Tuning the simulated response of the CMS detector to b-jets using Machine learning algorithms

Krunal Gedia
ETH Zürich

On behalf of the CMS Collaboration

SPS Annual Meeting 2018
31st August 2018
EPFL, Lausanne

Motivation

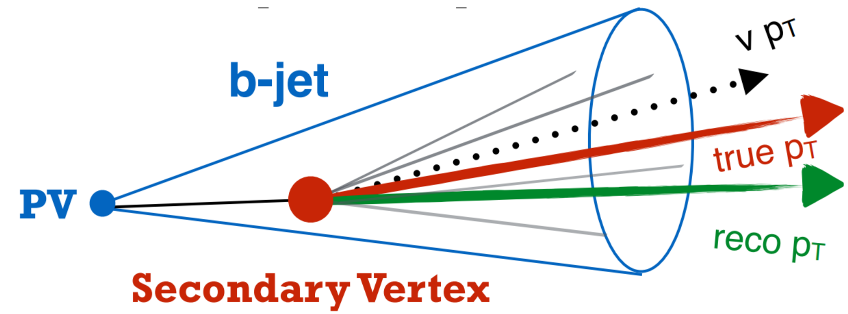
bjets → semi-leptonic decays
→ wider than light jets

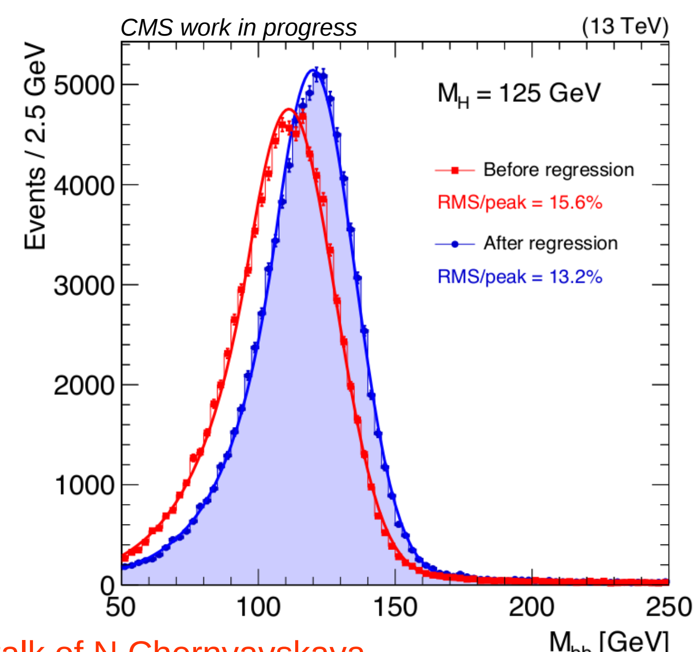
Neutrino escapes detection!

reco $p_T \neq$ gen p_T → Reconstruction of $\boldsymbol{X \rightarrow bb}$ difficult

Energy correction (gen p_T / reco p_T) ← o/p of NN-based bjet energy regression*

Uses **42 jet variables as inputs**





In general,
MC variable distribution ≠ data variable distribution

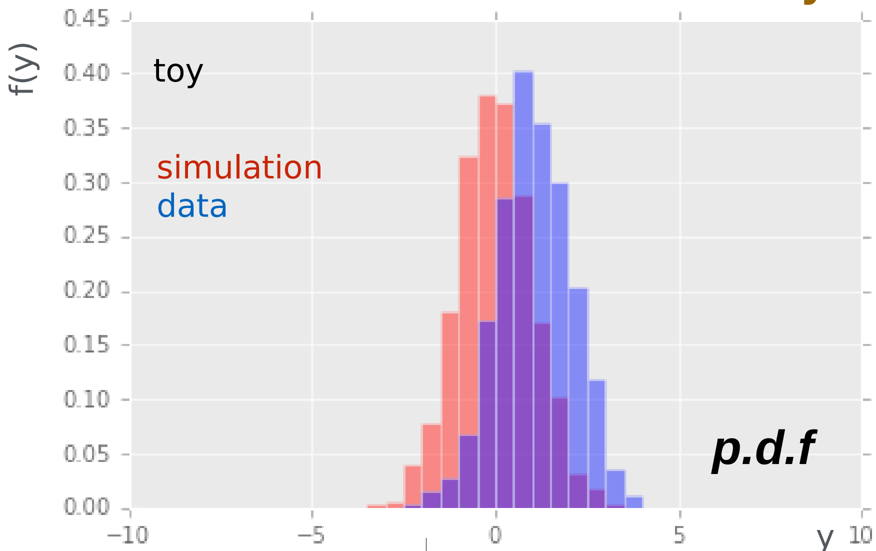
- Mis-modelling of detector.
- Mis-modelling of jet fragmentation.

Good data/MC matching is required!

*see talk of N.Chernyavskaya

Direct Quantile Morphing

Corrections by matching cumulative distributions!

