Minimal Supersymmetric Higgs Search at CMS

Determining the Optimal Control Region using MVA

Ewen L. Gillies
DESY Zeuthen, 22/8/2013

HELMHOLTZ ASSOCIATION

DESY

MSSM at the CMS Detector: The Basics

> MSSM = **M**inimal **S**uper**S**ymmetric **M**odel

- Predicts “super partners” for all fundamental
- Bosons ←→ Bosinos
- Fermions ←→ Sfermions

> Higgs Field

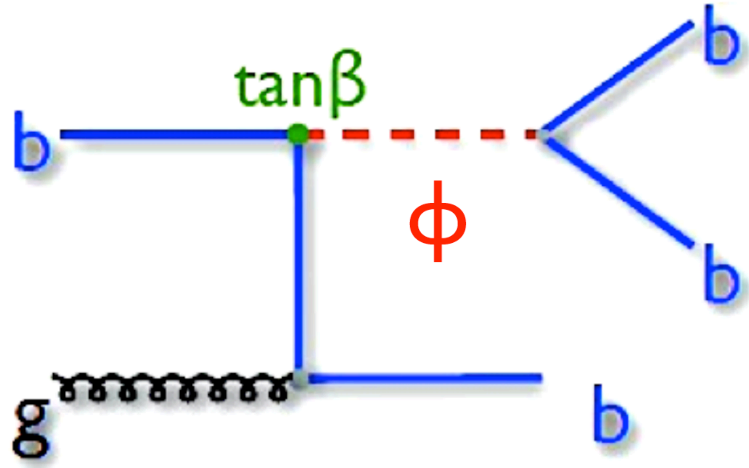
- MSSM demands *two* Higgs **complex** doublet fields:

$$H_u = \begin{pmatrix} H_u^+ & H_u^0 \end{pmatrix}; \qquad H_d = \begin{pmatrix} H_d^0 & H_d^- \end{pmatrix}$$

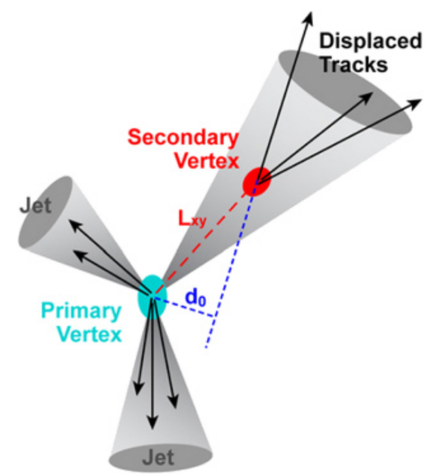
- Results in three neutral Higgs: A^0 h^0 H^0

Event Tagging and Cut Selection

- Characteristics of b-hadrons decays
  - Long lived hadrons yielding distinct primary and secondary vertices.
  - Lots of semileptonic decays → Missing transverse *p*



- CMS is looking for energy resonance resulting from more than one neutral Higgs
- Pre-selections
  - Only look at channels with at least three b-tagged jets
  - Combined Secondary Vertex (CVS) > 0.898 [= 1 for b-jet]
  - Medium mass scenario: $180\ GeV < M_{\phi} < 350\ GeV$

