Feature ranking

Showcase of subtraction method for ttH(H->bb) classification

Paul Glaysher (DESY)
Sitong An (DESY summer student), Judith Katzy (DESY)

18 April 2019
3rd IML Workshop

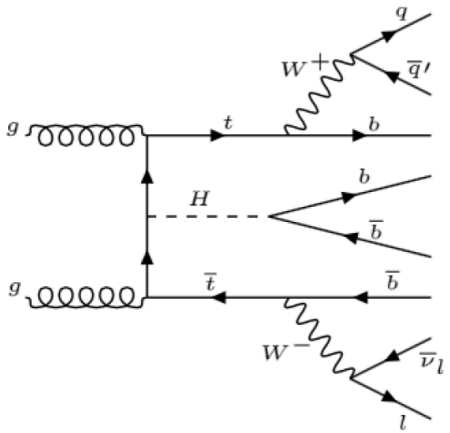
HELMHOLTZ ASSOCIATION

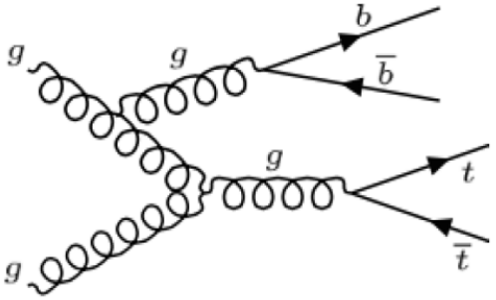
ATLAS EXPERIMENT

DESY

Intro

- For the example of a classification BDT for ttH(H->bb) vs ttbar+b-jets we will present different methods of ranking the relative importance of training features.
  - Test subtractive method: start with all variables and remove the least important
- The classification problem and event selection are inspired by the ATLAS ttH(H->bb) 1712.08895 paper.
  - Using open-data MC with Delphes simulation





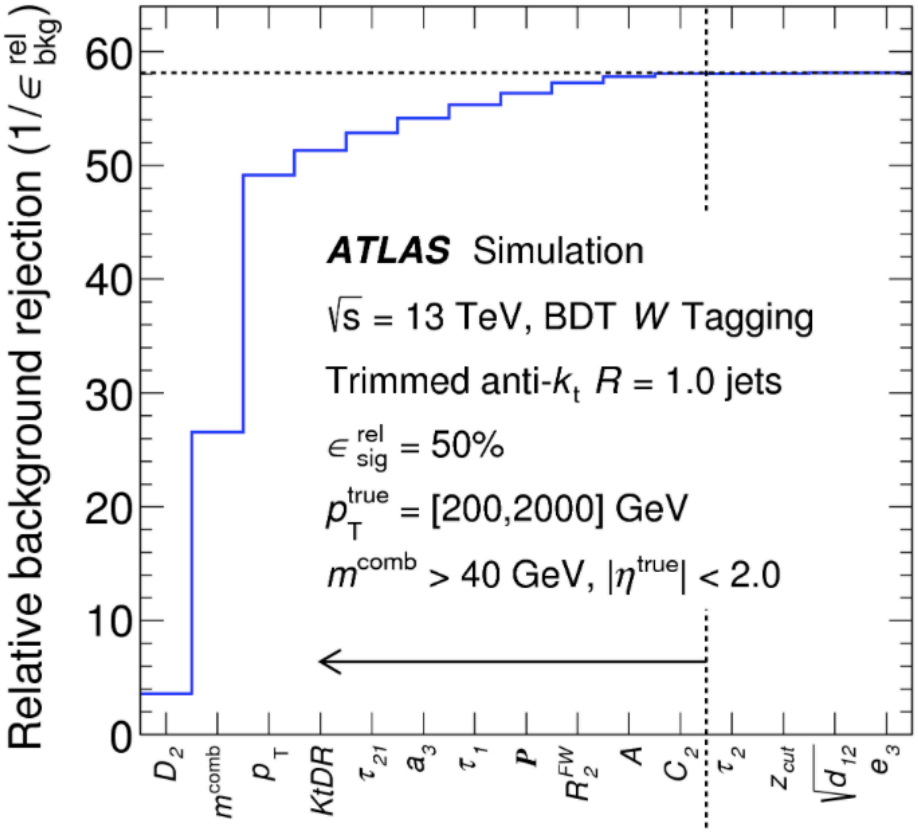
Motivation

- Feature ranking can help reduce unnecessary dimensionality
- Quicker run time and optimisation
- Improve insight into physical importance of the few selected variables
- Focus efforts of validating the modelling of inputs (time consuming)
- Help understand model response to different MC generators
- Not an issue for BDTs, but arbitrarily large number of inputs can compromise other learning algorithms.

- How to select the best N variables to use in the training?
  - Question of which training features are most important in the classification has no unique answer, particularly when they are highly correlated.

Example of additive ranking

- ATLAS Top and W tagger CERN-EP-2018-192
- Sequentially add variable that gives largest increase in performance
- The set variables that reaches saturation in performance (within stat. uncertainty) is selected
- Ranking complexity scales with number of variables n as $O(n^3)$ (assuming product of #trainings and #variables/training)
- Not clear if additive method correctly ranks correlated variables, e.g two individually useless variable that have separating power only when used together



