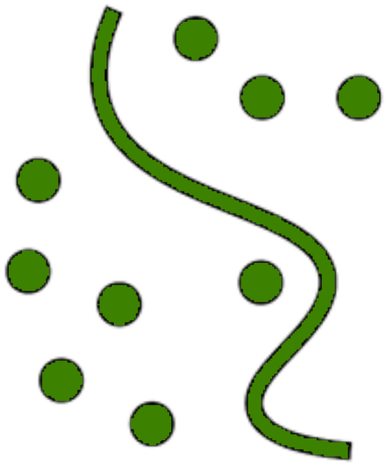
Argonne NATIONAL LABORATORY

# Transformation of collision data to rapidity-mass matrices for event classification using machine learning

S. Chekanov

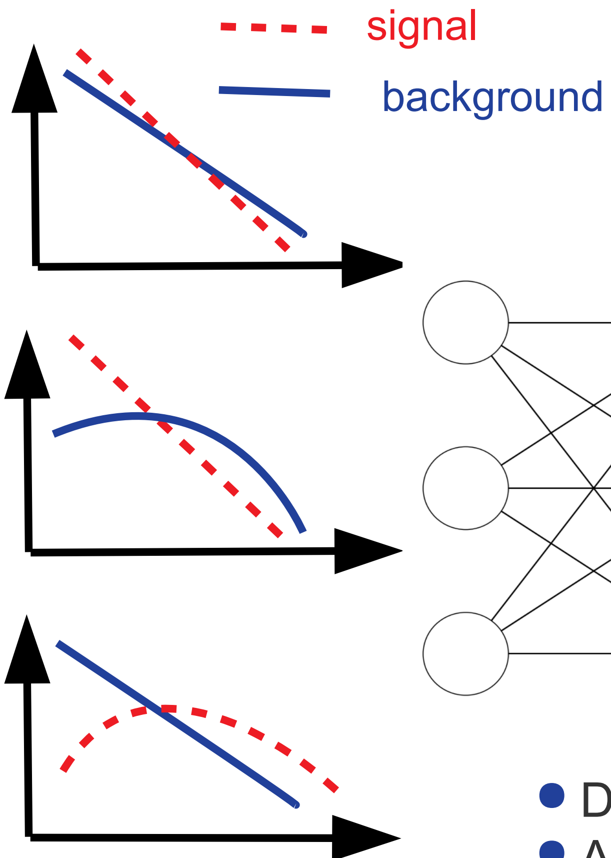
**ATLAS Machine Learning Workshop**

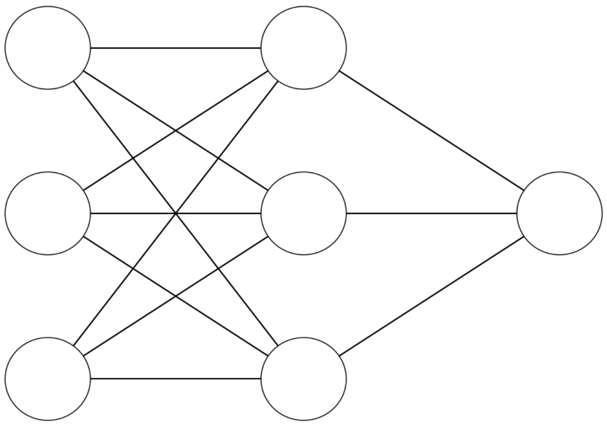
October 15-17, 2018



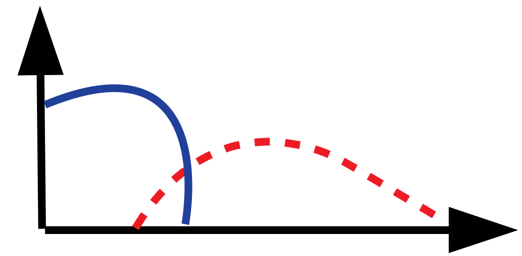
# Artificial Neural Networks (ANN) in HEP

Extensively used in HEP in the last ~25 years



Better separation of signal and background in ANN output space

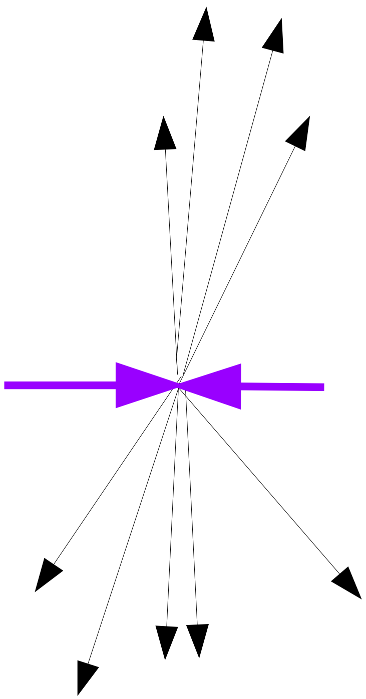


- Different studies require different feature space
- Ambiguous, reproducibility issues, time consuming

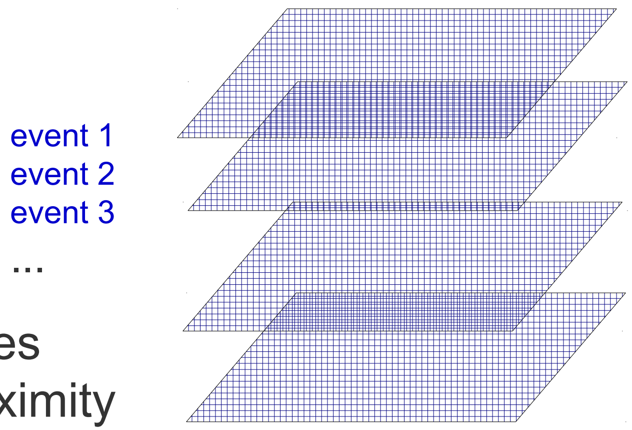
"feature space"

Can we find a "standard" feature space which is representative of many signatures used in BSM searches?

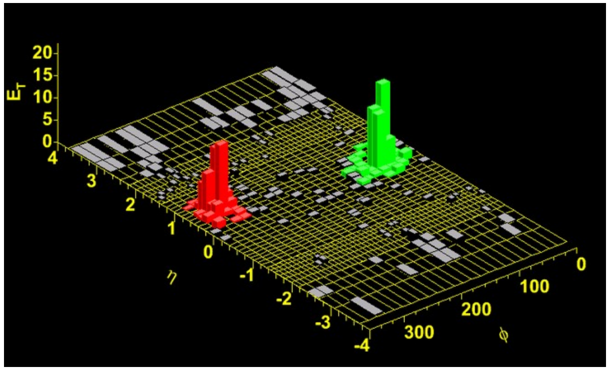
# Desirable requirements for ML feature space



- Fixed size arrays
- Dimensionless
- Lorentz invariant
- Fixed range of values
- Single and 2-particle densities
- Small correlations between variables
- Image like. Cells connected by proximity due to a well-defined hierarchy
- Easy to visualize for humans




**NOT GOOD for our goal**



# Rapidity-mass matrix (RMM)

https://arxiv.org/abs/1805.11650

Lorenz factors

Missing energy and missing transverse masses

$$
\begin{pmatrix}
\mathbf{e}_{\mathbf{T}}^{\mathbf{miss}} & m_T(j_1) & m_T(j_2) & \ldots m_T(j_N) & m_T(\mu_1) & m_T(\mu_2) & \ldots m_T(\mu_N) \\
h_L(j_1) & \mathbf{e_T}(\mathbf{j_1}) & m(j_1,j_2) & \ldots m(j_1,j_N) & m(j_1,\mu_1) & m(j_1,\mu_2) & \ldots m(j_1,\mu_N) \\
h_L(j_2) & h(j_1,j_2) & \delta\mathbf{e_T}(\mathbf{j_2}) & \ldots m(j_2,j_N) & m(j_2,\mu_1) & m(j_2,\mu_2) & \ldots m(j_2,\mu_N) \\
\ldots & \ldots & \ldots & \ldots; & \ldots & \ldots & \ldots \\
h_L(j_N) & h(j_1,j_N) & \ldots & \ldots\delta\mathbf{e_T}(\mathbf{j_N}) & m(j_N,\mu_1) & m(j_N,\mu_2) & \ldots m(j_N,\mu_N) \\
h_L(\mu_1) & h(\mu_1,j_1) & h(\mu_1,j_2) & \ldots h(\mu_1,j_N) & \mathbf{e_T}(\mu_\mathbf{1}) & m(\mu_1,\mu_2) & m(\mu_1,\mu_N) \\
h_L(\mu_2) & h(\mu_2,j_1) & h(\mu_1,j_2) & \ldots h(\mu_2,j_N) & h(\mu_1,\mu_2) & \delta\mathbf{e_T}(\mu_\mathbf{2}) & m(\mu_2,\mu_N) \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\
h_L(\mu_N) & h(\mu_N,j_1) & h(\mu_N,j_2) & \ldots h(\mu_N,j_N) & h(\mu_N,\mu_1) & h(\mu_N,\mu_2) & \delta\mathbf{e_T}(\mu_\mathbf{N})
\end{pmatrix}
$$

