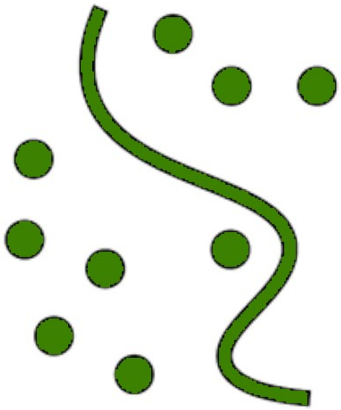
Argonne NATIONAL LABORATORY

Machine learning and anomaly detection using rapidity-mass matrices

S. Chekanov (ANL)

ISMD2021

50th International Symposium on Multiparticle Dynamics (ISMD2021)

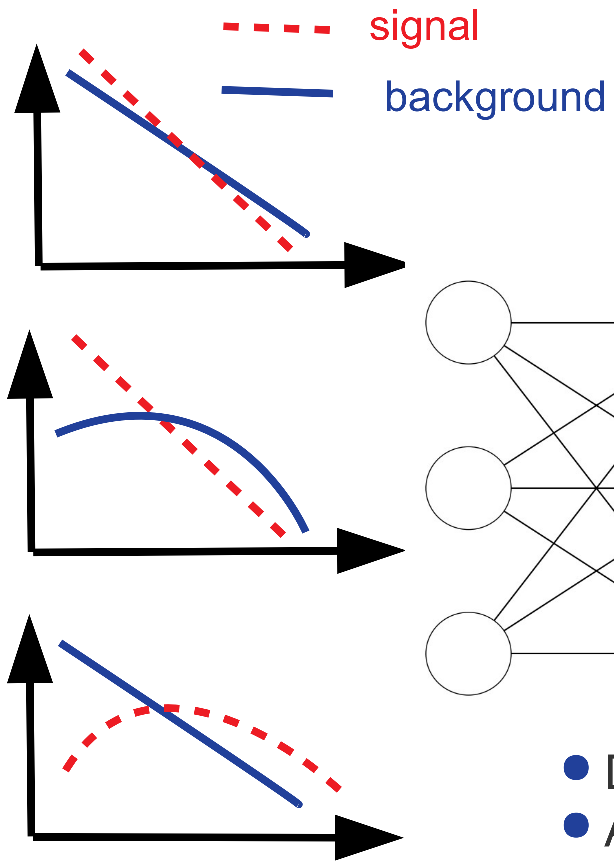


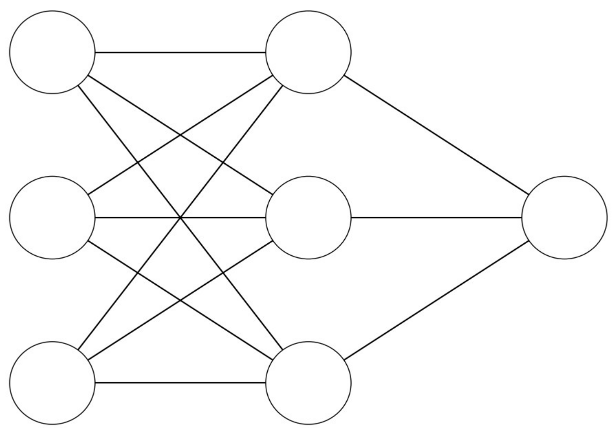
Also included in

arXiv:1805.11650 (NIMA, A931 (2019) 92)
arXiv:1810.06669 (Universe (2021) 7(1), 19)

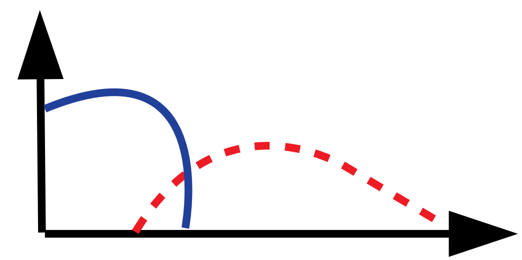
Machine learning (ML) and Artificial Neural Networks (ANN)

Extensively used in HEP in the last ~25 years





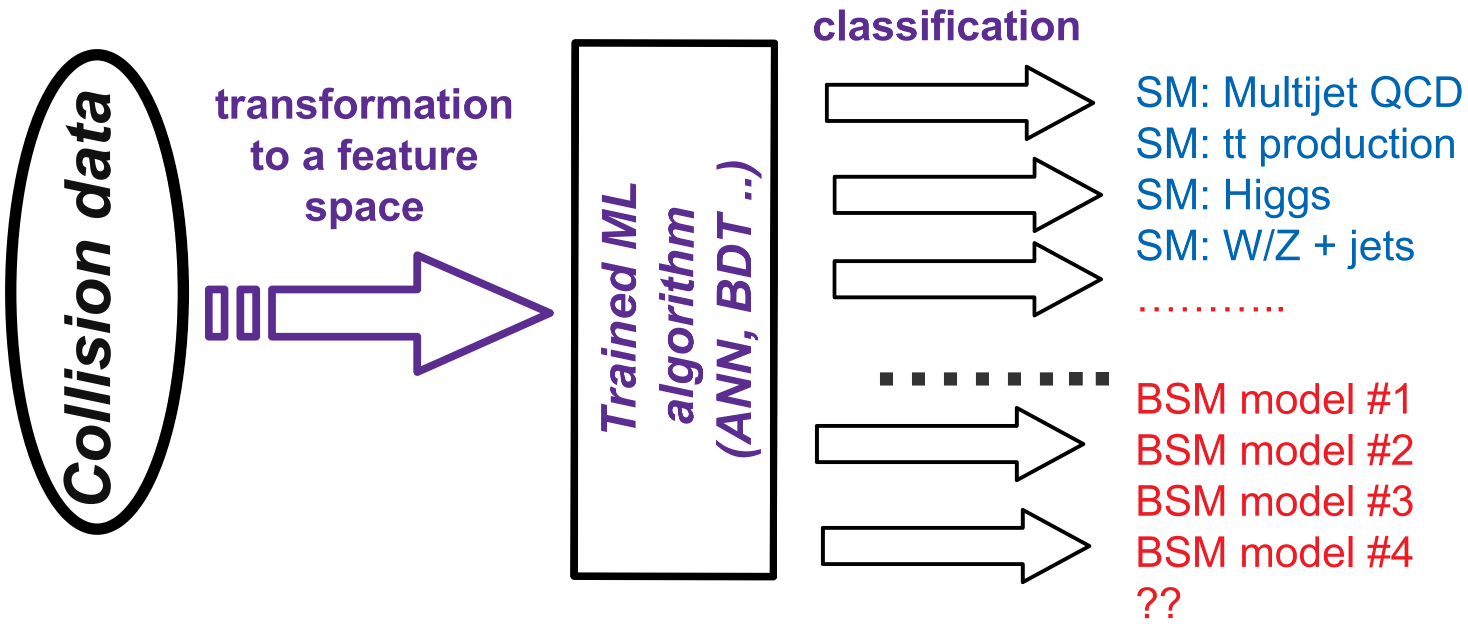
Better separation of signal and background in output space



"feature space"

- Different studies require different feature space
- Ambiguous, reproducibility issues
- Time consuming (> 80% spent for input preparation)

General feature space for supervised classification?