Data sets and BDT setup

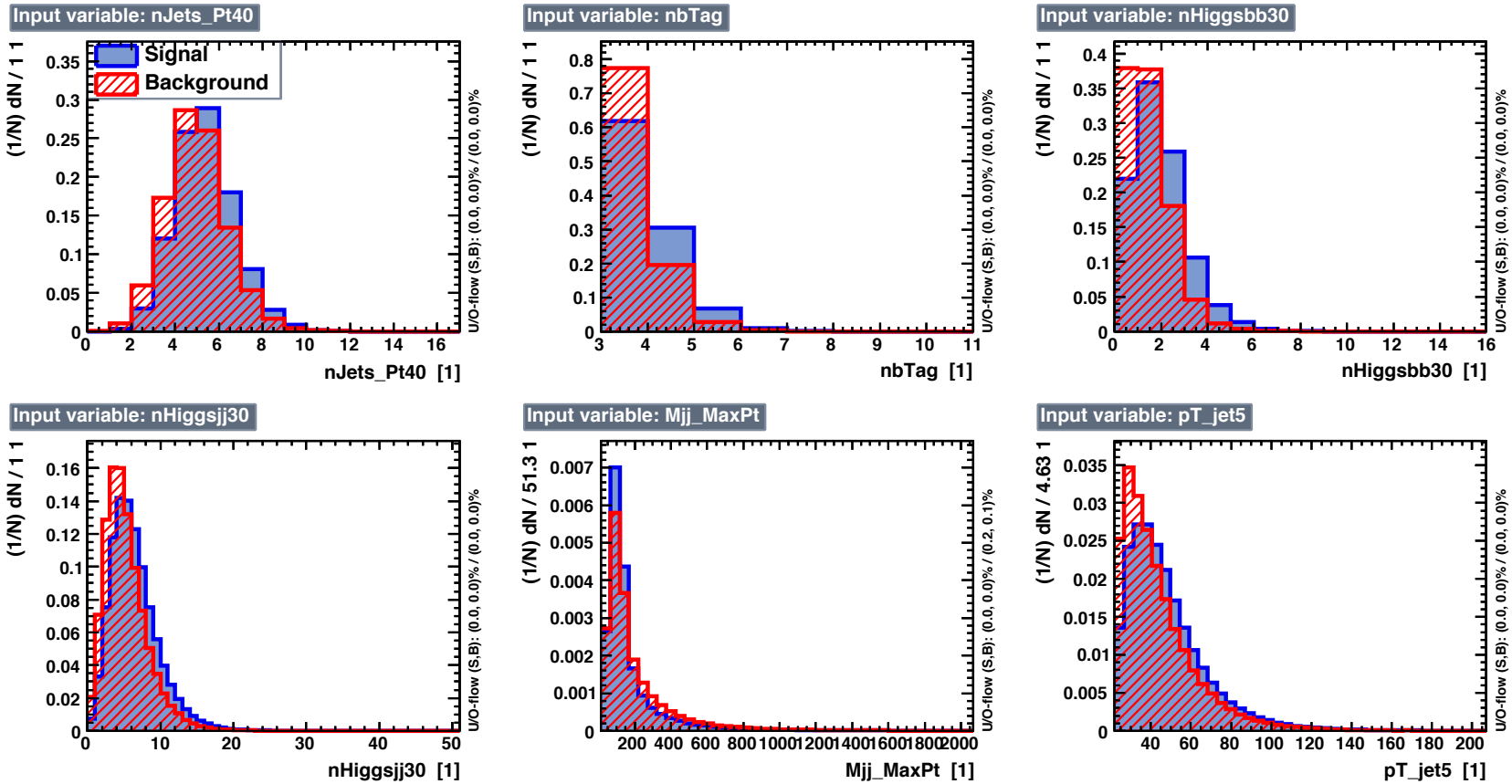
- Open data MC:

  tev13pp_mg5_ttbar_jet_ MadGraph/HW6. 2M events
  tev13pp_mg5_ttbar_bjet_ MadGraph/P6 10M events
  tev13pp_mg5_ttH MadGraph/HW6 13M events
  from https://hepsim.jlab.org
  The ttbar+jet and ttbar+bjet background samples are orthogonal, weighted by their cross section.

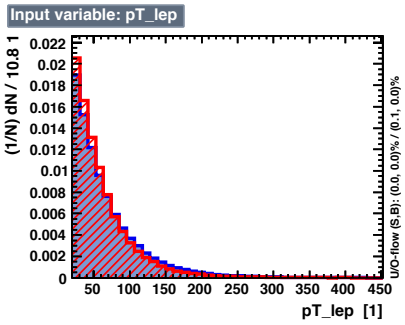
- Delphes simulation with atlas-like geometry https://cp3.irmp.ucl.ac.be/projects/delphes/
- Event selection, type of variables chosen similar to single lepton channel of ATLAS-CONF-2016-080 (some cuts loosened to gain stats)
  - 1 lepton with pt> 20 GeV and ≥5 jets with pt>25 GeV
  - ≥3 b-jets, with 70% WP, b-efficiency, light/c-rejection is parameterised according to JHEP08(2018)089
  - selects: 700k ttH signal and 275k tt+jets background events
  - Train on 2/3, test on 1/3 of events
- TMVA implementation of BDT:
  - 400 trees, MaxDepth=5, AdaBoostBeta=0.15, nCuts=80

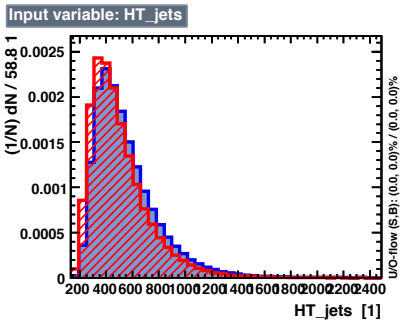
Training variables

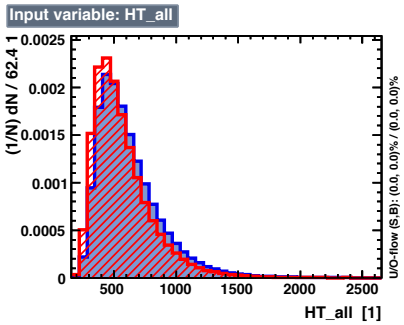
- Choice of variables inspired by reference paper, but also added additional ones and left those out that could not easily be reproduced
- 39 variables are computed, from which 26 are considered which have at least 1% separation in signal vs background shapes