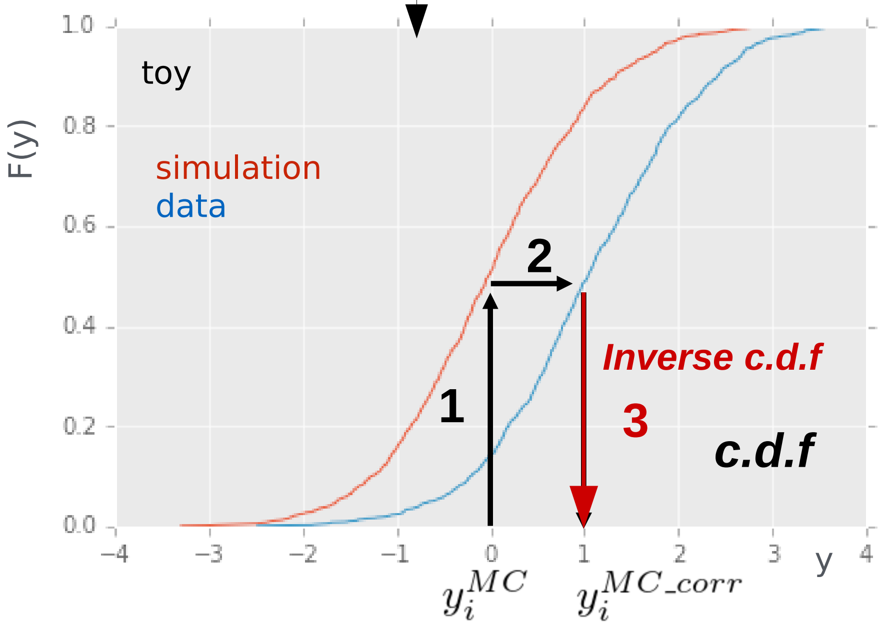


Requirements:
1. Continuous, differentiable, non-constant pdf
2. Pure control sample

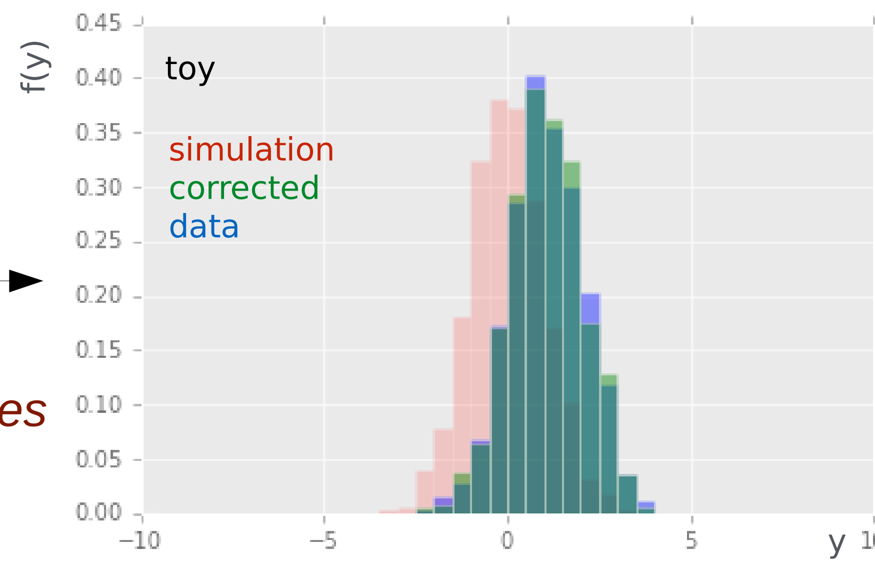
Disadvantages:
1. Correction not differential in kinematics and pile-up.

Goal: Differential corrections!

Integrate



Match Quantiles



Quantile Regression Morphing

Differential corrections by matching conditional cumulative distributions

Direct Quantile Morphing :	*cdf*	Match quantiles →	*Corrected MC*
Quantile Regression Morphing:	*conditional cdf* $F(y_i \mid \overline{x})$	Match quantiles at fixed $\overline{\mathbf{x}}$ →	*Corrected MC*

In practice:

Discretize cdf → estimate discrete quantiles $q_\tau(x_i)$ → linearly interpolate to get cdf

↓

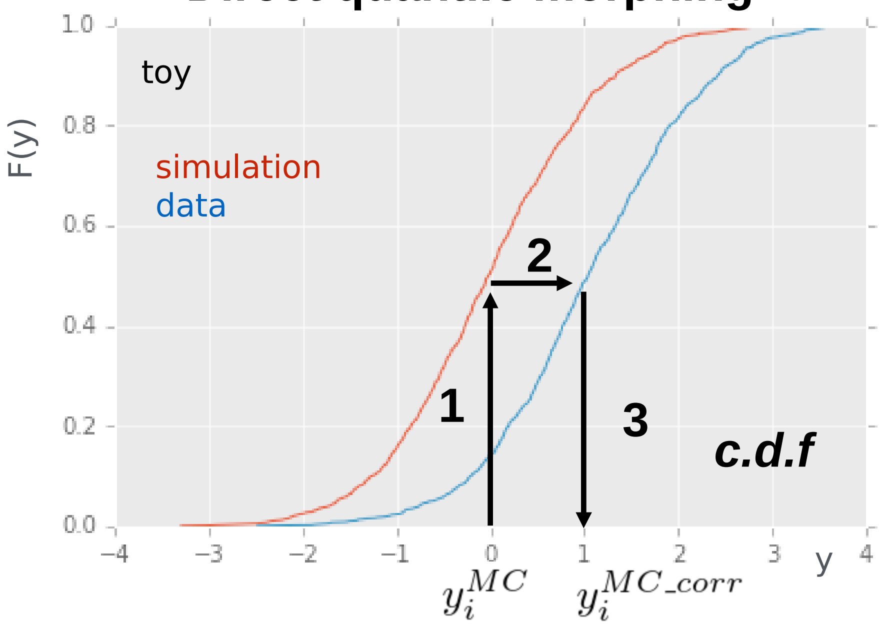
Train grad-BDT to minimize quantile loss function...scikit-learn package

Dependent variables: **Regression i/p variables** $y = [\text{secondary vertex, soft lepton ...40 variables}]$