Find a standard feature space for ML that represents many event signatures of particle collisions (without "handpicking" variables for every event topology)

Input feature space for anomaly detection

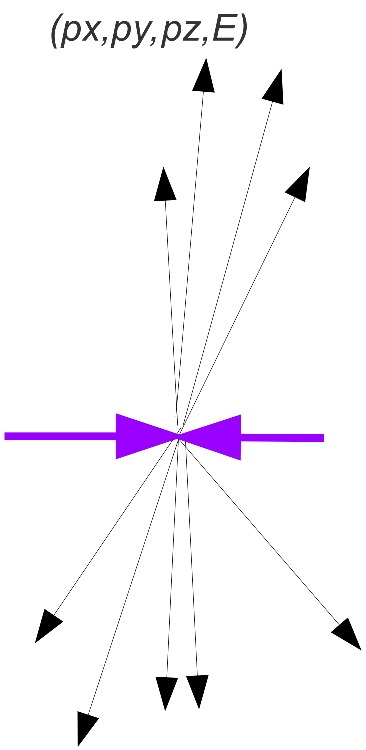
- Traditionally uses variable-size list of particles (Lorentz vectors) for different types
  - photons, jets, muons, electrons, taus etc.
- One can derive more complex Lorentz-invariant variables from this list, such as invariant masses, rapidity-difference etc. → Also *variable sizes*!
- Difficult input space for:
  - traditional ML (ANN, BDT etc) where certain neurons/nodes are typically mapped to particular (fixed) feature
  - visual inspection / debugging
- Example: Consider a single event with 6 particles/jets:
  - 1 photons
  - 2 jets
  - 1 b-jet
  - 1 muon
  - 1 electron

15 two-body invariant masses!

This number changes from event to event depending number of particles/jets!

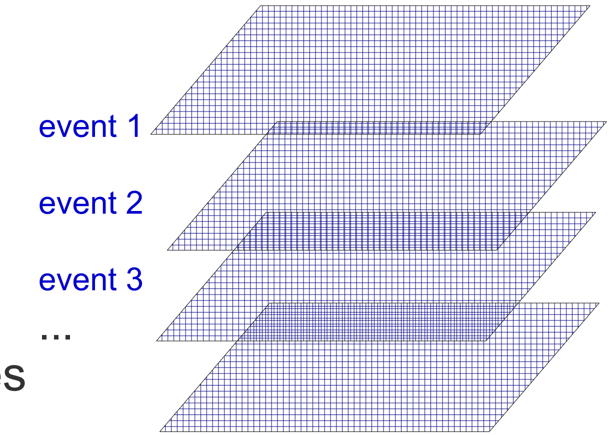
- Some LHC events have up to 300 two-particle invariant masses* for bump hunting (*depends on pT cuts*)

"Imaging kinematics" of particle collision events



from variable-size list with particles → fixed-size matrices

- Fixed size
- Dimensionless
- Lorentz invariant
- Fixed range of values
- Single and 2-particle densities
- Small correlations between variables
- Similarity with images



Organizes variable-size list in compact fixed-size data structures
Convenient input to ML & easy to visualize (similar to images)

arXiv:1805.11650 (NIMA, A931 (2019) 92)
arXiv:1810.06669 (Universe (2021) 7(1), 19)

Rapidity-mass matrix (RMM)

https://arxiv.org/abs/1805.11650
NIMA, A931 (2019) p92

Missing momentum and transverse masses

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

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

E_T / E_{CM} scaled transverse energies and energy imbalances

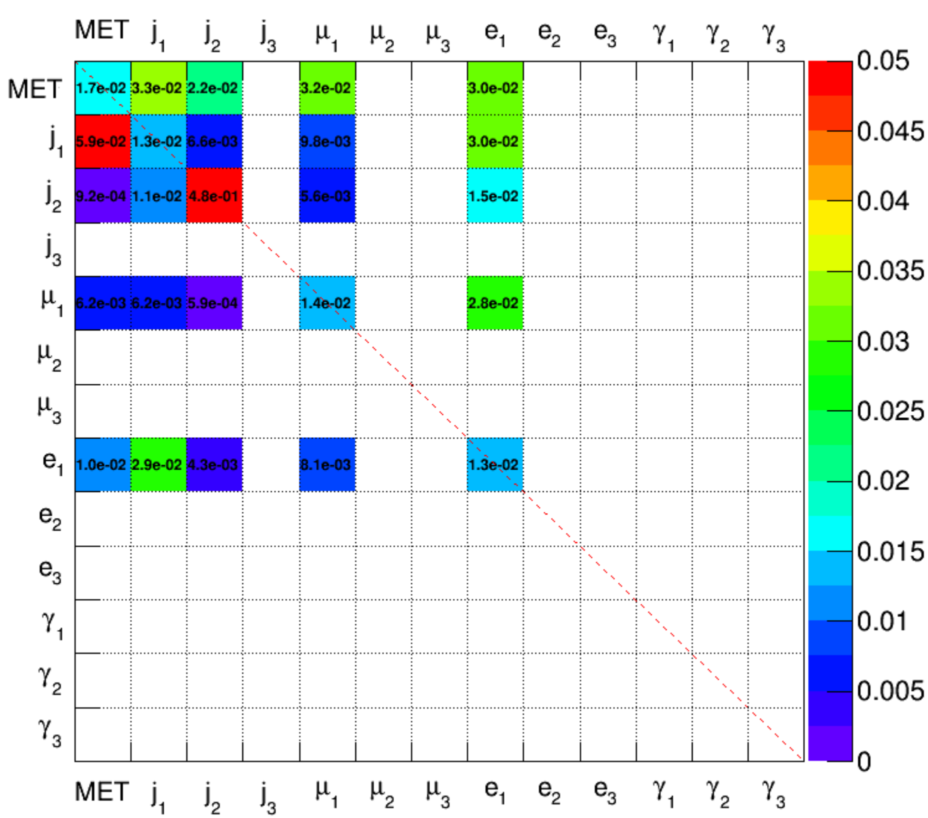
Lorentz factors

Rapidity difference: h(i,j) ~ cosh (y_j - y_i)

- Dimensionless, Lorentz invariant (1st column are Lorentz factors themselves)
- Single and two-particle densities for each identified particle or jet
- Cell values are ~ independent for SM processes → decorrelation by construction
- Re-scaling and normalization by construction
- Fixed sizes, well-defined mapping to input nodes
- Cells connected by proximity → good for visualization