Invariant masses of pairs $m(i,j)/E_{CM}$

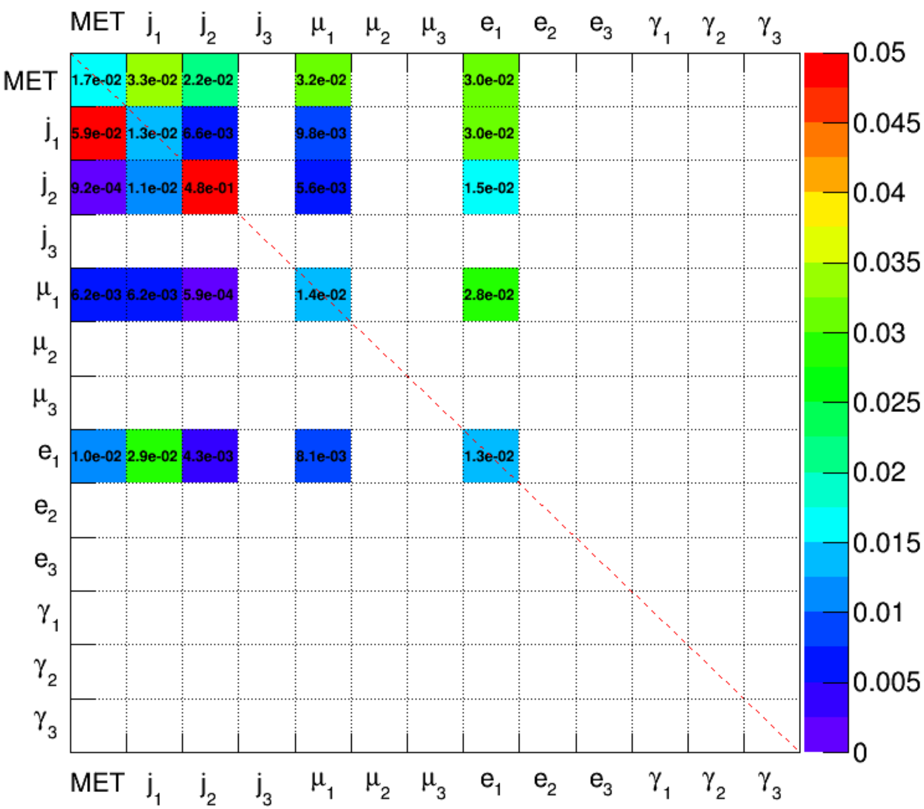
Transverse energies of leading objects and energy imbalances

Rapidity difference: $h(i,j) \sim \cosh(y_j - y_i) - 1$

- **Dimensionless, Lorentz invariant (1st column are Lorentz factors themselves)**
- **Single and two-particle densities for each identified jet/objects**
  - Covers many aspects of invariant masses, forward physics, DM searches etc.
- **Cells are almost independent for SM processes (*)**
- **Re-scaling and normalization by construction**
- **Fixed sizes with well-defined mapping to input nodes → "Natural language" for ANN**
- **Cells connected by proximity → good for visualization**

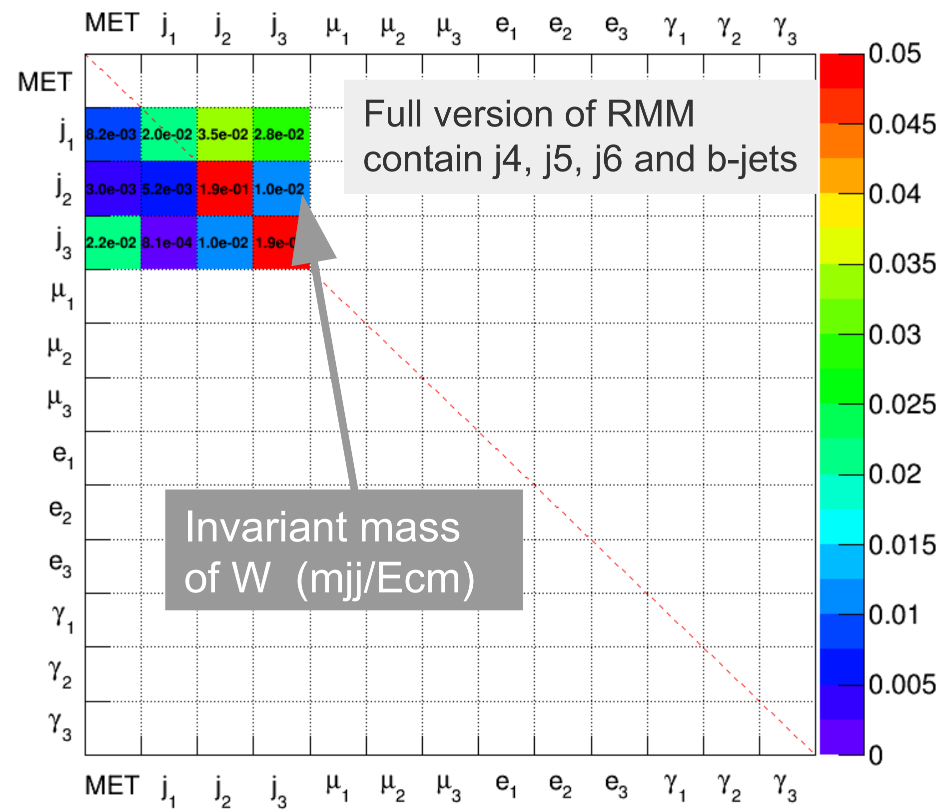
# Example: Two PYTHIA8 events with $t\bar{t}$

$t\bar{t} \rightarrow Wb\,W\bar{b} \rightarrow e\ \nu\ b\ \mu\ \nu\ \bar{b}$




**Cell with MET, μ and e leptons activated**

$t\bar{t} \rightarrow Wb\,W\bar{b} \rightarrow$ 6 jets

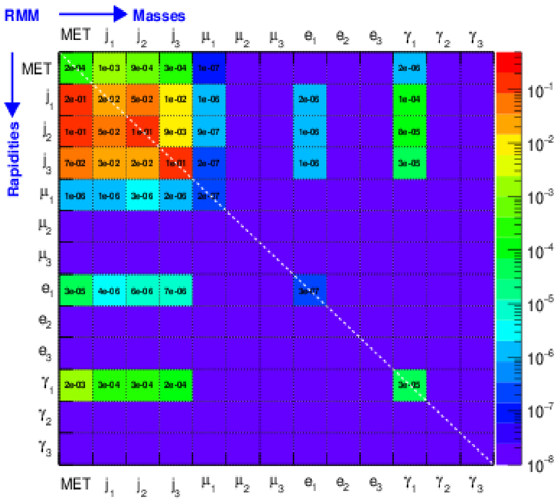



**Many jets, no MET and leptons**

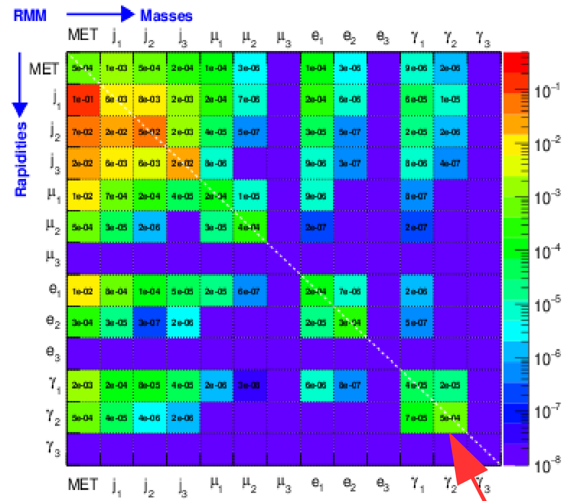
**Each cell maps to an input neuron: Use ANN for image identification from leading industries (or even simple backpropogation or BDT)**

