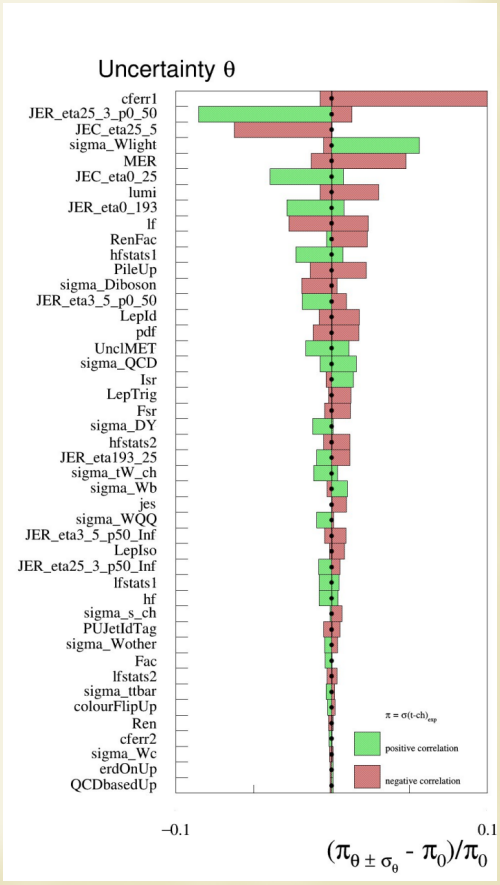


$$p_m(d|\vec{p}) = \prod_{i=1}^{N} \prod_{l=1}^{b_i} \text{Poisson}(d_{i,l}|m_{i,l}(\vec{p}))$$

$$p(\vec{\mu}_s|d) = \int p(d|\vec{\mu}_s, \vec{\mu}_b, \vec{\theta}) \frac{\pi(\vec{\mu}_s)\pi(\vec{\mu}_b)\pi(\vec{\theta})}{\pi(d)} \mathrm{d}\vec{\mu}_b \mathrm{d}\,\vec{\theta},$$