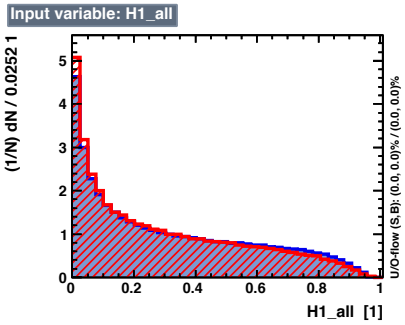


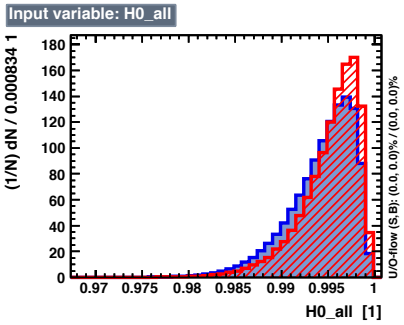
Training variables

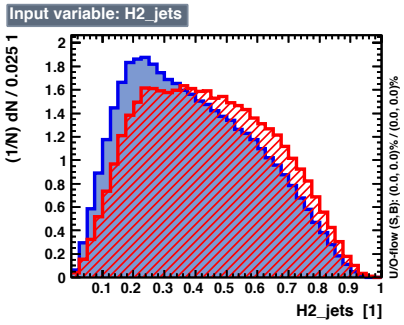




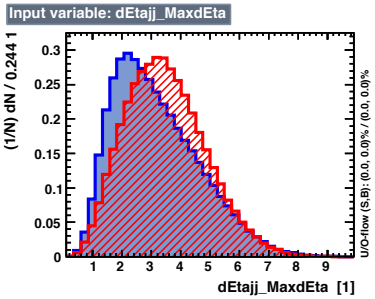


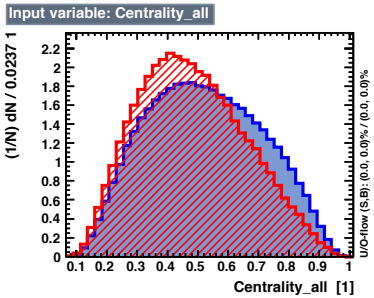


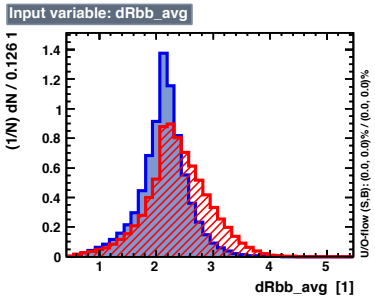


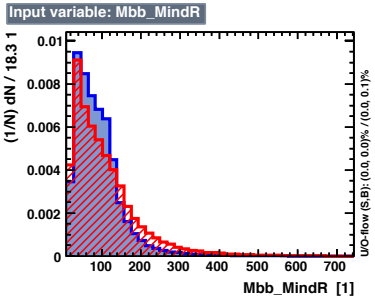


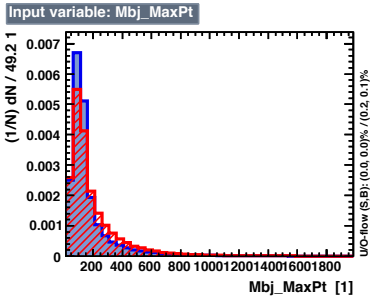
Mainly high level quantities.
Invariant mass and dR e.g., probably high redundancy among them.