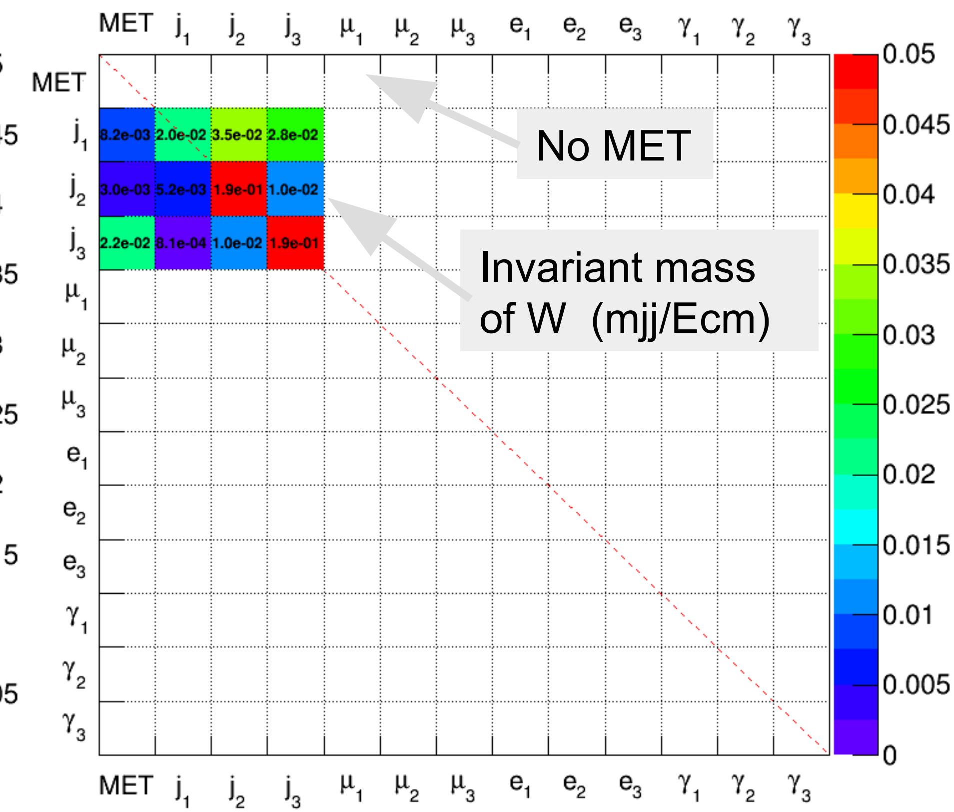
Example: Two PYTHIA8 events with $t\bar{t}$ as RMMs

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



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 → "natural" language for ML (even for simple backpropogation ANN or BDT)

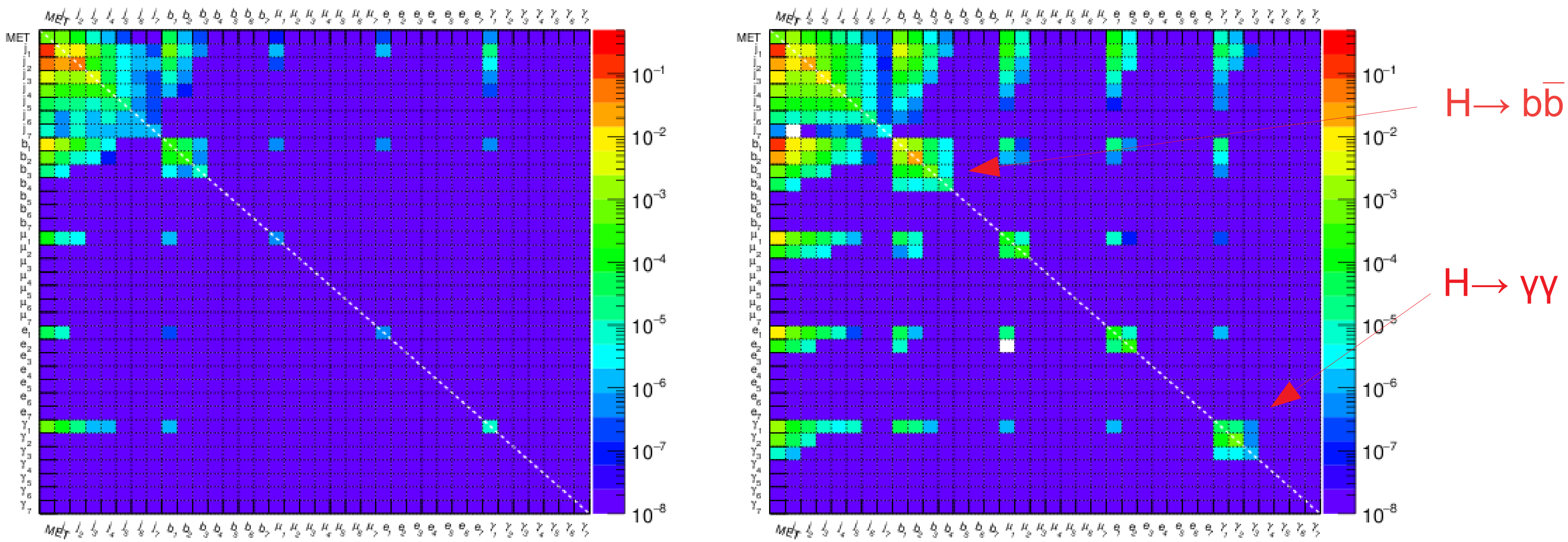
RMM for general event identification problem

- RMM includes all single & two-particle (+jet) densities
- Good choice for general event classifiers

Example based on Pythia8:

- SM QCD and Higgs (all decays)
- RMM using Np=7 and 6 objects using b-jets

Average values of RMM cells for 50k events

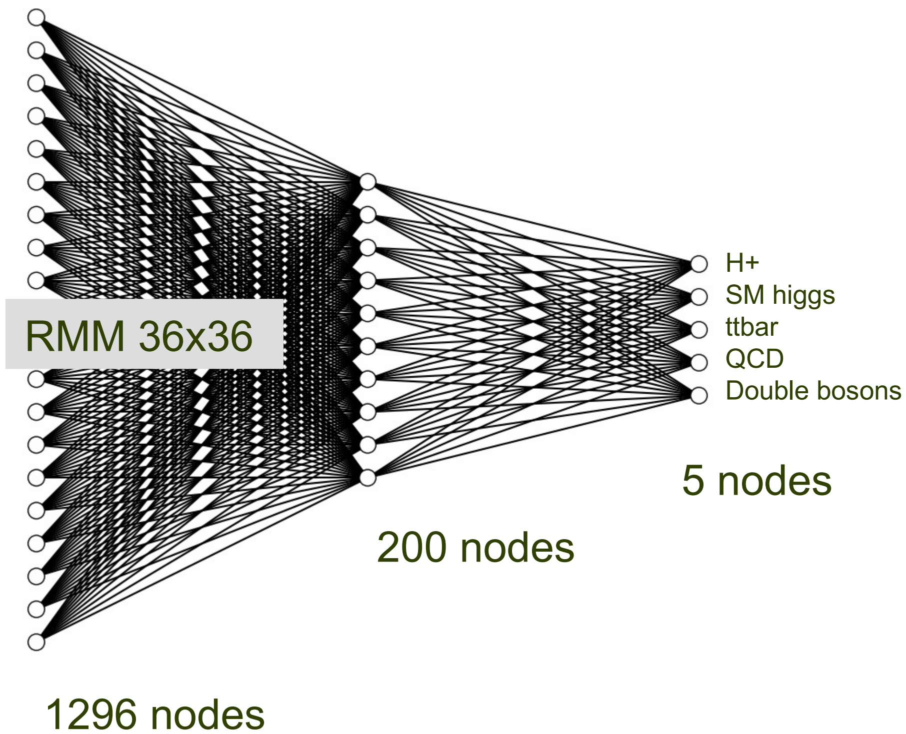


Multi-jet QCD

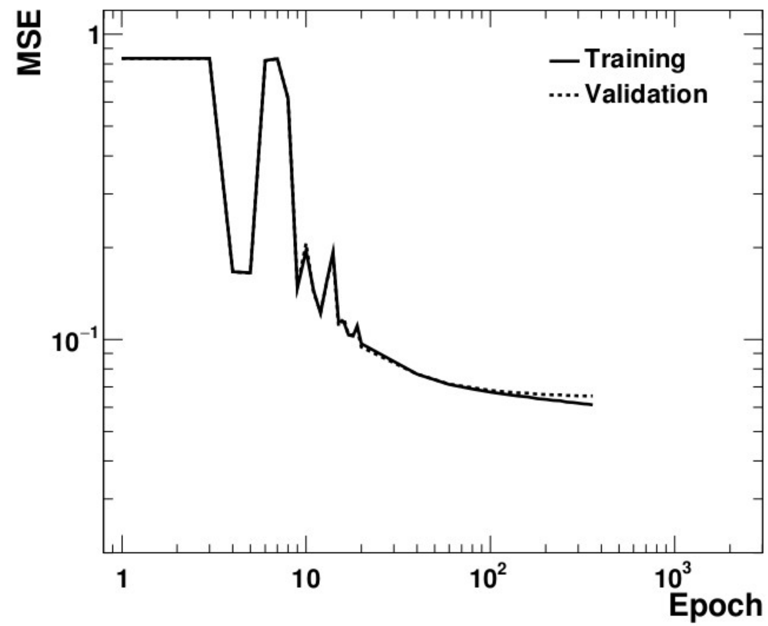
Higgs productions (all decays)

