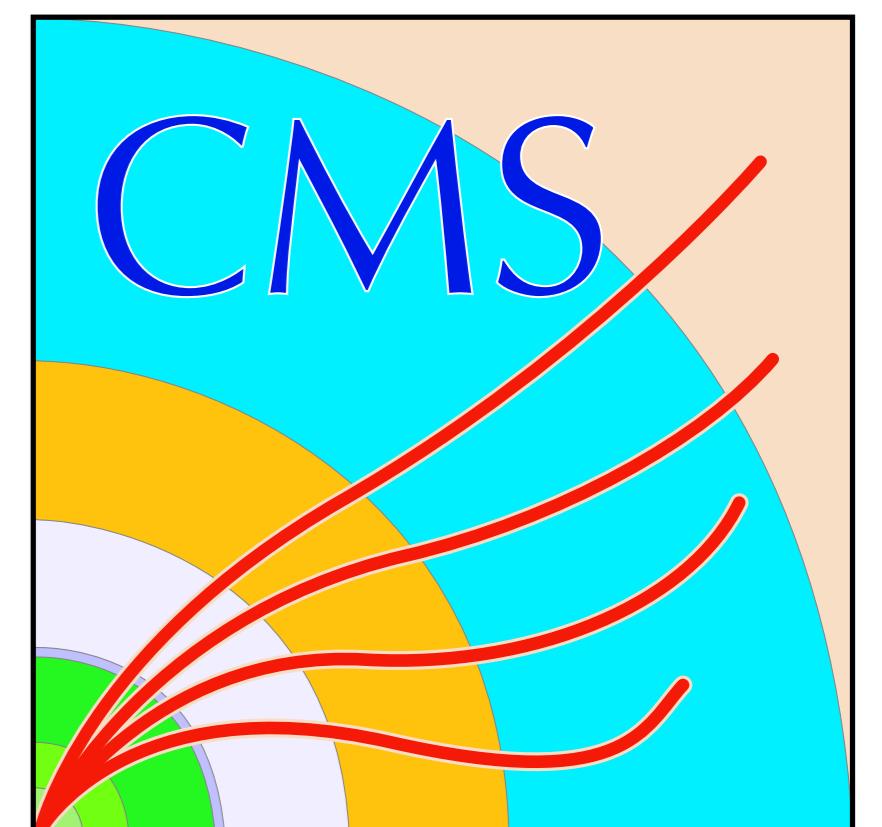


# Testing TMVA software in b-tagging for the search of MSSM Higgs bosons at the LHC

T. Lampén, F. García, A. Heikkinen, P. Kaitaniemi, V. Karimäki,  
M. J. Kortelainen, S. Lehti, T. Lindén, and L. Wendland



Helsinki Institute of Physics, P.O.B. 64, FIN-00014 University of Helsinki, Finland

We demonstrate the use of a ROOT Toolkit for Multivariate Data Analysis (TMVA) in tagging b-jets associated with heavy neutral MSSM Higgs bosons at the LHC. The associated b-jets can be used to extract Higgs events from the Drell-Yan background, for which the associated jets are mainly light quark and gluon jets. Here we use b jets from tt events as signal jets. Background discriminating power is shown for several TMVA classifiers.

## Introduction

We demonstrate TMVA in tagging b-jets associated with heavy neutral MSSM Higgs bosons. Background discriminating power is demonstrated for several TMVA classifiers. TMVA working in transparent factory mode guarantees an unbiased performance comparison, as all classifiers are evaluated with the same training and test data.

## TMVA Key Features

TMVA is an easy-to-use machine learning environment for sophisticated multivariate classifiers.