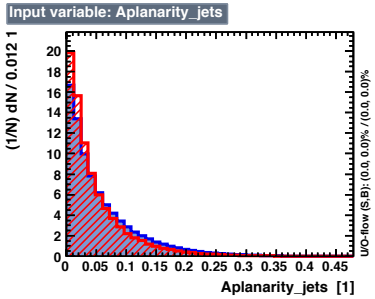




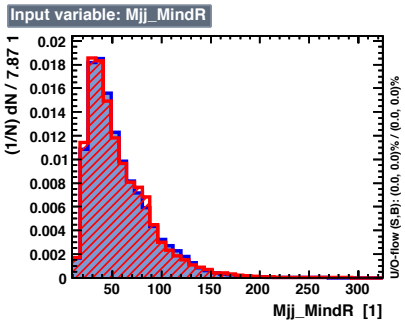


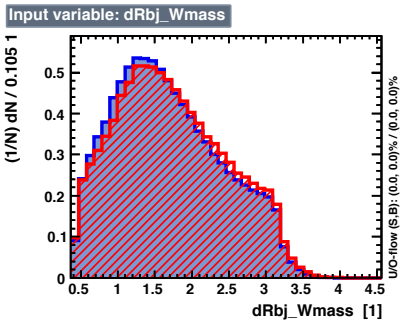


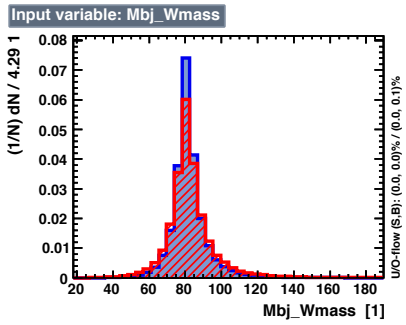


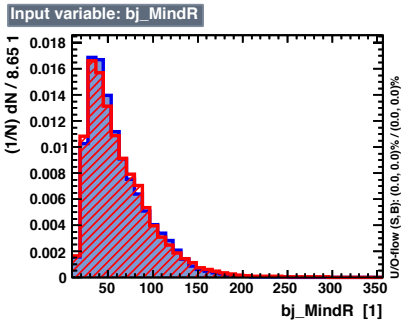


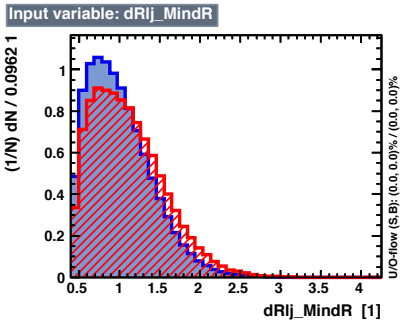
Training variables

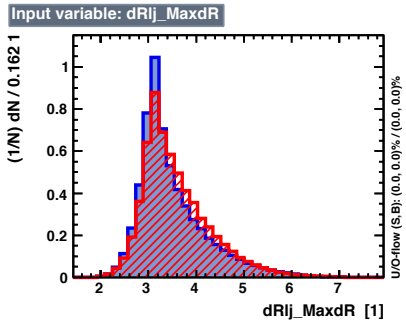


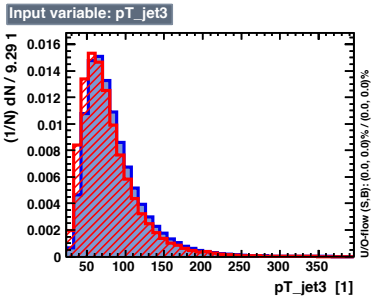


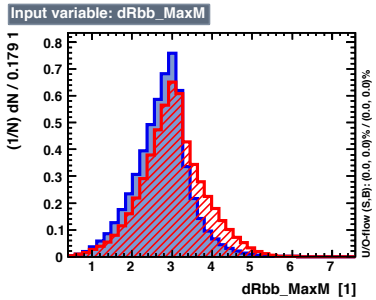


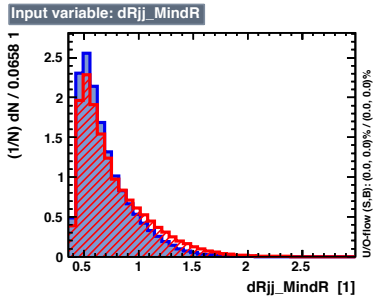


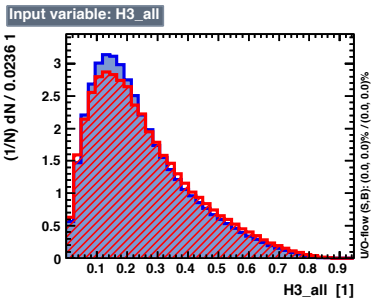


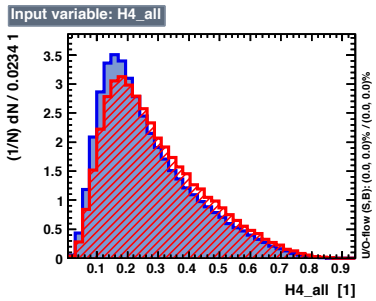


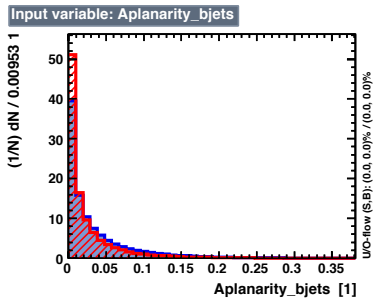




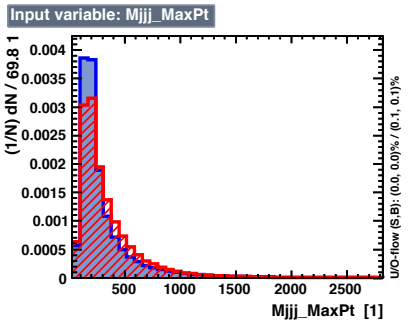


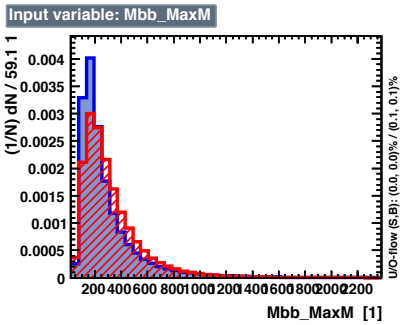


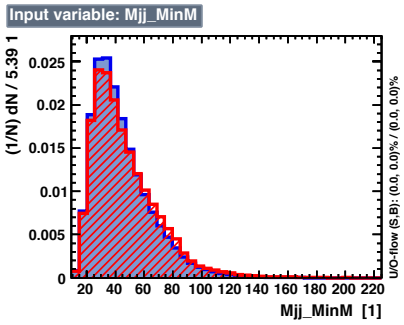


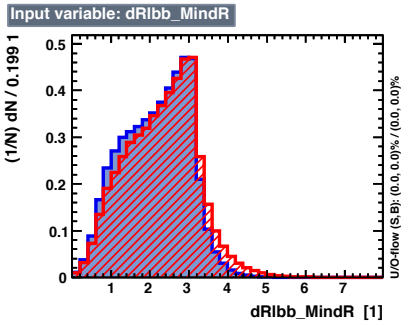


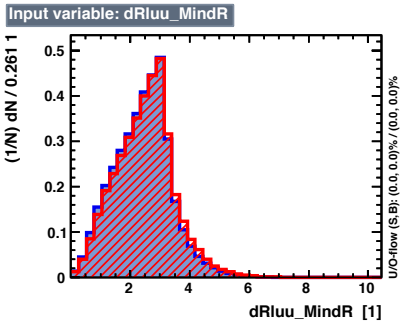
Training variables

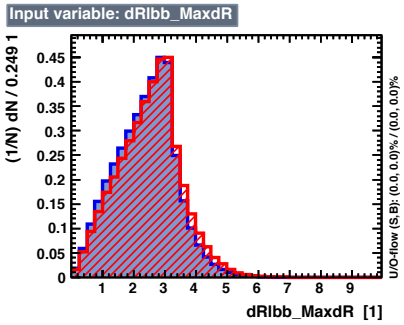


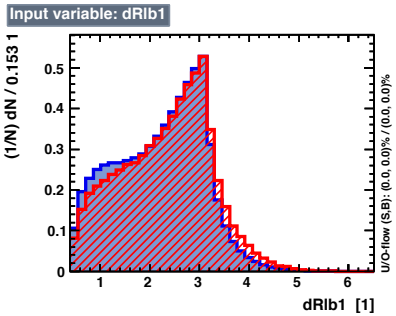


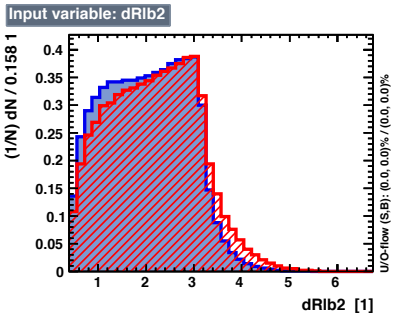


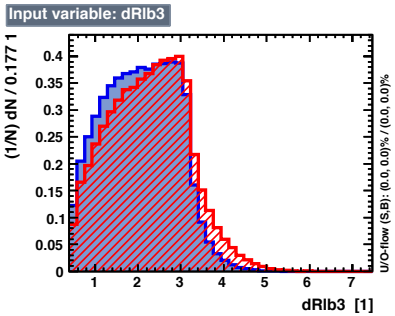










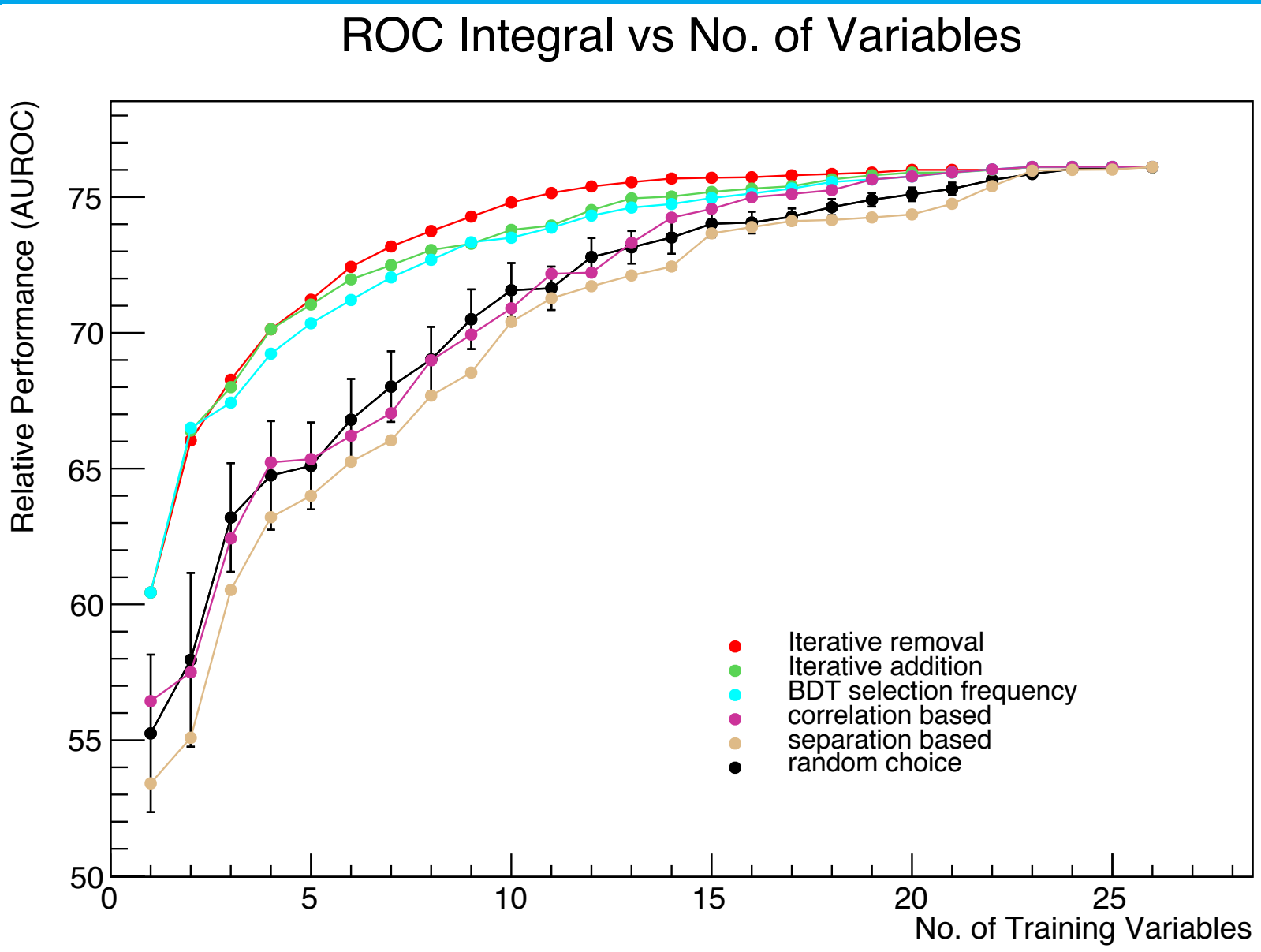


Ranking Methods tested

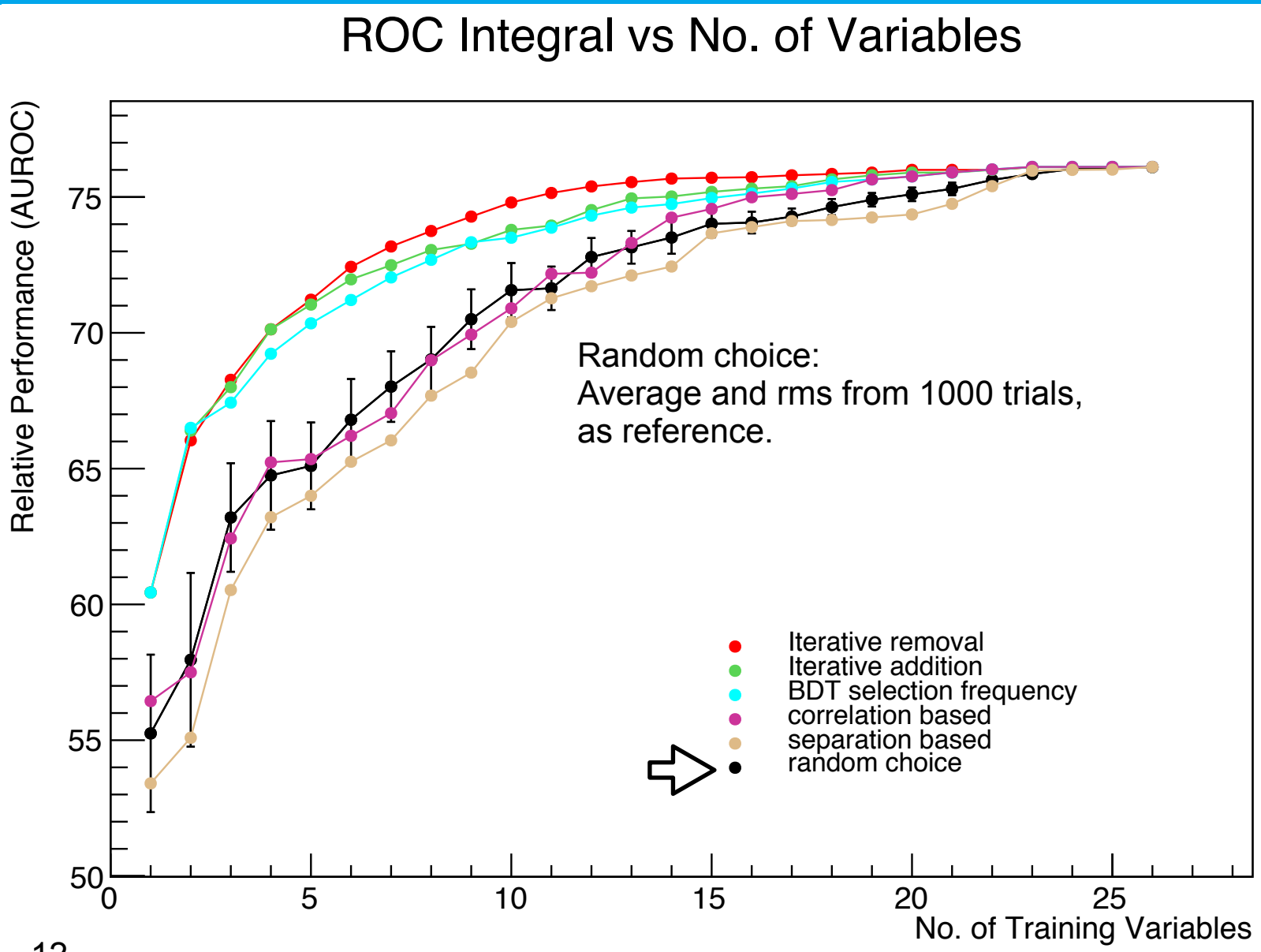
- Iterative addition: start with n=1, take best training of all n options. Then take best option of adding one more from remaining n-1 variables, etc. Ranking complexity scales as $O(n^3)$.
- Iterative removal: start with training on all variables and remove iteratively remove the one that degrades the performance the least, scales as $O(n^3)$.
  - Hypothesis: better consideration of variables that only add to performance in combination with others.
- Correlation based: rank the variables based on their correlation to the BDT score computed with all variables. Computationally cheap, scales as O(n).
- BDT selection frequency 'TMVA ranking': train once on all variables, rank by how often a variables provided the optimal decision in the BDT, scales as O(n).
- Separation based: rank by overlap of signal vs background shapes. Only method that establishes ranking without performing any training.
- Random choice: serves as reference, use a random subset of the variables. Repeat and average over 1000 trials.

Here performance is measured as integral of the ROC curve.

