MASS UNSPECIFIC SUPERVISED TAGGING (MUST) FOR BOOSTED JETS

João Seabra

Departamento de Física and CFTP, Instituto Superior Técnico, Lisboa

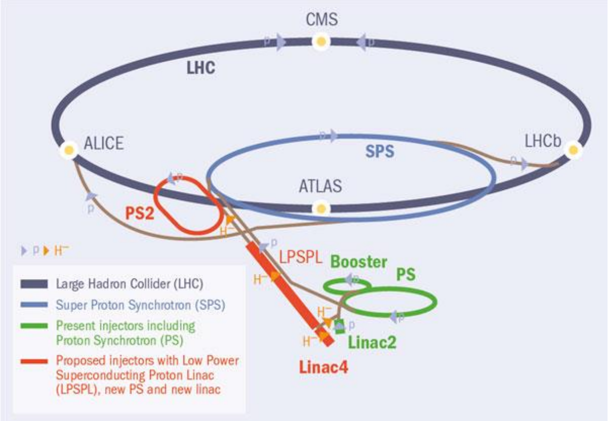
Beyond Standard Model: From Theory to Experiment (BSM-2021)

2nd of April 2021

Based on work made in collaboration with: **Juan A. Aguilar-Saavedra** and **Filipe R. Joaquim**

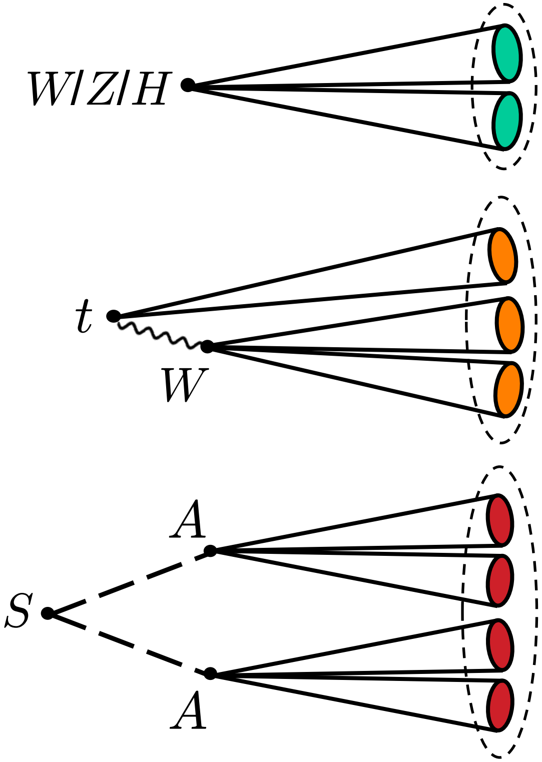
JHEP 03 (2021) 012

MOTIVATION



- In the decades to come, the high-energy frontier of particle physics will continue to be explored at the Large Hadron Collider (LHC);
- Most jets stem from Quantum Chromodynamics (QCD) processes...
- ...but when sufficiently boosted, the hadronic decays of Standard Model (SM) particles like the W, Z and Higgs bosons and the top quark also yield jets;
- Hadronic decays of new particles can produce jets too;
- A lot of theoretical frameworks beyond the SM predict multi-jet signals originated from direct or cascade decays of yet unseen particles.

e.g. J. A. Aguilar-Saavedra, F. R. Joaquim; JHEP 01 (2016) 183
K. S. Agashe *et al.*; JHEP 05 (2017) 78



Therefore...

MOTIVATION

Jet identification tools are crucial for new physics searches at the LHC.

Searches for new gauge-bosons, scalars and spin-2 particles

A. M. Sirunyan *et al*. [CMS]; JHEP 08 (2017) 29

A. M. Sirunyan *et al*. [CMS]; JHEP 09 (2018) 148

M. Aaboud *et al*. [ATLAS]; Phys. Lett. B 781 (2018) 327

M. Aaboud *et al*. [ATLAS]; Phys. Lett. B 788 (2019) 316

M. Aaboud *et al*. [ATLAS]; Phys. Lett. B 783 (2018) 392

M. Aaboud *et al*. [ATLAS]; Phys. Rev. D 98, 3 (2018) 32015

A. M. Sirunyan *et al*. [CMS]; Phys. Rev. D 99, 1 (2019) 12005

A. M. Sirunyan *et al*. [CMS]; Eur. Phys. J. C 80, 3 (2020) 237

A. M. Sirunyan *et al*. [CMS]; Phys. Rev. D 100, 11 (2019) 112007

G. Aad *et al*. [ATLAS]; Eur. Phys. J. C 80, 12 (2020) 1165

Searches for vector-like quarks

A. M. Sirunyan *et al*. [CMS]; Phys. Lett. B 781 (2018) 574

A. M. Sirunyan *et al*. [CMS]; Eur. Phys. J. C 79 (2019) 90

M. Aaboud *et al*. [ATLAS]; JHEP 05 (2019) 41

A. M. Sirunyan *et al*. [CMS]; Eur. Phys. J. C 79, 3 (2020) 36

Searches for dark-matter

A. M. Sirunyan *et al*. [CMS]; Eur. Phys. J. C 79, 3 (2019) 280

JET IDENTIFICATION

It requires:

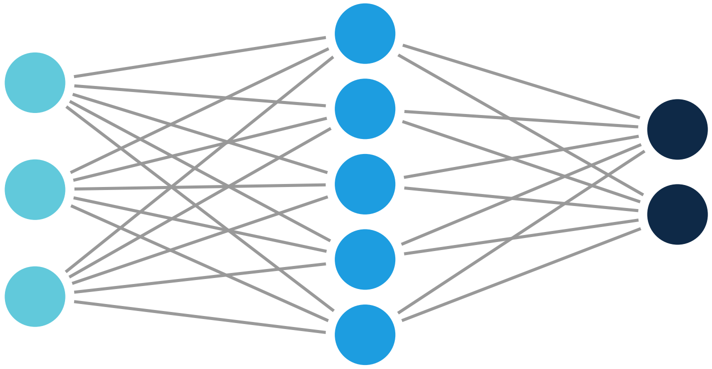
- Quantifying its mass, usually after applying some *grooming*;
- Infering the number of quarks and gluons clustered inside it (*prongs*).

Tagging

Examples:

Processes	Prongness	Classification
QCD	One-pronged (1P)	**Background**
$W/Z/H \to q\bar{q}$	Two-pronged (2P)	**Signal**
$t \to W^+b \to q\bar{q}b$	Three-pronged (3P)	
$S \to AA \to q\bar{q}q\bar{q}$	Four-pronged (4P)	

Here, this procedure relies on the training of Neural Networks (NNs).

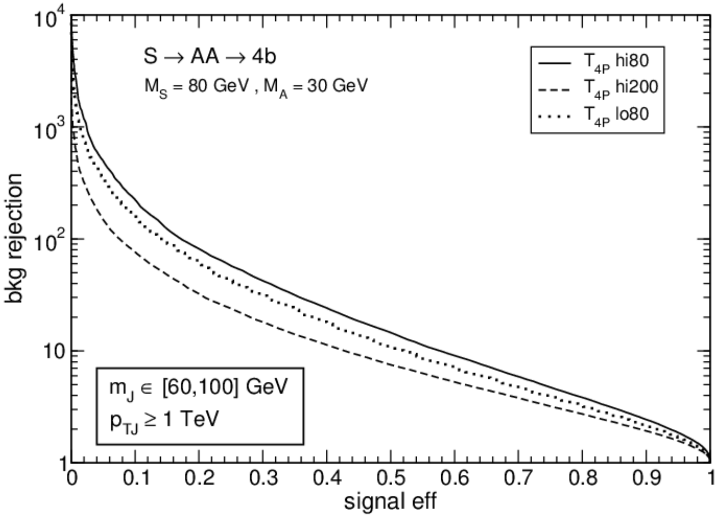


MASS DECORRELATION

- The mass of a jet and the variables that encode its substructure are usually correlated;

- The mass decorrelation methods employed so far in supervised taggers leave a residual dependence of the results on the jet mass and transverse momentum training ranges. **Consequently...**

Their performance drops when applied to kinematical regions different from those used to train them.



J. A. Aguilar-Saavedra, B. Zaldívar, Eur. Phys. J. C 80, 6 (2020) 530

Solution ??

OUR GOALS

- Build generic taggers, sensitive to any kind of jets
- Excellent performance for all jet masses
- Mass decorrelation

MASS UNSPECIFIC SUPERVISED TAGGING

Considering the jet mass and its transverse momentum varying over wide ranges, we make them input variables of a multivariate tool, together with jet substructure observables.

Mass Unspecific Supervised Tagging (MUST) for boosted jets

JHEP03(2021)012

J.A. Aguilar-Saavedra,[a] F.R. Joaquim[b] and J.F. Seabra[b]

[a] *Departamento de Física Teórica y del Cosmos, Universidad de Granada, E-18071 Granada, Spain*

[b] *Departamento de Física and CFTP, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal*

Our MUST-inspired jet taggers have **19** input variables:

- 17 N-subjettiness observables which characterise jet substructure,

$$\left\{\tau_1^{(1/2)},\tau_1^{(1)},\tau_1^{(2)},\cdots,\tau_5^{(1/2)},\tau_5^{(1)},\tau_5^{(2)},\tau_6^{(1)},\tau_6^{(2)}\right\};$$

- Jet's mass, m_J ;
- Jet's transverse momentum, p_T .