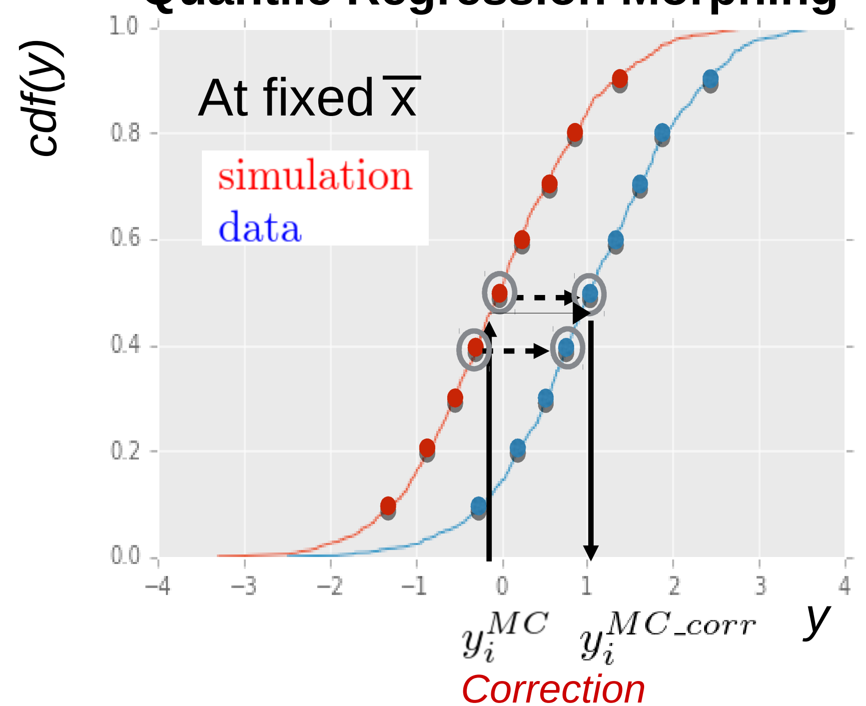
Independent variables: **Kinematics and pile-up** $\bar{x} = [p_T,\ \eta,\ \phi,\ \rho]$

Quantile levels: [19 levels] τ = [0.05, 0.10, 0.15,….,0.90, 0.95]

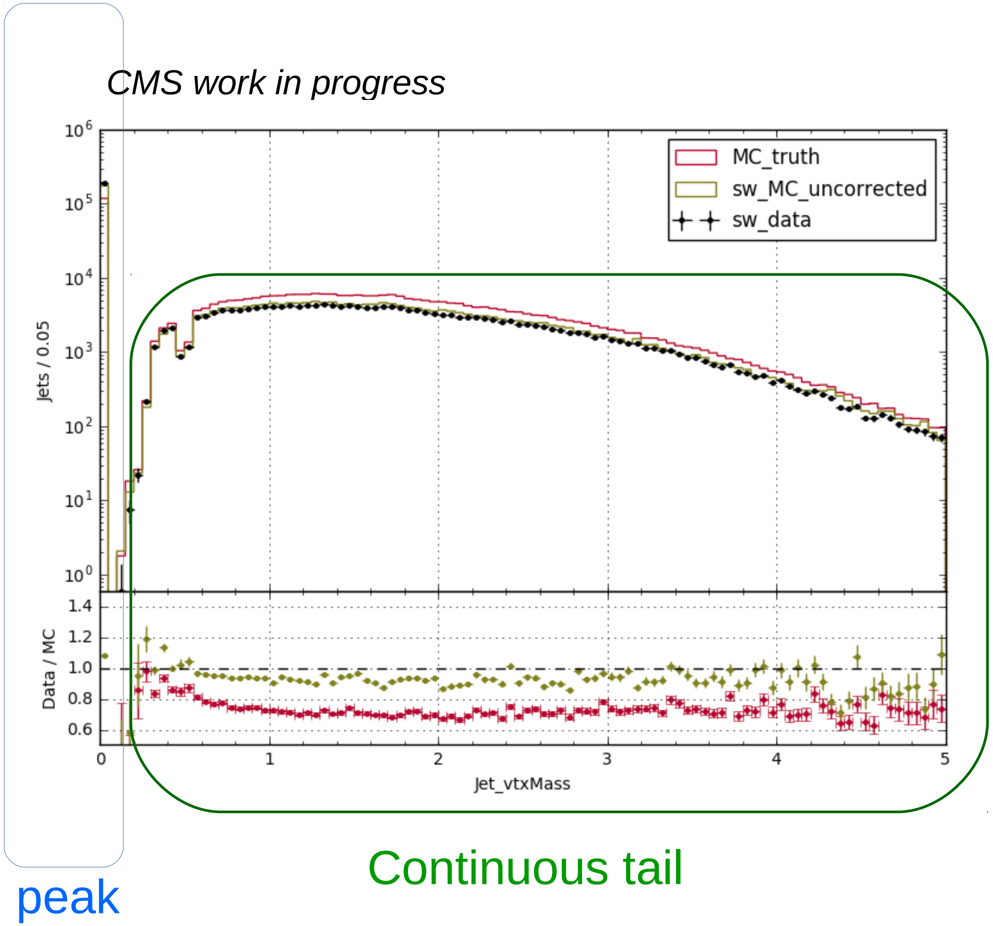
Direct quantile morphing



Quantile Regression Morphing



Limitations of Quantile Morphing:
pdf should be continuous and differentiable!



Solution: Stochastic Morphing

Train a binary classifier and get prediction of probabilities of the MC event to be in peak and tail for MC and data.

$$P_{peak}^{MC}, P_{tail}^{MC}, P_{peak}^{data}, P_{tail}^{data}$$

Move the MC event from peak to tail if

$$P_{tail}^{data}(y_i) > P_{tail}^{MC}(y_i)$$

Or move MC event from tail to peak if

$$P_{peak}^{data}(y_i) > P_{peak}^{MC}(y_i)$$

Then perform quantile morphing for the tail!