ANN training using RMM as input

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



Wide and shallow ANN for sparse input RMM data

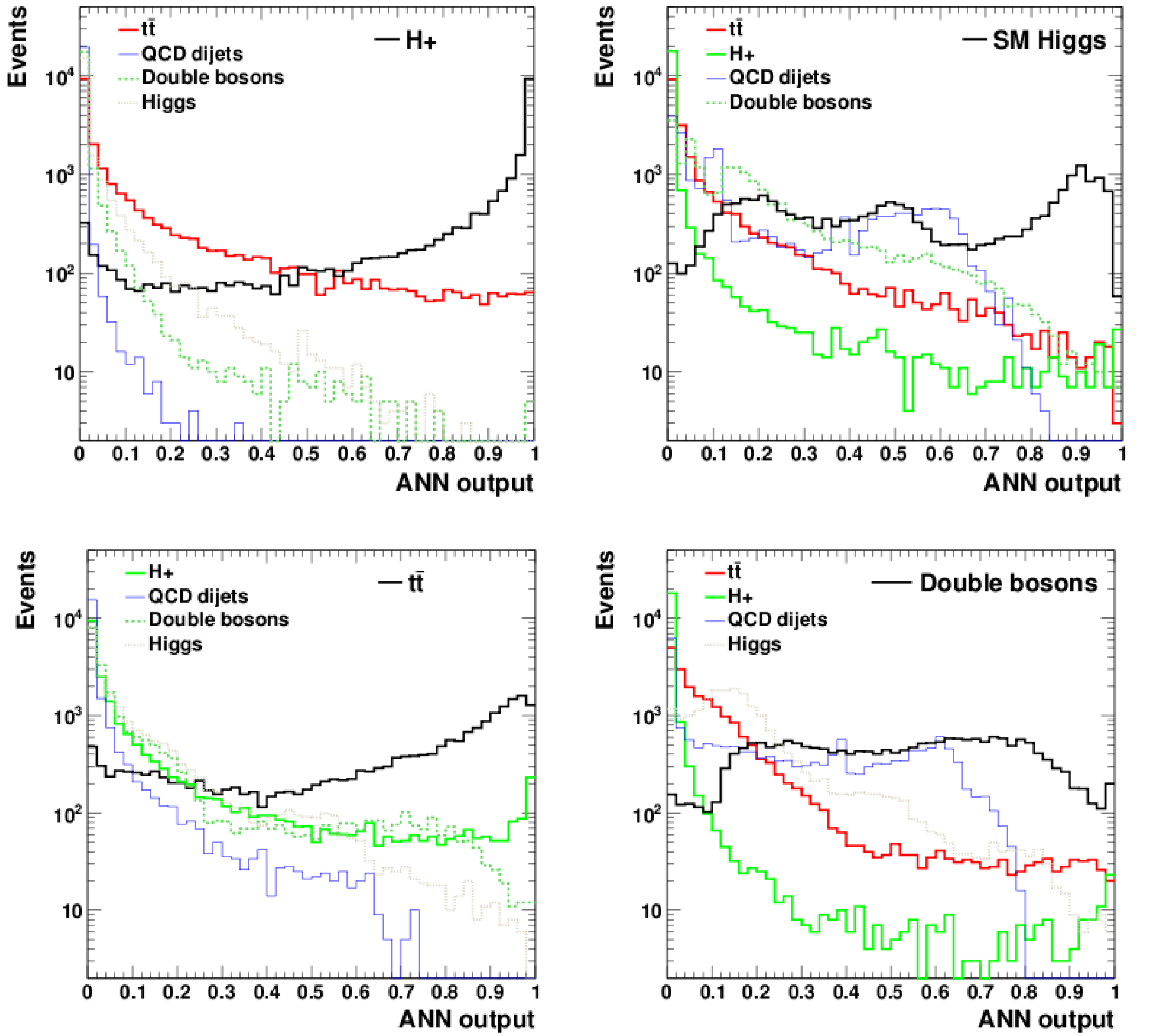


Well trained after 100 epochs:
Mean Squared Error (MSE) decreases from 0.8 to 0.07 (~ 1h training for 200k RMM)

Output scores after ANN training using RMM

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

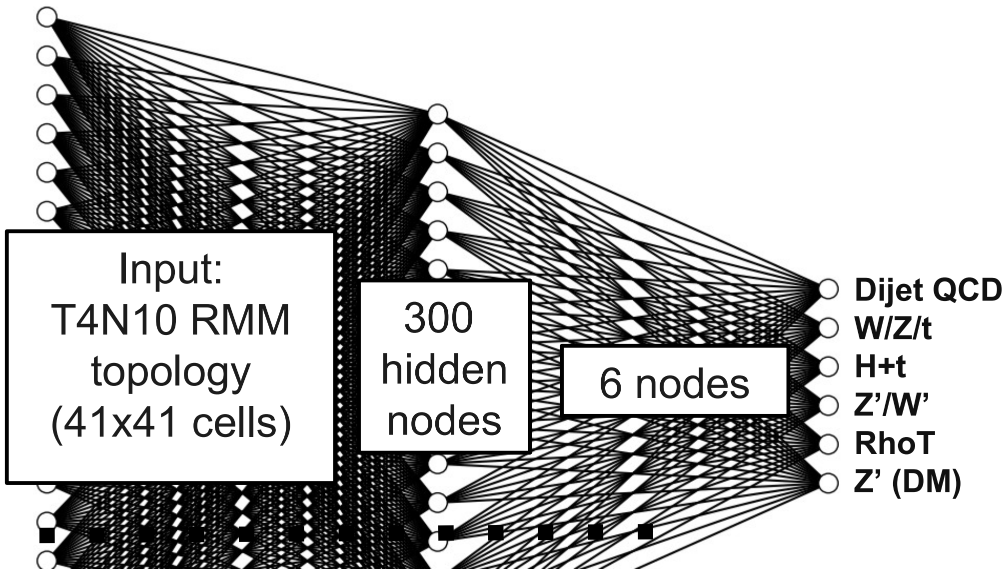
Purity of event classification is 80%-90% assuming 0.8 cut on output nodes