Note:
Ratios of N-subjettiness variables will be denoted as

$$\tau_{mn} \equiv \frac{\tau_m^{(1)}}{\tau_n^{(1)}} .$$

All those variables should be standardised according to the SM background distributions.

TRAINING SET GENERATION

	Background	**Signal**
Processes	$pp \to jj$	$pp \to ZS$ • All signal types: $Z \to \nu\nu$ • 2P: $S \to u\bar{u}$, $S \to b\bar{b}$ • 3P: $S \to F\nu$; $F \to udd$, $F \to udb$ • 4P: $S \to u\bar{u}u\bar{u}$, $S \to b\bar{b}b\bar{b}$
p_T **range**	[200, 2200] GeV	
Mass ranges	$m_J \in$ [50, 250] GeV $M_{S,F} \in$ [30, 400] GeV ($M_S \leq p_T R/2$, R=0.8)	

The decays of S and F are implemented with a flat matrix element (**to achieve generic taggers).**

TAGGER PROPERTIES

Name	Types of events used in training	NN architecture	Output Layer
GenT	Background + 2P + 3P + 4P	2048 x 128	Sigmoid
GenT_{2P}	Background + 2P	1028 x 64	Sigmoid
GenT_{3P}	Background + 3P	1028 x 64	Sigmoid
GenT_{4P}	Background + 4P	1028 x 64	Sigmoid
Prongness selection tagger	2P + 3P + 4P	2048 x 128	Softmax

To evaluate the performance of GenT, GenT_{2P}, GenT_{3P} and GenT_{4P} we use the Area Under the ROC curve (AUC) whereas the Prongness selection tagger is evaluated by measuring its accuracy.

- All our NNs use the Rectified Linear Unit (ReLU) activation function;
- The optimisation of GenT, GenT_{2P}, GenT_{3P} and GenT_{4P} (Prongness selection tagger) rely on the binary (categorical) cross-entropy;
- The Adam optimiser is applied to all NNs.

BENEFIT OF MUST TAGGERS

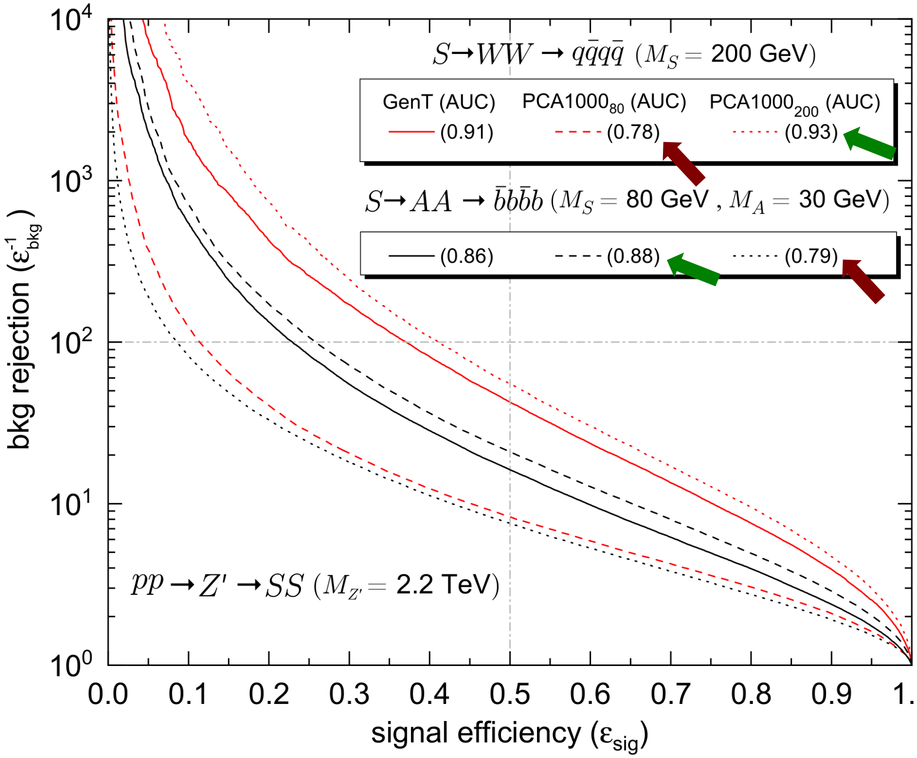
Non-MUST taggers:

PCA1000$_{80}$: Trained for $p_T \geq 1.0$ TeV and on the mass interval $m_J \in [60,100]$ GeV.

PCA1000$_{200}$: Trained on the same region of momentum but in a different mass interval, $m_J \in [160,240]$ GeV.

Principal Component Analysis (PCA) is used in both taggers to perform mass decorrelation.

- ✓ These taggers perform slightly better on a mass region close to the one where they were trained...
- ✗ ... but are much less efficient when applied to masses out of the training region.

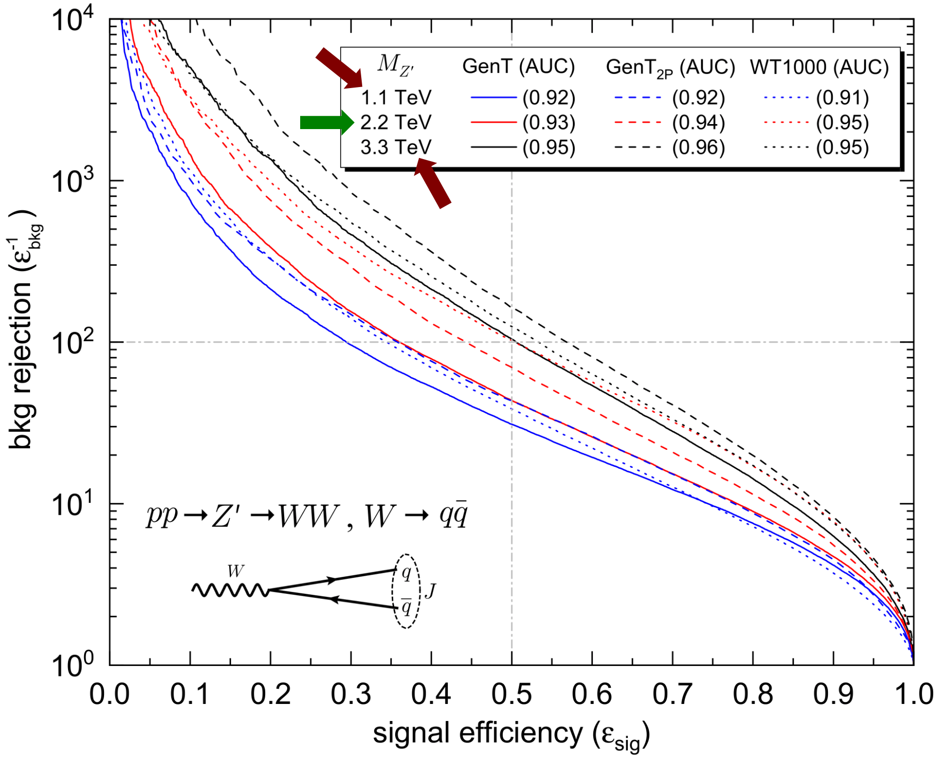


BENEFIT OF MUST TAGGERS

Non-MUST tagger:

WT1000: PCA-decorrelated tagger trained with W jets obtained from $Z' \rightarrow WW$ ($M_{Z'}$ = 2.2 TeV) and QCD jets with $p_T \geq 1$ TeV and $m_J \in [60,100]$ GeV.

- ✓ It performs slightly better than GenT_{2P} for $p_T \geq 1$ TeV and $m_J \in [60,100]$ GeV.
- ✗ It performs slightly worse than GenT_{2P} for $p_T \geq 500$ GeV and $p_T \geq 1.5$ TeV.

