technische universität dortmund

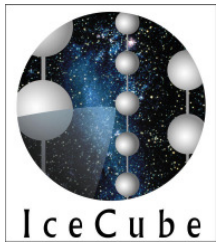
Fakultät Physik
Experimentelle Physik V

Event Selection with the Random Forest for IceCube-22

Tim Ruhe, TU Dortmund

AT-Schule Obertrubach

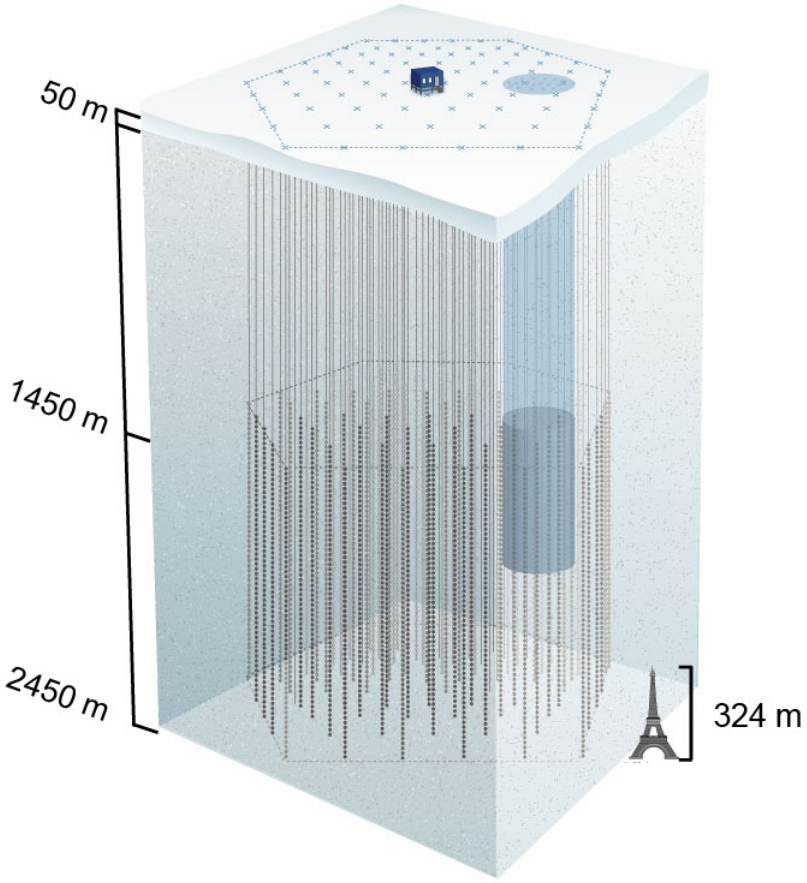
Oktober 2009



bmb+f - Förderschwerpunkt
Astro-Teilchenphysik
Großgeräte der physikalischen Grundlagenforschung

Outline:

- motivation
- the RF method
- training procedure
- application on MC
- summary and outlook



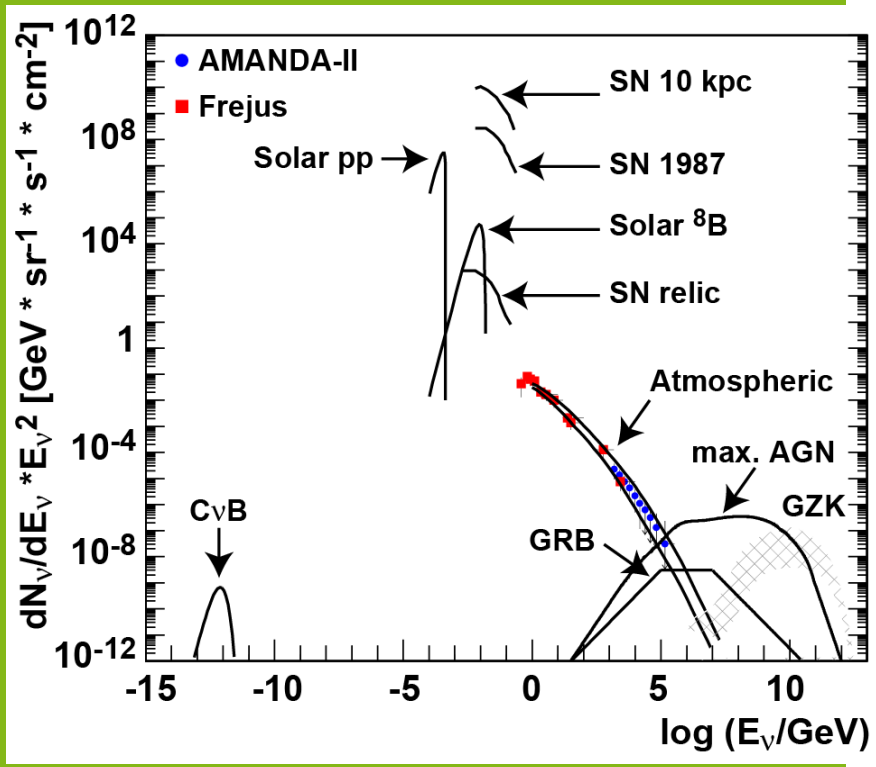
Motivation:

- understand atm. ν-spectrum
- Atm. μ form background for atm. ν-spectrum

→select upgoing events only!

- BUT: Still some μ misreconstructed!

→ reject: Cuts, multivariate methods (BDTs, RF)



The Random Forest:

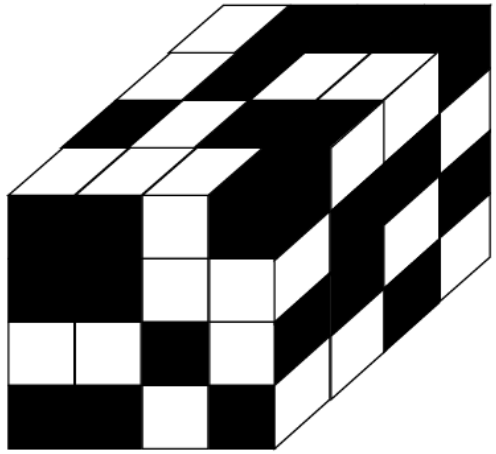
- Random Forest is a multivariate classification tool
- developed by Leo Breiman
- uses a collection of decision trees (few hundred)
- no boosting between individual trees

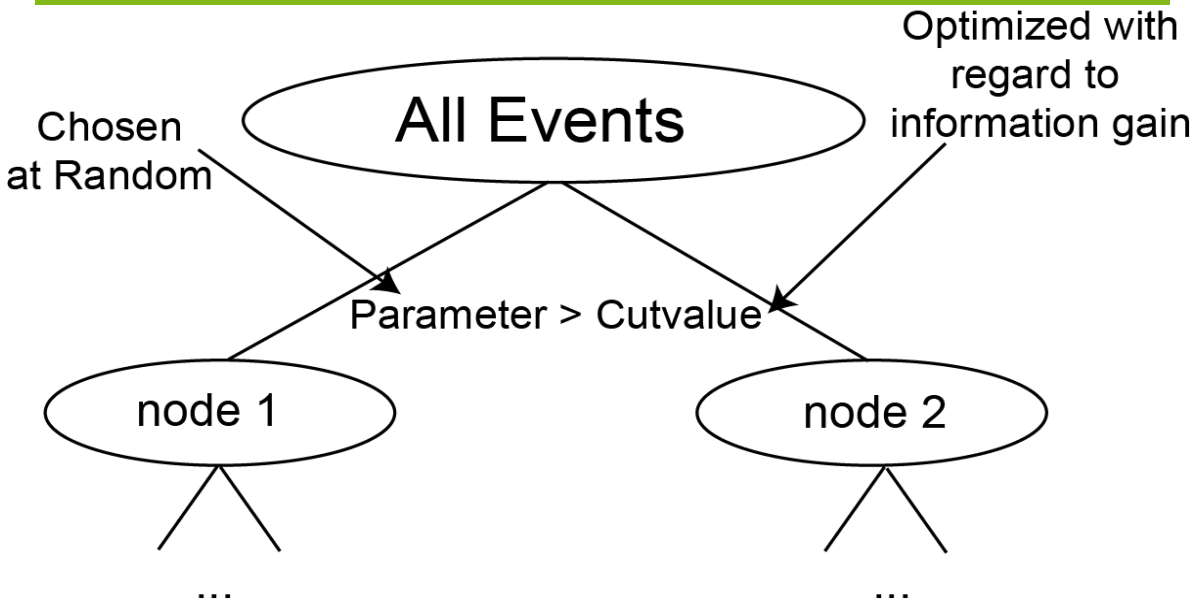
- Training process (Step 1):