See arXiv:1810.06669 for details

Supervised ANN architecture

Backpropogation ANN with Signoid function, 6 outputs for each process

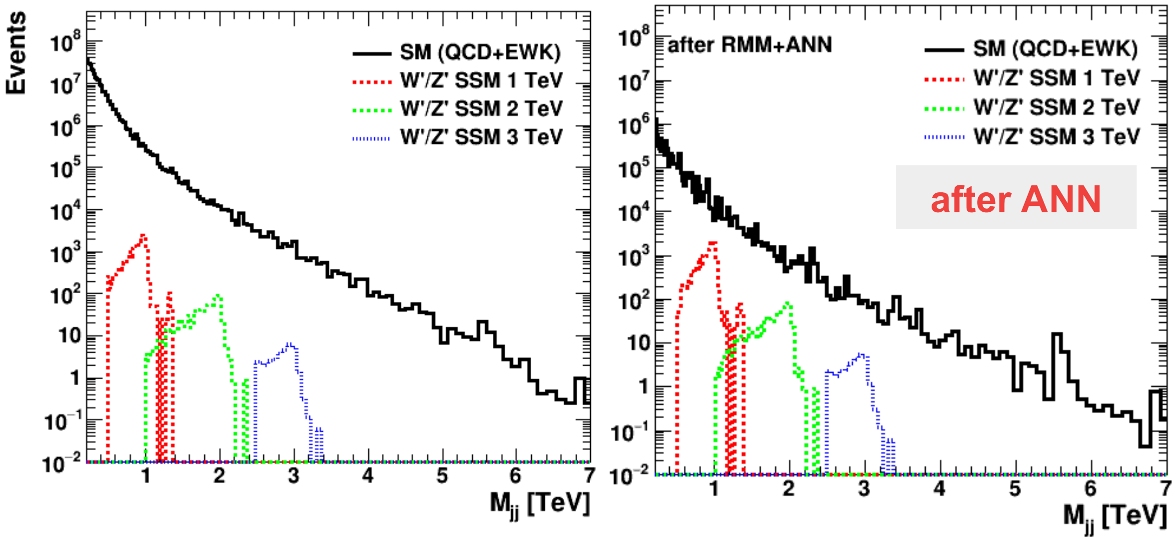


Single neural network can be trained on many BSM and SM processes using RMM feature space (6 processes in this example)

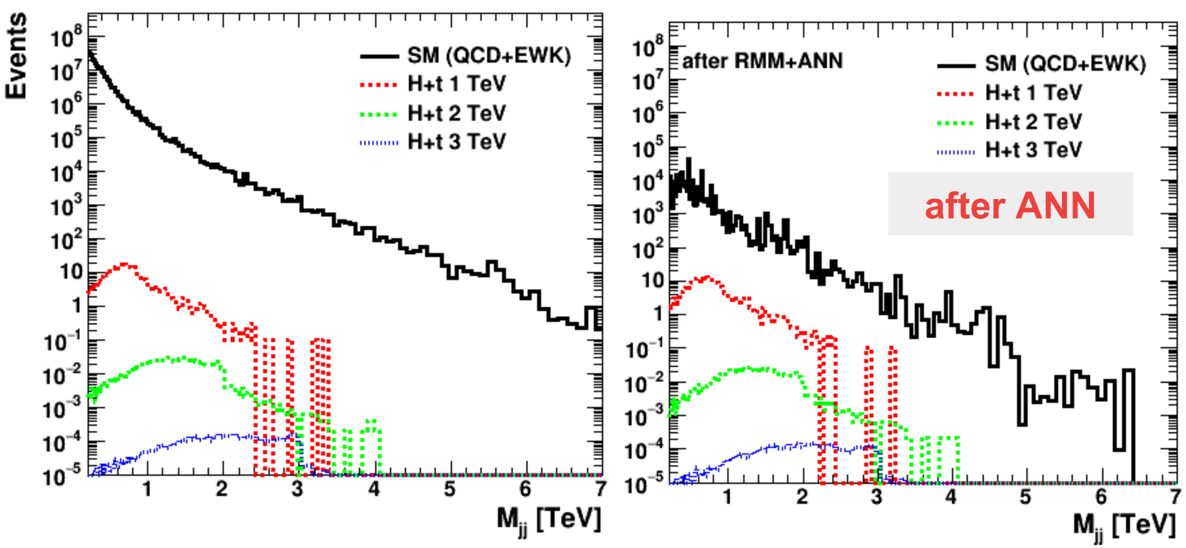
- Event conversion to RMM:
  - 2000 events for each mass
  - 10k events from SM processes
- Each RMM is associated with a vector with 6 values that define event type:
  - Examples:
    - (1,0,0,0,0,0) - SM dijets
    - (0,0,0,1,0.0) - Z'/W'
- Mix events and feed to ANN
- After training:
  - **Model specific selection**:
    - Require a value close to 1 for a neuron associated to a specific process
  - **Model agnostic:**
    - Reject events that have large values for dijet QCD and W/Z/t

W'/Z' and H+ signals before and after ANN

W'/Z' bosons



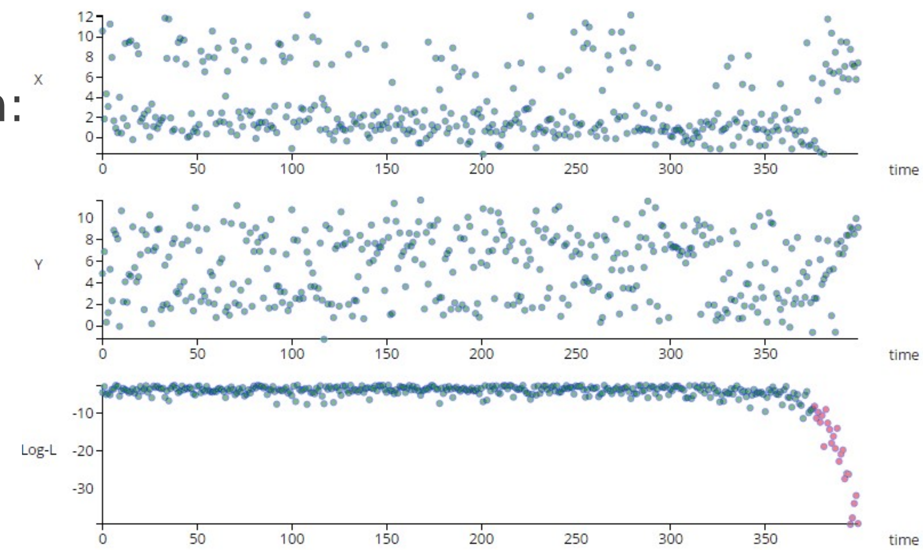
H+t bosons



- ✔ Dijet invariant masses M_{jj} for:
  - SM processes (black line)
  - BSM: W'/Z' (3 masses)
  - BSM: H+t (3 masses)
- ✔ M_{jj} before and after cut 0.5 on ANN score for each output
- ✔ Signal-over-background ratios for BSM models increased by 2-3 orders of magnitude after applying ANN
- ✔ Similar background reduction for all other BSM models
- ✔ No distortions of background shapes after ANN score cuts

Anomaly detection

- Finding unusual pattern in data:
  - Outlier detection
  - Fault detection
  - Novelty detection
  - Event detection
  - Deviation detection