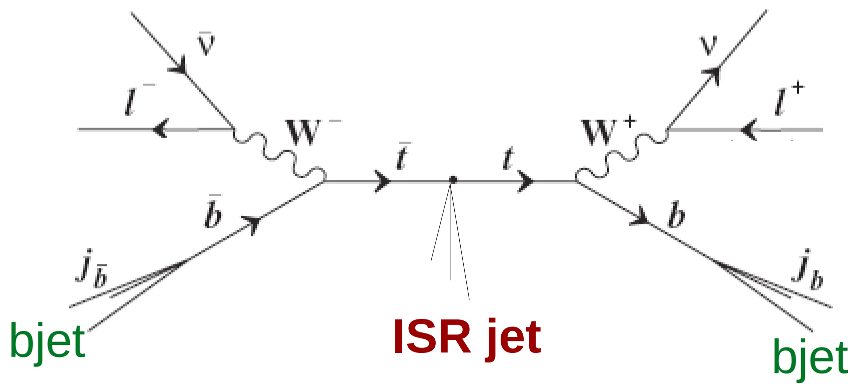
Limitations of Quantile Morphing:
Pure sample of bjets required!

Partial solution: bjets from **leptonic eμ decay channel of ttbar**.

(To avoid non-bjets from hadronic decay)

Did we avoid background? …………….. NO!

Presence of **ISR jets**



Solution 2: A statistical tool named sPlot*

*arXiv:physics/0402083

Basic idea of sPlot technique:

Reweight data set in unbiased way such that signal-like events get higher weight than background-like events.

No b-tagging variables, **only kinematics**

$$m_{(nl,j)}, \Delta\phi_{(nl,j)}, \Delta\eta_{(nl,j)}, \Delta\phi_{(nl+j,ll)}, \Delta\eta_{(nl+j,ll)}$$

$$m_{(fl,j)}, \Delta\phi_{(fl,j)}, \Delta\eta_{(fl,j)}, \Delta\phi_{(fl+j,ll)}, \Delta\eta_{(fl+j,ll)}$$

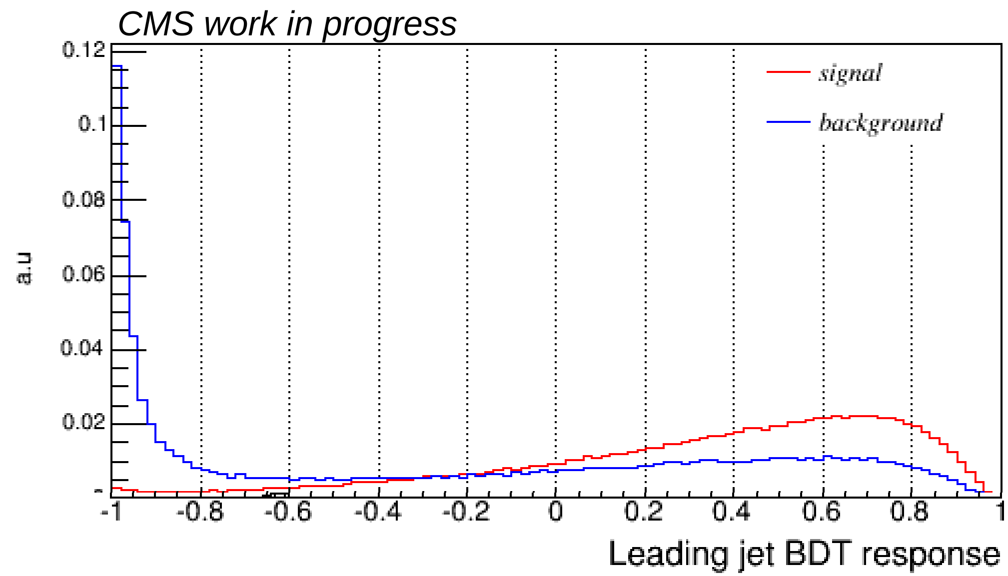
$$\Delta\phi_{(ll,j)}, \Delta\eta_{(ll,j)}$$

BDT classifier response*

TMVA (ROOT) package

If $\Delta R(l_1, j) < \Delta R(l_2, j)$,
then l_1 = near lepton (nl) and l_2 = far lepton (fl) for jet j in an event.