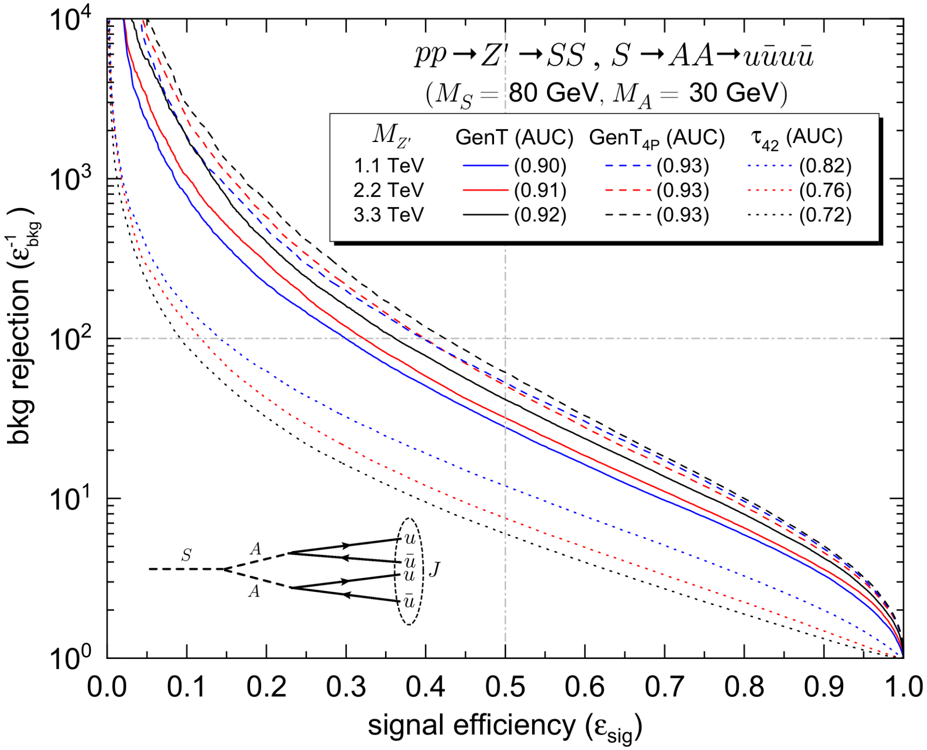


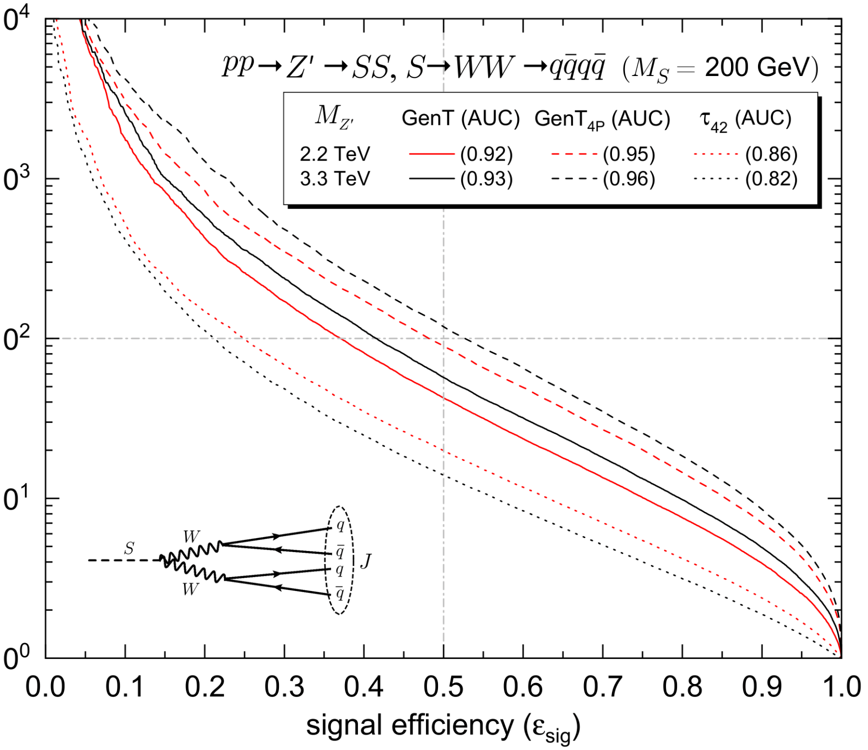
GenT_{2P} is nearly optimal for W jets

TAGGER PERFORMANCE (4P SIGNALS)

Background: Quark and gluon jets generated in $pp \to Zq$, $pp \to Zg$, with $Z \to \nu\nu$

- $p_T \geq 0.5$, 1.0, 1.5 TeV for $M_{Z'} = 1.1$, 2.2, 3.3 TeV, respectively;

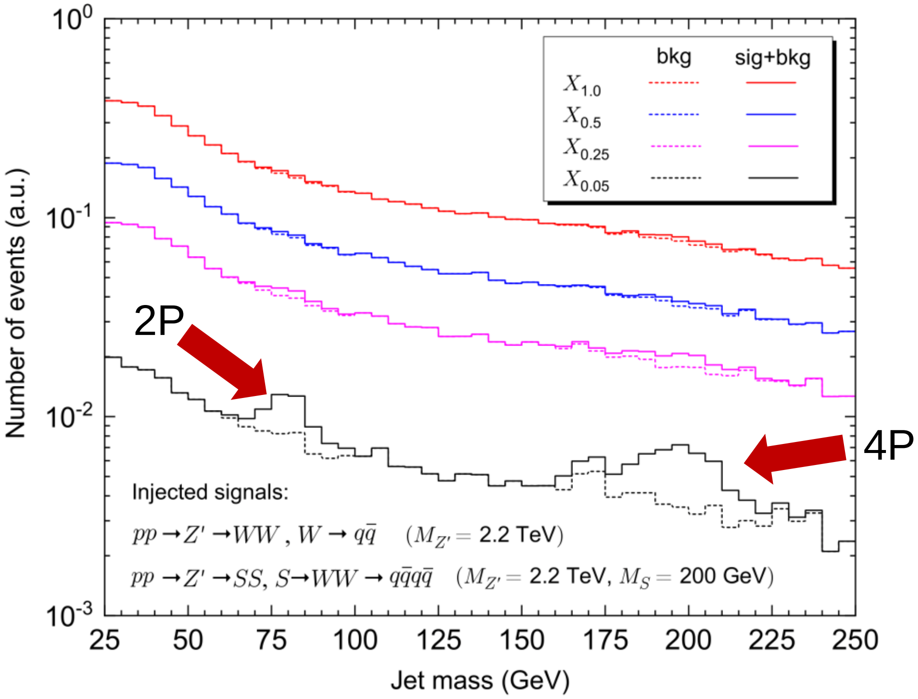




- The performance of GenT and GenT$_{4P}$ is significantly better than that of τ_{42}.
- The performance improves as $M_{Z'}$ increases

MASS DECORRELATION

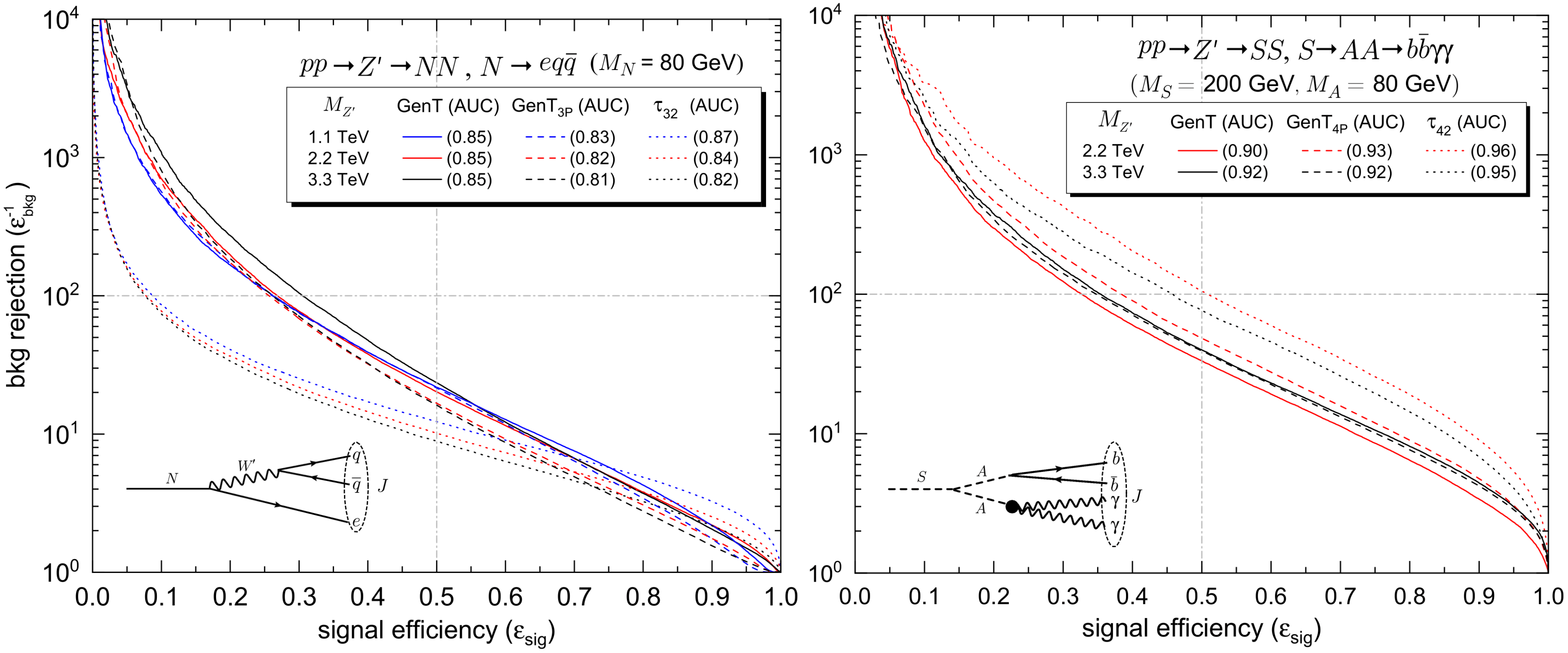
- Defining $\rho = 2\log(m_J / p_T)$, we compute at each bin of a 2D grid (ρ, p_T) the 5%, 25% and 50% percentiles of the NN score ($X_{0.05}$, $X_{0.25}$ and $X_{0.5}$ respectively);



- This varying threshold preserves the SM background distribution and the injected signals show up when the cut is sufficiently tight.

Our generic taggers also provide a perfect solution to the mass correlation problem.