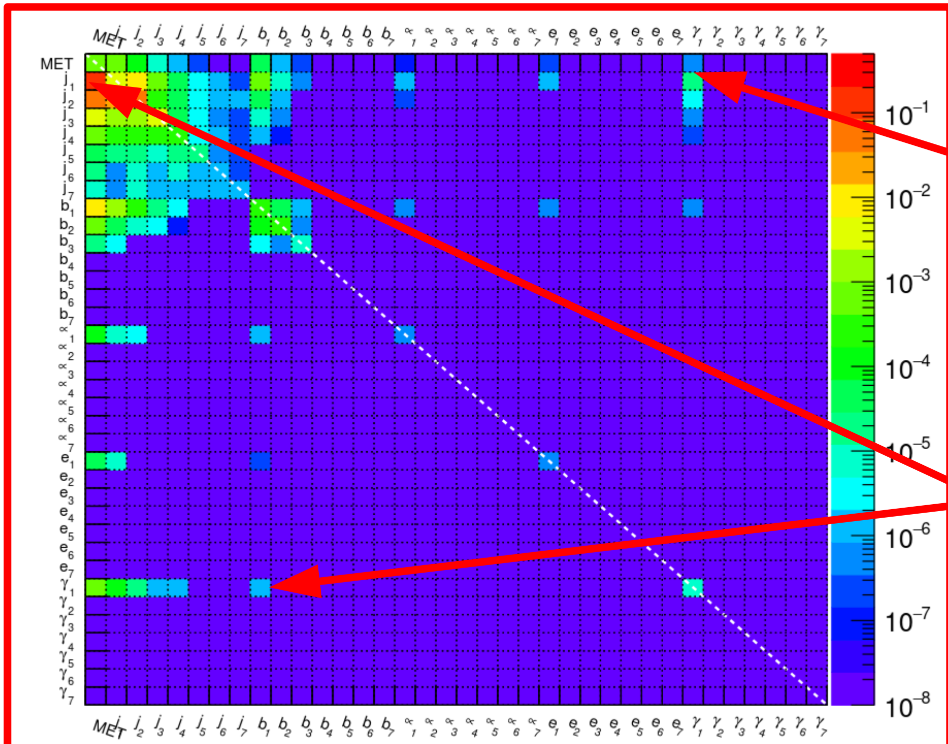
New physics may produce unexpected signatures (like peaks in invariant masses) hidden in large SM background. To find such BSM events, select uncharacteristic SM events ("outliers") and look at their signatures.
Note: Anomaly detection algorithm must not bias the signatures themselves (i.e. artificial peaks etc)

1) Use RMM as input space
2) Apply an anomaly detection algorithm using statistical methods (or ML).
3) Define anomalous events (outliers)
4) Study physics distributions of outliers (bumps in invariant masses etc.)

Advantages: No complex ANN training & Monte Carlo simulations

Anomaly detection in RMM

- Typical RMM has ~2000 cells, and about 400 non-zero values (distributions)
- Each cell has a distribution of values in the range 0-1
- Types of anomalies:
  - **Outlier detection:** Multiple cells with values above some threshold
  - **Novelty detection:** Appearance of new active cells (new objects)



(a)Multi-jets QCD

presence of photons?
("novelty detection")

Large MET? Rapidity gap?
("outlier detection")

Anomaly detection using Z-score

- Popular statistical method applied to RMM:

$$Z = \frac{\overset{\text{Score}}{x} - \overset{\text{Mean}}{\mu}}{\underset{\text{SD}}{\sigma}}$$

(1) Stouffer's Z-scores for ***"Outlier" detection.*** *Sensitive to values of activated cells, rather than to the number of active cells.* Calculate Z-scores for each cell and then combine them using Stouffer's method (x 1 / √N of the number of cells)

$$Z_S = \sum_{i,j}^{N} Z_{ij} / \sqrt{N} \qquad Z_{ij} = \frac{(X_{ij} - \overline{X}_{ij})}{\sigma(X_{ij})}$$

X_{ij} – value of RMM cell

Sum runs over all active cells *(N)*

$\overline{X}_{ij}$ and $\sigma(X_{ij})$ calculated for all events

(2) Event Z-score for "**Event Novelty" detection**. Sum all matrix values for a given event to get "X", and then calculate Z-score for "X" using all events

$$Z = \frac{(X - \overline{X})}{\sigma(X)} \qquad X = \sum_{i,j}^{N} X_{ij}$$

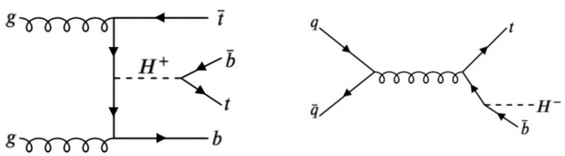
X_{ij} – value of RMM cell

Sum runs over all active cells *(N)*

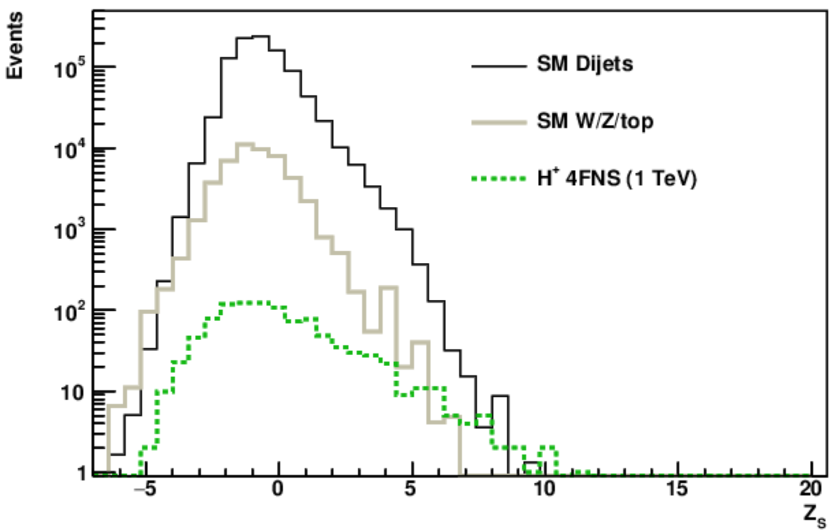
$\overline{X}$ and σ calculated for all events

Anomaly detection using truth-level MC

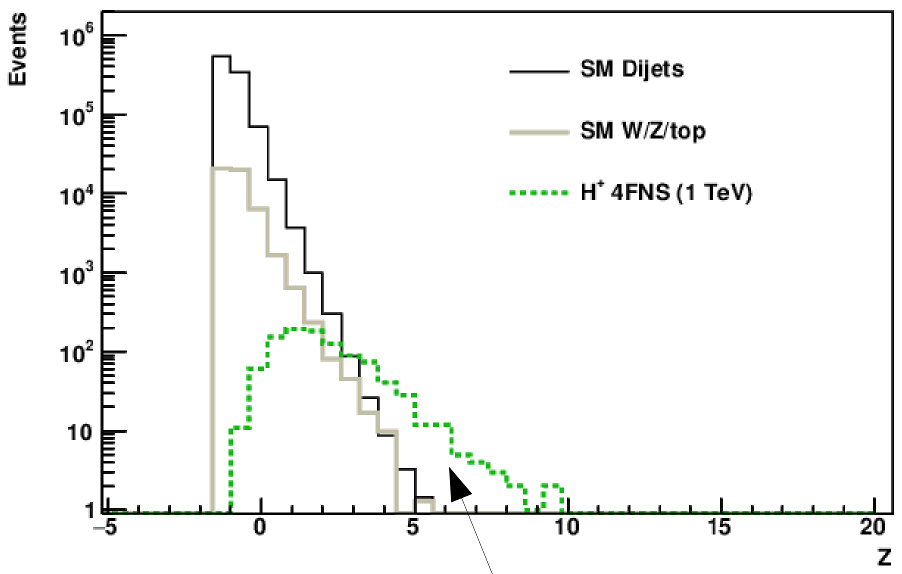
- Monte Carlo samples using Pythia8 as in arXiv:1810.06669 (Universe (2021) 7(1), 19)
- **Standard model:** Dijet QCD (770k events)+W/Z/$t\bar{t}$ (200k),
- **BSM:** 1000 H+ Higgs with mass 1 TeV
  - ($H^{+} \rightarrow t\bar{b}$, all decays of top)
- All events pre-selected with at least 1 lepton (“fake” for Dijet QCD)