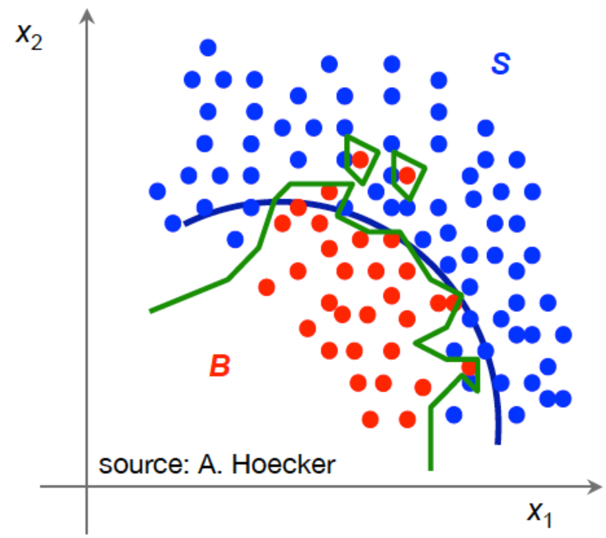


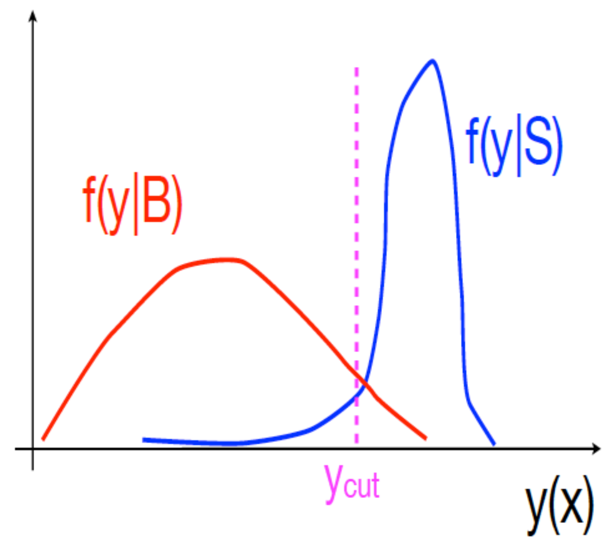
Using TMVA in Control Regions

- Motivation: Determine control region for blinding
- Goal: Determine optimal control regions and cuts (in phase space)
  - Control Region = background to signal ratio maximised
  - Optimal Cuts = Signal and background separation
  - Automate this process in TMVA
- TMVA: **T**oolkit for **M**ulti**V**ariate **A**nalysis in ROOT
  - Optimization software
  - Classification techniques
- Method: Boosted Decision Trees in TMVA
- My Project: Check the results of the automated process
  - Adjust input settings in classification analysis
  - Compare S/B ratio and purity

Classification

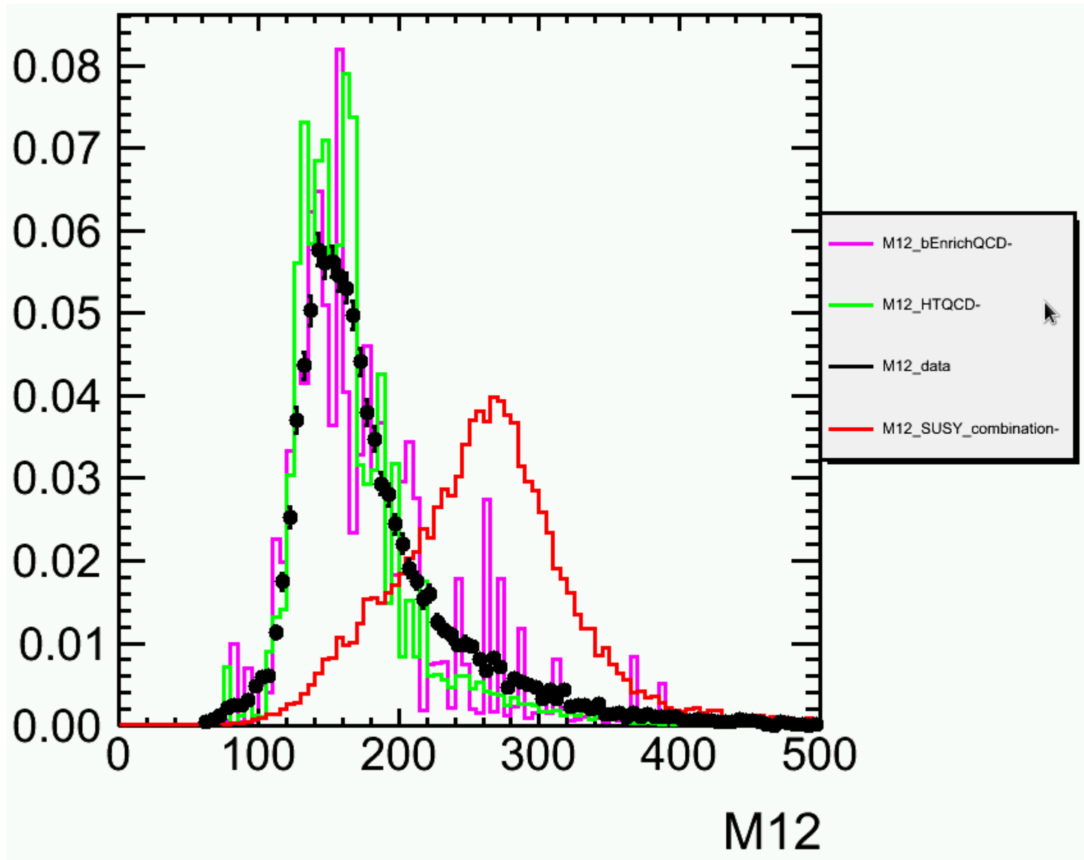
- Used to determine optimal cut region
  - Creates a discriminating function between background and signal
  - Input variables are final state kinematics
  - Output is MVA metric: [-1,1]
  - Goal: Maximum separation in distribution
- Two step process
  - Training: Maps functional dependence to classifier
  - Testing: Iterates classifier form on remainder of sample or new sample
  - Samples for each must be *statistically independent*