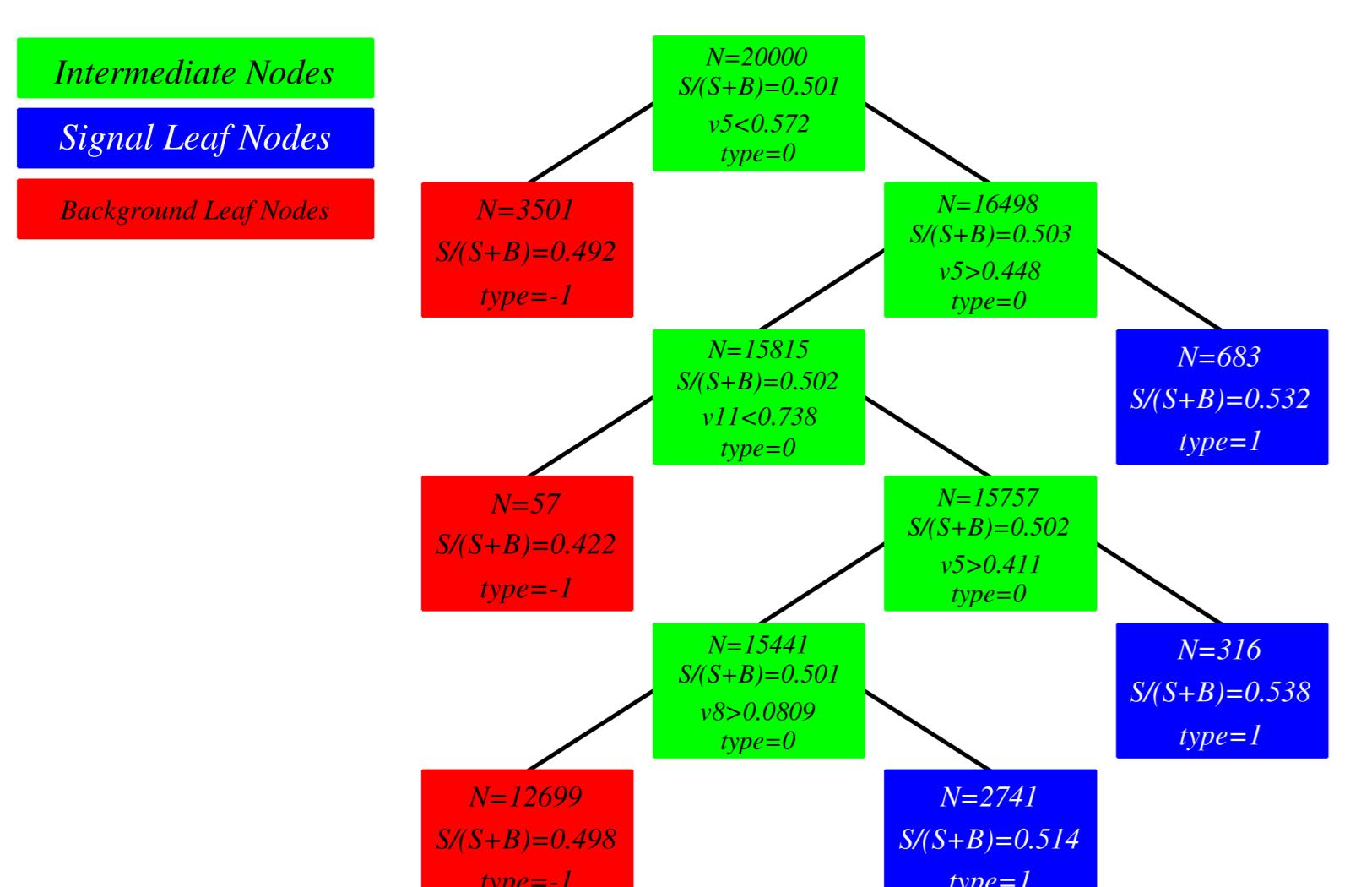
Key features:

- Training and testing of classifiers with user data
- Performance of classifiers evaluated with same training and test data within the same execution job
- Visualization scripts with GUI
- Individual data pre-processing for each classifier (decorrelation, principal component analysis)
- Printout of tabulated benchmark results
- Smooth efficiency vs. background rejection curves (ROC curves) in ROOT output file
- Generates C++ code independent of ROOT and TMVA for standalone use of trained classifiers