- use MCs as input
- signal to background ratio bad

→ data files can be used instead of background MC

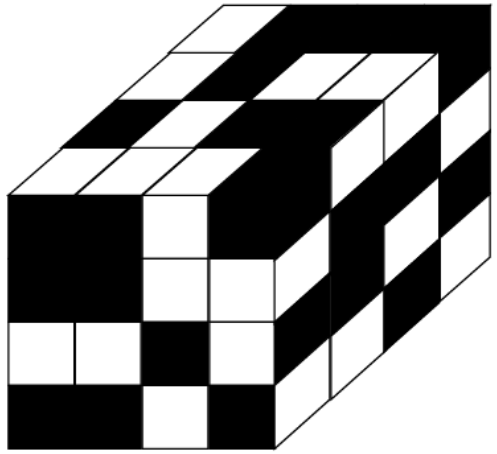
(1) Single Tree



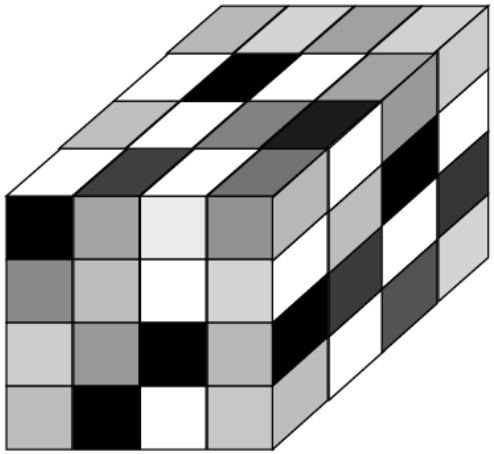


- **Training process (Step 1):**
  - use MCs as input
  - signal to background ratio bad
    → data files can be used instead of background MC
- **Testrun (Step 2):**
  - use MCs (again)
  - statistically independent from Step 1!
- **Application on data (Step3)**

(1) Single Tree



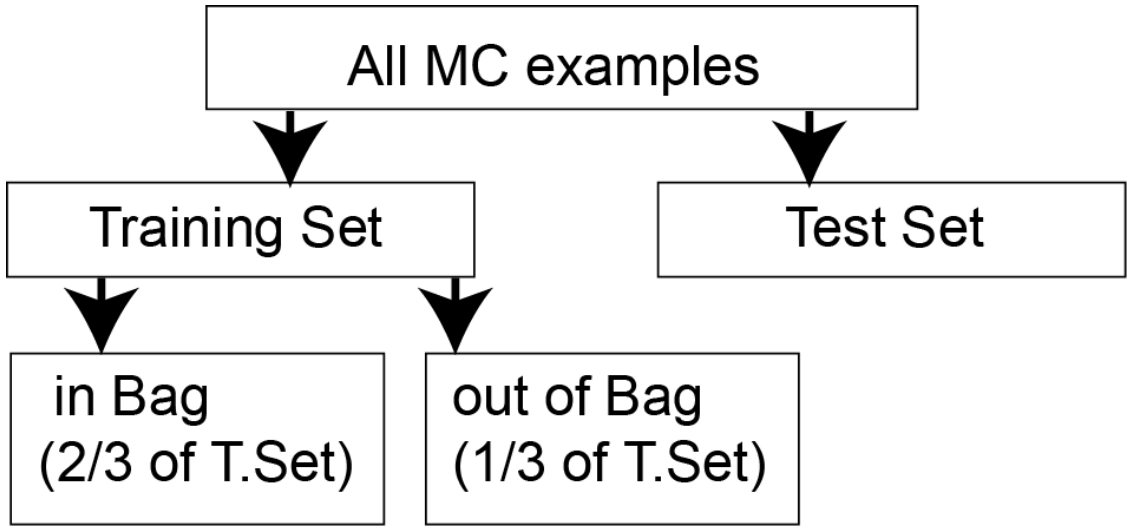
(2) Averaging over n Trees



The Rapidminer toolkit (YALE):

-data mining toolkit developed at TU Dortmund (K. Morik)

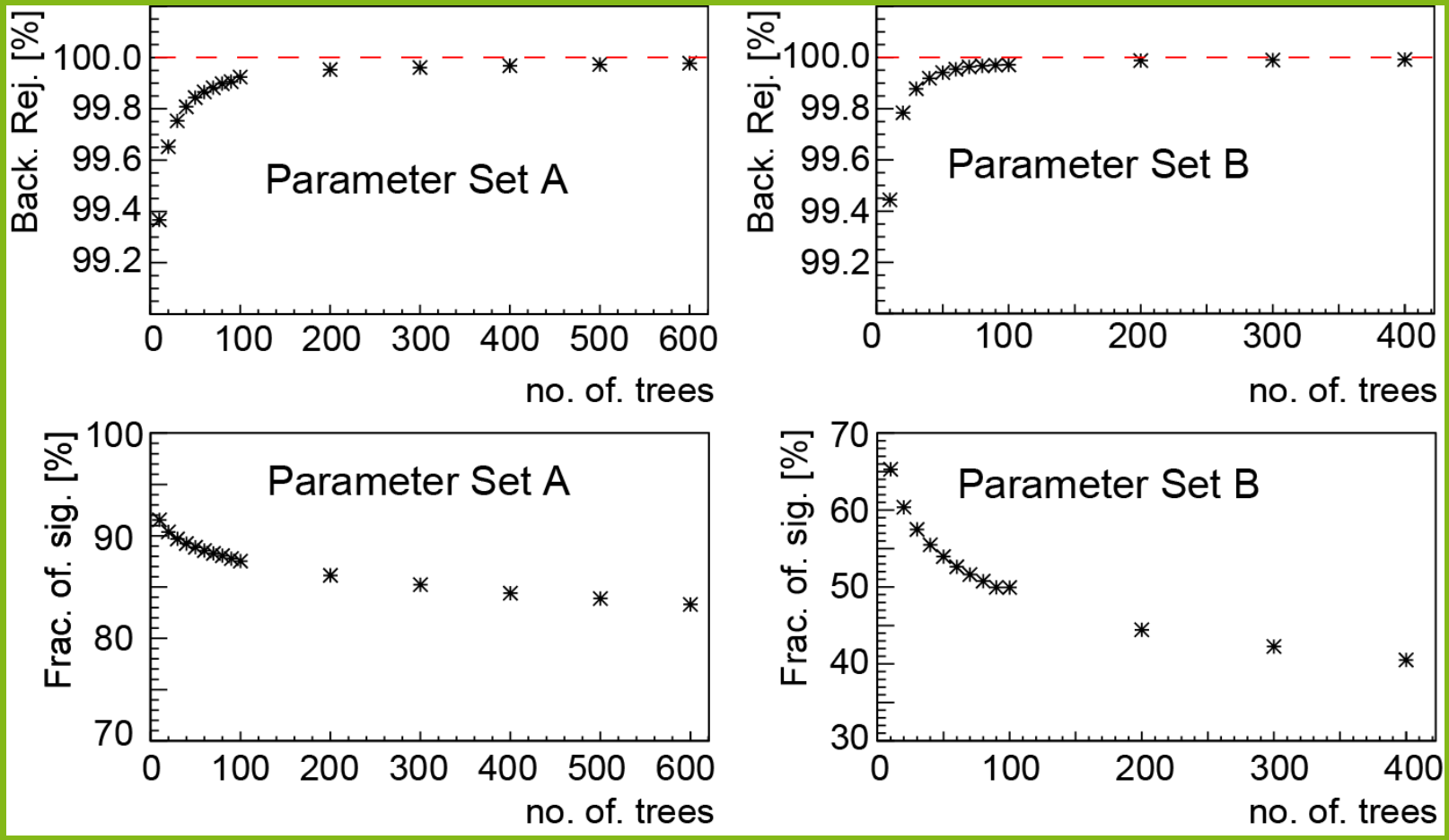
- Weka (data mining toolkit) fully implemented