# Visualization of the RMM feature space

https://arxiv.org/abs/1805.11650




(a)multijets QCD

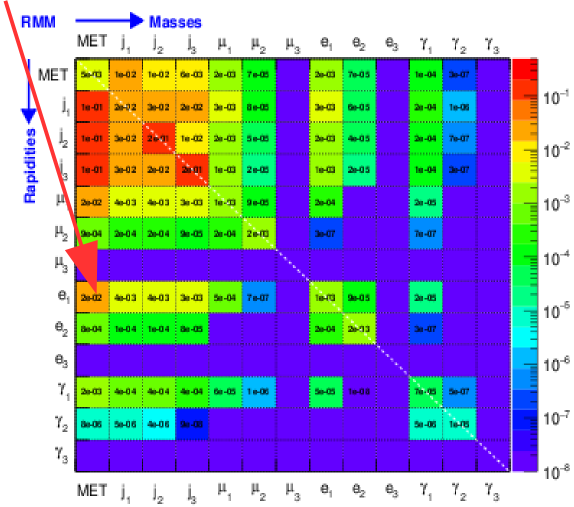



(b)Higgs processes

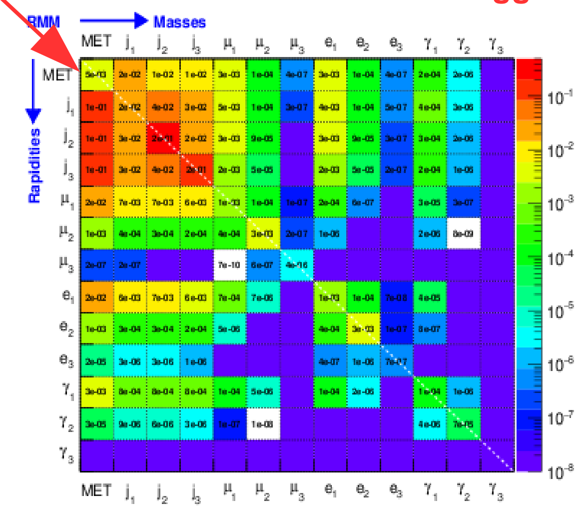
Muons

large MET

Higgs mass (γγ)




(c)Top production




(d)$H^+t$ production

**Average RMM for PYTHIA8:**

- Multijet QCD
- SM Higgs production
- Top production
- H+t production

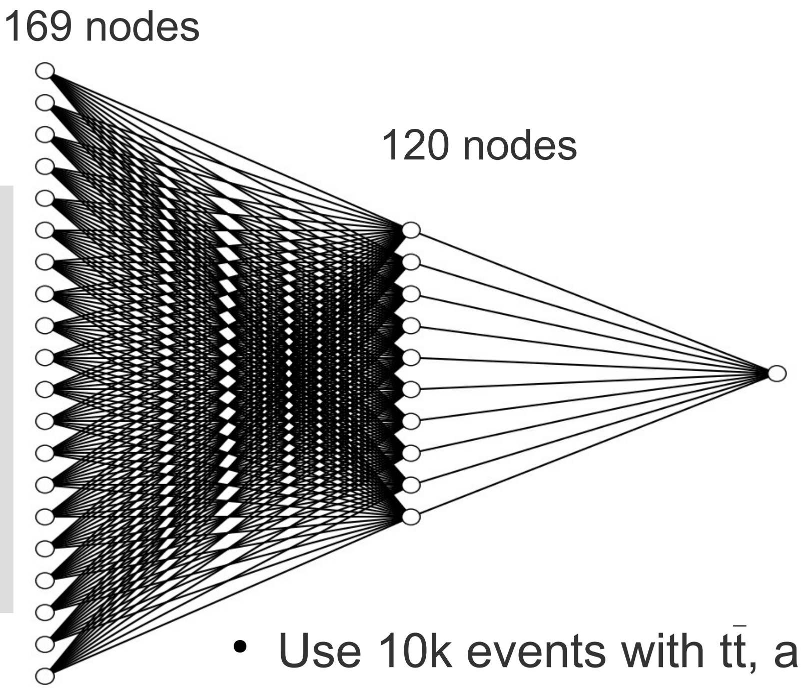
All allowed decays of W/H/t
Averaged over 50k events
(for each process)

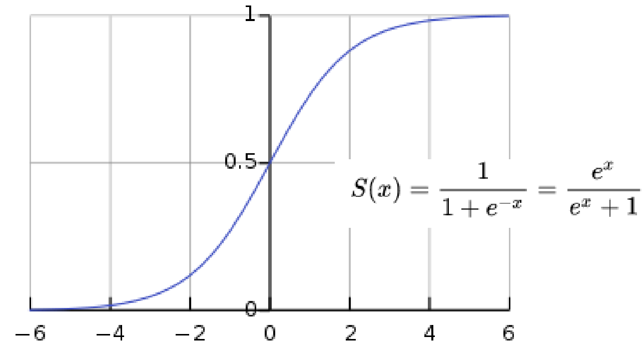
**Considered:**
- jets, mu, e, photons
- up to 3 objects

- **$t\bar{t}$ and H+t are similar**
- **Apply RMM to identify H+t**

# Using RMM for Charged Higgs searches

10k Pythia8 events used to create 10k RMM (13x13) for H+ and ttbar processes

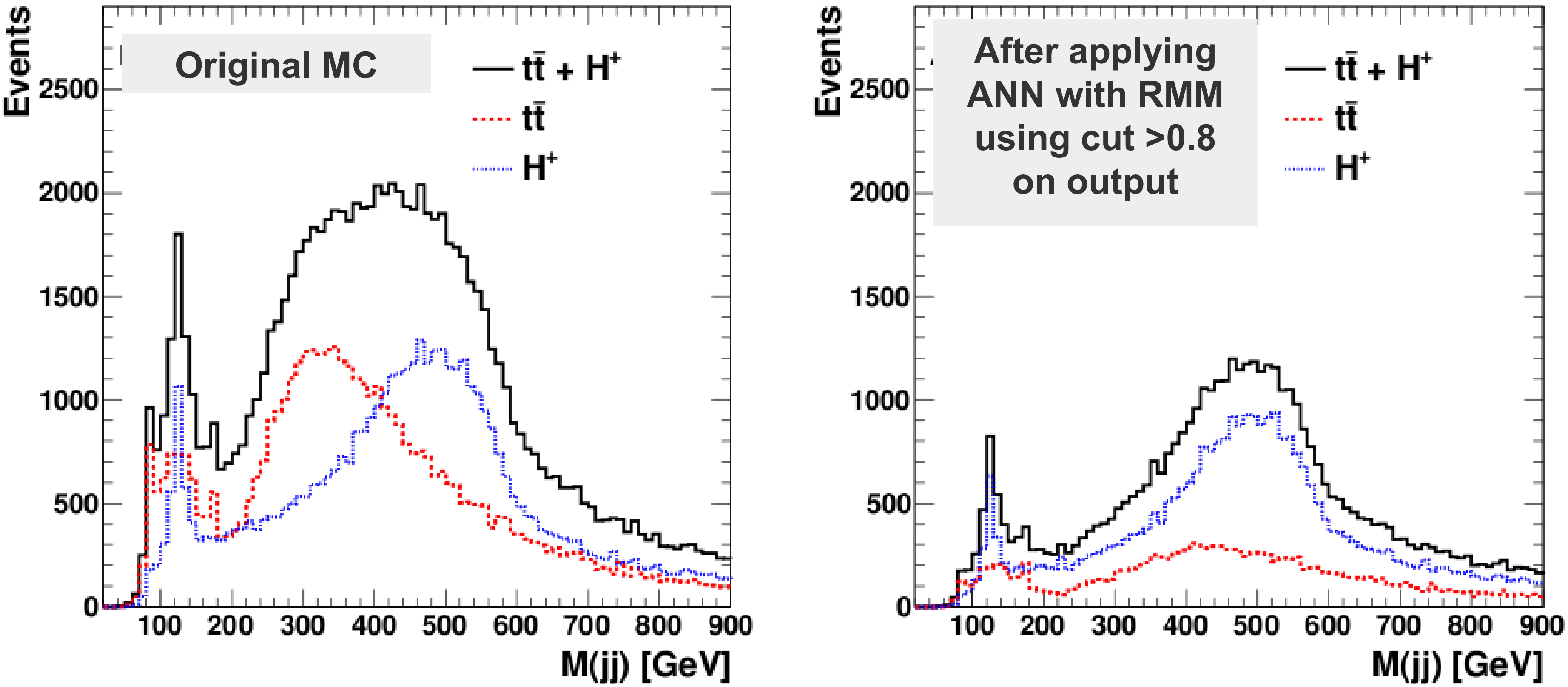



output: 0 ($t\bar{t}$) or 1 (H+)