## Classifiers

Following classifiers were used:

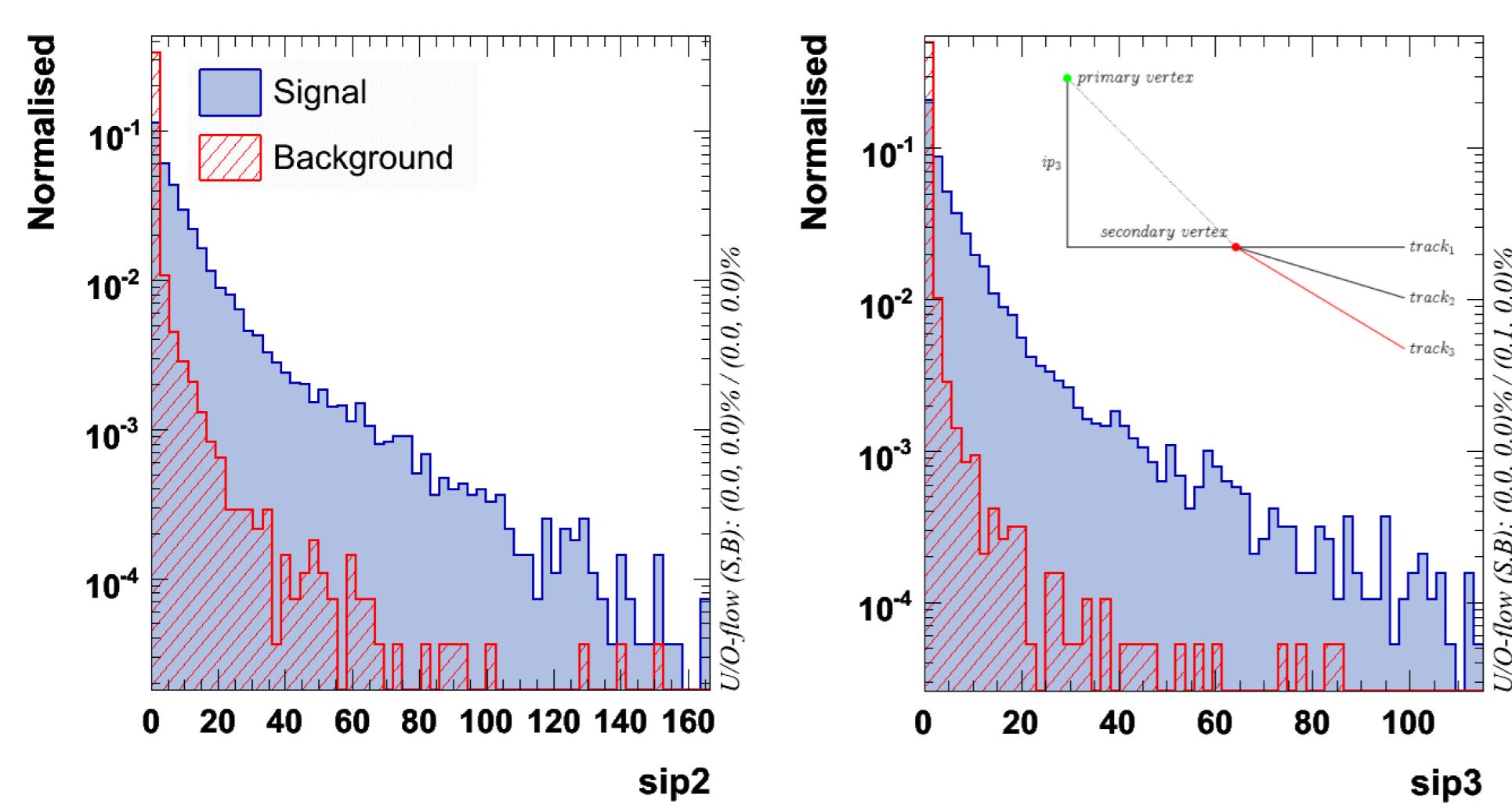
- Rectangular cuts
- Projective likelihood estimator (PDE)
- Multidimensional likelihood estimator with range search (PDERS)
- K-nearest neighbours (KNN)
- H-matrix
- Fisher linear discriminants
- Function discriminant analysis (FDA)
- Artificial neural network (ANN)
- RuleFit
- Support Vector Machine (SVM)
- Boosted/bagged decision tree (BDT)



Example of a decision tree of the BDT classifier.