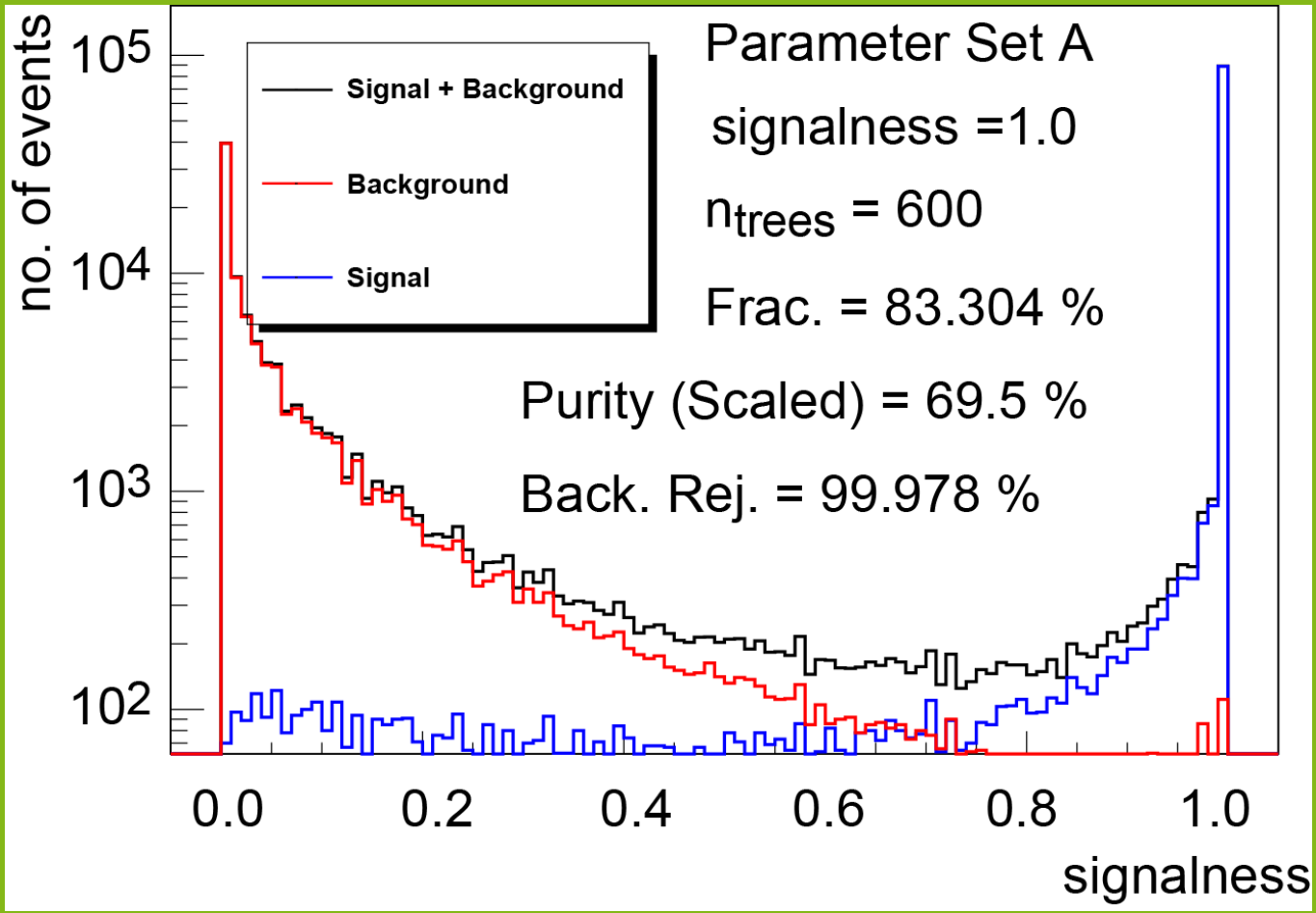
Parameters for training:

- 2 forests (different parameter Sets)
- 10^5 background events (CORSIKA-in-ice)
- 10^5 signal events (nu-gen)
- 10^5 data events (if trained with data)
- RapidMiner toolkit with Weka RF
- minimal node size: 1
- IceCube-22 data and MC (Level 3)

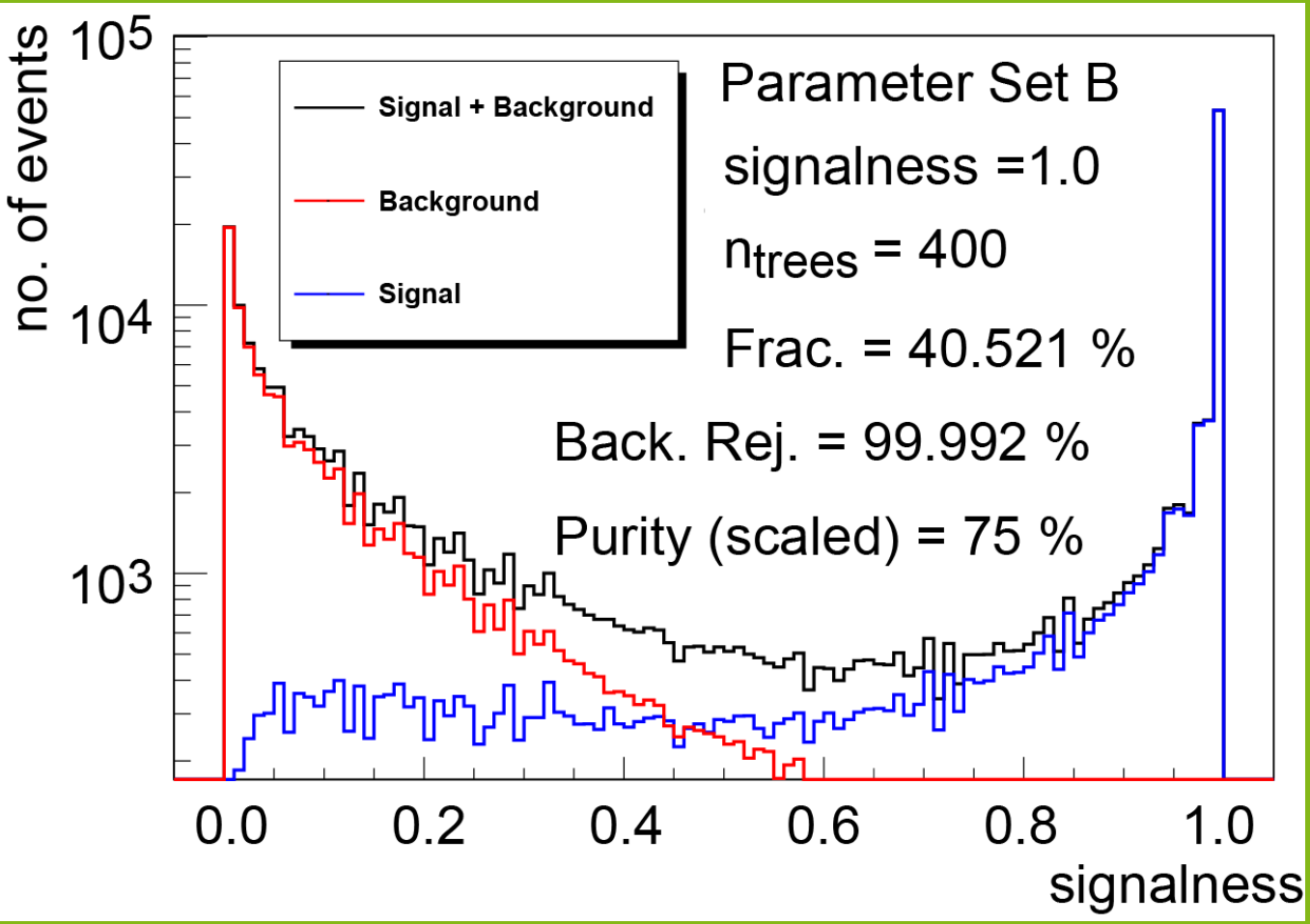
Performance of the forests:



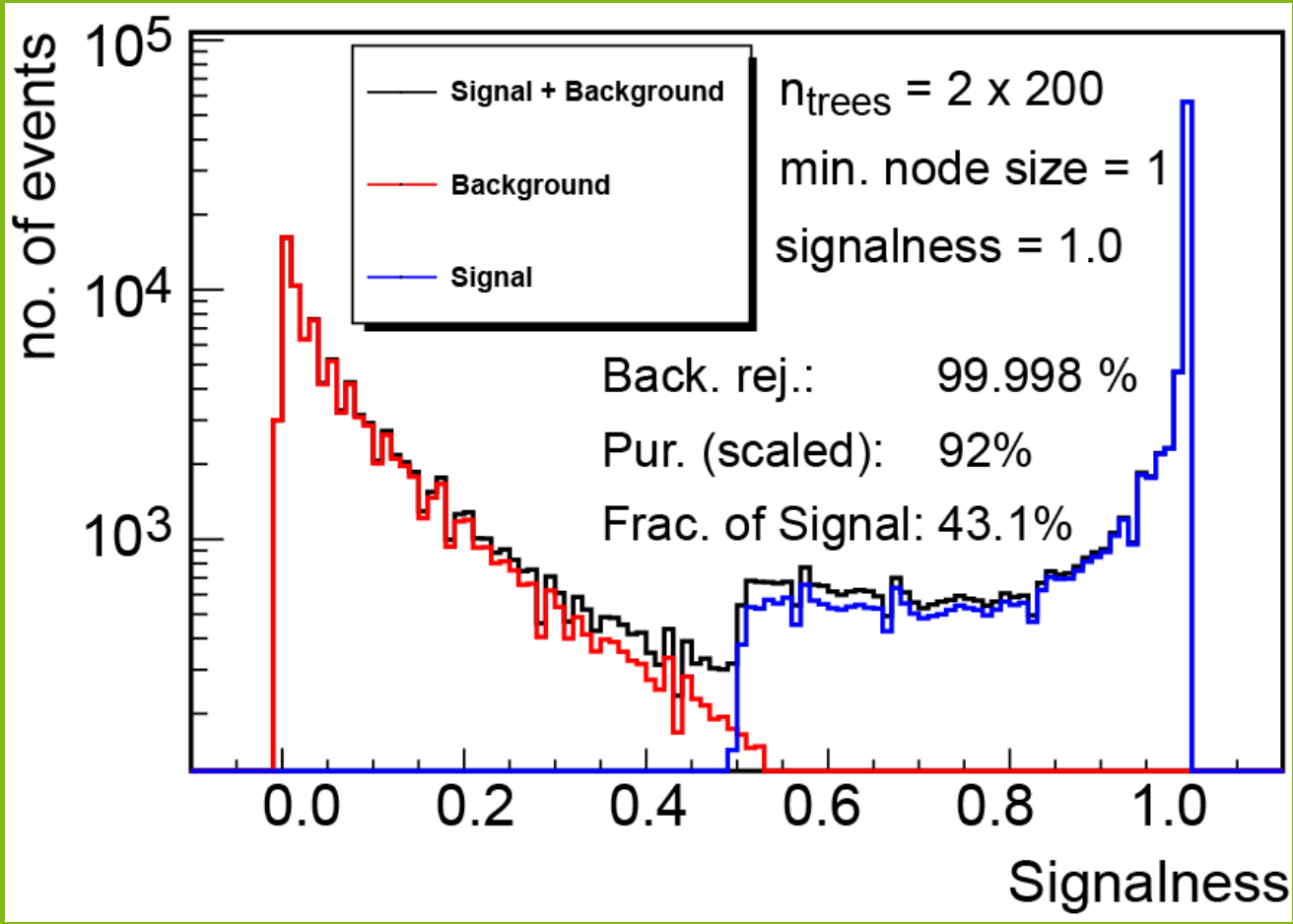
Parameterset A: Application on Monte Carlo



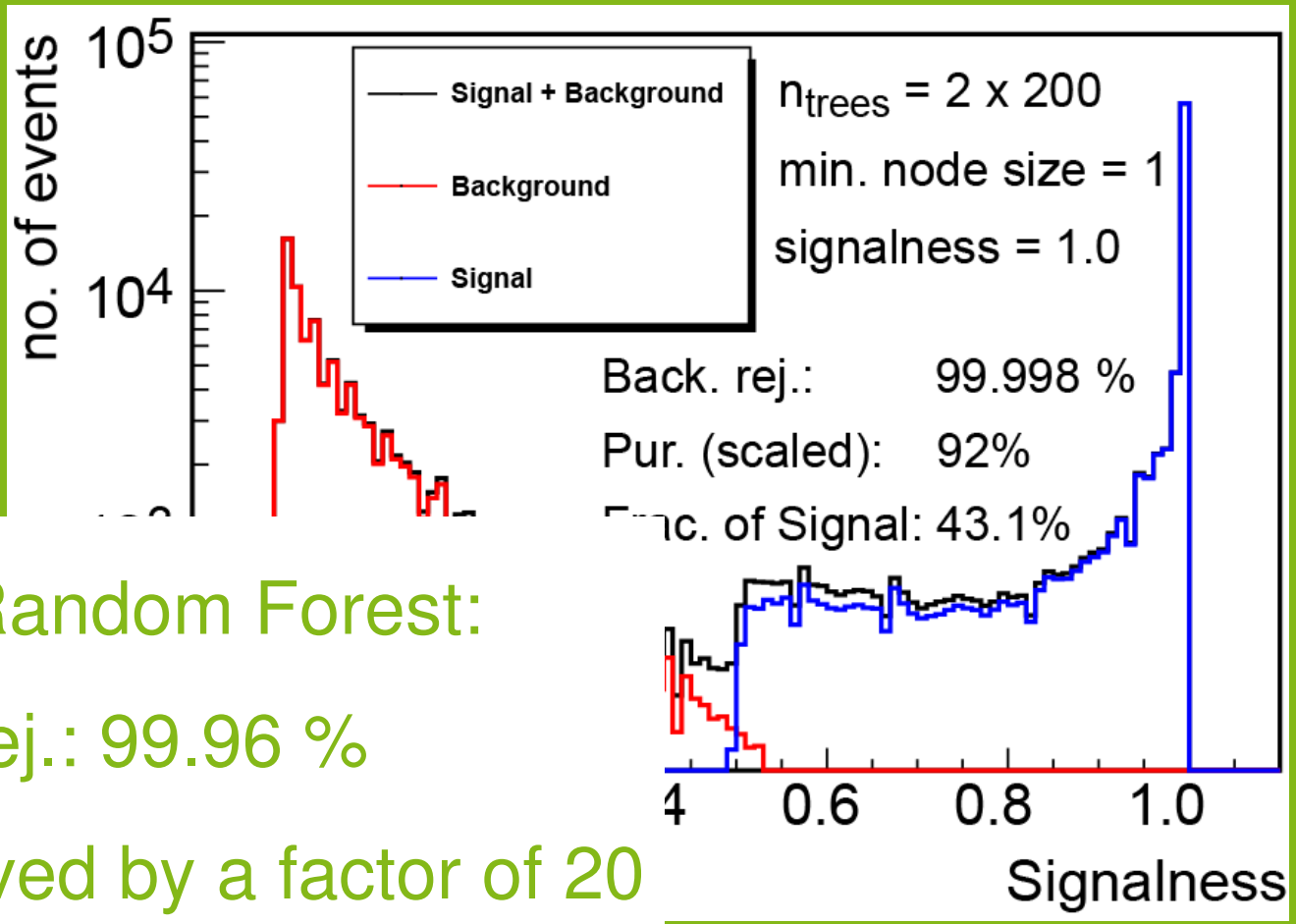
Parameterset B: Application on Monte Carlo