- Use 10k events with $t\bar{t}$, and 10k with H+
- Assume 600 GeV mass for H+
- Create cross validation sample for ANN
- Stop training when MSE < than for cross validated ANN
- Compare M(jj)  (RMM cell (2,1)) for H+ and top processes before after applying cut > 0.8 on output
- Disable cell (2,1) during training (avoid Mjj biases!)

# Separation of H+ from $t\bar{t}$ background before and after ANN



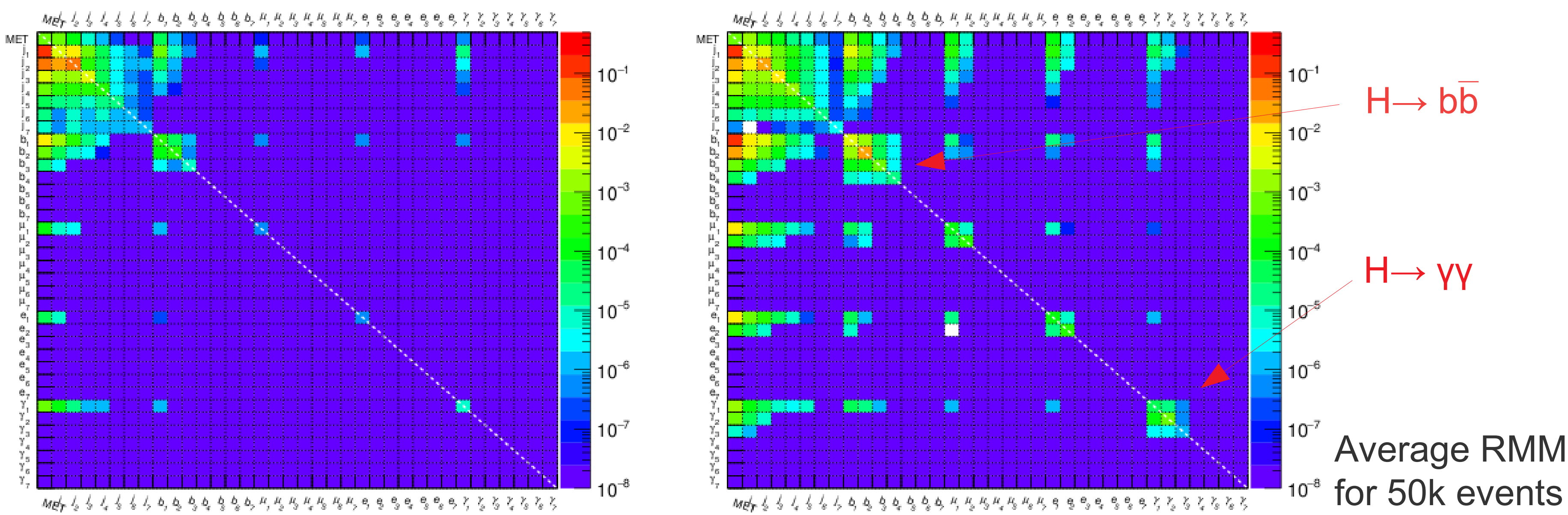

- H+ mass at 600 GeV. Look at invariant mass of 2 leading jets ((2,1) cell)
- ANN with RMM inputs increases the S/B by a factor 3.
  - Signal efficiency reduced by 30%
- Small shift for $t\bar{t}$ (may require better tuning of disabled RMM links)

# RMM for general event identification problem

- RMM includes all single & two-particle (jet) densities
- No “handpicking” input variables for every topology/decay
- Good choice for general event classifiers?

**Example:**

- 5 processes: **(1) SM QCD (2) Higgs (3) H+ (4) ttbar (5) Double bosons**
- Create RMM using Np=7 and 6 objects using b-jets

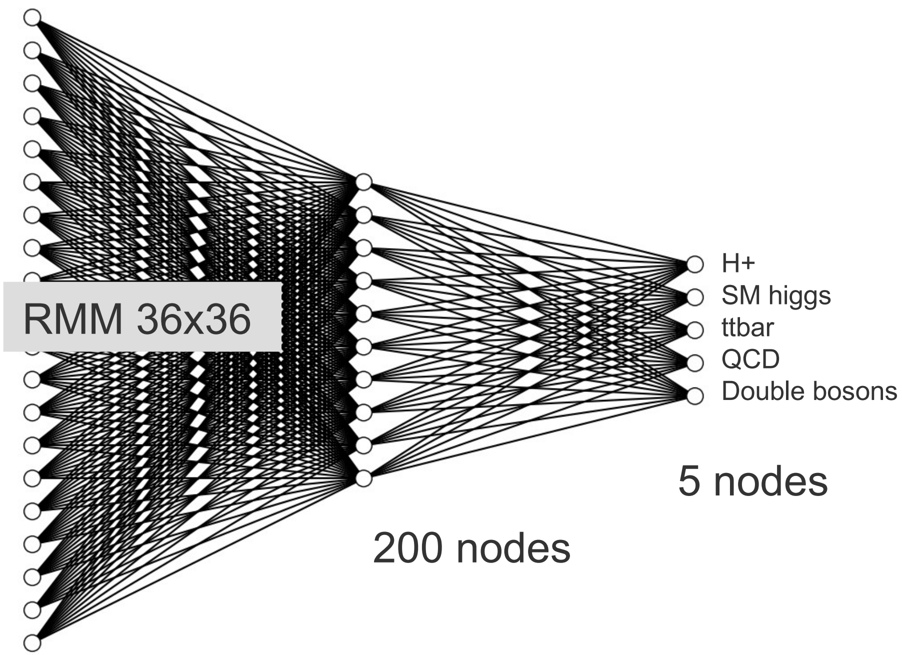



Multi-jet QCD

Higgs productions (all decays)

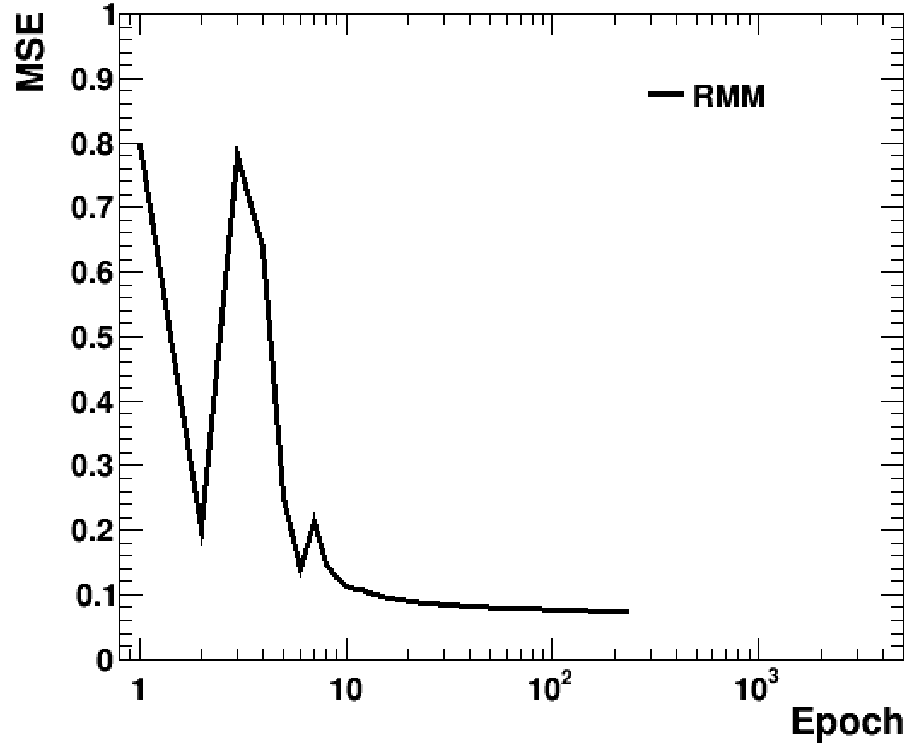
# ANN training using RMM as input

Backpropogation NN with Signoid function, 5 outputs for each process (0-1 values)