## Data Description

- Signal and background jets generated with TopREX (tt events) and with PYTHIA (Z/ $\gamma^*$  events)
- Events selected with generator level Monte-Carlo truth
- Signal consists of 162k b jets from tt and 588k light



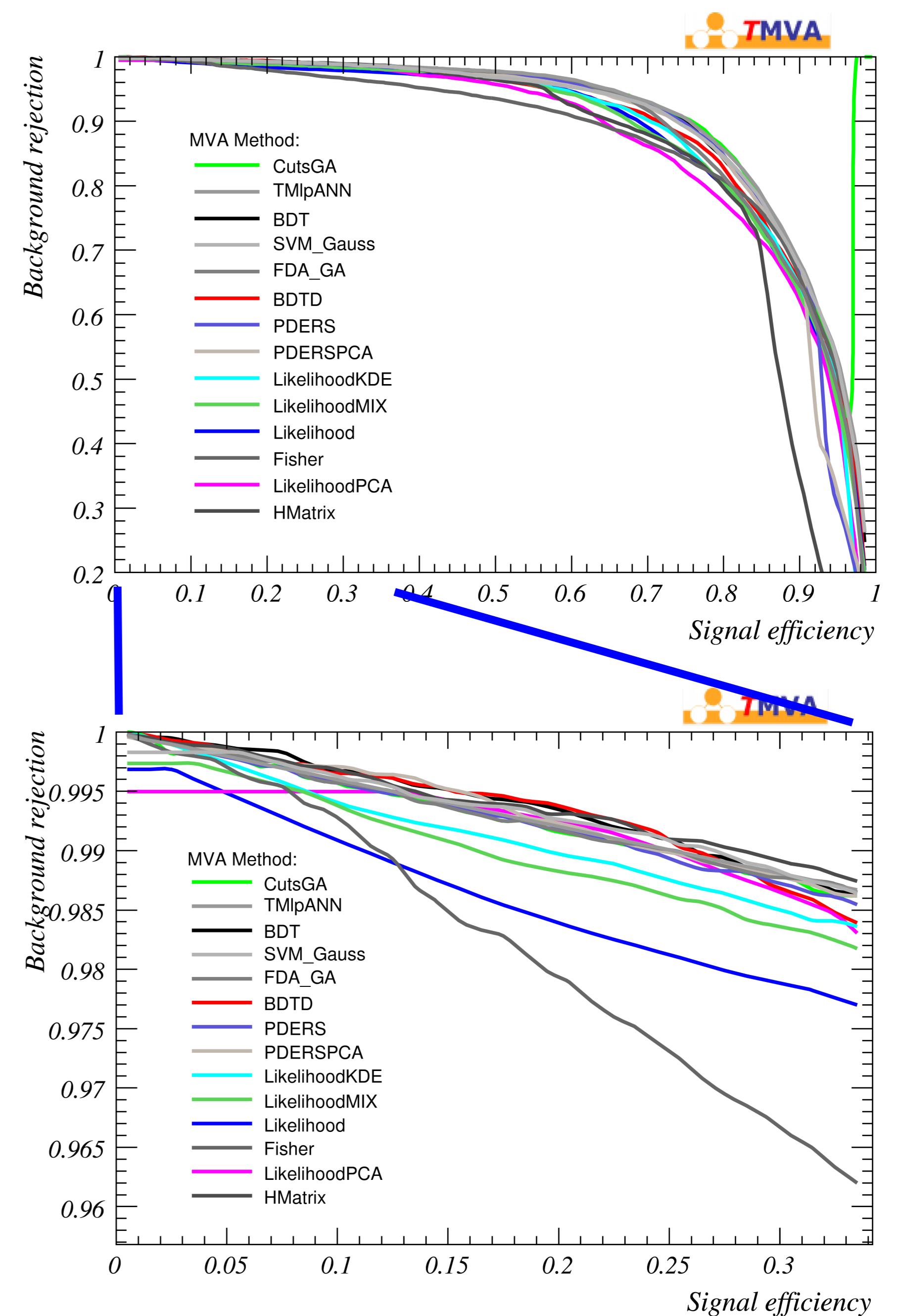
Distributions of impact parameter significances  $\sigma_{ip,2}$  and  $\sigma_{ip,3}$  for two leading tracks.