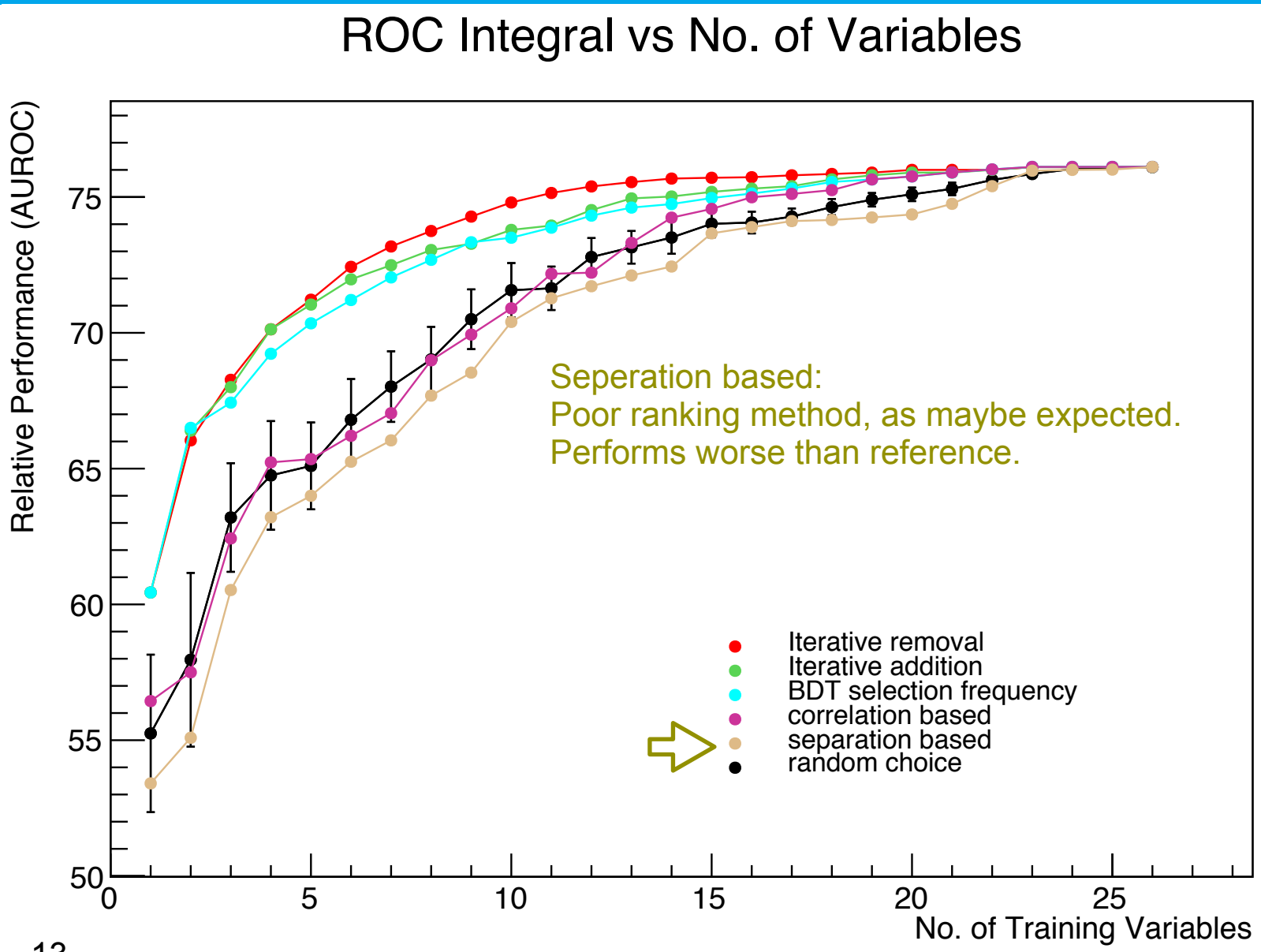
Ranking results



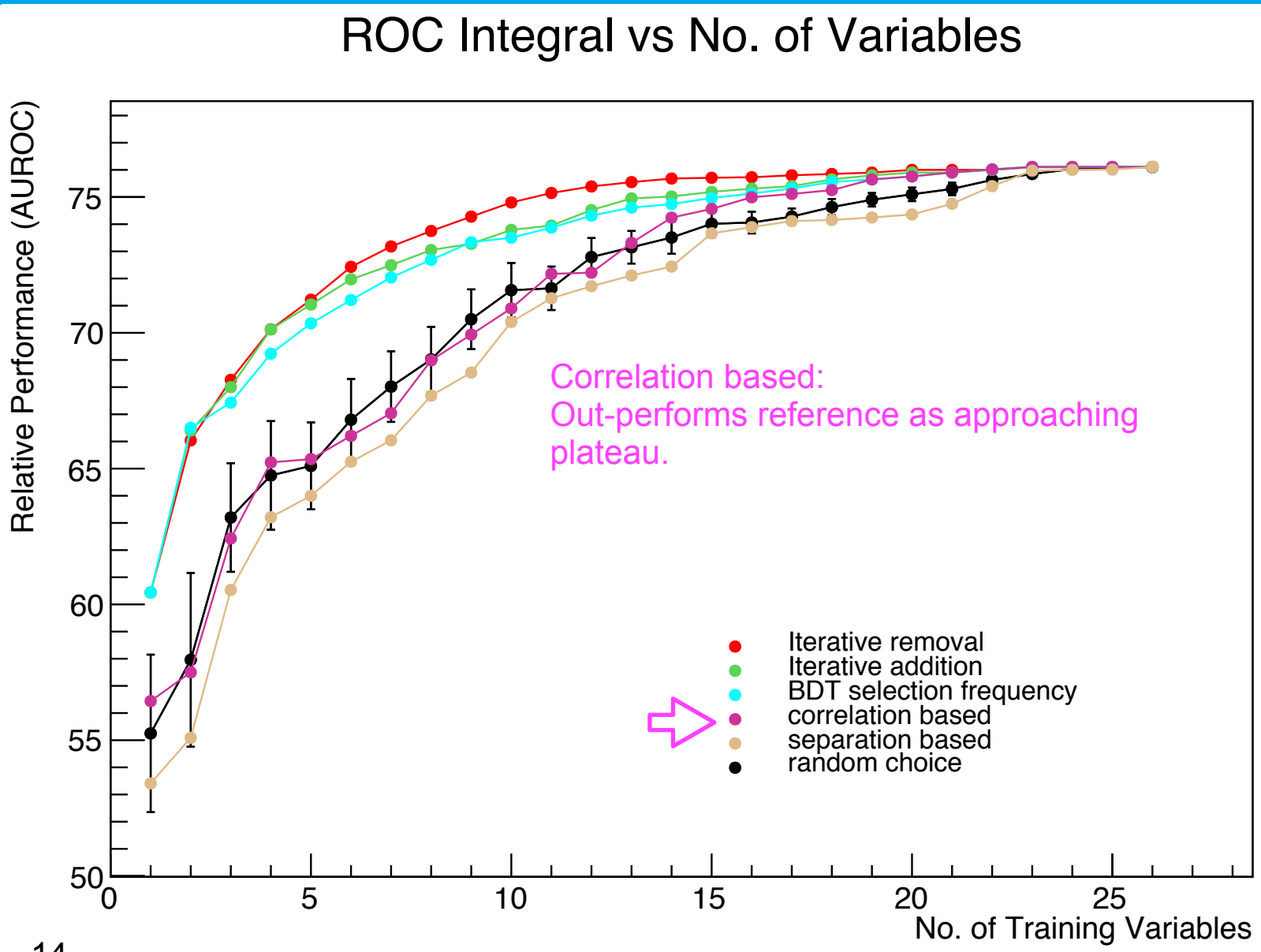
Ranking results



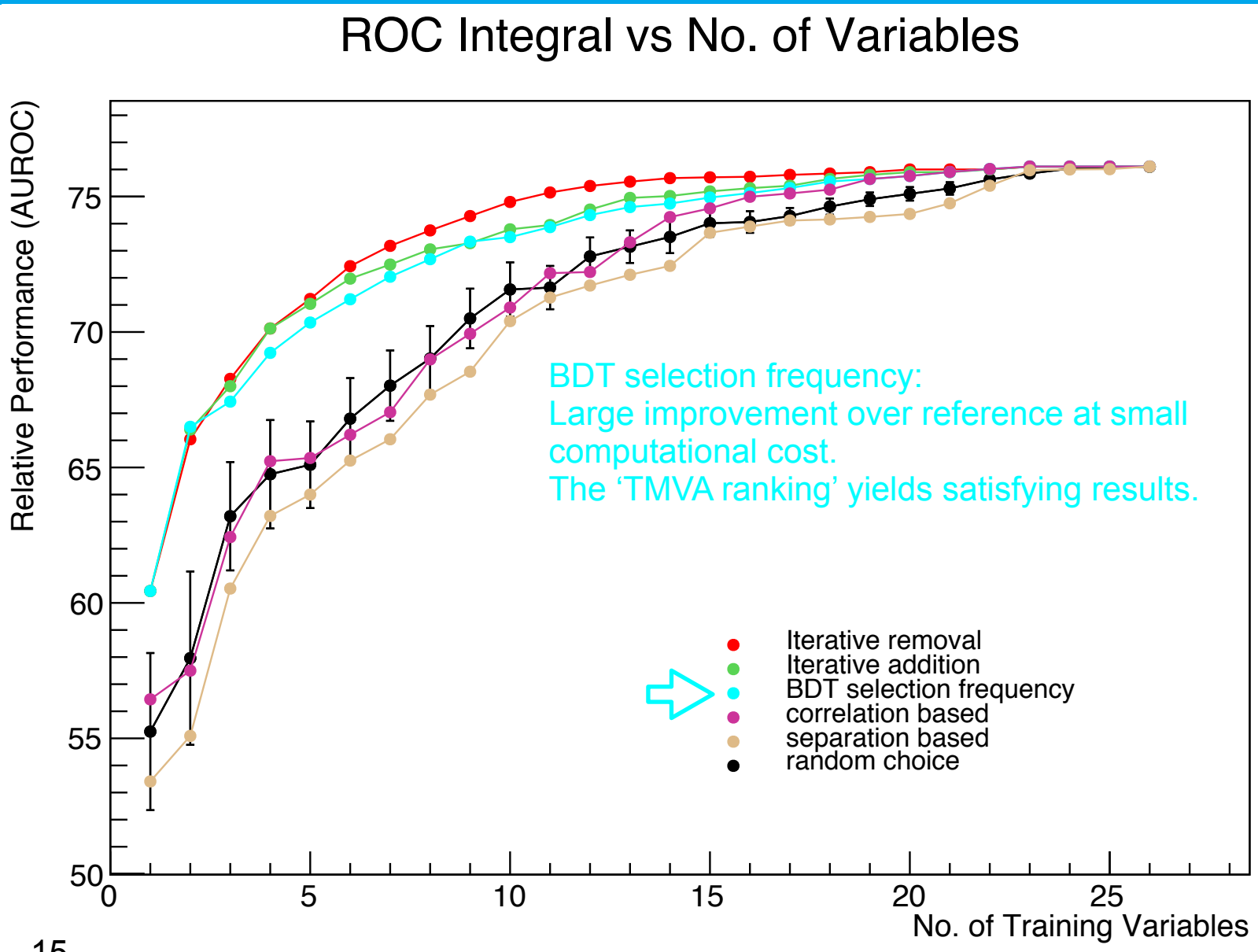
Ranking results



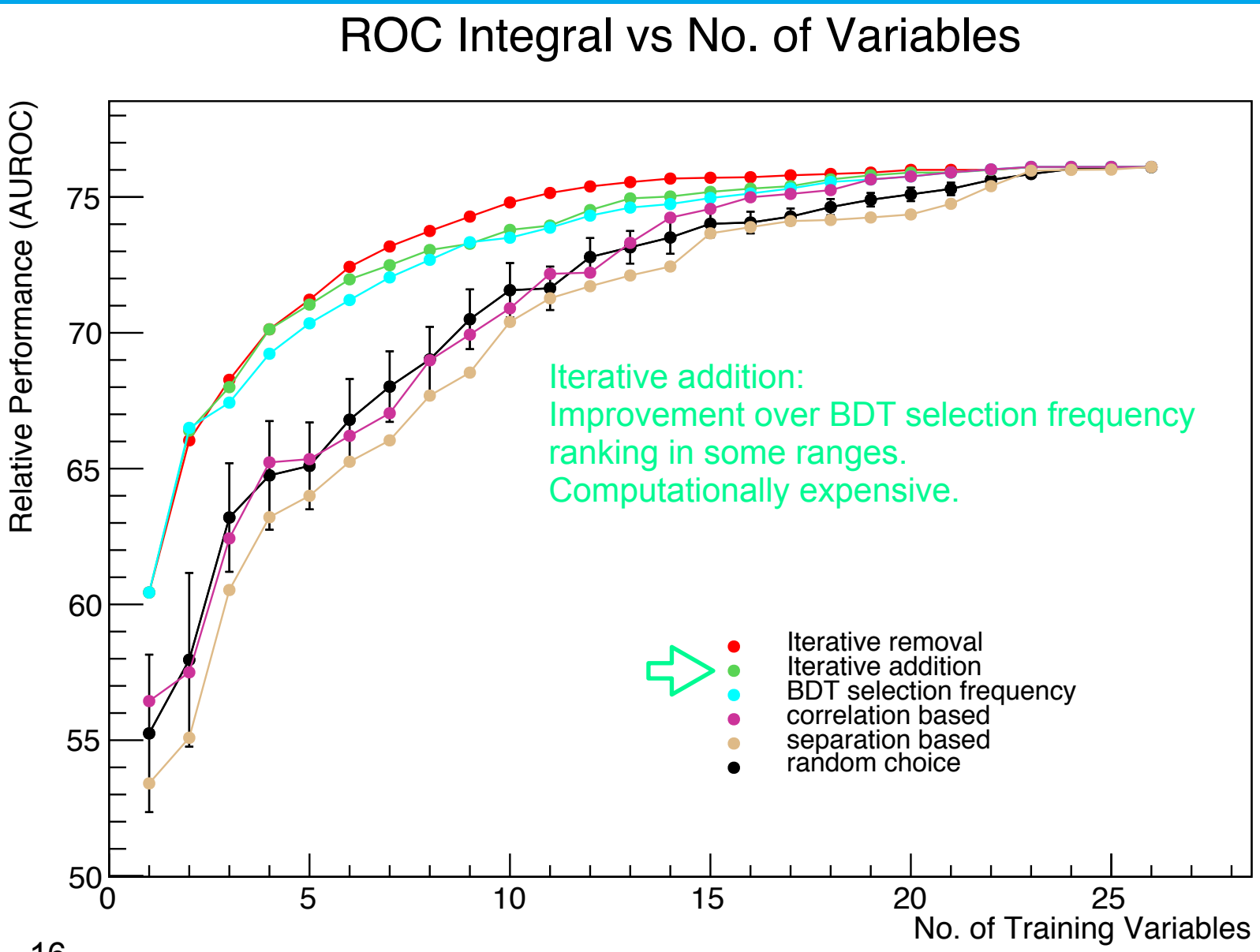
Ranking results



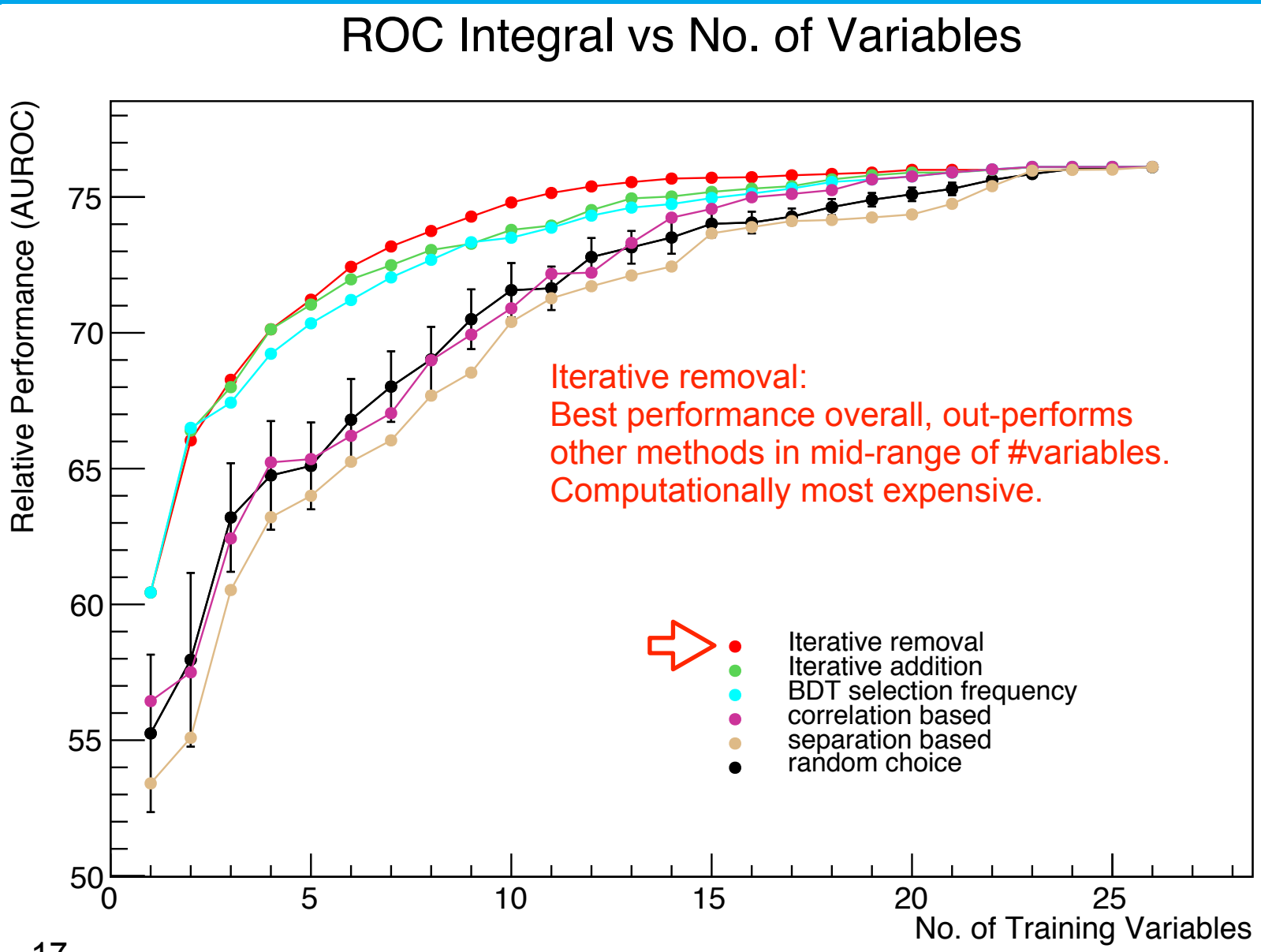
Ranking results



Ranking results



Ranking results



Variable ranking

- Top 10 variables selected by most promising methods
- Should only serve as indication, would not make physics conclusions on the basis of the sample and simulation used