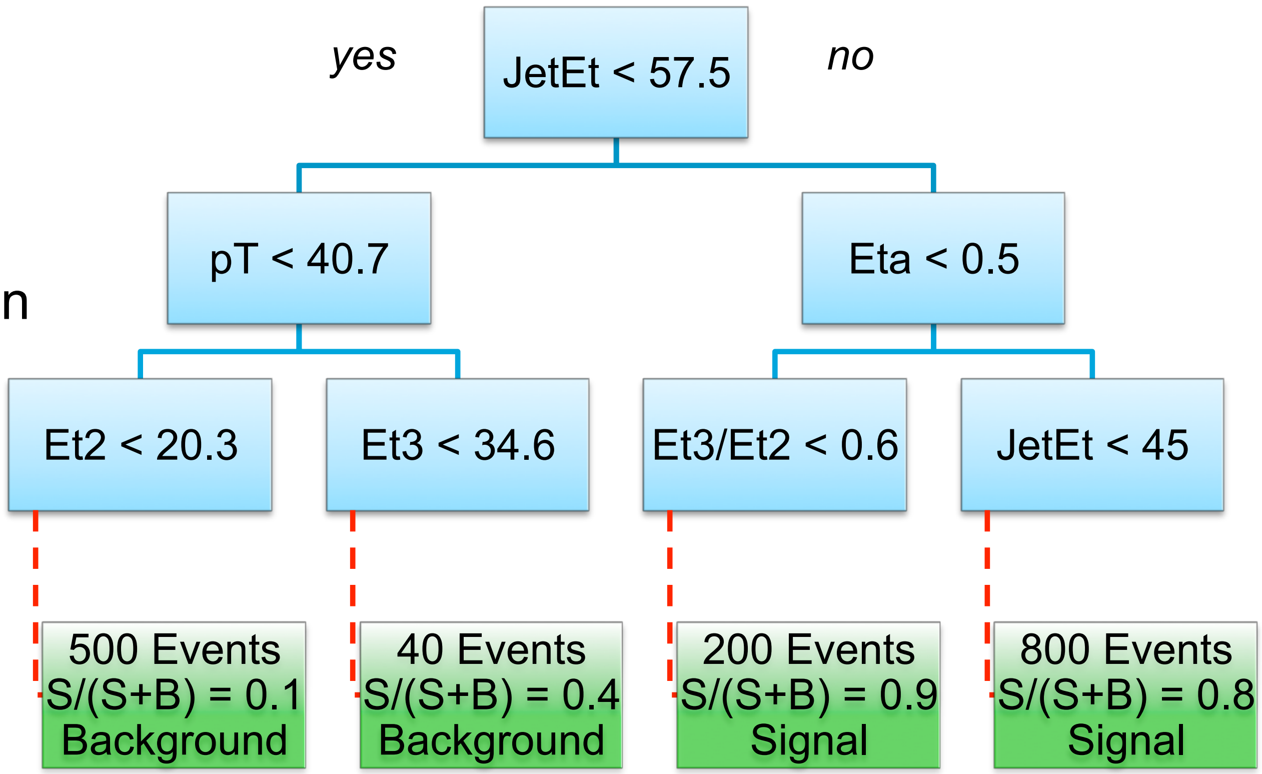
Kinematic Variables

- Events described by final state kinematics
  - Jet Transverse Momenta : Pt1, Pt2, Pt3
  - Number of Constituent Jets
  - Total Energy: sumEt
  - Jet Transverse Energy Ratios: Et3byEt1
- Usually select ~10 in multivariable cut analysis