quark and gluon jets from Z/ $\gamma^*$  events, 10k+10k events used for training

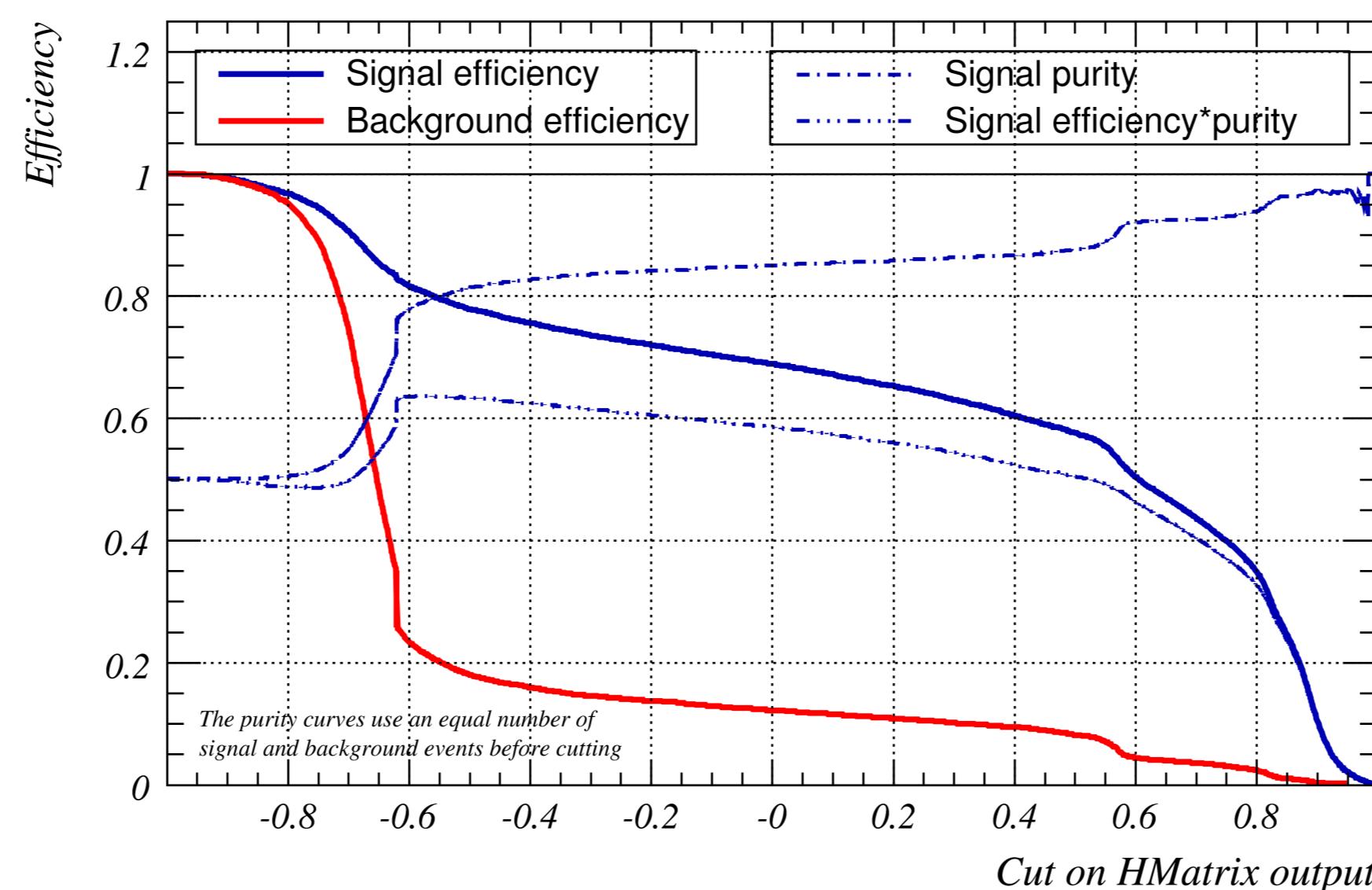
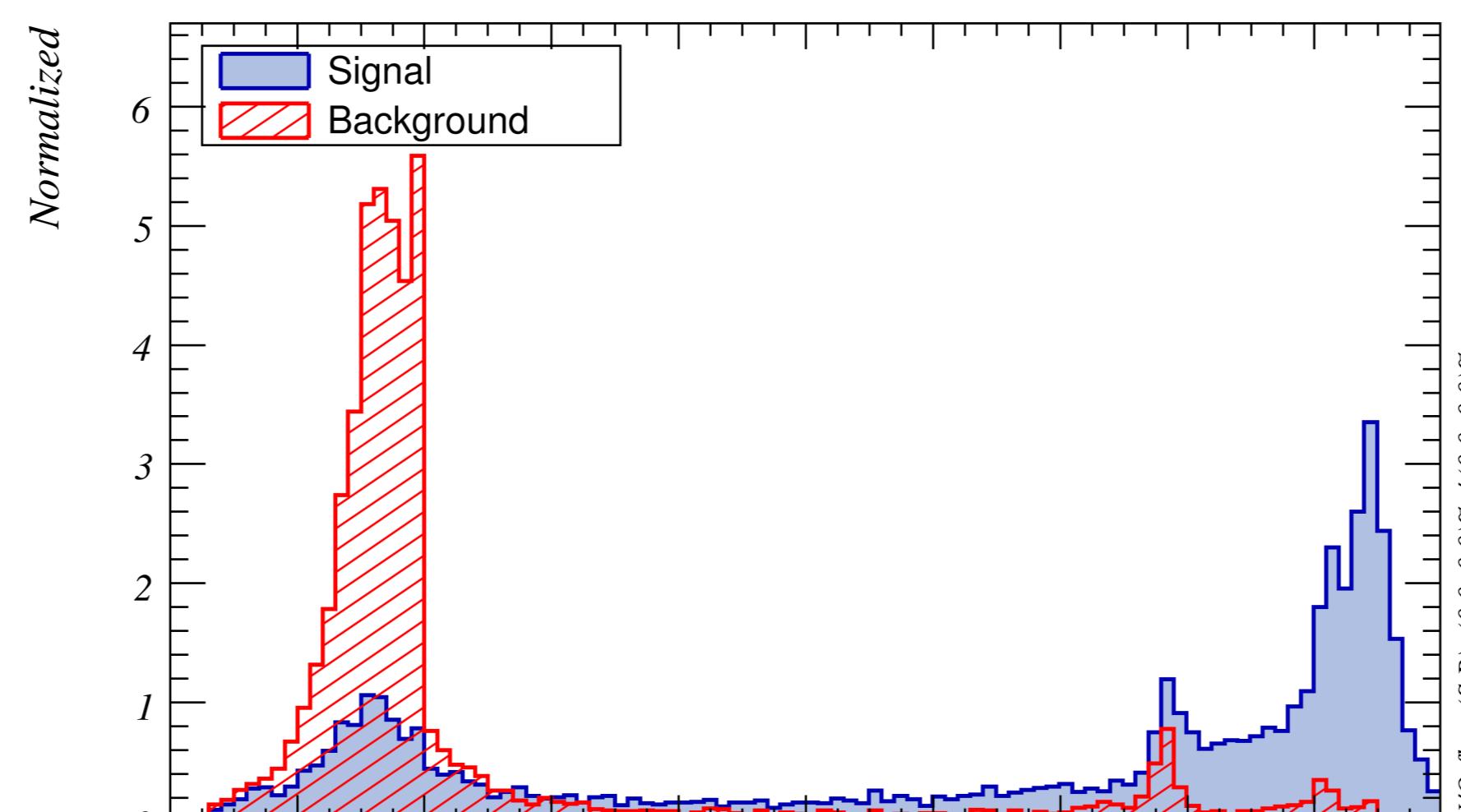
- Event reconstruction based on official [4] CMS digitized datasets with pile-up (3.4 min. bias events per crossing for  $L=2 \times 10^{33} \text{ cm}^{-2} \text{s}^{-1}$ )