JETS NOT USED TO TRAIN MUST TAGGERS



- **MUST taggers can detect unseen signals with good efficiency;**
- AUC is not good to evaluate the performance of taggers for neutrino jets;
- Simpler multivariate methods like logistic regression may achieve better performance for stealth boson jets with two photons in the final state.
  J. A. Aguilar-Saavedra, B. Zaldívar; Eur. Phys. J. C 80, 6 (2020) 530

IDENTIFICATION OF NEW PHYSICS SIGNALS

Using the prongness selection tagger, we apply the following classification criteria in the four benchmark examples below:

$$\begin{cases} \text{2P}\,, & \text{if } P_{\text{2P}} \geq 0.5 \\ \text{3P}\,, & \text{if } P_{\text{3P}} \geq 0.5 \\ \text{4P}\,, & \text{if } P_{\text{4P}} \geq 0.5 \\ \text{Undefined}\,, & \text{otherwise} \end{cases}$$

Benchmark 1 (4P)

$$Z' \to SS\,,\quad S \to AA \to b\bar{b}b\bar{b}\,,\quad M_{Z'} = 2.2\,\text{TeV}\,,\quad M_S = 80\,\text{GeV}\,,\quad M_A = 30\,\text{GeV}$$

Benchmark 2 (2P)

$$Z' \to AA\,,\quad A \to b\bar{b}\,,\quad M_{Z'} = 2.2\,\text{TeV}\,,\quad M_A = 80\,\text{GeV}$$

Benchmark 3 (4P)

$$Z' \to SS\,,\quad S \to WW \to q\bar{q}q\bar{q}\,,\quad M_{Z'} = 3.3\,\text{TeV}\,,\quad M_S = 200\,\text{GeV}$$

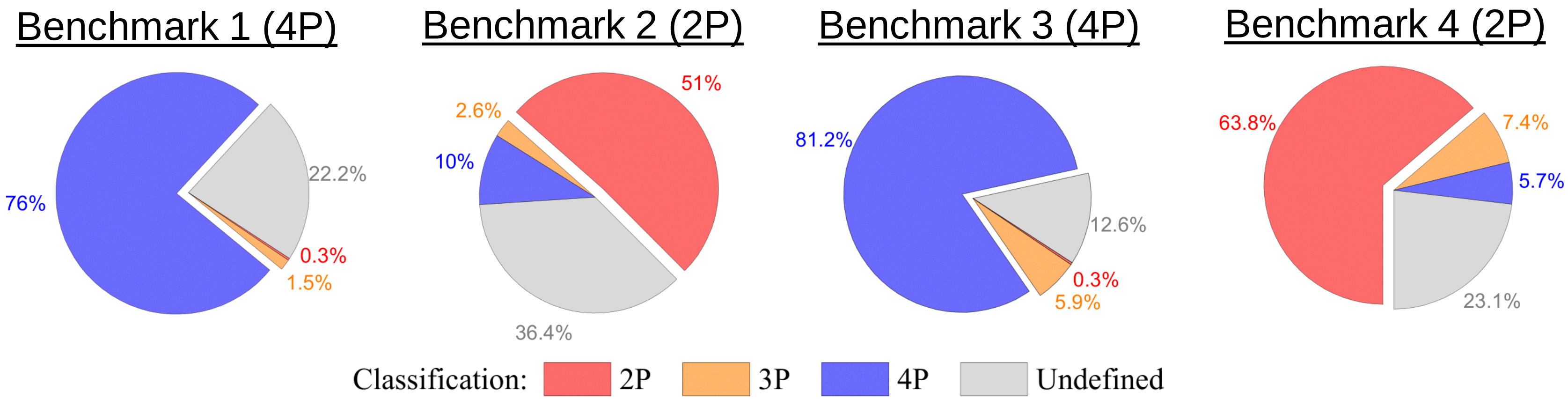
Benchmark 4 (2P)

$$Z' \to AA\,,\quad A \to u\bar{u}\,,\quad M_{Z'} = 3.3\,\text{TeV}\,,\quad M_A = 200\,\text{GeV}$$

IDENTIFICATION OF NEW PHYSICS SIGNALS

Using the prongness selection tagger, we apply the following classification criteria in the four benchmark examples below:

$$\begin{cases} 2\mathrm{P}\,, & \text{if } P_{2\mathrm{P}} \geq 0.5 \\ 3\mathrm{P}\,, & \text{if } P_{3\mathrm{P}} \geq 0.5 \\ 4\mathrm{P}\,, & \text{if } P_{4\mathrm{P}} \geq 0.5 \\ \text{Undefined}\,, & \text{otherwise} \end{cases}$$



- The fraction of correctly identified jets is several times larger than that of misidentified ones;
- Mistag rates can be further reduced by raising the value of the threshold that separates undefined jets from the classified ones.

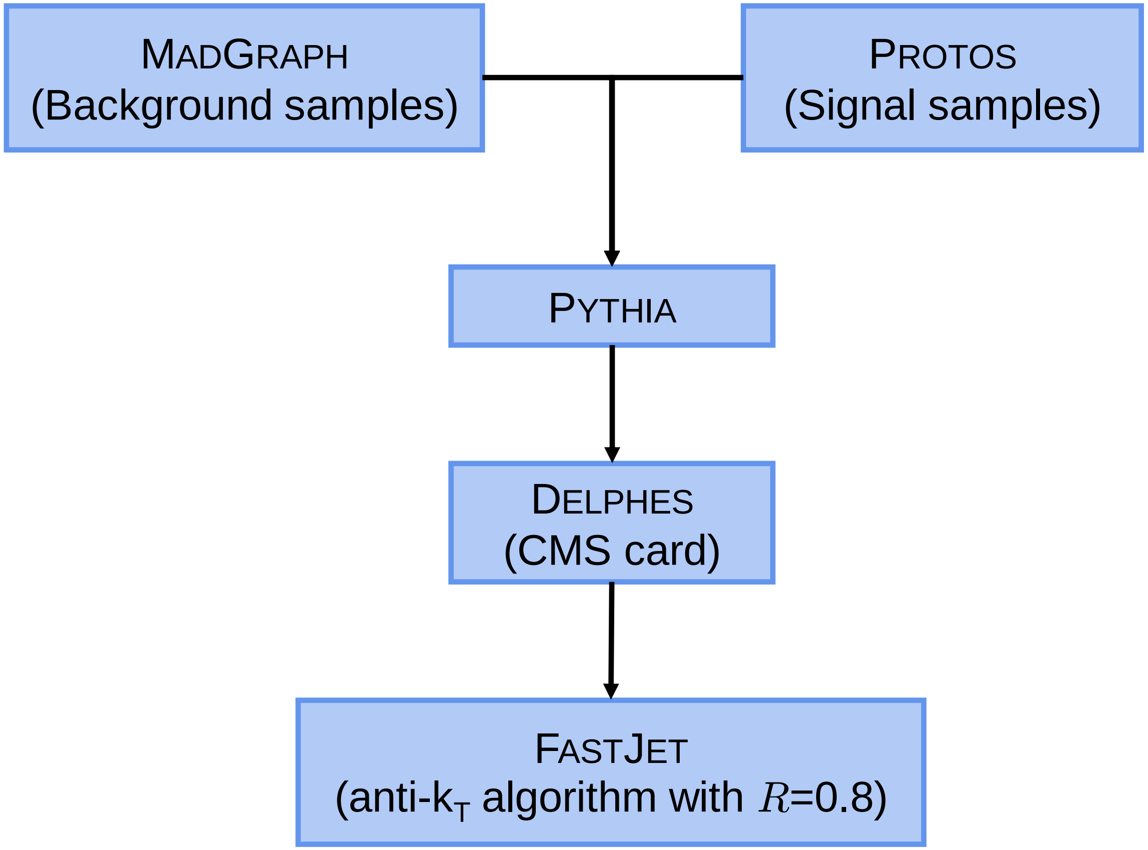
CONCLUDING REMARKS

- We introduced the method of **MUST** for multi-pronged jets;
- Taggers built upon MUST keep an excellent performance across a very wide m_J and p_T range;
- Our taggers are sensitive to any kind of multi-pronged jets, outperforming simple variables;
- Mass decorrelation can easily be implemented using the varying threshold method;
- MUST taggers can achieve good performances on signals for which they were not trained;
- The MUST concept can also be applied to selection taggers that can determine the prongness of signal jets.

Thank you!

Backup slides

EVENT SIMULATION AND RECONSTRUCTION



Parton-level event generation

MadGraph
(Background samples)

Protos
(Signal samples)

Hadronisation

Pythia

Detector simulation

Delphes
(CMS card)

Jet reconstruction

FastJet
(anti-k_T algorithm with R=0.8)

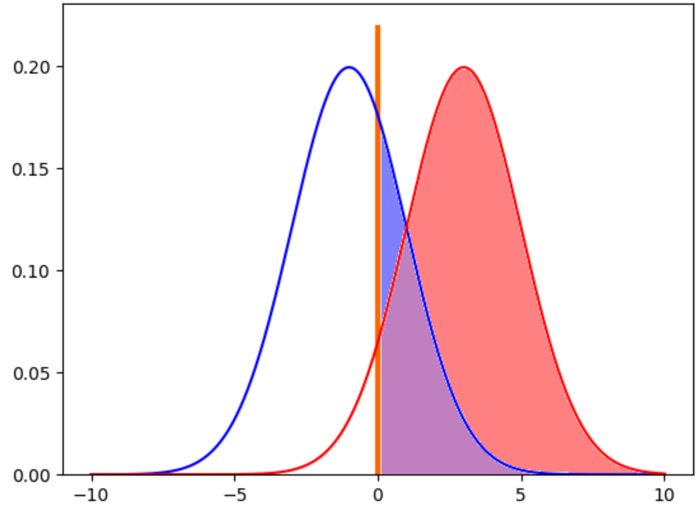
ROC CURVES

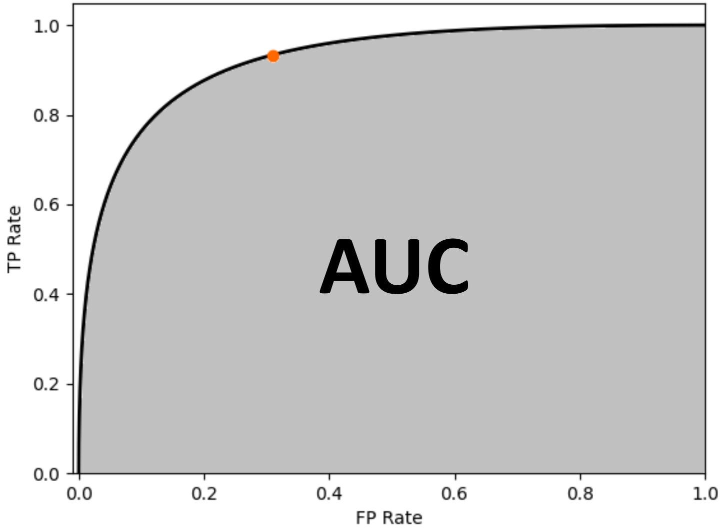
- In a binary classification task, there is always a **threshold** separating the two classes*;*