Combination of two forests: sig = (sig(1) + sig(2))/2



Combination of two forests: sig = (sig(1) + sig(2))/2

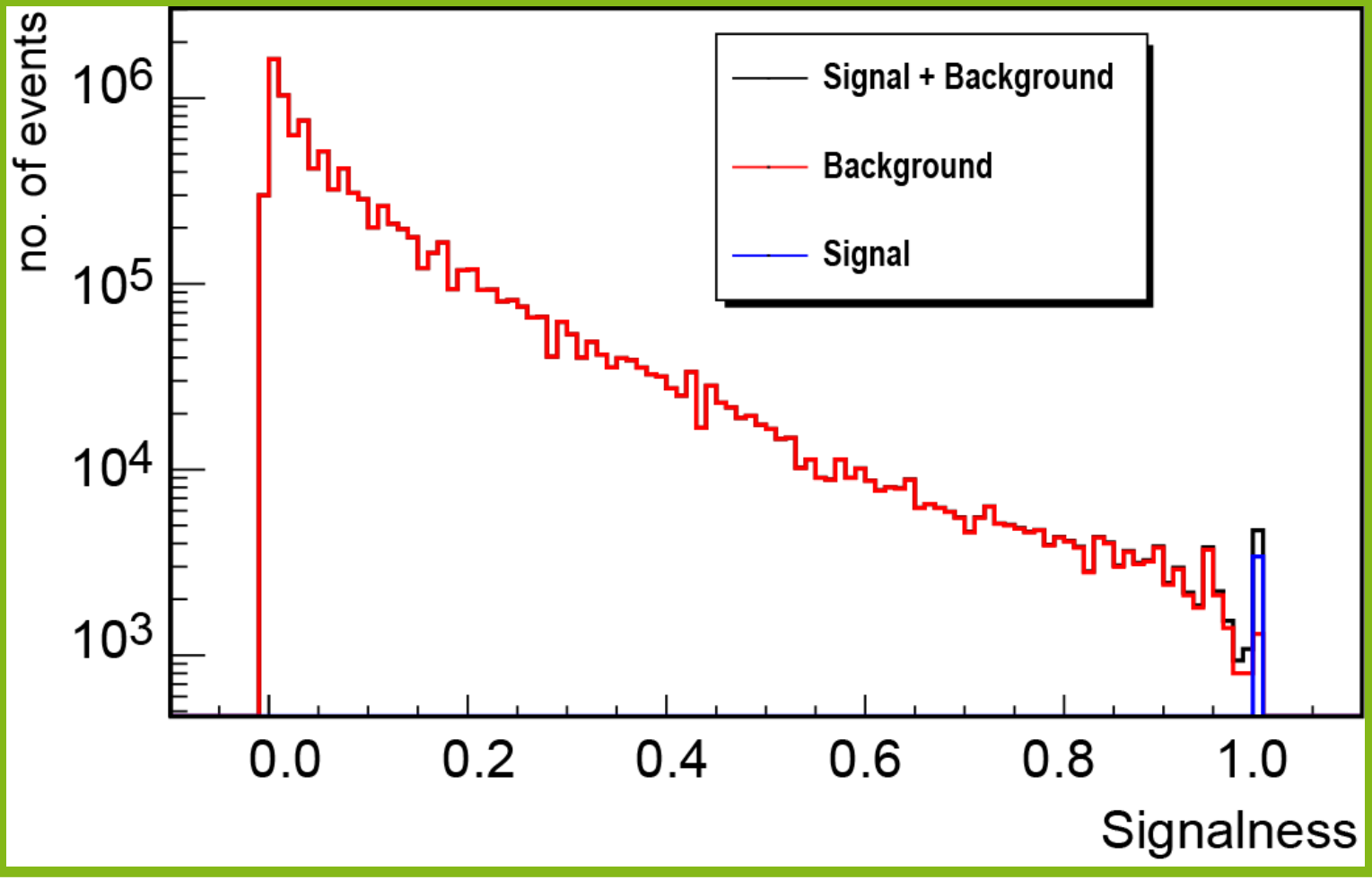


MARS Random Forest:

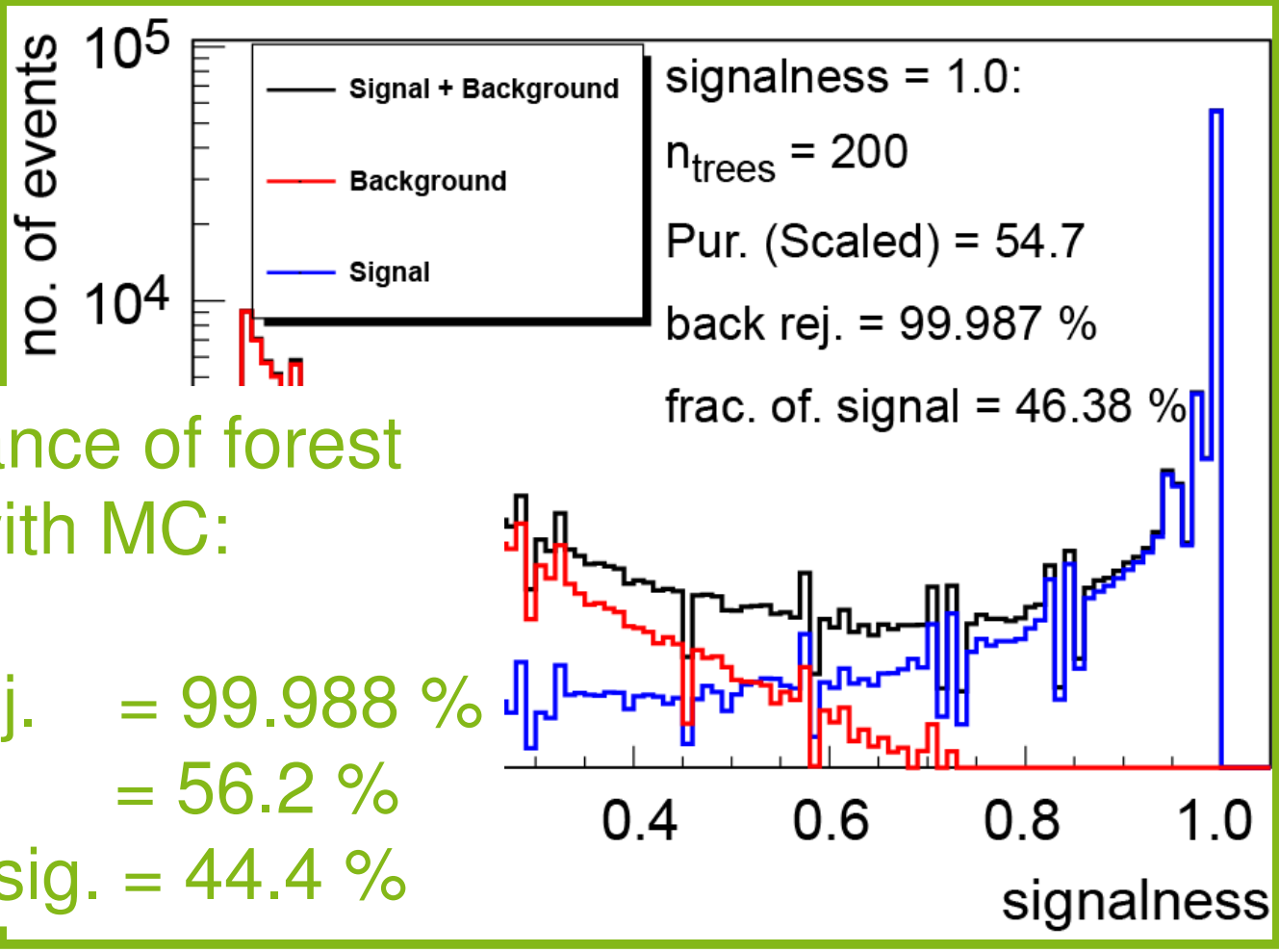
Back. Rej.: 99.96 %

→improved by a factor of 20

Combination: Scaled to data



Trained with data: Parameter Set B



Performance of forest trained with MC:

Back. Rej. = 99.988 %
Purity = 56.2 %
Frac. Of sig. = 44.4 %

Summary & Outlook:

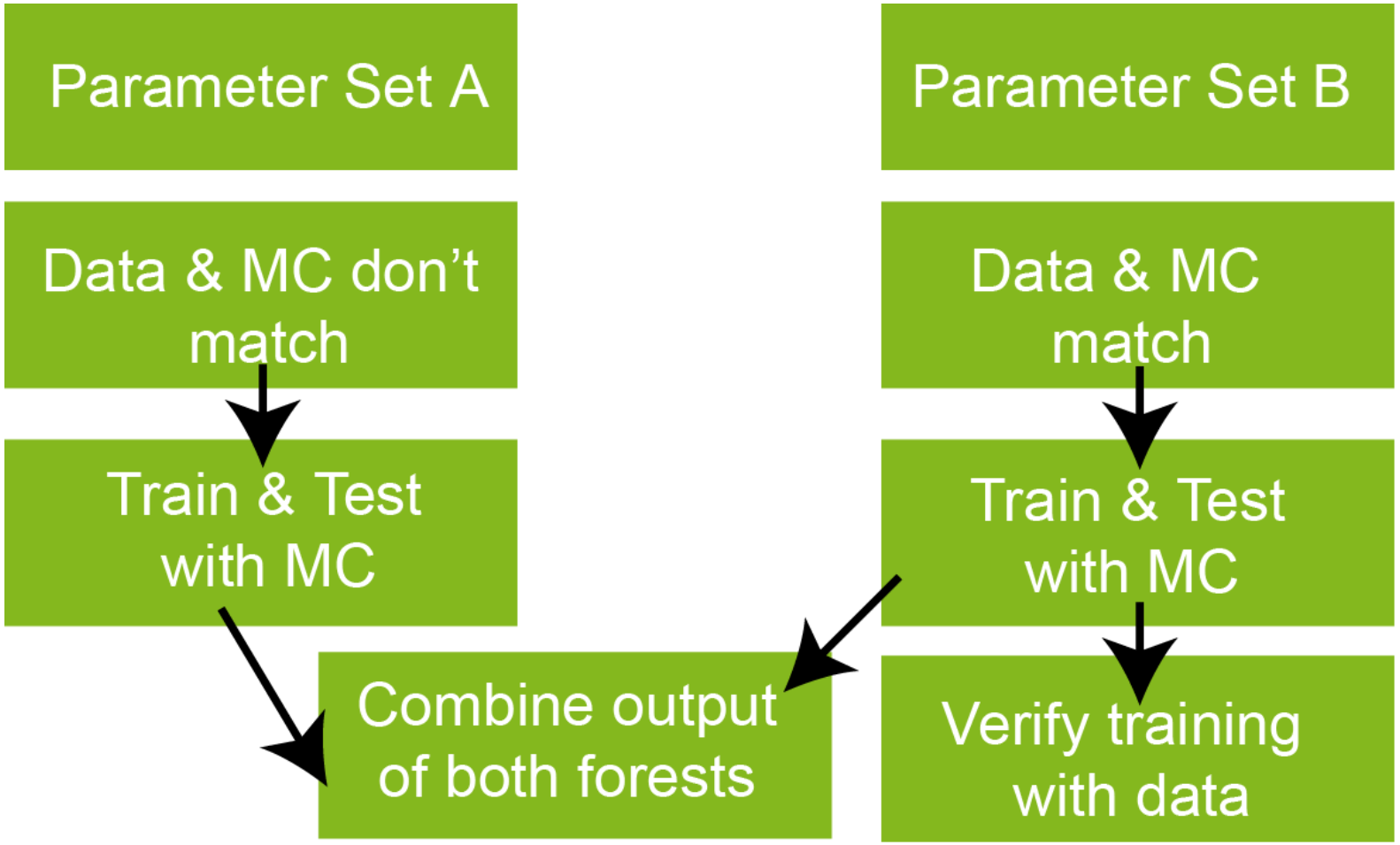
-background rejection improved by a factor of 20 compared to first studies on the RF