- Fast Ethernet network for remote management
- Cluster management with NPACI Rocks Cluster Distribution v 4.1 software
- Sun Grid Engine (SGE) batch queue
- ROOT version 5.14/00d and TMVA 3.8.5 were used

## Results



| Background rejection level: | Parameters $\sigma_{ip,1-3}$ |                | Optimal set of parameters |                |
|-----------------------------|------------------------------|----------------|---------------------------|----------------|
|                             | 1%                           | 10%            | 1%                        | 10%            |
| Cuts (GA)                   | $27 \pm 1$                   | $74.4 \pm 0.5$ | $49 \pm 1$                | $74.4 \pm 0.5$ |
| PDE                         | 23                           | 70.0           | 40                        | 75.3           |
| PDERS                       | 27                           | 74.2           | 40                        | 76.5           |
| HMatrix                     | 26                           | 62.6           | 34                        | 73.7           |
| Fisher                      | 12                           | 61.3           | 37                        | 73.7           |
| FDA (GAMT)                  | 17                           | 74.1           | 39                        | 77.0           |
| nTMpANN                     | 28                           | 74.9           | 40                        | 78.9           |
| BDT                         | 29                           | 70.0           | 51                        | 80.1           |
| RuleFitJF                   | 32                           | 74.9           |                           |                |
| RuleFitTMVA                 | 31                           | 74.4           |                           |                |
| SVM                         | 18                           | 74.4           | 35                        | 76.7           |
| Track counting b-tagging    | 27                           |                |                           |                |



a) Output of H-Matrix classifier for test data  
b) H-Matrix classifier efficiencies for different cut values

## Computing

- Ametisti, a 64-bit 1.8/2.2 GHz AMD Opteron cluster with 260 CPUs in 130 computational nodes with 2/4 GB RAM was used
- A Gb/s network for nodes communication
- Dedicated Gb/s network for traffic of shared NFS system

## Conclusions

- Our preliminary results, at  $Bg=1\%$ , indicate improved classification power compared with [4, 5] and reference algorithm
- TMVA makes a large number of MVA methods accessible for easy comparison
- Particularly we like the C++ code generation feature
- Some classifiers suffer still from instabilities, yet TMVA shows potential for LHC data analysis

## References

- [1] J. Stelzer, TMVA - Toolkit for Multivariate Data Analysis, CHEP 2007
- [2] A. Hocker et al., TMVA - Toolkit for Multivariate Data Analysis, arXiv:physics/0703039
- [3] TMVA homepage, <http://tmva.sourceforge.net>
- [4] CMS Physics Technical Design Report, Volume II, CERN/LHCC 2006-021 CMS TDR 8.2 June 2006
- [5] A. Heikkinen and S. Lehti, Tagging b jets associated with heavy neutral MSSM Higgs boson, Nuclear Instruments and Methods A 559 (2006) 195-198