- The fraction of **True Positives** (**TP**) and **False Positives** (**FP**) for all possible thresholds defines the classifier's **ROC curve***;*

- The **Area Under the ROC curve** (**AUC**) is often used to evaluate the performance of the classifier (it assumes a value between 0 and 1).

Note: In our results, we represent the ROC curves on the plane (ε_{sig} , ε^{-1}_{bkg}). Considering Background and Signal events as being Negative and Positive, respectively, ε_{sig} = TP Rate and ε_{bkg} = FP Rate.

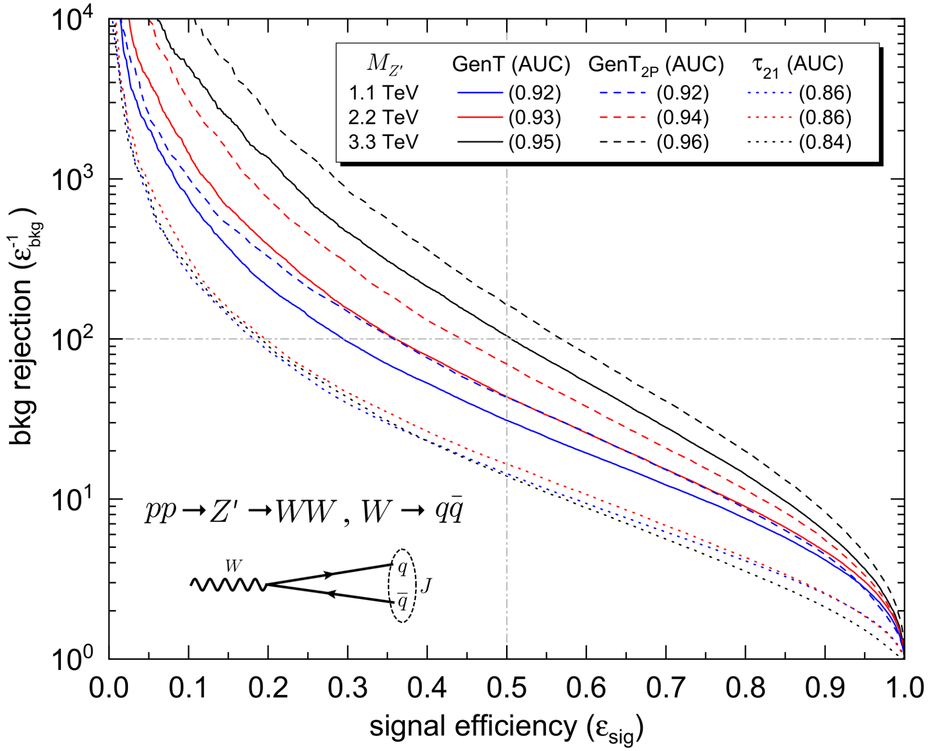


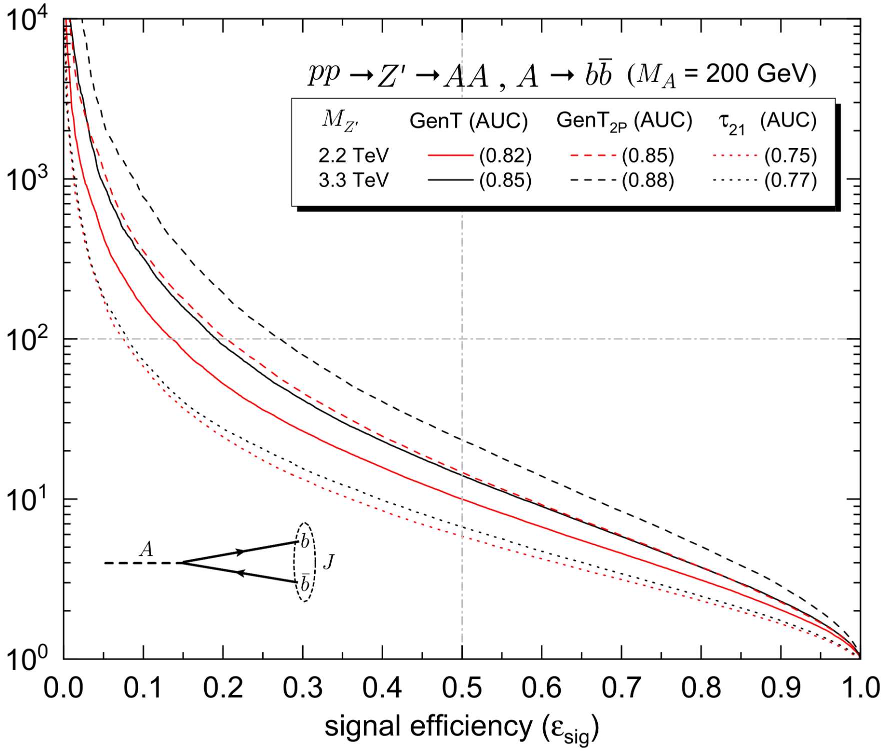


TAGGER PERFORMANCE (2P SIGNALS)

Background: Quark and gluon jets generated in $pp \to Zq$, $pp \to Zg$, with $Z \to \nu\nu$

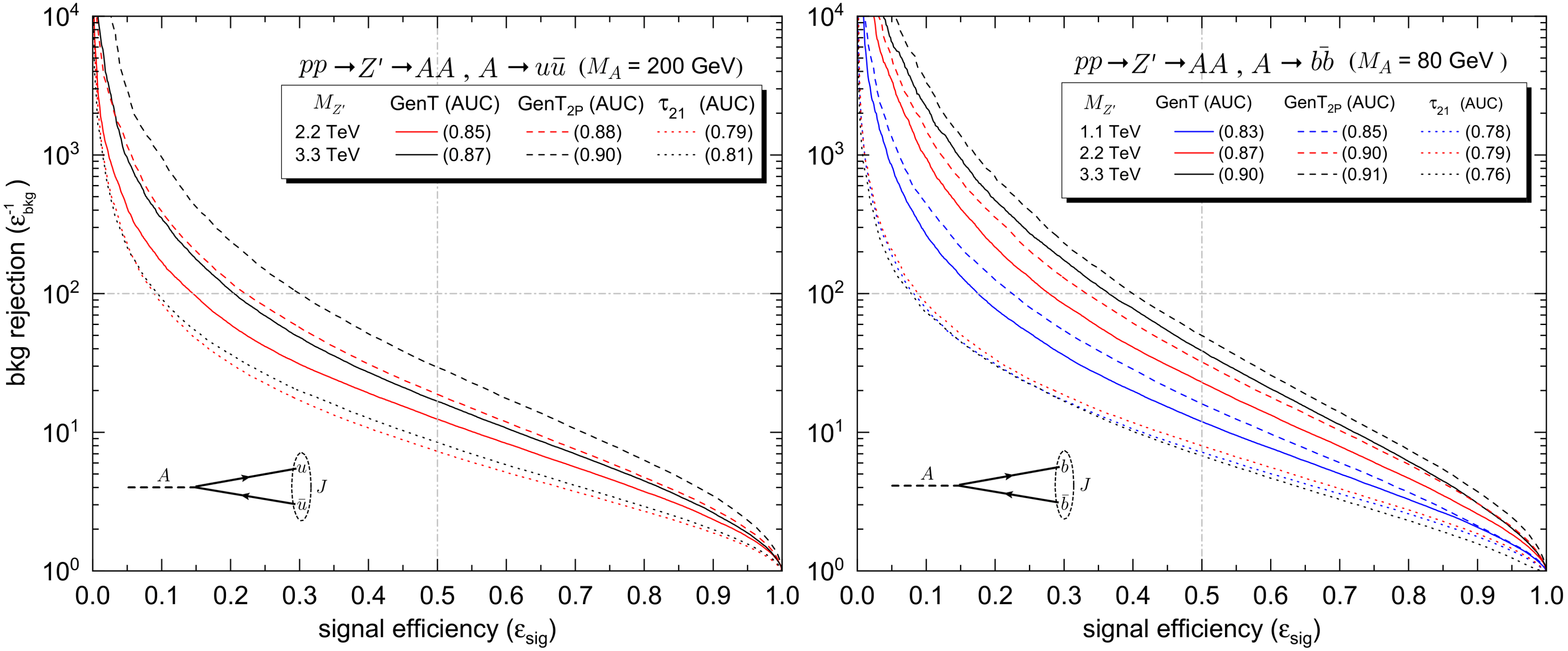
- $p_T \geq 0.5, 1.0, 1.5$ TeV for $M_{Z'} = 1.1, 2.2, 3.3$ TeV, respectively;