- Difference in performance at most ~1% on AUROC

Rank	Iterative removal	Iterative addition	BDT selection freq.
1	dRbb_avg	dRbb_avg	dRbb_avg
2	HT_jets	Mbb_MaxM	Mbb_MaxM
3	nHigssbb30	nbTag	HT_jets
4	Mbb_MaxM	dRlb2	H0_all
5	nbTag	Mjjj_MaxPt	nJets_Pt40
6	Mbb_MinR	Pt_lep	dRlb2
7	dRlb3	dRbb_MaxM	Mjjj_MaxPt
8	H2_jets	dRlbb_minR	Pt_lep
9	H0_all	HT_all	Max_dEtajj
10	Mjjj_MaxPt	Mbb_MinR	dRlb1

Conclusions

- The example of the ttH(H->bb) vs tt+jets BDT classification was shown to demonstrate the difference of selected feature ranking methods.
- Identifying the top 5 or 10 most important variables is not straightforward in this case, given the high correlation among the variables.
- The computationally cheap BDT selection frequency ranking was found to be an adequate rough estimate
- The computationally costly (greedy) iterative addition and removal methods were compared, where the removal method yields the highest performance for any subset of variables.
- We recommend the iterative removal method for analogous cases where you want to prune the list training variables.

https://github.com/sitongan/vSearch