Discriminate between bjets and non-bjets based on their correlation with leptons

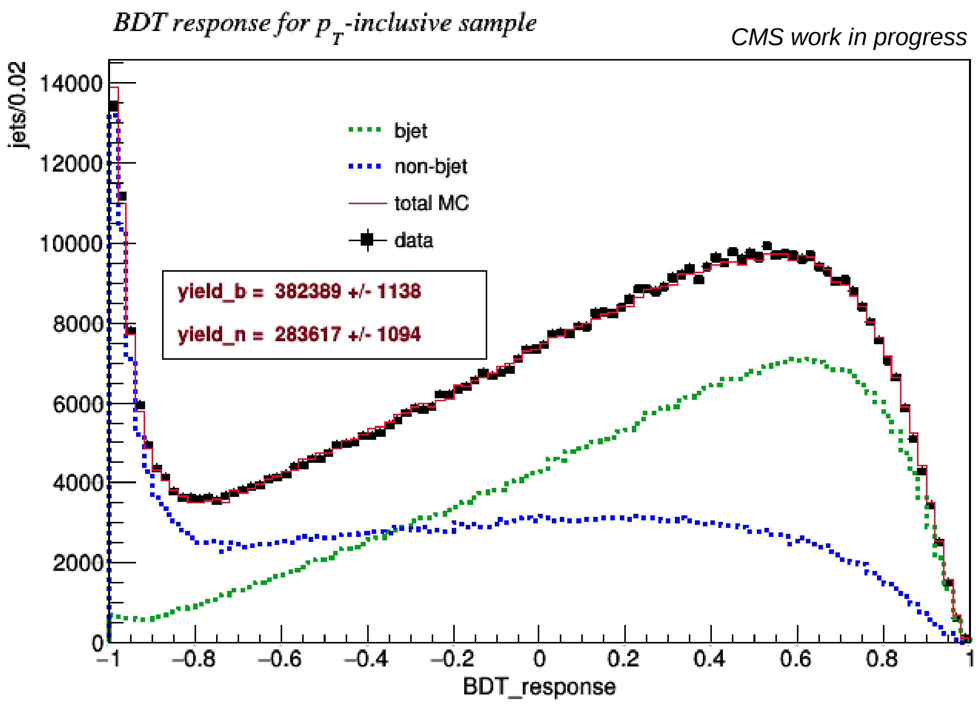


Done for all p_T ranked jets:
(i.e. leading jet, sub-leading jet and other jets)

Compute ‘sWeights’ for both data and MC

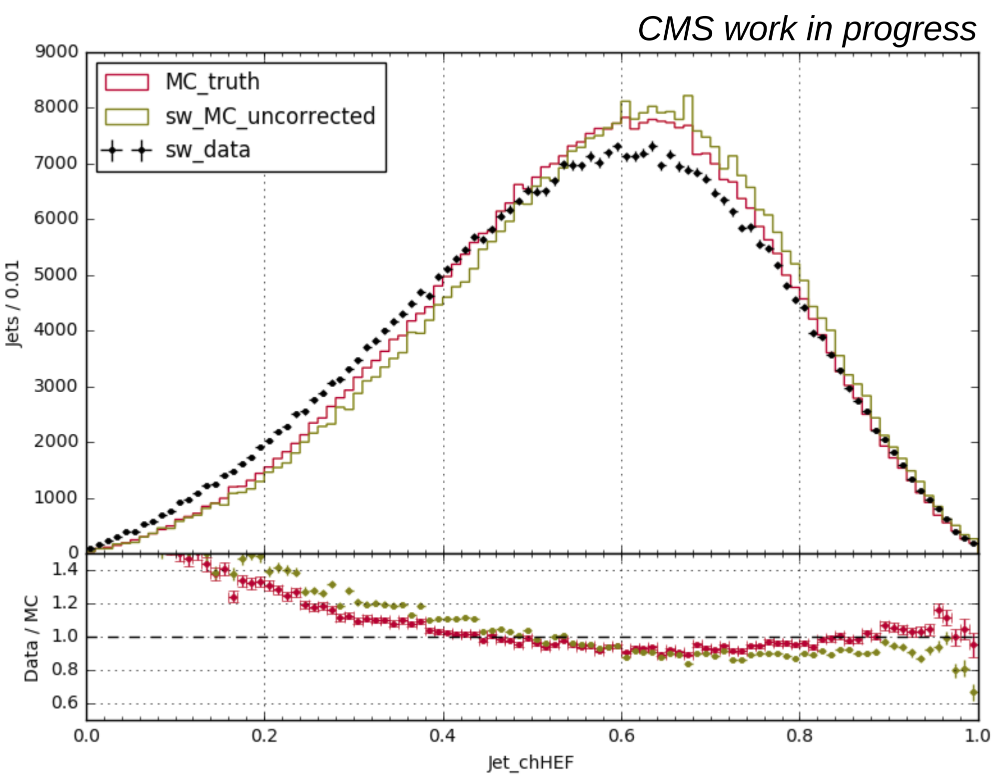
↓

Parameters for ‘sWeights’ are obtained using **Likelihood fit to BDT**

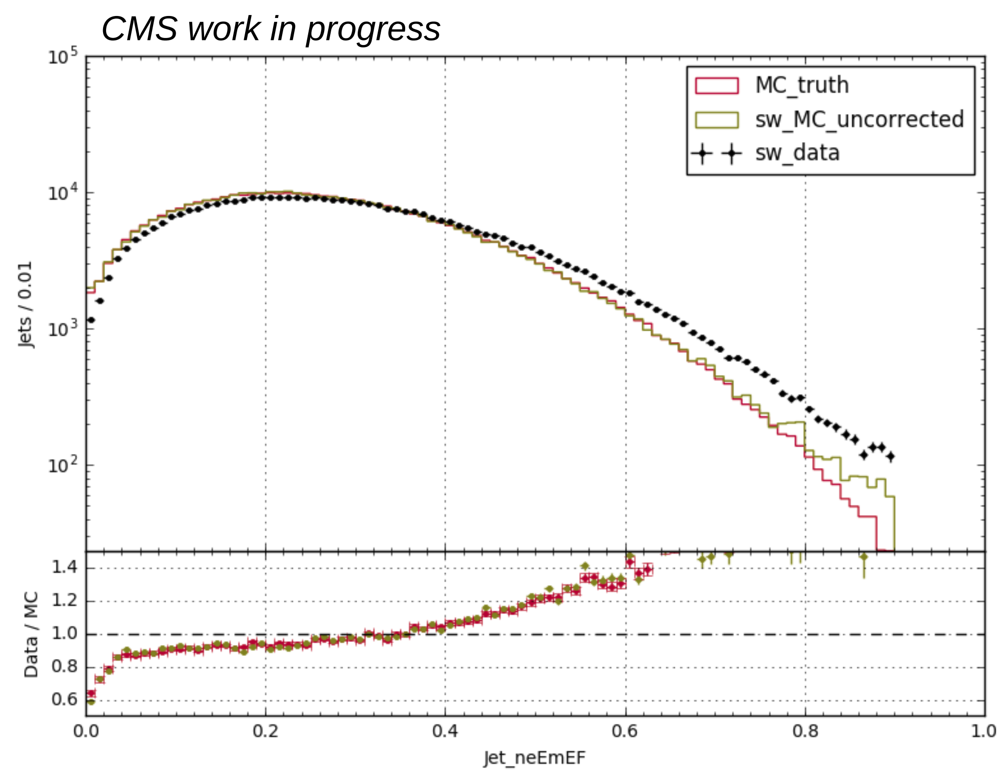


$$ {}_s\mathcal{P}_{sig}(bdt_i) = \frac{V_{ss} f_s(bdt_i) + V_{sb} f_b(bdt_i)}{N_s f_s(bdt_i) + N_b f_b(bdt_i)} $$