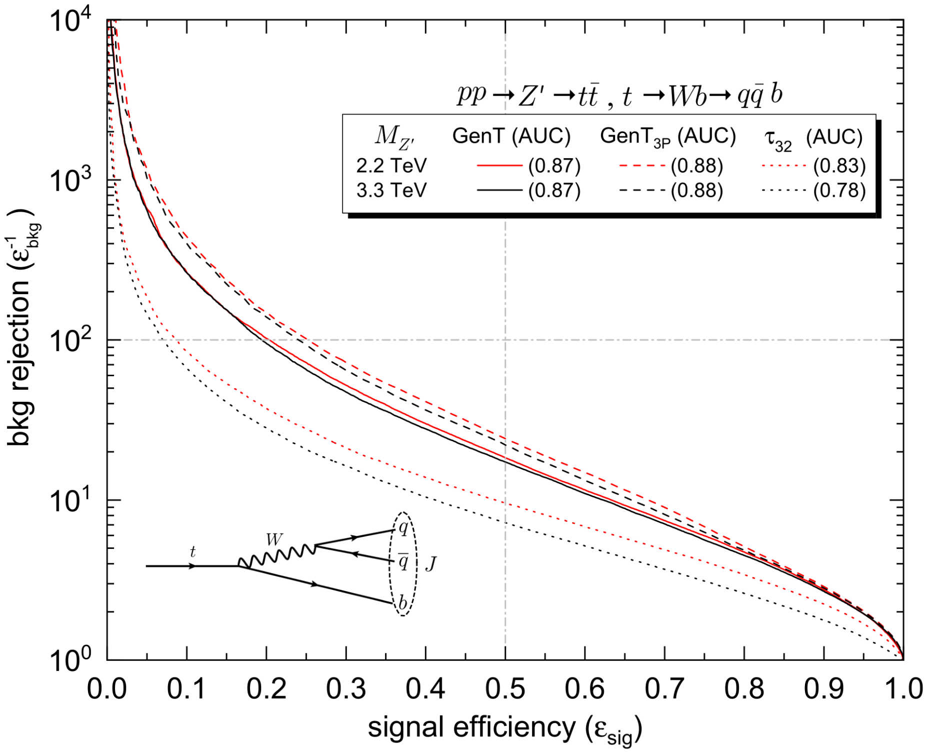


- In general, GenT and GenT$_{2P}$ perform better than the commonly used ratio τ_{21};
- The performance improves as $M_{Z'}$ increases.

TAGGER PERFORMANCE (2P SIGNALS)



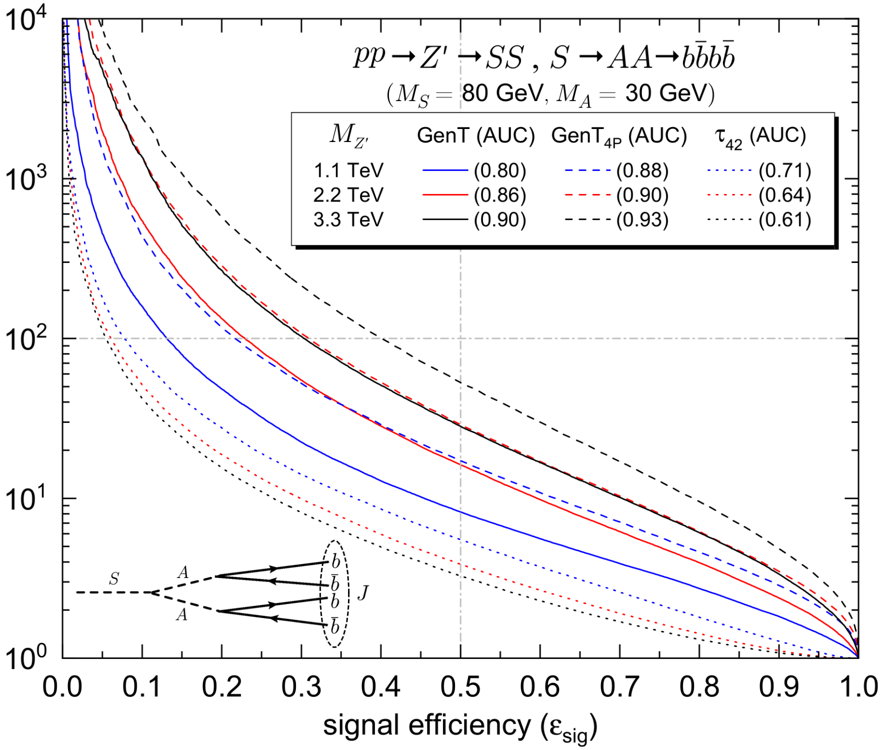
TAGGER PERFORMANCE (3P SIGNAL)

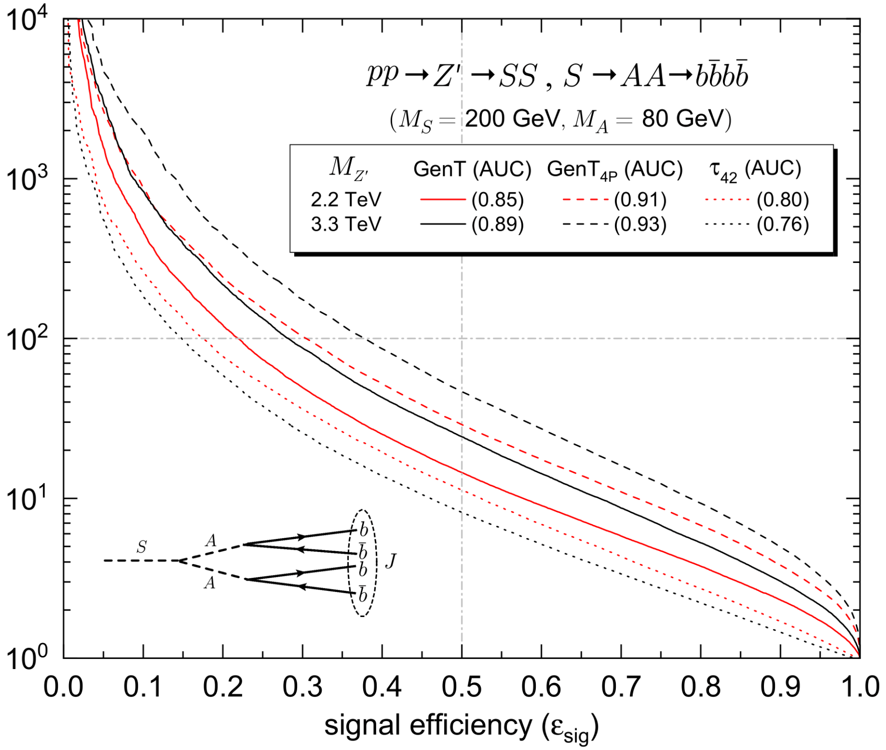


- Although GenT and GenT$_{3P}$ perform well on top quark jets and would not miss those signals, fully-dedicated top taggers perform better.

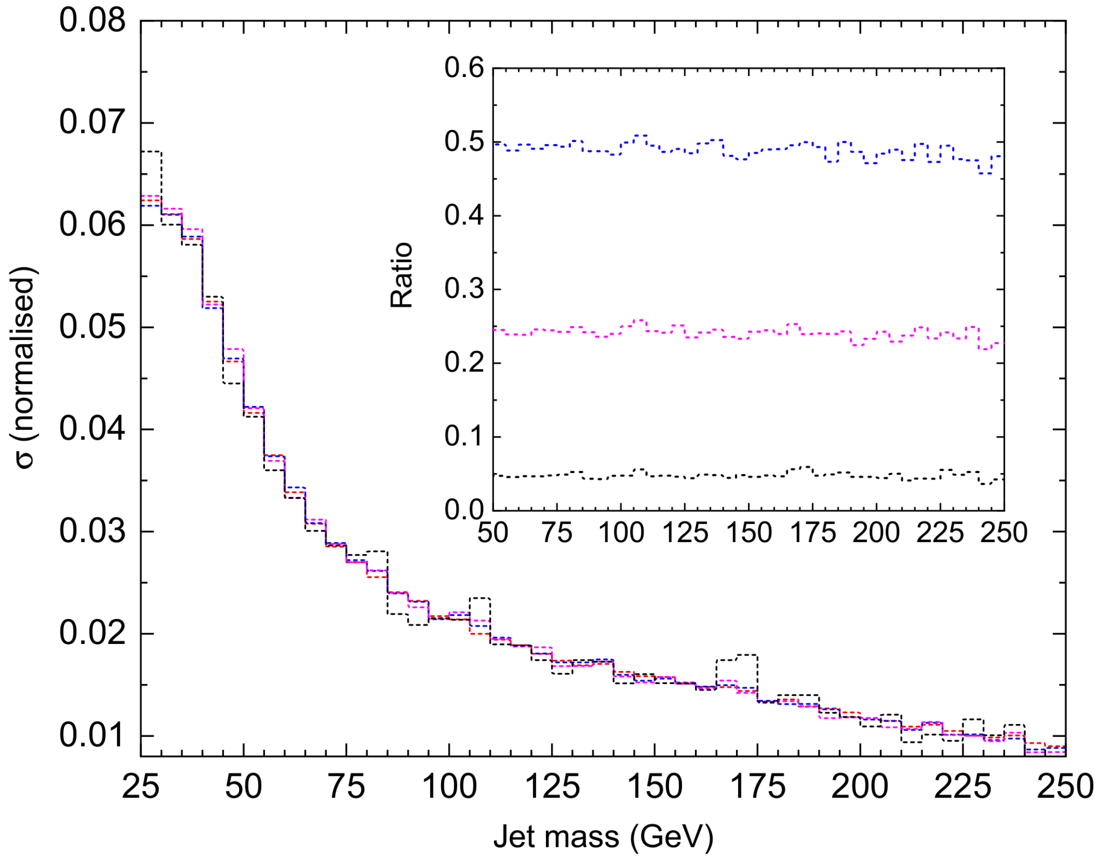
e.g. S. Macaluso, D. Shih; JHEP 10 (2018) 121