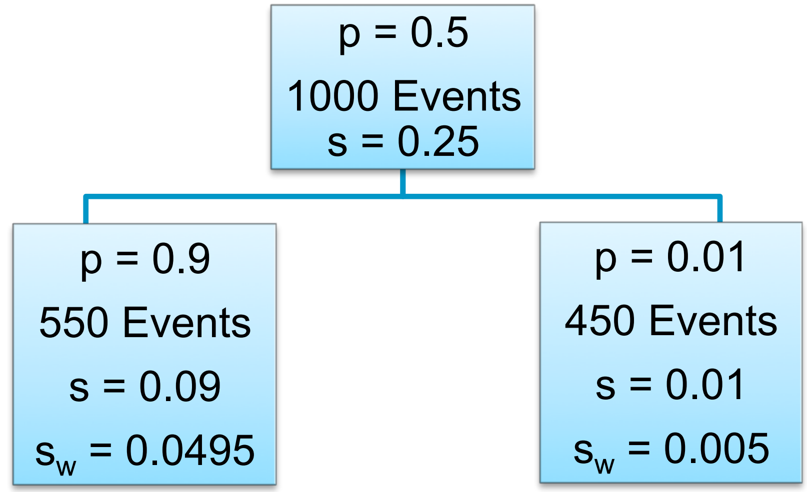


Decision Tree Training

- Training done with Monte Carlo data
- Root node is cut by
  - Best variable
  - Optimal Value
- Process is repeated on resulting node
  - Never backtracks
  - Iterates until nodes reach minimal event content
- Leafs are classified as background or signal



Decision Tree Training: Nodes and Cuts

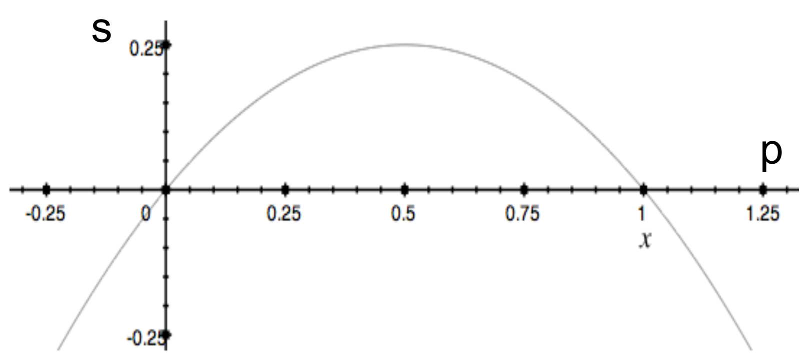


> Node Properties

- Number of events
- Purity of events: $p = \dfrac{S}{S+B}$

> Cutting Method

- At each node, TMVA determines best variable to cut on, and best value
- Scans with preset granularity



> Separation Index

- Metric with maximum at 0.5
- *p, 1-p* symmetric
- Examples:

Gini Index: $s = p \bullet (1-p)$

Cross Entropy: $s = -p \bullet \ln(p) - (1-p) \bullet \ln(1-p)$

Decision Tree Training: Boosting

- Adaptive Boosting: "Averaging process"
  - Used to stabilize BDT classifier against fluctuations in input variables
- Train many decision trees
  - Uses a weighted resample of data
  - Subsequent re-samples give higher weights, α, to misclassified events

$$err = \frac{misclassified}{total}; \quad \alpha = \frac{1 - err}{err}$$

- Results in classification "forest"
  - Pro: Better separtion power, more stable
  - Con: Lose simplicity of single tree