1296 nodes

**Wide and shallow ANN for sparse input RMM data**

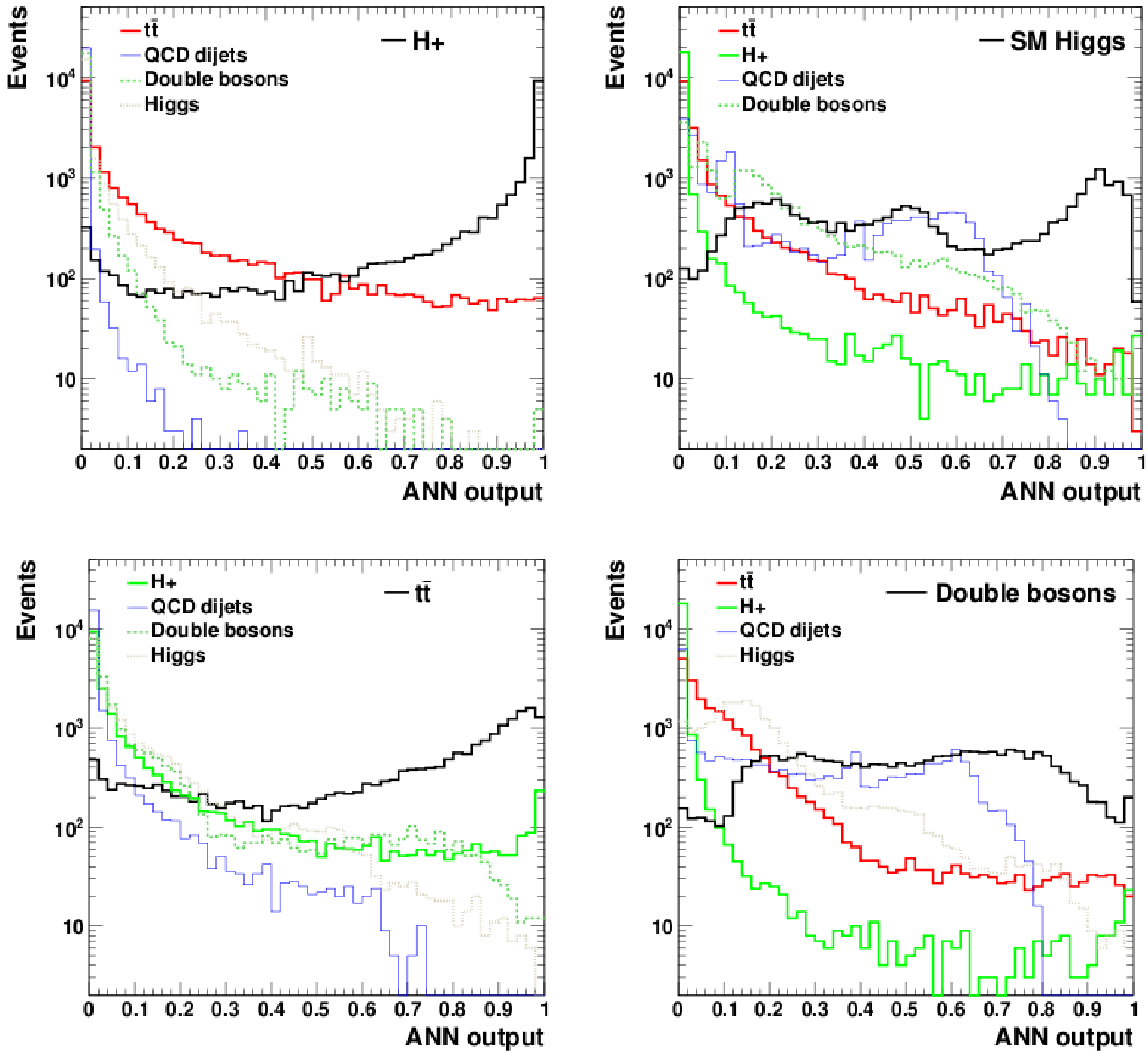



<u>Well trained after 200 epochs:</u>
Mean Squared Error (MSE) decreases from 0.8 to 0.07 (~ 1h training on a desktop for 200k RMM)

# Result of ANN training using RMM

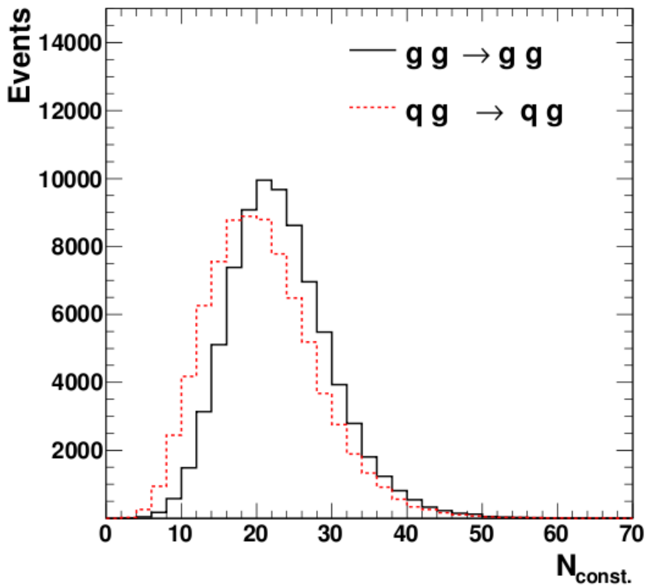
Good event separation of signal events (black lines) from other processes

Purity of event classification is 80%-90% assuming 0.8 cut on output nodes (see backup slide 22)



# Challenging case: QCD dijets

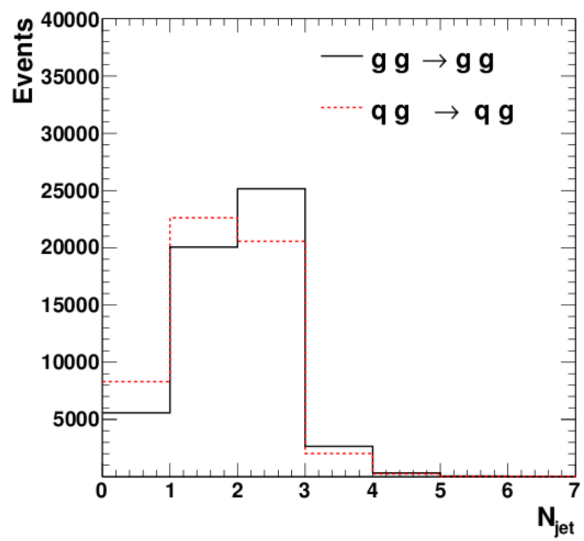
Separate **gg** from **qg** final states (dijets) → Distributions are nearly identical.
Presence of **g** instead of **q** leads to broader jets and changes in jet kinematics / shape

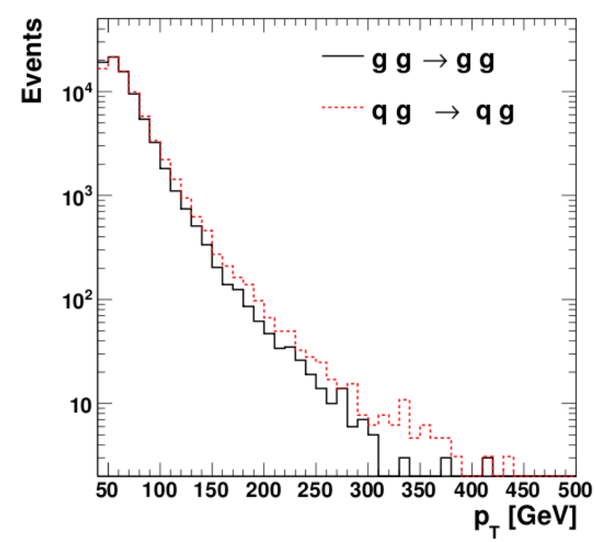



Well-known difference: Number of jet constituents is larger for gluon jets than for quark jets due to difference in color factors ($C_A$ =3 vs $C_F$ = 3/4)

But there are many other distributions that can be used for ANN. How to choose them?