Weight dataset by sWeights to get signal distribution!



Charged energy fraction in HCAL

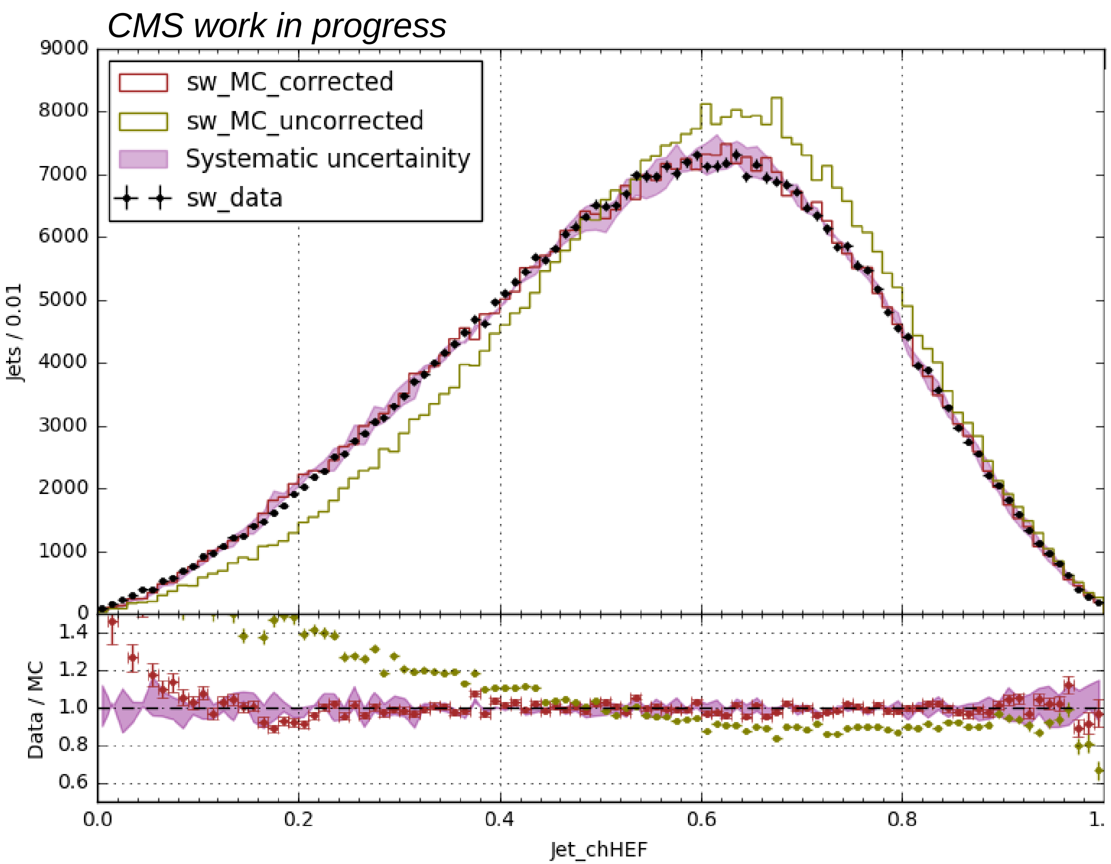


Neutral energy fraction in ECAL

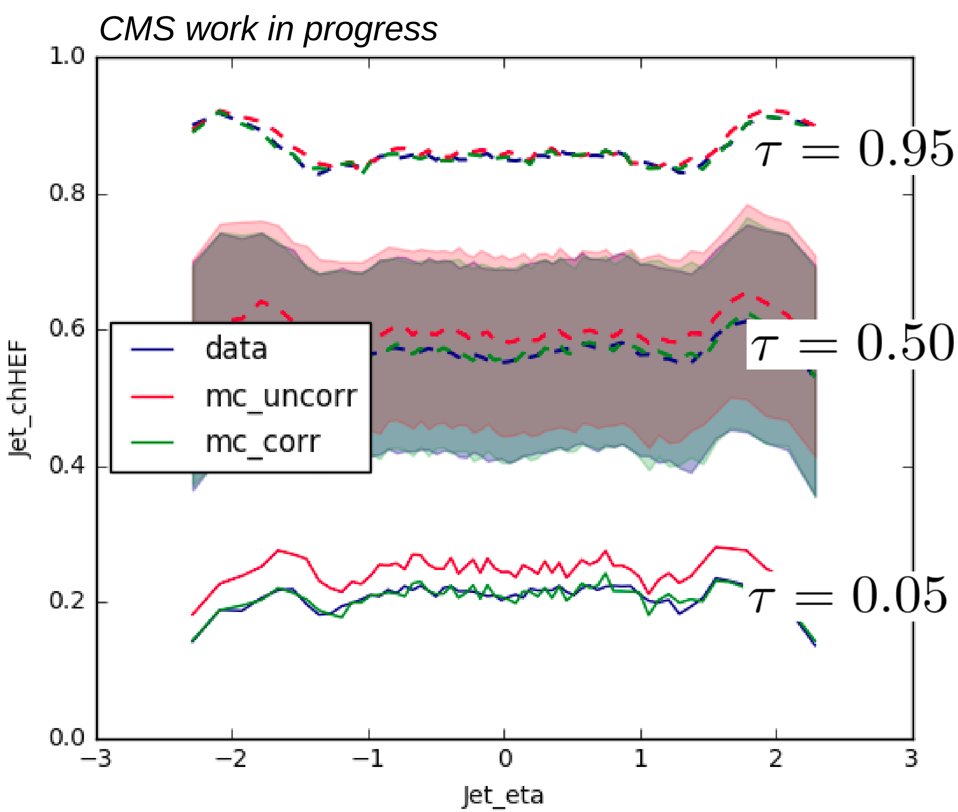
<u>We compute sPlots of MC to study the bias produced by sPlot technique.</u>