Decision Tree Testing

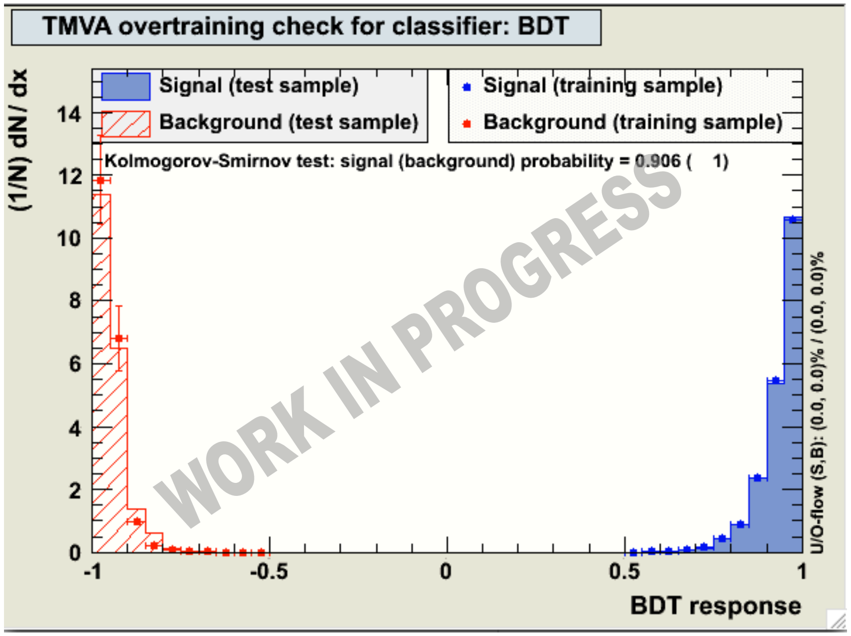
> Testing sample now run through "forrest"

- Each tree gives each event a value of h(x) = +1 (signal) or h(x) = -1 (background)
- Adaptive boosting yields the following classification

$$y_{ada} = \frac{1}{NTrees} \sum_{i=1}^{NTrees} \ln(\alpha_i) \bullet h_i(x)$$

> This quantity is the MVA output!!

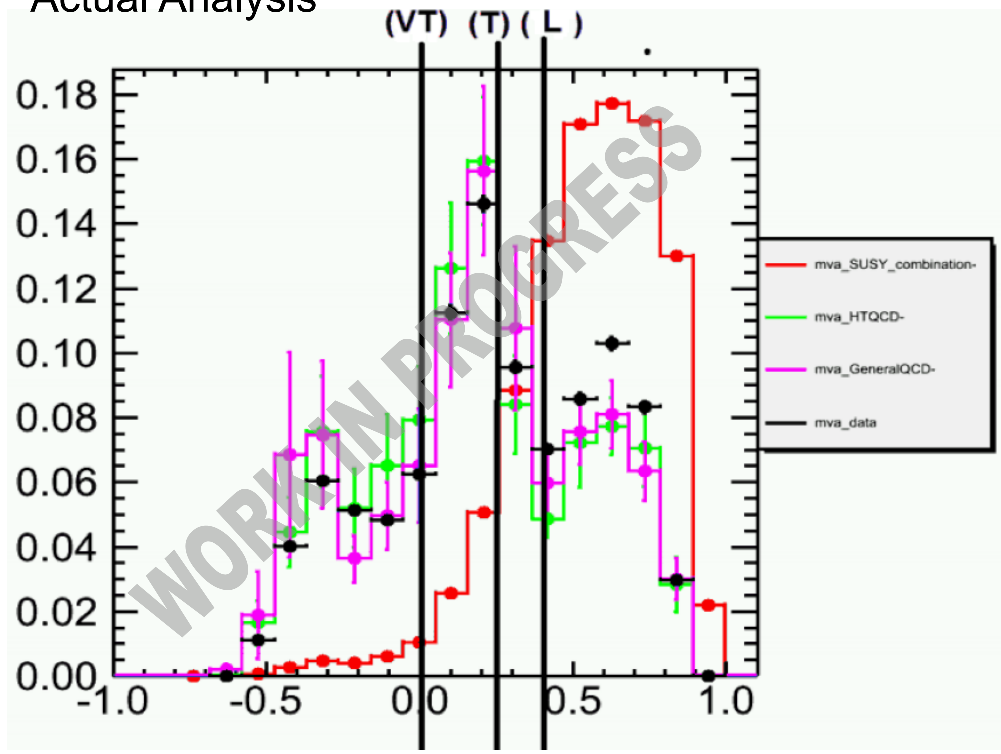
Ideal Plot



Current Progress

- **Cuts**
  - Very Tight: *mva < 0.45*
  - Tight: *mva* < 0.25
  - Loose: *mva* < 0.0
- **Key**
  - Red = Signal
  - Magenta = General QCD
  - Green = QCD as sum of two bins (50-100 GeV, 100-250 GeV)
  - Black = **Data**
- Working to isolate control region