TAGGER PERFORMANCE (4P SIGNALS)





ANOTHER PLOT FOR MASS DECORRELATION



- Main plot – Normalised background distributions before and after cuts
- Inner plot – Ratios of distributions after/before cuts

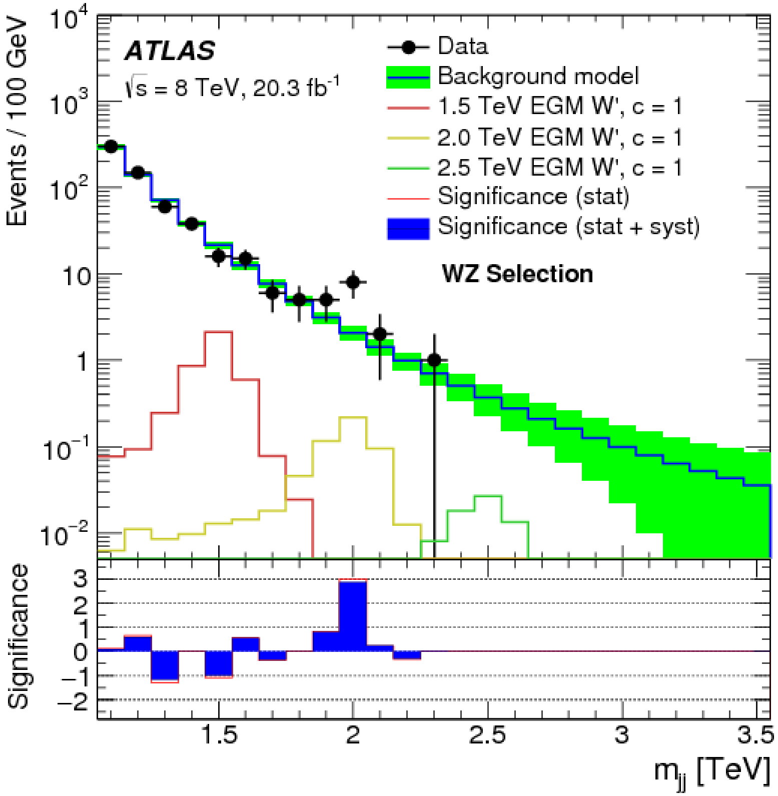
MORE ABOUT STEALTH BOSONS

- Stealth bosons are relatively light boosted particles with a cascade decay:

$$S \to AA \to q\bar{q}q\bar{q}$$

A particles can be weak bosons W, Z, a Higgs boson or new relatively light (pseudo-)scalars.

- Heavy resonances decaying into two such stealth bosons, or one plus a W/Z boson, may offer an explanation for small excesses found in hadronic diboson resonance searches near an invariant mass of 2 TeV (example on the right).

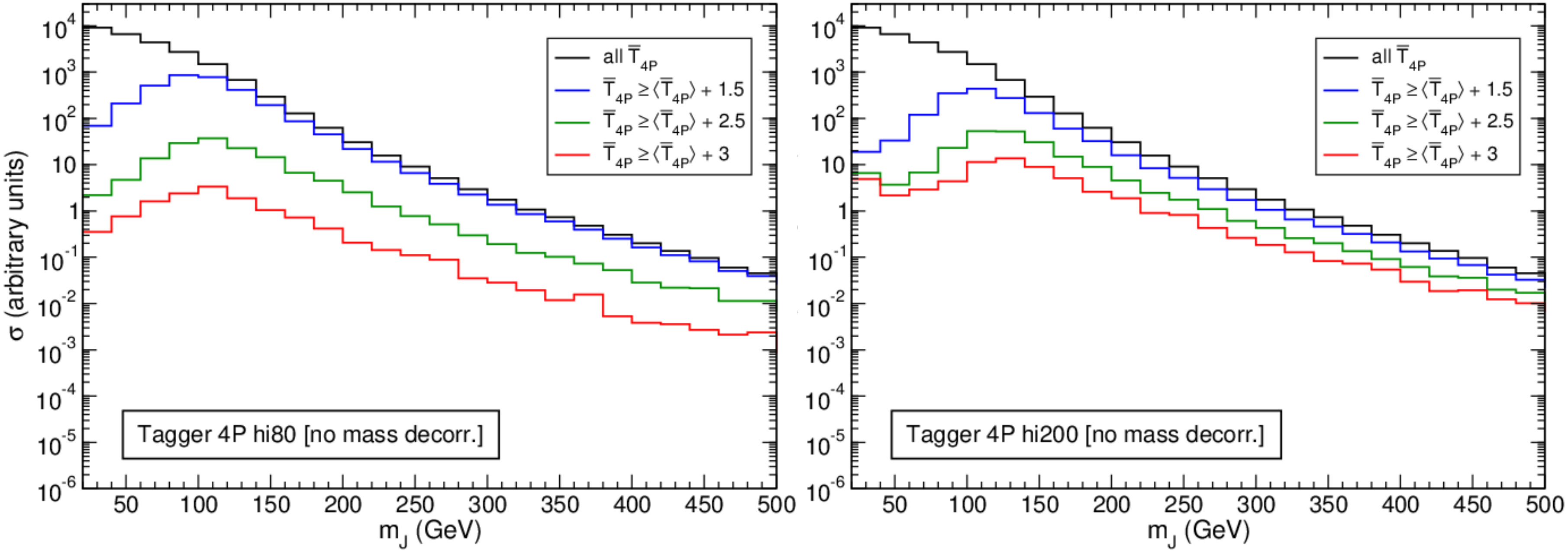


(For more information about this topic, see:
J. A. Aguilar-Saavedra; Eur. Phys. J. C 77, 10 (2017) 703)

G. Aad *et al*. [ATLAS]; JHEP 12 (2015) 55

TAGGERS WITHOUT MASS DECORRELATION

Jet mass spectrum for QCD background produced by 4P taggers with no prior mass decorrelation



J. A. Aguilar-Saavedra, B. Zaldívar, Eur. Phys. J. C 80, 6 (2020) 530

- The peak-like structure produced near 100 GeV is not in any case related to the design mass interval.

N-SUBJETTINESS OBSERVABLES

- The N-subjettiness observable $\tau_N^{(\beta)}$ is a measure of the radiation about N axes in the jet, specified by an angular exponent $\beta > 0$,

$$\tau_N^{(\beta)} = \frac{1}{p_T} \sum_{i \in \text{jet}} p_{T\,i} \min\left\{R_{1i}^{\beta}, R_{2i}^{\beta}, \cdots, R_{Ni}^{\beta}\right\}$$

Transverse momentum of particle i in the jet

Angular distance between particle i and axis N in the jet

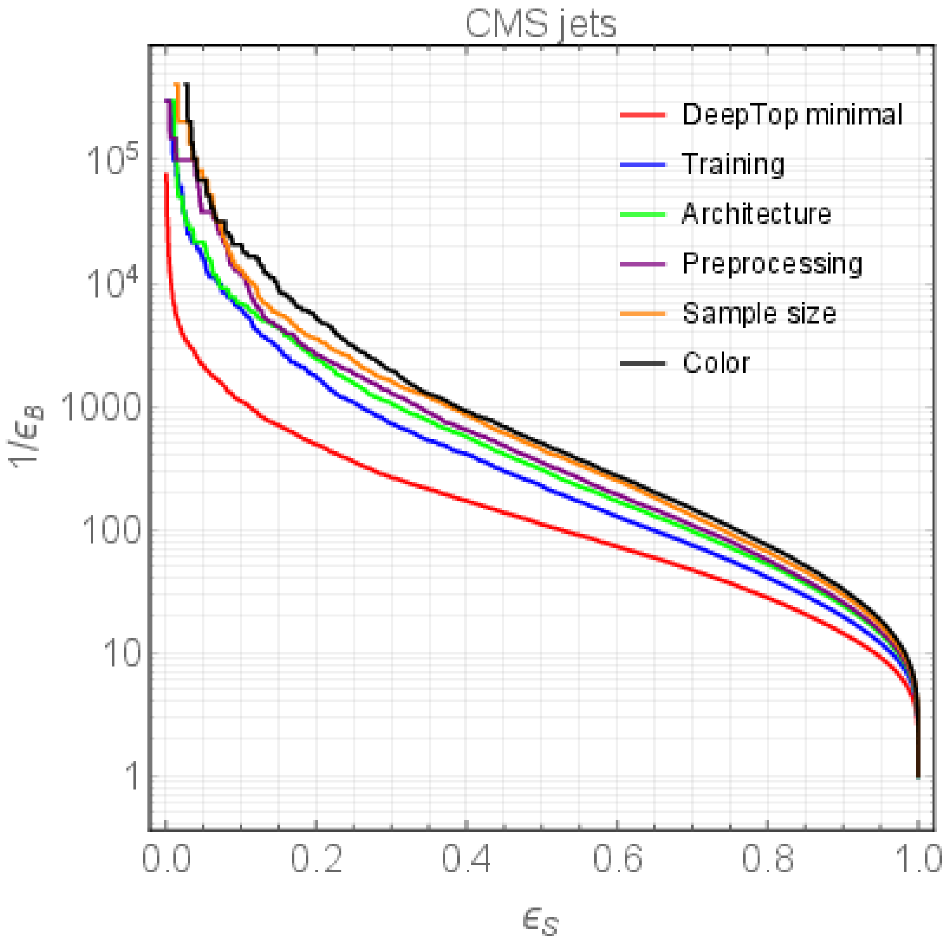
- The coordinates of the M-Body phase space can be defined by (M – 1) transverse momentum fractions and (2M – 3) angles, so we need **(3M – 4)** N-subjettiness observables to completely specify the coordinates of that space:

$$\left\{\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(0.5)}, \tau_2^{(1)}, \tau_2^{(2)}, \cdots, \tau_{M-2}^{(0.5)}, \tau_{M-2}^{(1)}, \tau_{M-2}^{(2)}, \tau_{M-1}^{(1)}, \tau_{M-1}^{(2)}\right\}$$

(For more information about this topic, see K. Datta, A. Larkoski; JHEP 06 (2017) 73)

TOPDEEP TAGGER

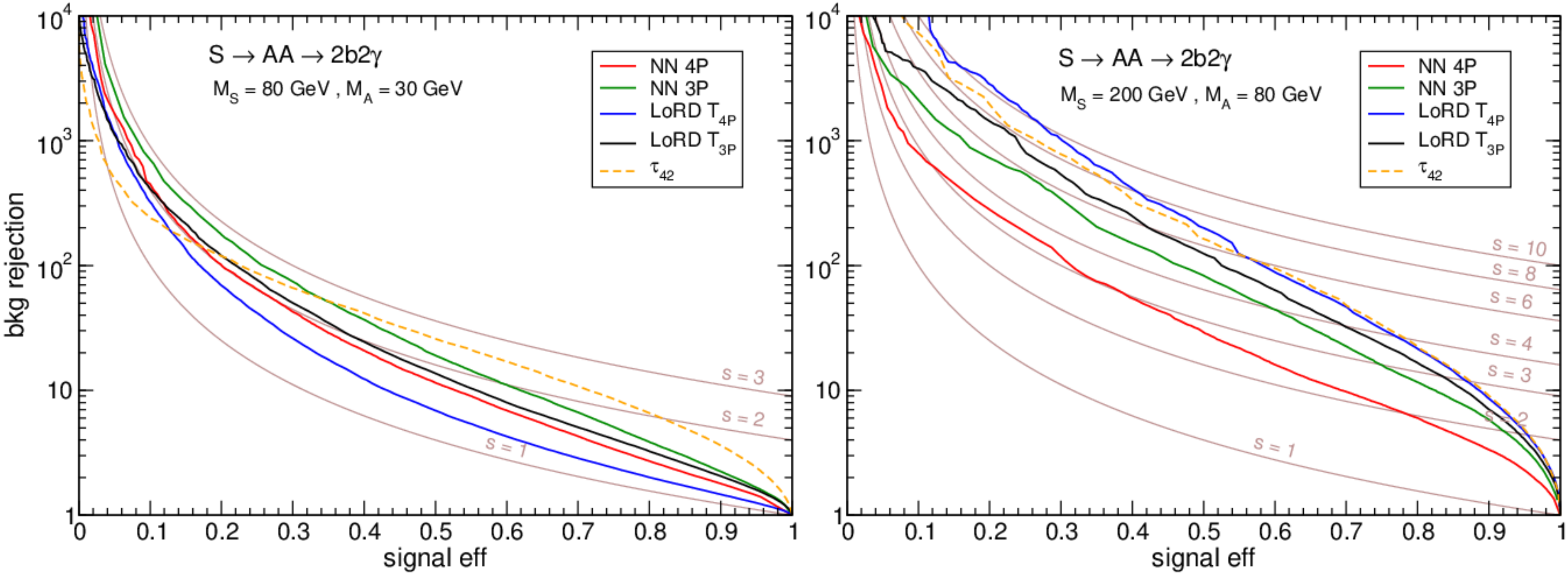
Performance of DeepTop tagger (after several improvements) discriminating top quark jets



S. Macaluso, D. Shih; JHEP 10 (2018) 121

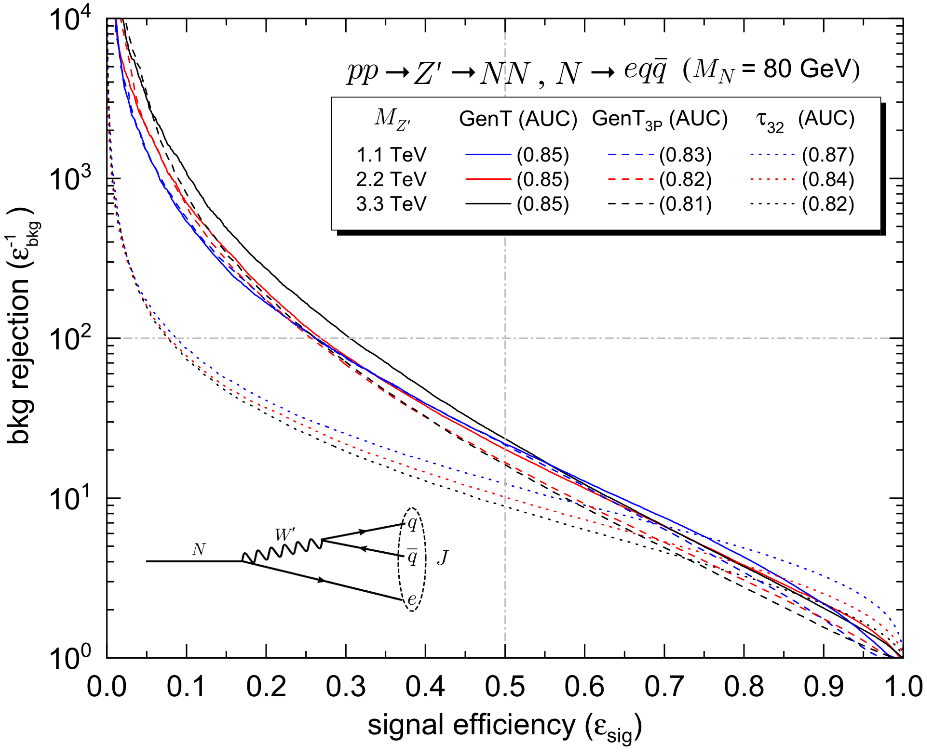
LoRD

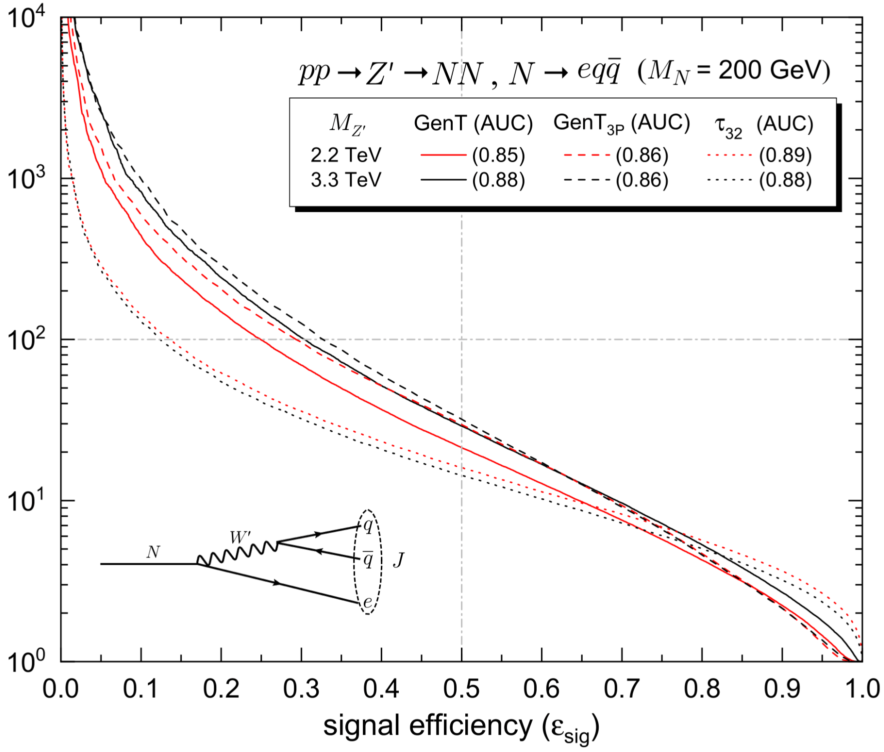
Performance of Logistic Regression Design (LoRD) classifying jets with two hard photons



J. A. Aguilar-Saavedra, B. Zaldívar, Eur. Phys. J. C 80, 6 (2020) 530

JETS NOT USED TO TRAIN MUST-TAGGERS





JETS NOT USED TO TRAIN MUST-TAGGERS