- RF can be trained with data

Ongoing work:

- find preprocessing
- check parameter combinations
- application on data
- migrate to IC40 & IC59

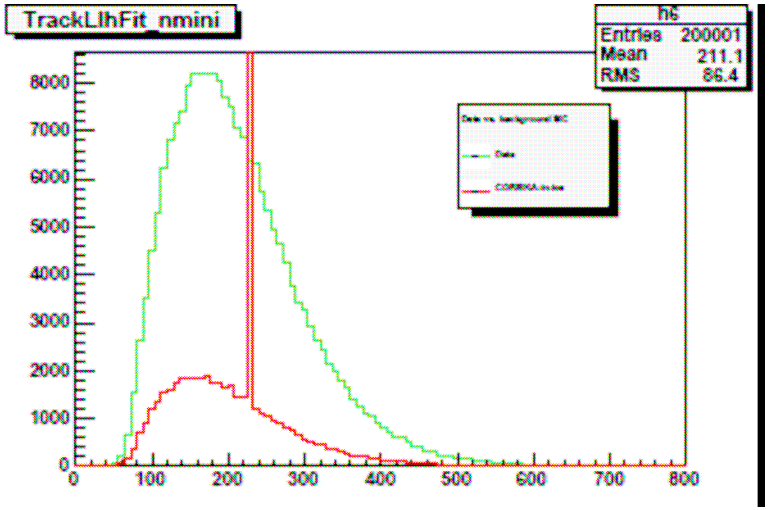
Backup Slides

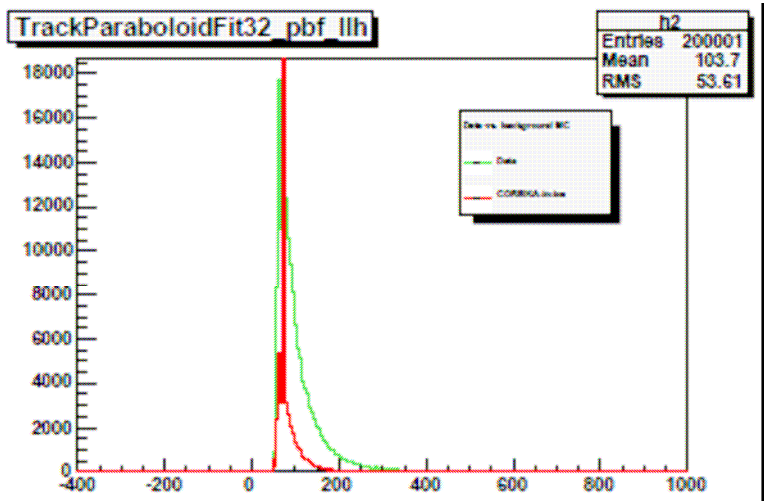

Parameter Set A
Parameter Set B
Data & MC don't match
Data & MC match
Train & Test with MC
Train & Test with MC
Combine output of both forests
Verify training with data

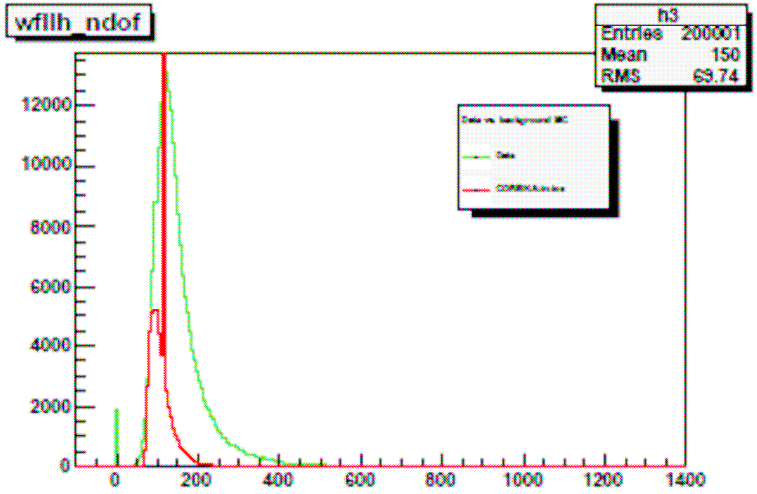
Monte Carlo Sets used:

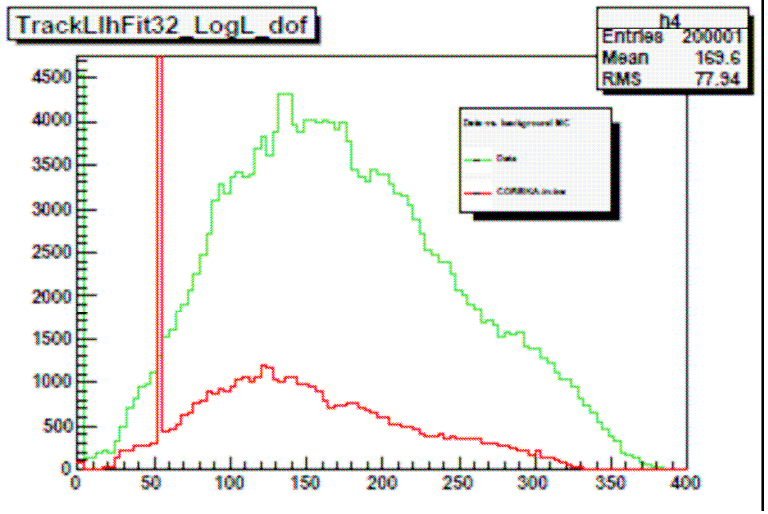
Signal	Background
neutrino-generator_000651	CORSIKA-in-ice_000618
neutrino-generator_000753	CORSIKA-in-ice_000629
neutrino-generator_000768	CORSIKA-in-ice_000630
	CORSIKA-in-ice_000642
	CORSIKA-in-ice_000645
	CORSIKA-in-ice_000861

Parameter Set A: Data (green) vs. CORSIKA (red)