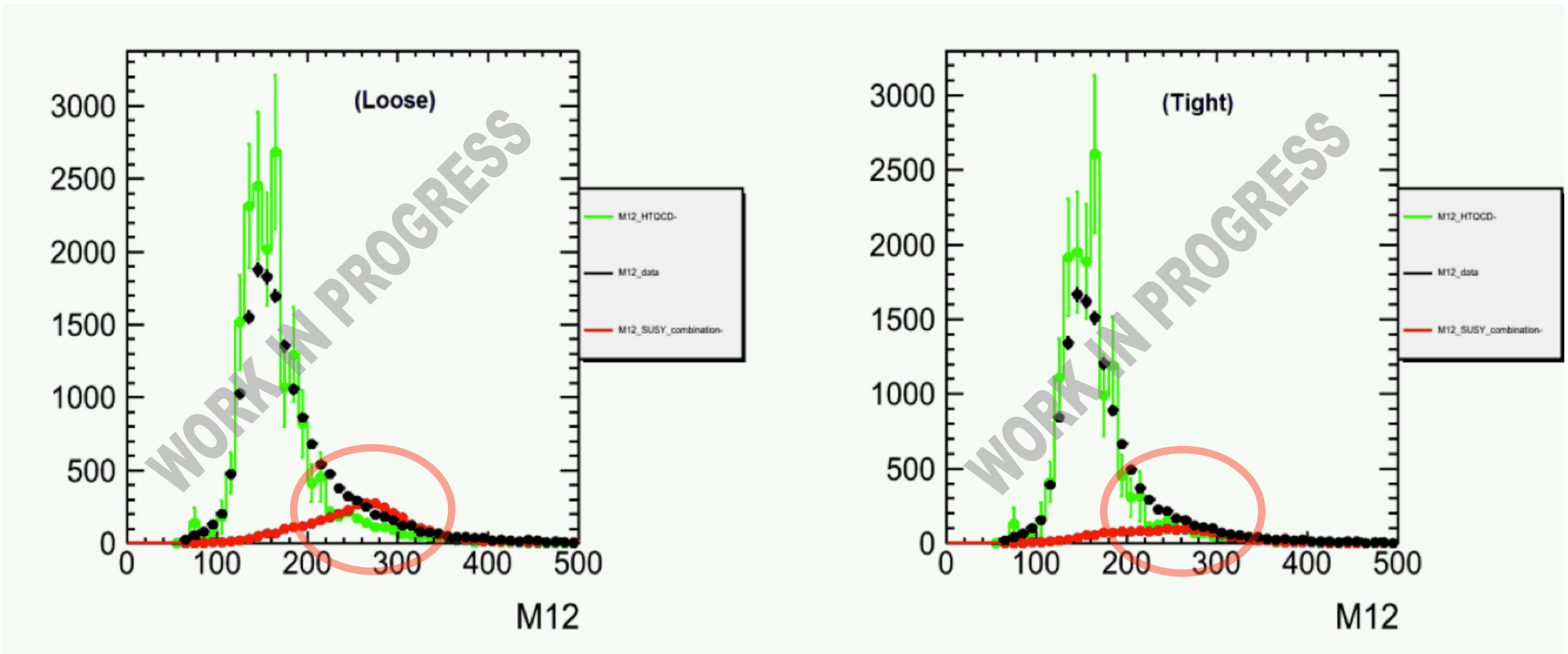
Actual Analysis



Suppression of Signal in Cut Region

> Results:

- Suppression of signal in di-jet mass distribution
- Background more or less unaffected



Et Voila

Questions?