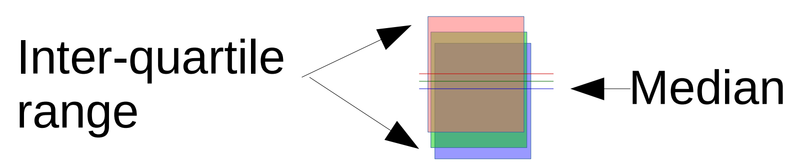
Results (1)

Charged energy fraction in HCAL



Differential in η?

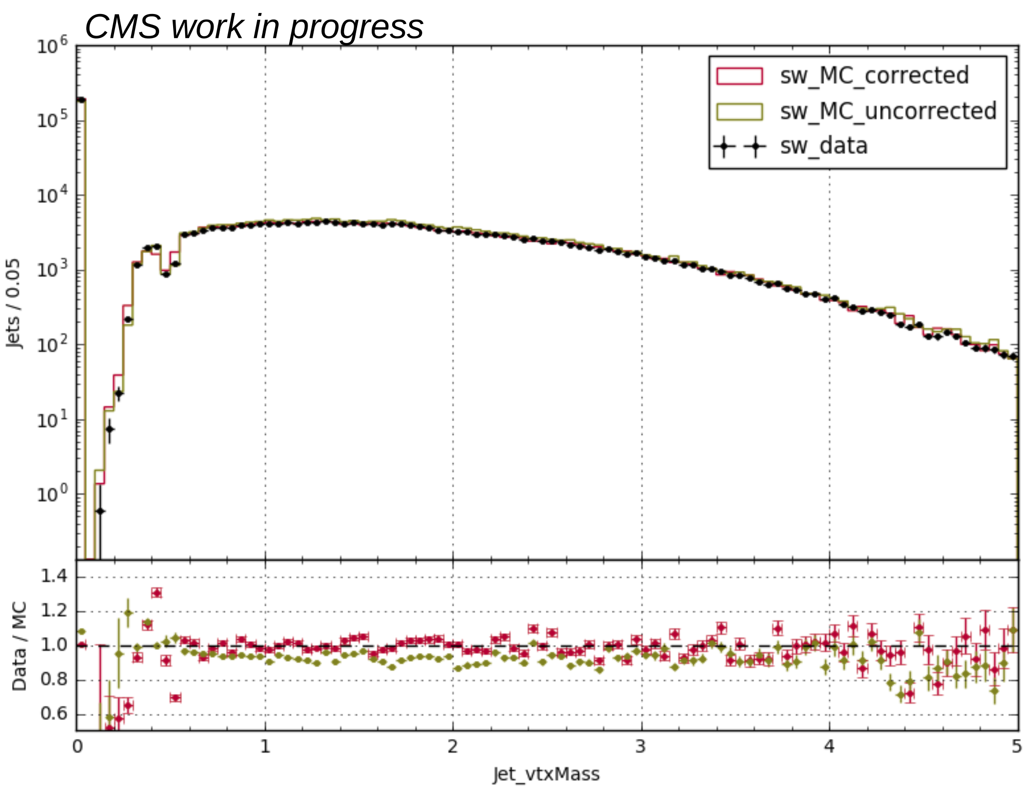




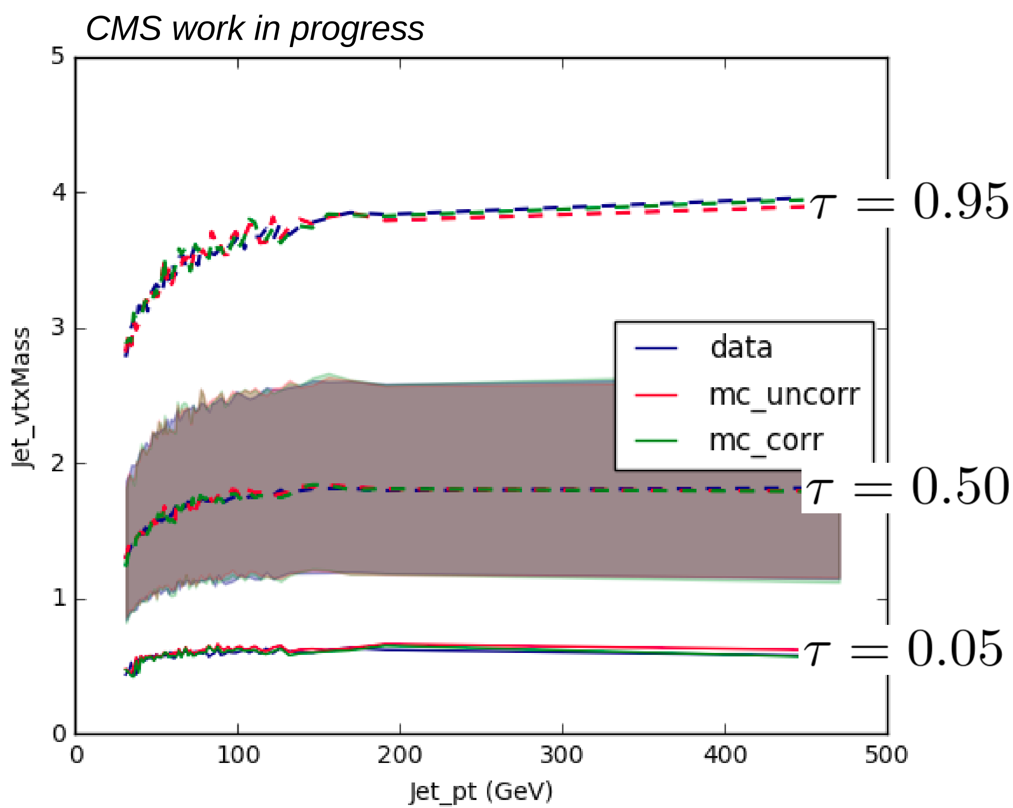
Results (2)