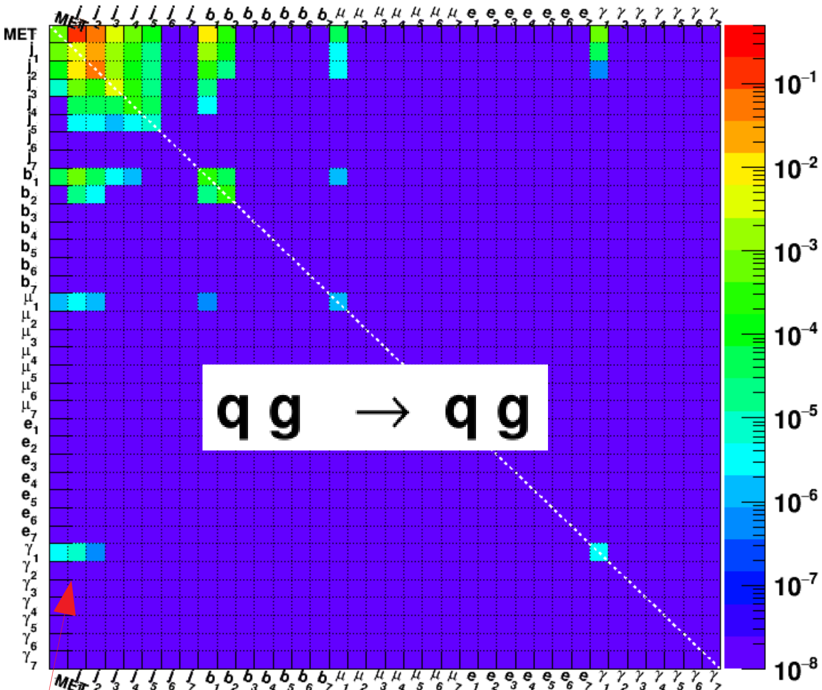
Use hand-crafted variables using Pick-and-Use approach?

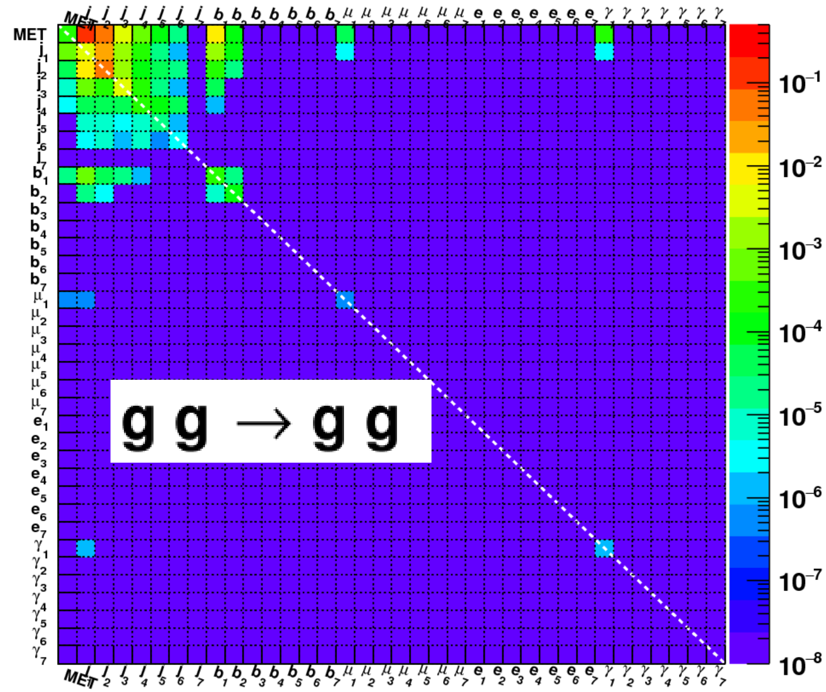


# RMM for gg and qg events (example)

Average RMM
for 100k events




photons

**gg** process compared to **qg** has:

- softer pT
- more jets
- reduced photon rate

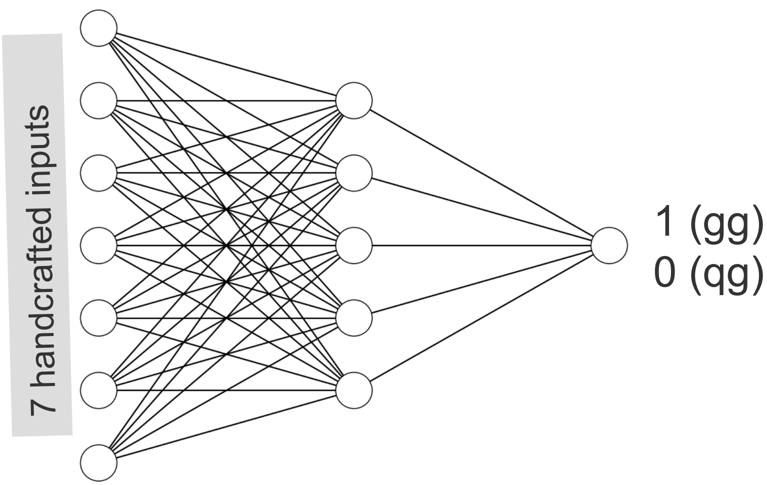
..

# gg and qg separation: PaU vs standard RMM

Two approaches for ML:

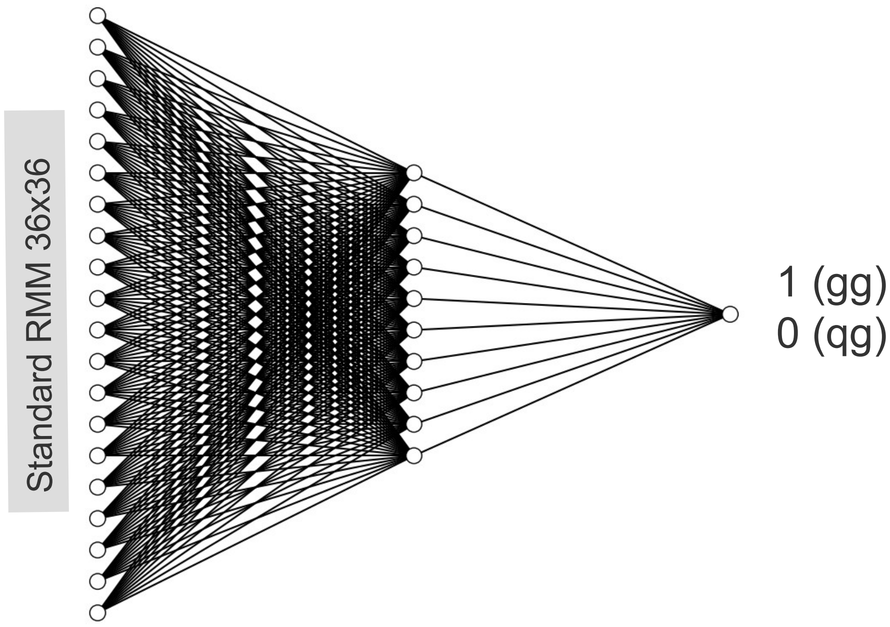
## Traditional PaU

- handcrafted input variables (7 nodes)
- hidden layer (5 nodes)
- output with 1 (gg) or 0 (qg)