Summary

Regression Analysis Method

> Step 1: Select a target

- $GenP_t^*$ = Event Generator
  $DetectedP_t^*$ = Detector Simulation
  $DetectedP_t$ = Experiment
  P_t = Result!

$$y = \frac{GenP_t *}{DetectedP_t *} \approx \frac{P_t}{DetectedP_t}$$

> Step 2: Use half of sample to map this target's function dependence on a list of variables

$$y = f\left(x_1, x_2, x_3 \ldots\right)$$

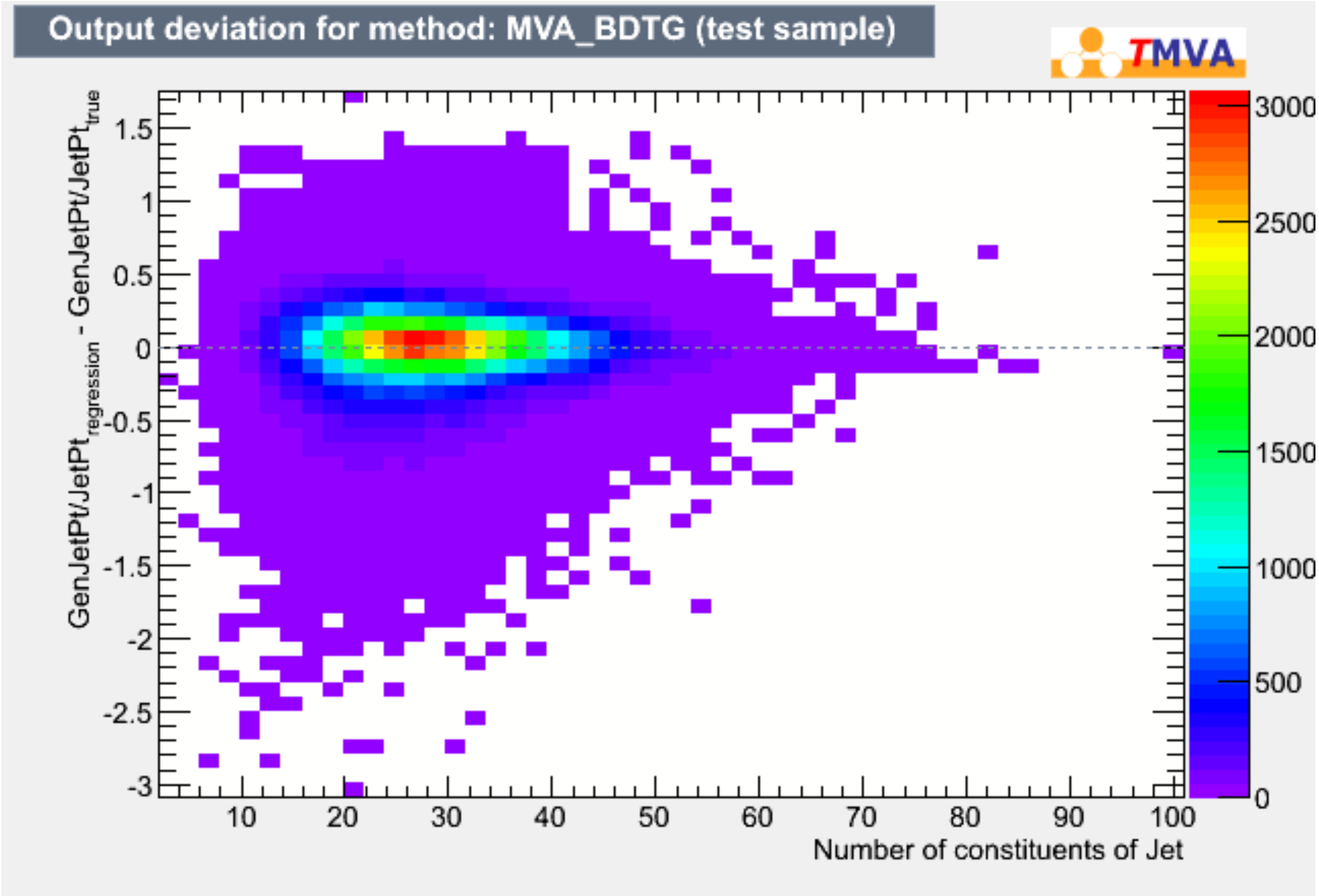
> Step 3: Use other half to calculate P_t

$$y \bullet DetectedP_t = P_t$$

KEY: Can use functional dependence to **limit sample** to region where generation/simulation is closest to reality/data.