Stouffer's Z



Event Z-score

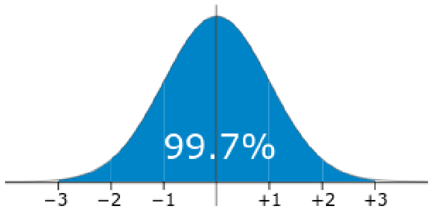


Large values of Z-scores dominated by BSM

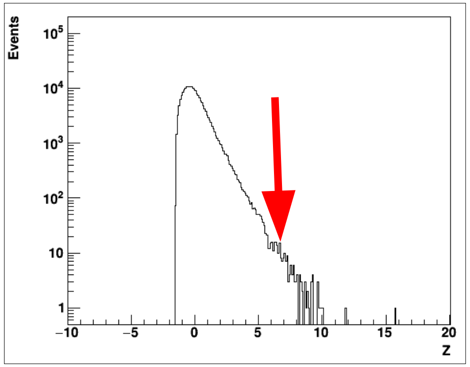
H+ model has more leptons than SM due to top-quark decays

How to define outliers in statistical z-score approach?

- **Model independent:**
  - z-score is "significance". |z-score| > 3 corresponds to more than 3 σ deviation. For a normal distribution, outlier can be defined with probability ~0.3% for events to deviate from their mean
  - Frequent choice for anomaly detection, agnostic to BSM



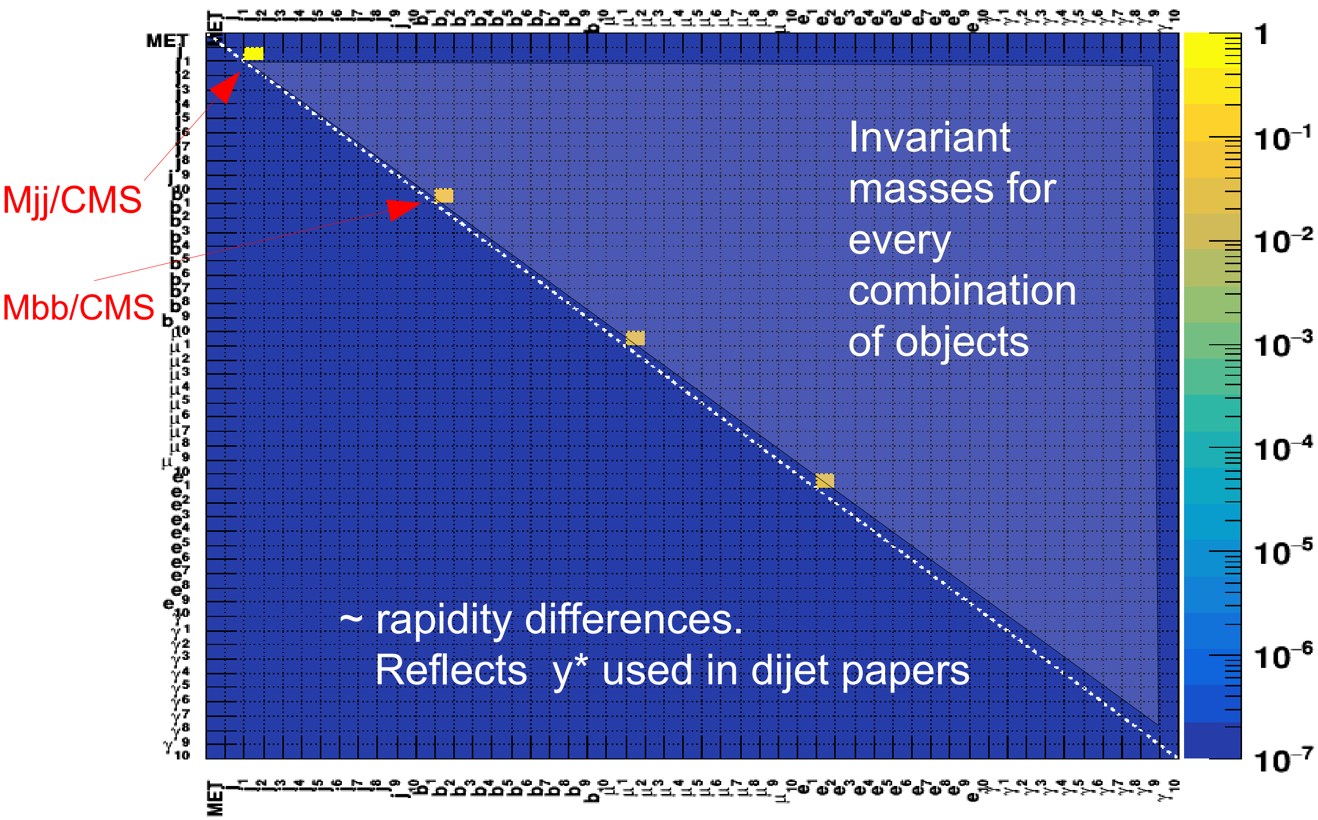
- **Based on BSM simulations?**
  - Region where z-scores have more than 50% contamination from most common BSM models?
  - Model dependent and requires simulations of BSM



- .. ?

Physics distributions for outlier events

- **Select events for outliers with Z > 3**
- **Look at invariant masses stored in RMM**
  - Example: Mjj, Mbb, Mee, Mμμ are shown using yellow color:



Most dijet studies can be reproduced from cells:

- (2,1) → Mjj
- (1,2) → y*

B-jet papers from:

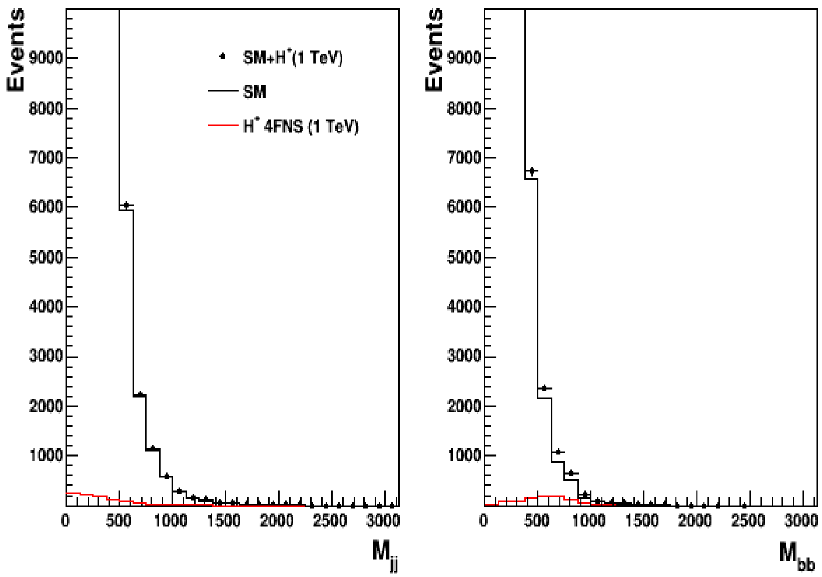
- (11,12) → Mbb
- (12,11) → y*

Leptoquark papers from cells corresponding to jets and leptons

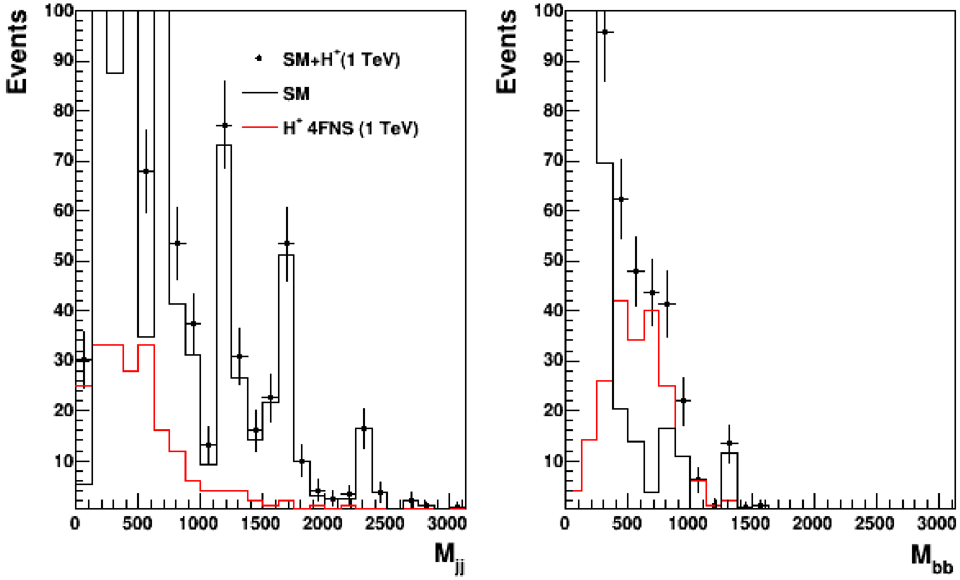
- .. etc. etc.

2-jet distributions (Mjj and Mbb) from cell (2,1) and (7,8)

All Z_s



outlier Z >3

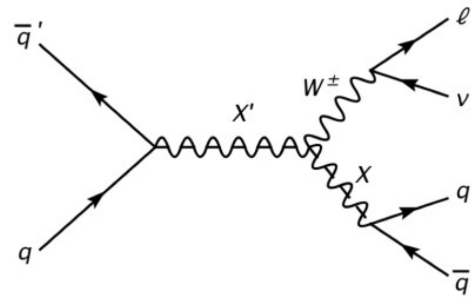


Large spikes in SM are due to low statistics in weighted Monte Carlo simulations

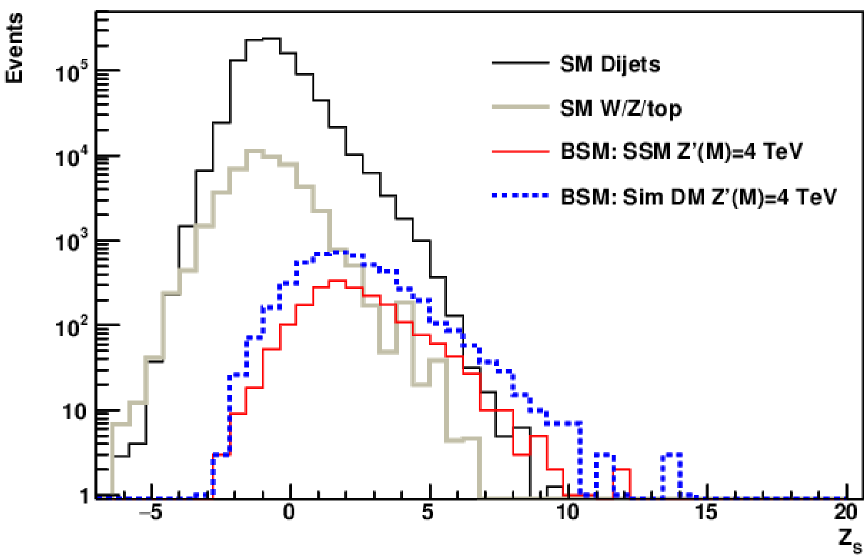
Improvements in S/B for outlier Z>3

Anomaly detection: M(Z')=4 TeV

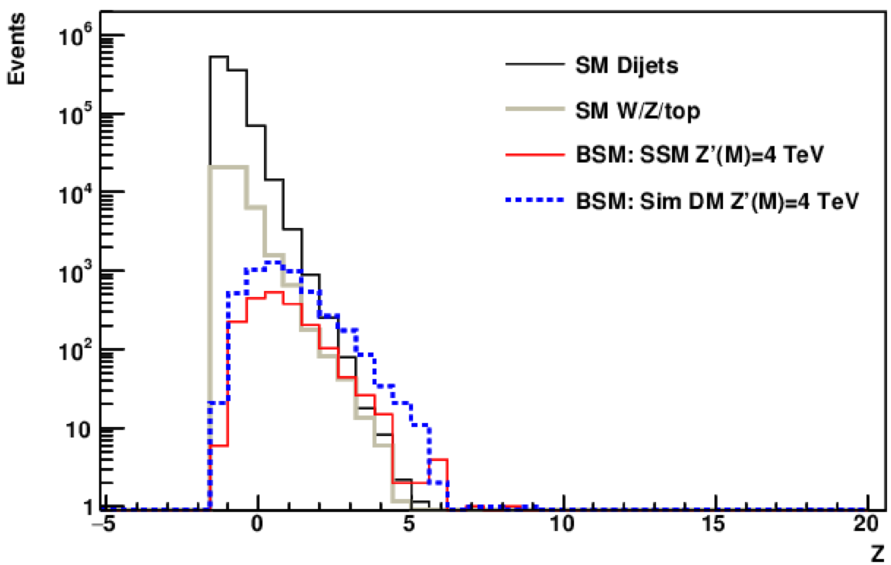
- Monte Carlo samples using Pythia8 as in arXiv:1810.06669 (Universe (2021) 7(1), 19)
- **Standard model:** Dijet QCD (770k events)+W/Z/$t\bar{t}$ (200k),
- **BSM:** 5k events with Z' at M=4 TeV
  - (1) SSM (2) Simplified DM
- All events pre-selected with at least 1 lepton ("fake" for Dijet QCD)



Stouffer's Z



Event Z-score



Final-state for SM & BSM are similar, but BSM have harder spectra

Summary

- **RMM is useful feature space for ML for collider experiments**
  - Simple-to-use and well-defined sparse matrices
  - Works even for simplest ANN/BDT → "natural language" for ML
  - Easy visualization for humans ("image-like")
- **Easy to apply to anomaly detection (like statistical Z-score method)**

- **RMM is well suited for general event classification problems due comprehensive (nearly independent) single and two-particle densities**
  - Same RMM transformation can be plugged into different BSM searches to produce results with minimal tweaking using single ML algorithm
  - But large input if no pruning is done (> 1,000 input variables)

Program library on GIT: https://github.com/chekanov/Map2RMM