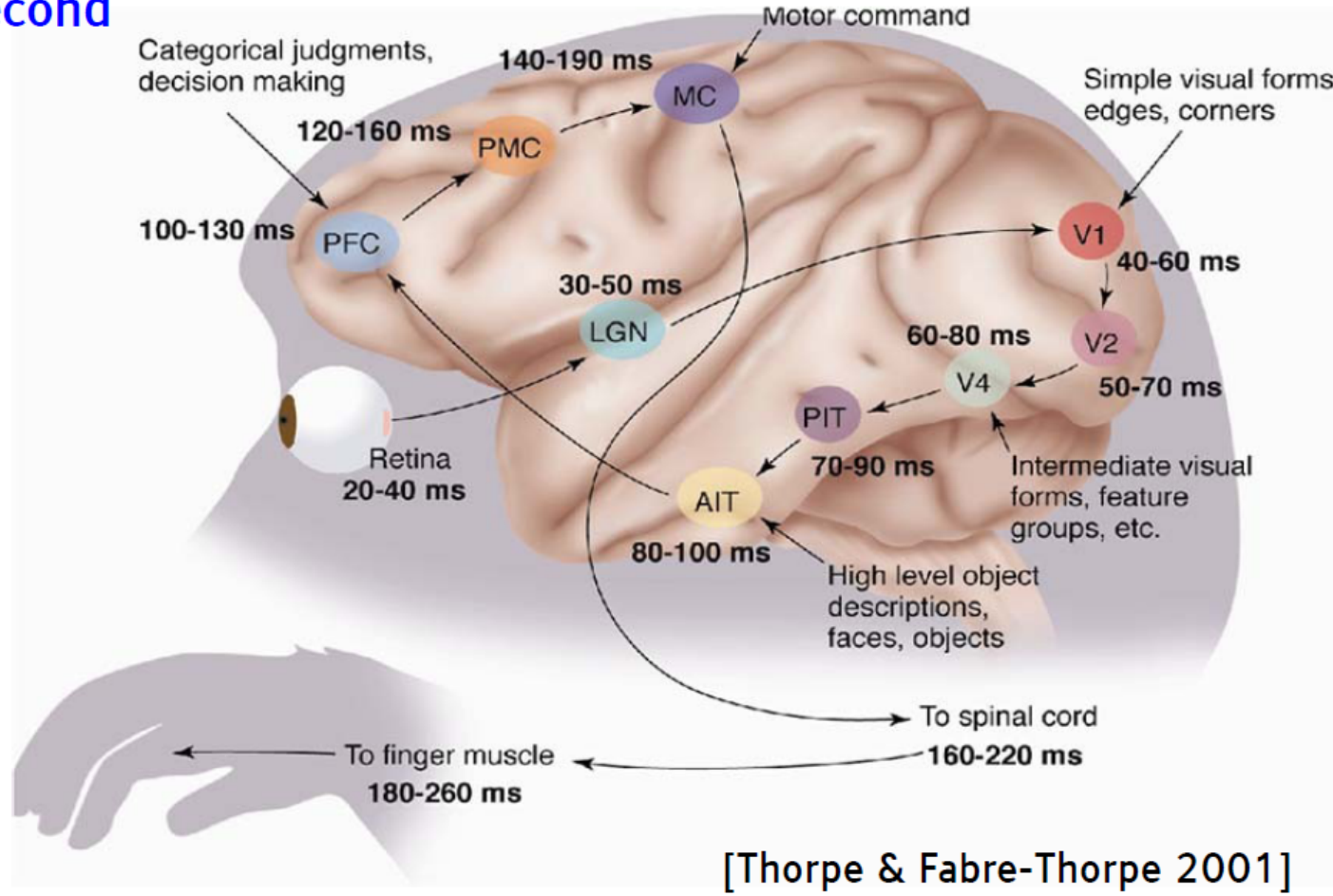




The Brain: an Amazingly Efficient "Computer"

- 10^{11} neurons, approximately
- 10^{4} synapses per neuron
- 10 "spikes" go through each synapse per second on average
- 10^{16} "operations" per second

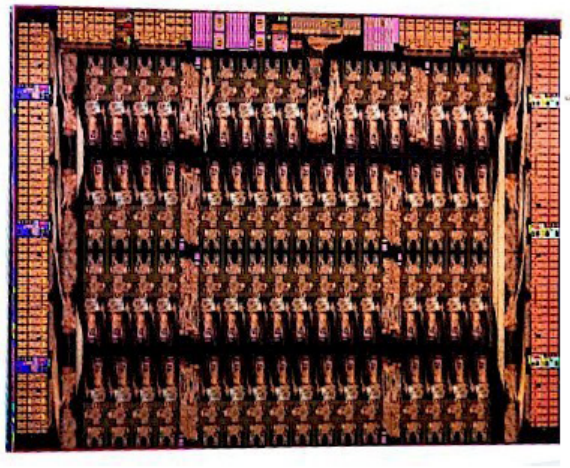
- 25 Watts
  - Very efficient
- 1.4 kg, 1.7 liters
- 2500 cm2
  - Unfolded cortex

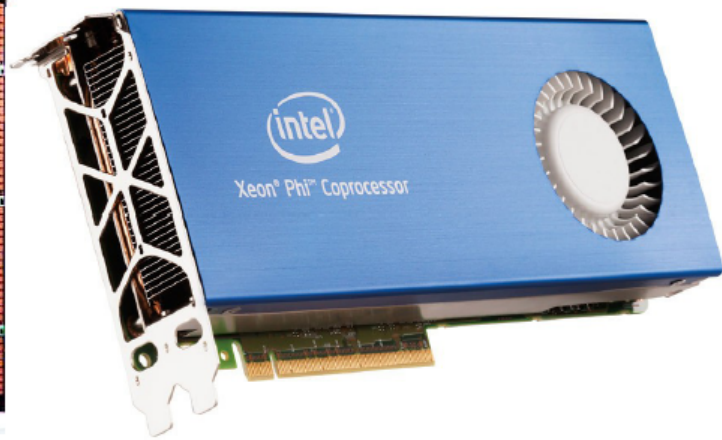


[Thorpe & Fabre-Thorpe 2001]

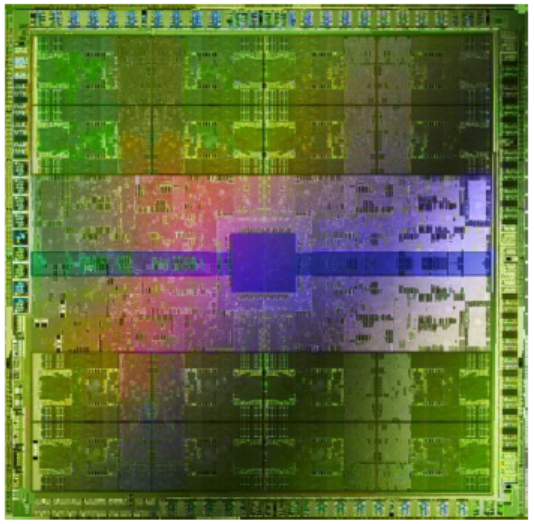
Fast Processors Today

- Intel Xeon Phi CPU
  - $2x10^{12}$ operations/second
  - 240 Watts
  - 60 (large) cores
  - $3000





- NVIDIA Titan-Z GPU
  - $8x10^{12}$ operations/second
  - 500 Watts
  - 5760 (small) cores
  - $3000



- Are we only a factor of 10,000 away from the power of the human brain?
  - Probably more like 1 million: synapses are complicated
  - A factor of 1 million is 30 years of Moore's Law
  - 2045?