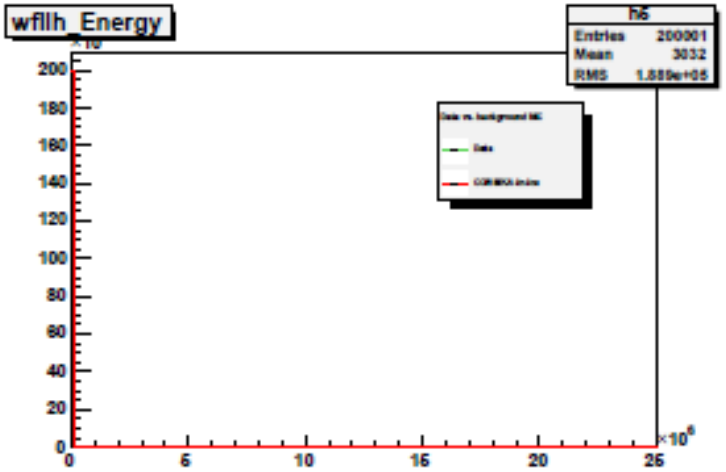


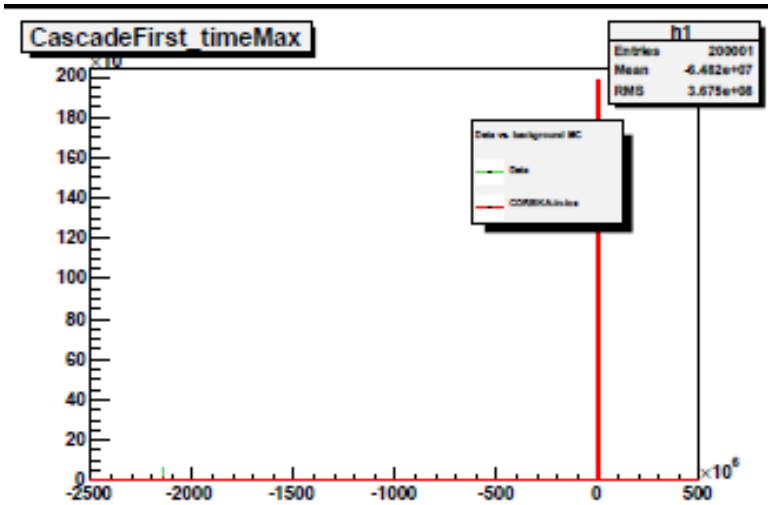


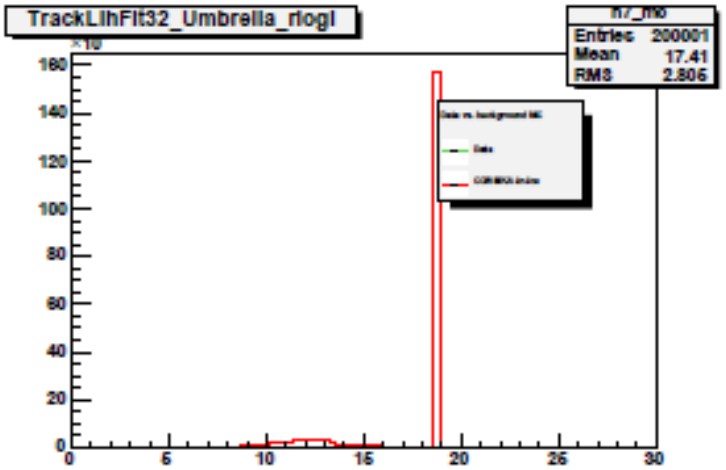




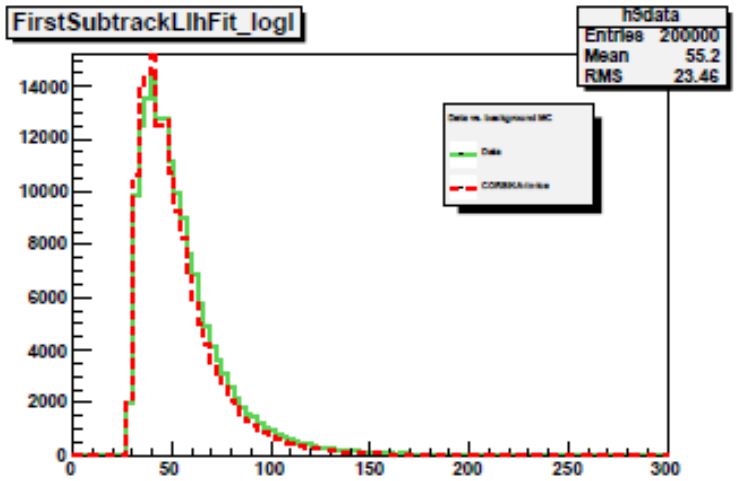
Parameter Set A: Data (green) vs. CORSIKA (red)

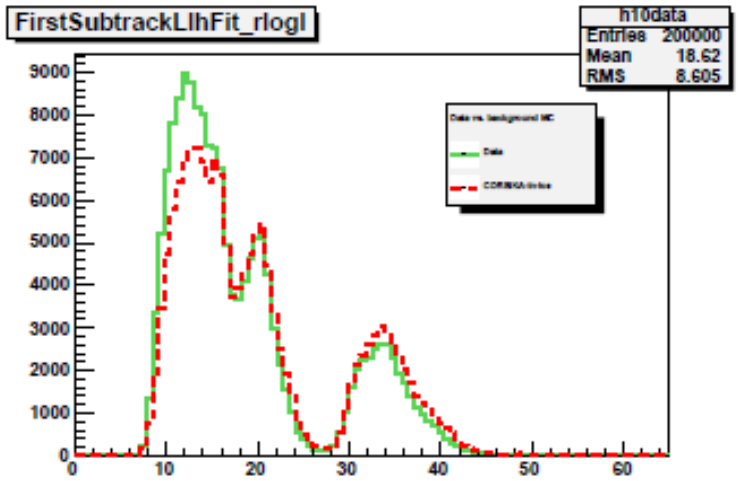


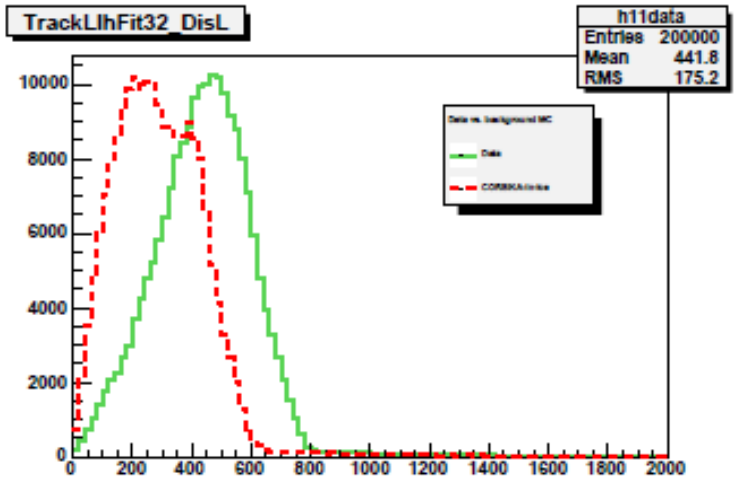




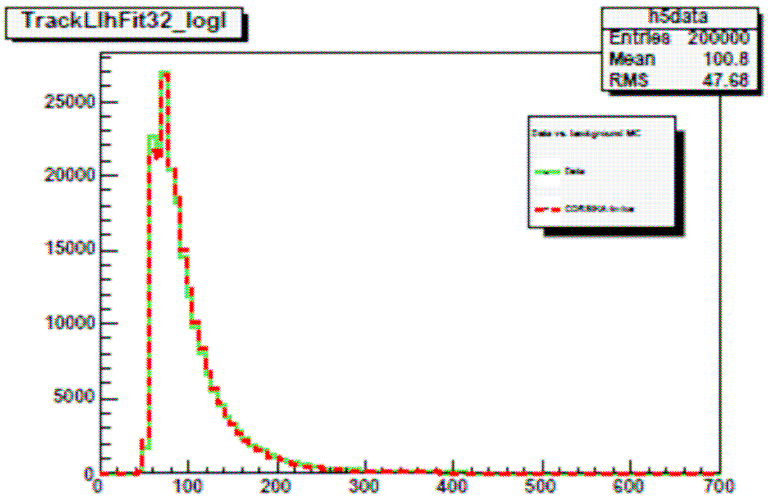
Parameter Set B: Data (green) vs. Corsika (red)