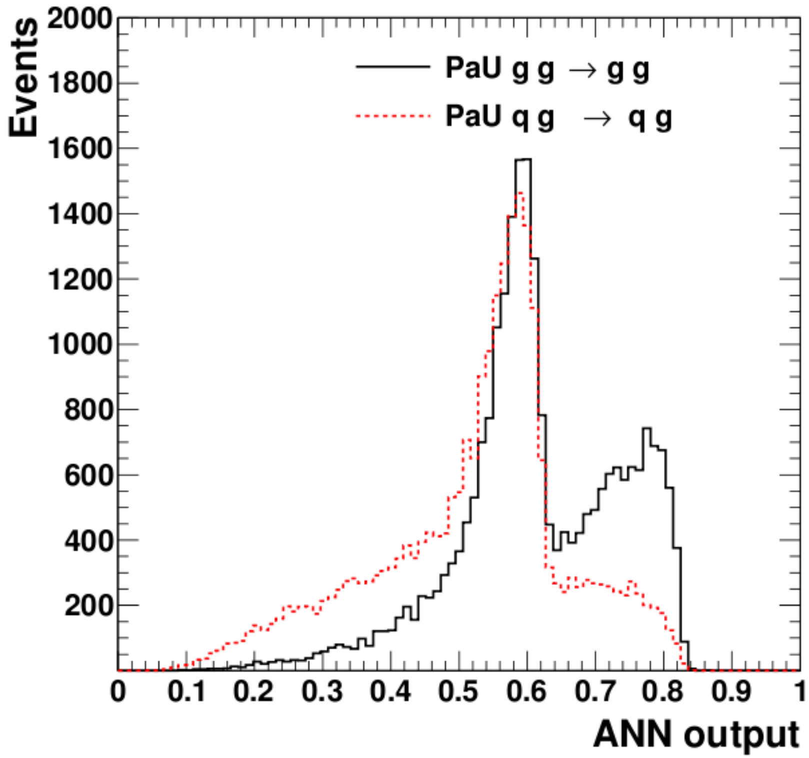

## RMM

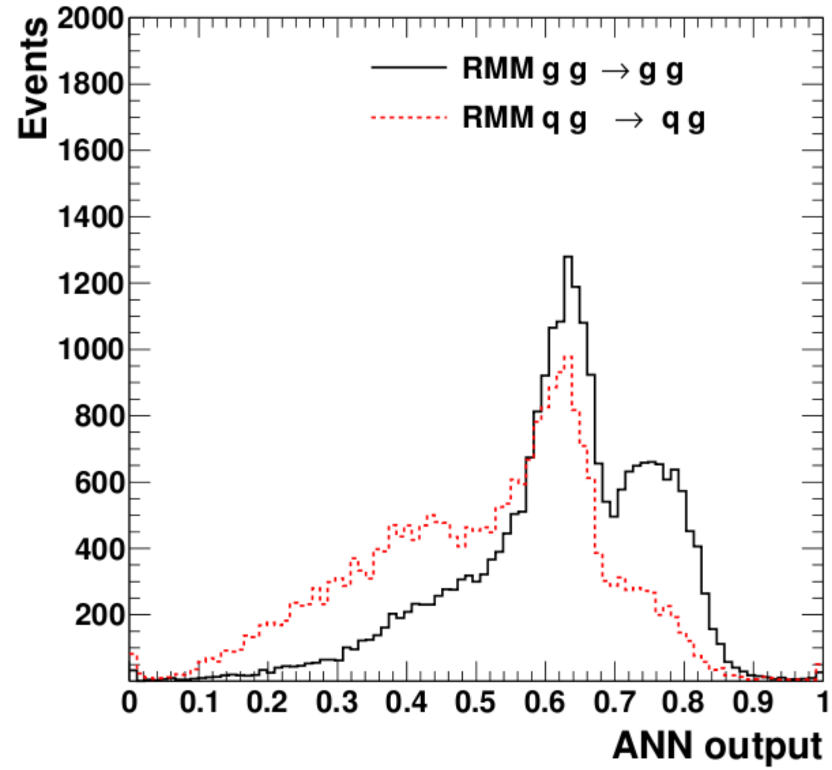
- RMM matrix as input (36x36+2)
- hidden layer (200 nodes)
- output with 1 (gg) or 0 (qg)




Alternatively: Boosted Decision tree (BDT) using PaU and RMM
100 trees, depth 7, stochastic gradient (arXiv:1609.06119)

# gg and qg separation: PaU vs standard RMM




- ANN output space shows separation of **gg** from **qg**
- RMM over-performs hand-crafted “pick-and-use” (PaU) method with 7 inputs
  - RMM has separation purity 68% vs 65% for PaU assuming ANN output cut 0.5
- BDT instead of backpropogation confirms this conclusion

# Conclusions

- RMM is well suited for general event classification problems due comprehensive (nearly independent) single and two-particle densities
  - Works even for simplest ANN/BDT
  - Requires a wide input if no pruning of RMM input is done

- Same RMM transformation can be plugged into different BSM searches to produce good results with minimal tweaking
  - Unless you care about jet substructure which are not covered by RMM

- RMM can identify events with rather unexpected features without much thinking about ML inputs
  - Different decay channels (and their kinematics) are taken into account automatically

- Will be applied to ATLAS searches for H+t in  dijet+lepton analysis using Run II data

# Backup

# Feature space for event classifications

- **Event classification depends on prepared inputs**
  - Identify variables with background and signal “features”
  - Data and dimensionality reduction
  - Data re-scale (the range between 0 and 1 is a popular choice),
  - Data normalization (to avoid cases when some of input values overweight others)
  - etc.
- **ANN are suppose to simplify analysis  but:**
  - Preparing analysis for NN  is time consuming
  - Need to hand-pick variables, study them etc.. No uniqueness of input variables.

- **Idea: create a general image-like transformation of lists with 4-momenta to data structures that reflect most significant features of hadronic-final state**
  - General representation of collision event. Single and double- particle densities
  - Natural language for machine learning → leverage algorithms from leading industries
  - Easy to visualize for humans
  - Leverage algorithms for image identification from  leading industries

# Rapidity-mass matrix (RMM)

https://arxiv.org/abs/1805.11650

jets — muons — .. electrons, photons

$$
\begin{pmatrix}
\mathbf{e}_{\mathbf{T}}^{\mathbf{miss}} & m_T(j_1) & m_T(j_2) & \ldots m_T(j_N) & m_T(\mu_1) & m_T(\mu_2) & \ldots m_T(\mu_N) \\
h_L(j_1) & \mathbf{e_T(j_1)} & m(j_1,j_2) & \ldots m(j_1,j_N) & m(j_1,\mu_1) & m(j_1,\mu_2) & \ldots m(j_1,\mu_N) \\
h_L(j_2) & h(j_1,j_2) & \delta\mathbf{e_T(j_2)} & \ldots m(j_2,j_N) & m(j_2,\mu_1) & m(j_2,\mu_2) & \ldots m(j_2,\mu_N) \\
\ldots & \ldots & \ldots & \ldots, & \ldots & \ldots & \ldots \\
h_L(j_N) & h(j_1,j_N) & \ldots & \ldots\delta\mathbf{e_T(j_N)} & m(j_N,\mu_1) & m(j_N,\mu_2) & \ldots m(j_N,\mu_N) \\
h_L(\mu_1) & h(\mu_1,j_1) & h(\mu_1,j_2) & \ldots h(\mu_1,j_N) & \mathbf{e_T}(\mu_1) & m(\mu_1,\mu_2) & m(\mu_1,\mu_N) \\
h_L(\mu_2) & h(\mu_2,j_1) & h(\mu_1,j_2) & \ldots h(\mu_2,j_N) & h(\mu_1,\mu_2) & \delta\mathbf{e_T}(\mu_2) & m(\mu_2,\mu_N) \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\
h_L(\mu_N) & h(\mu_N,j_1) & h(\mu_N,j_2) & \ldots h(\mu_N,j_N) & h(\mu_N,\mu_1) & h(\mu_N,\mu_2) & \delta\mathbf{e_T}(\mu_N)
\end{pmatrix}
$$

$e_T^{miss}$ – missing ET of events

$m_T(i)$ - transverse mass of object “i”