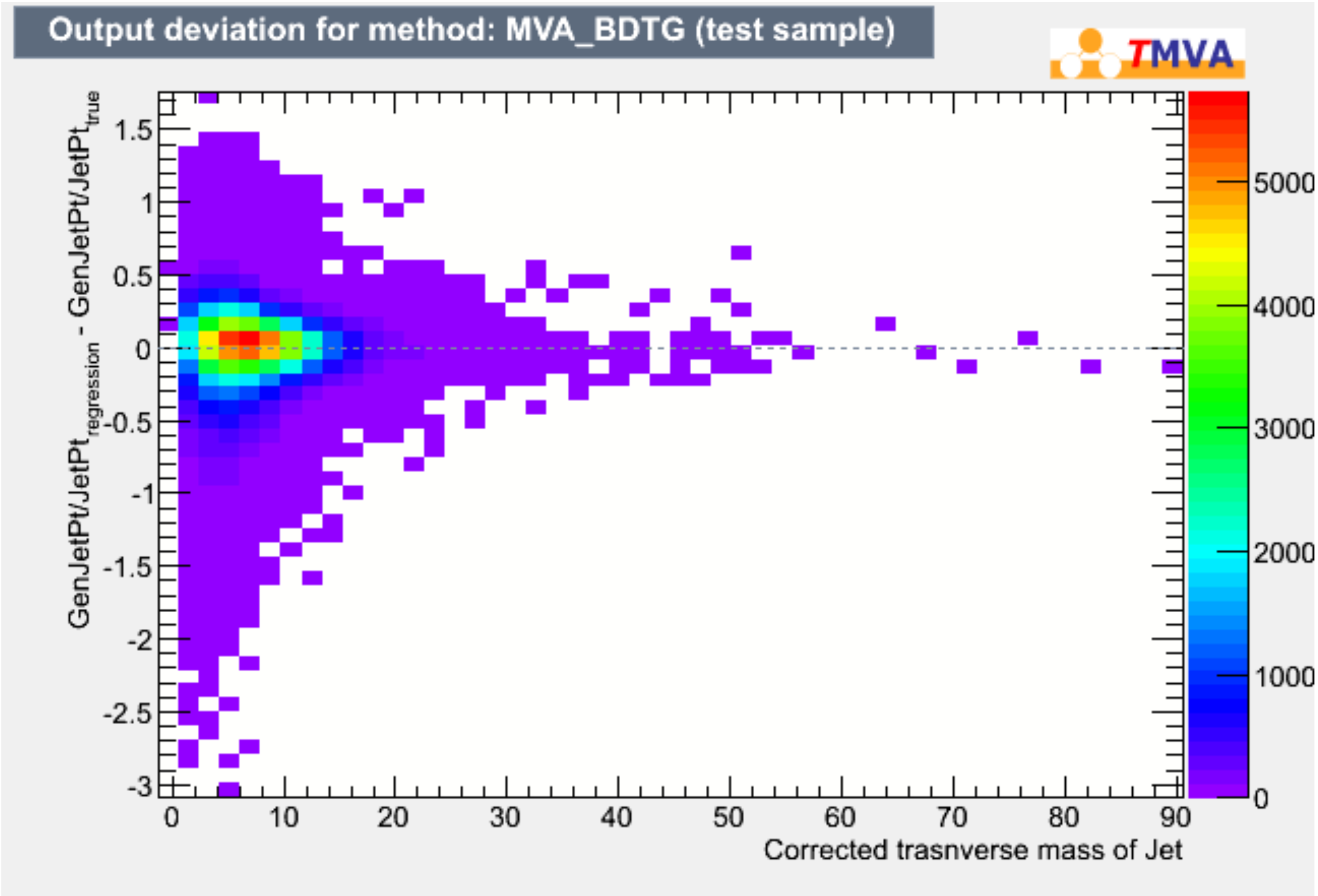
$$\frac{GenP_t *}{DetectedP_t *} - \frac{P_t}{DetectedP_t} = \Delta = smallest$$

The Good Plot



Not Final: Work in Progress

The Bad Plot



Not Final: Work in Progress