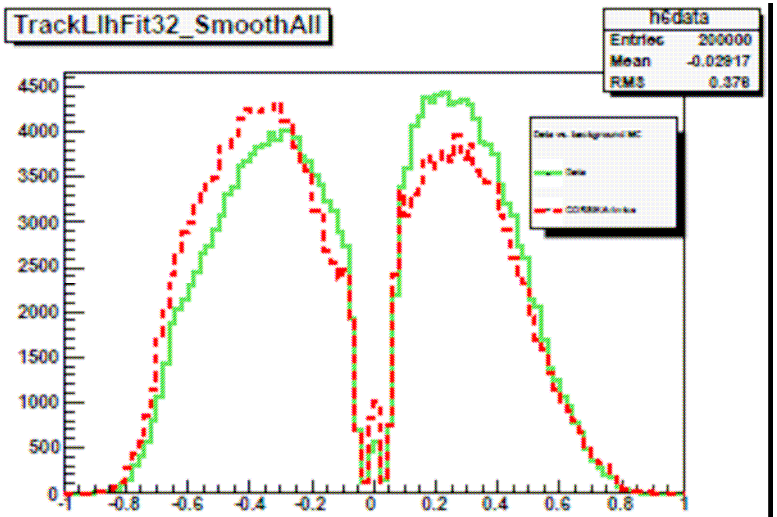


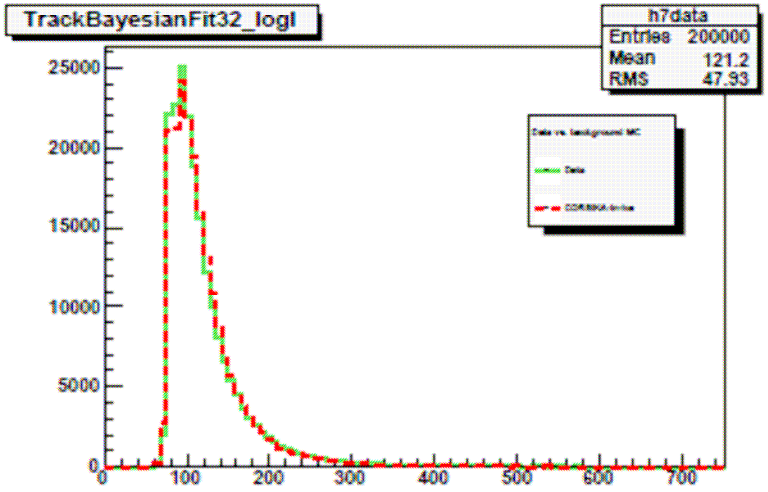


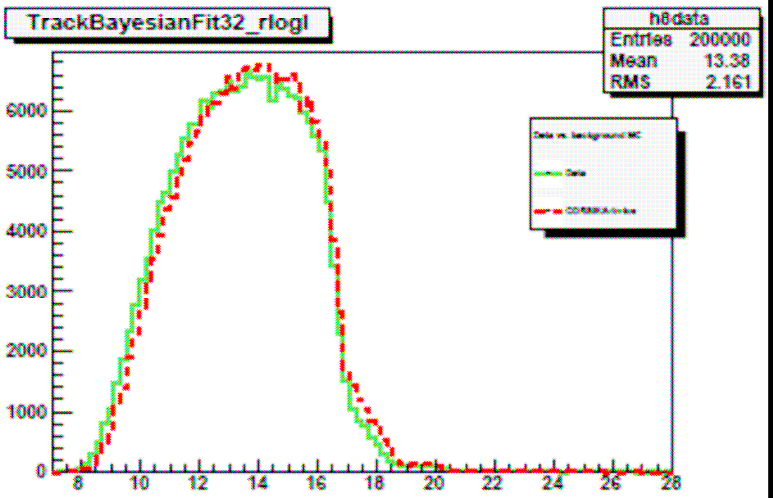


Parameter Set B: Data (green) vs. Corsika (red)

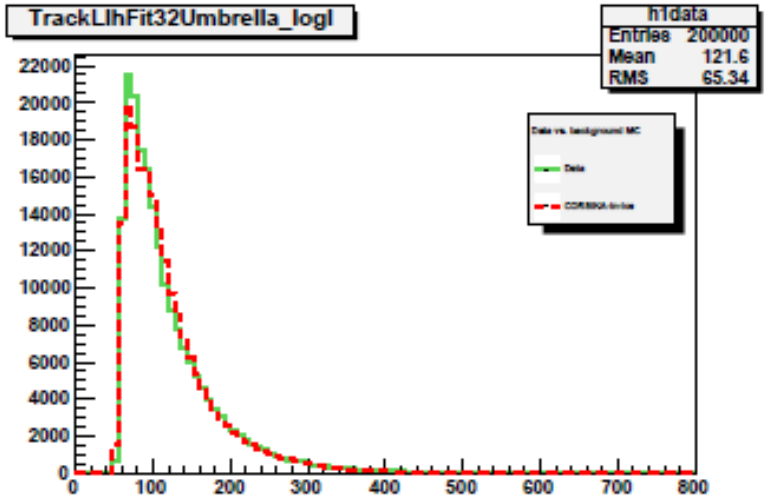


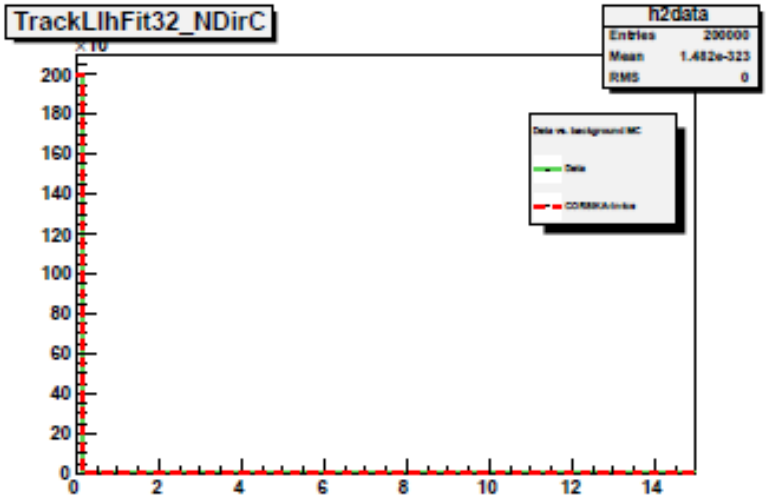


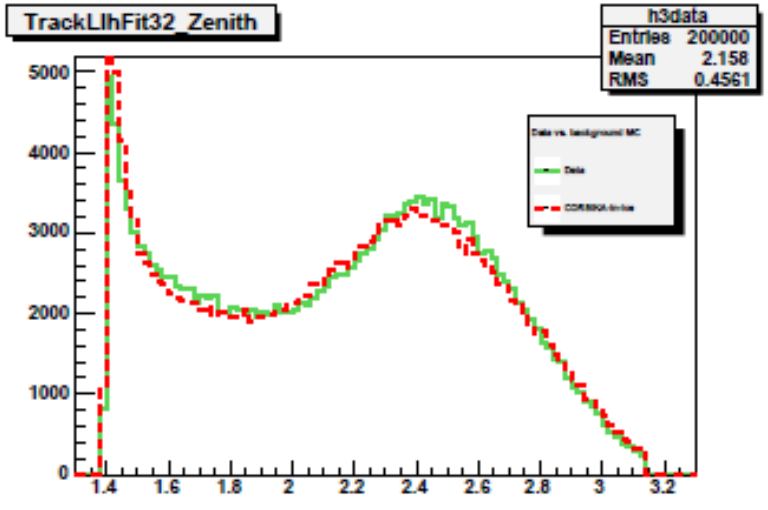


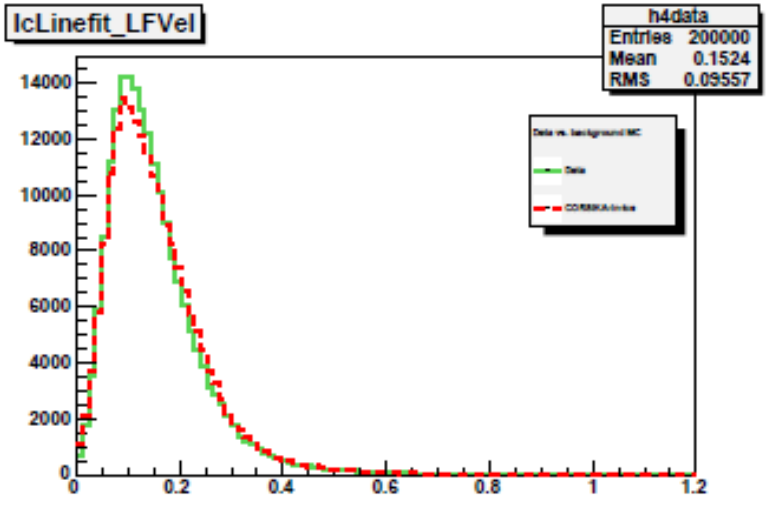


Parameter Set B: Data (green) vs. Corsika (red)









Combination of two forests: Parameter Sets

Set A	Set B
CascadeFirst_timeMax	TrackLlhFit32Umbrella_logl
TrackParaboloidFit_pbf_llh	TrackLlhFit32_NDirC
wfllh_ndof	TrackLlhFit32_Zenith
wfllh_Energy	IcLinefit_LFVel
TrackLlhFit32Umbrella_rlogl	TrackLlhFit32_logl
TrackLlhFit_nmini	TrackLlhFit32-SmoothAll
TrackLlhFit32_LogL_dof	TrackBayesianFit32_logl
	TrackBayesianFit32_rlogl
	TrackLlhFit_Zenith
	TrackLlhFit32_DisL

Control of Training process: Gini index