$e_T(i)$ - transverse energy (ordered)

$\delta e_T(i)$ – transverse energy imbalances

$m(i,j)$ – two-particle invariant masses

scaled by $1/\sqrt{s}$

$h_L(i)$ - cosh(y)-1 (y is rapidity) – Lorentz factor

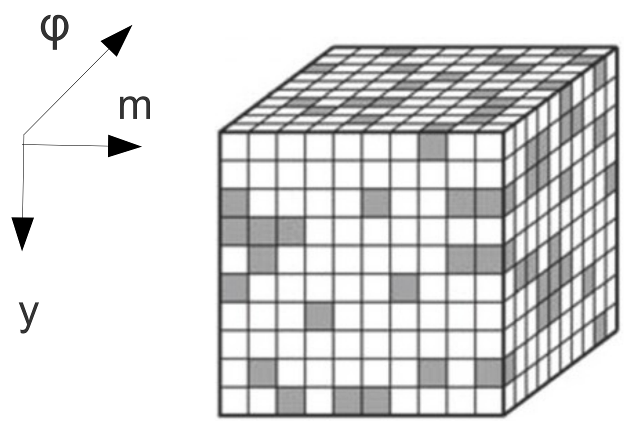
$h(i,j)$ - $\cosh(0.5(y_i - y_j)) - 1$ – rapidity difference

scaled by a constant

What does this matrix represent?

# Extending RMM

- RMM includes information on single and two-particle densities
  - but no phi due to rotational symmetry
- Can be extended to 3D matrices to include φ,  3-particle densities etc.




**Plus:**

- Add tau, leptons with + and – charges (separately), b-jets
- Increase multiplicity of each object to ~10-20 (empty cells are not stored)
- Add more complex (and well reconstructed) types: J/Phi, W, Z, Higgs

# Monte Carlo simulations

## Several processes from Pythia8 (LO+PS)

- **Dijet QCD:**
  - All 2→2 processes (10)
- **Top production:**
  - g g -> t tbar
  - q qbar -> t tbar
- **Charged Higgs production**
  - b g -> H+- t
- **Double boson production**
  - f fbar -> gamma*/Z0 gamma*/Z0
  - f fbar' -> Z0 W+-
  - f fbar -> W+ W-
- **SM Higgs production**

http://atlaswww.hep.anl.gov/hepsim/


Show all

p →← p

8 TeV

13 TeV

14 TeV

27 TeV

33 TeV

100 TeV

e+ →← e−

250 GeV

380 GeV

500 GeV

1 TeV

3 TeV

μ+ →← μ−

1 TeV

Get involved | Full Search | Experiments | Manual | Mirrors | Tools | About | Login

HEP.ANL.GOV

HepSim

Repository with Monte Carlo simulations for particle physics

- March 15 2018: Charged Higgs event samples
- Sep,22 2017: Z+Higgs → nunu+XX event samples
- Sep,15 2017: Higgs → mu+mu- event samples

Show 25 entries

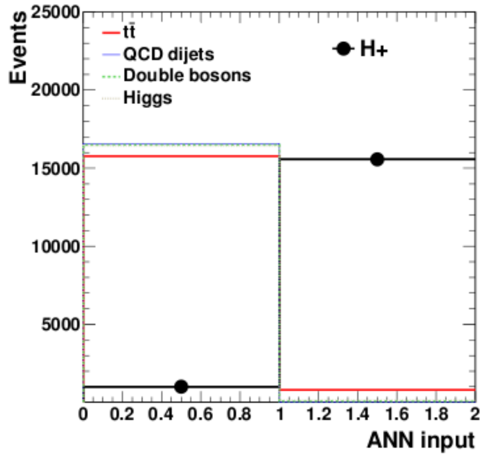
Previous 1 2 3 4 5 ... 13 Next

Search:

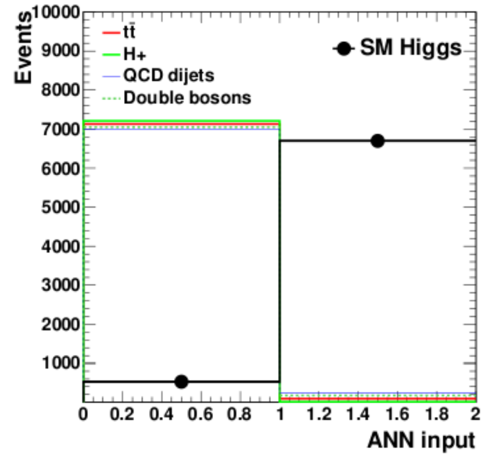
| Id | →← | E [TeV] | Dataset name | Generator | Process | Topic | Files | Created |
|---|---|---|---|---|---|---|---|---|
| 328 | pp | 13 | tev13pp_pythia8_rmm | PYTHIA8 | Various SM/BSM process for ML | SM | Info | 2018/09/16 |
| 327 | pp | 13 | tev13pp_qcd_pythia8_proio | PYTHIA8 | QCD dijets for ProIO tests | SM | Info | 2018/08/27 |
| 326 | pp | 13 | tev13pp_qcd_pythia8_proio_tests | PYTHIA8 | QCD dijets for tests of ProIO | SM | Info | 2018/08/20 |
| 325 | e-p | 0.035 | gev35ep_pythia8_dis1q2ct14lo | PYTHIA8 | DIS events at Q2>1 GeV2 | SM | Info | 2018/07/25 |
| 323 | pp | 13 | tev13pp_mg5_chaHT_tbeta_hw | MADGRAPH/PY8 | H- top with H- to HW and tan(beta)=1-7 | Exotics | Info | 2018/06/13 |
| 322 | pp | 13 | tev13pp_mg5_chaHT_tbeta_tb | MADGRAPH/PY8 | H- top with H- to tb and tan(beta)=1-7 | Exotics | Info | 2018/06/13 |
| 321 | pp | 13 | tev13pp_mg5_chaHW_tbeta_tb | MADGRAPH/PY8 | H+ W- with H+ decay to t-bbar tan(beta)=1-7 | Exotics | Info | 2018/06/06 |
| 320 | pp | 13 | tev13pp_mg5_chaHW_tbeta_hw | MADGRAPH/PY8 | H+ W- with H+ decay to HW for tan(beta)=1-7 | Exotics | Info | 2018/06/06 |
| 318 | pp | 13 | tev13pp_pythia8_gamgam | PYTHIA8 | Higgs to gamma gamma | SM | Info | 2018/04/20 |

**All LO processes and all top/W/H decays enabled**

# Results of the ANN training using RMM

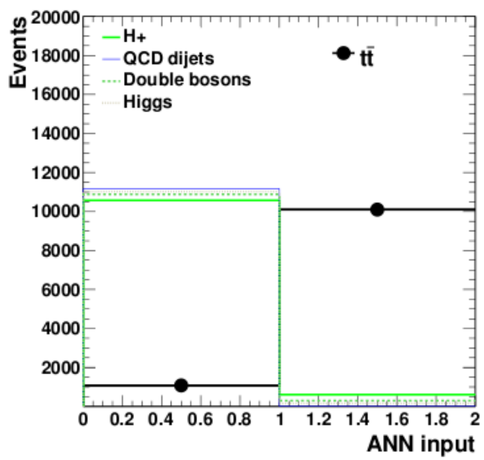



(a)Charged H+

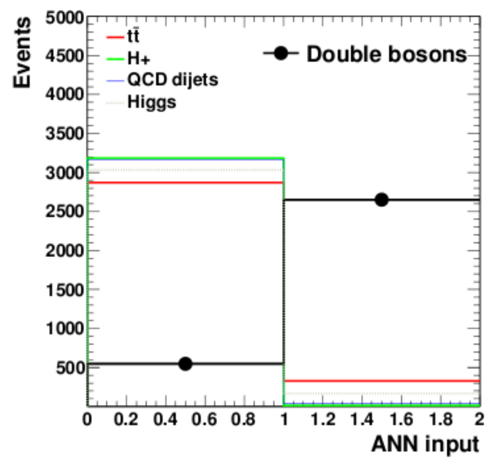



(b)SM Higgs

ANN output > 0.8




(c)$t\bar{t}$ production




(d)Double $W/Z$ production