Stochastic corrections:

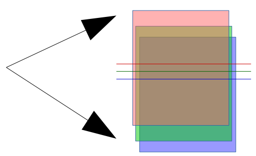
Secondary vertex mass



Differential in p_T?



Inter-quartile range

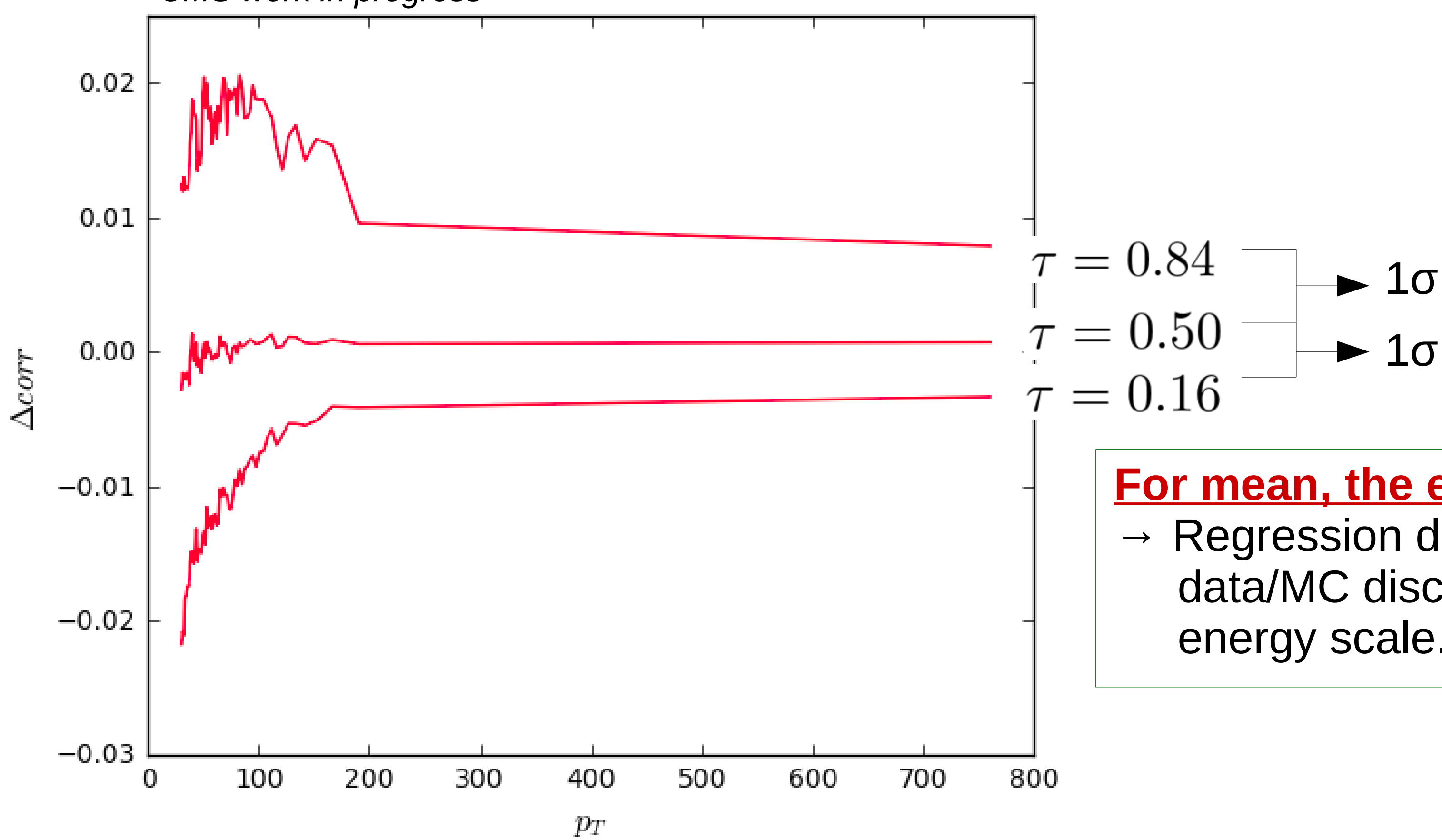


Median

Results (3)

Effect on jet energy corrections wrt to jet p_T due to data/MC corrections for NN regression is **<2% for 1σ deviation → Effect on resolution is <2%**.

$$\Delta corr = \left(\frac{p_T^{gen}}{p_T^{reco}}\right)_{MC_{uncorr}} - \left(\frac{p_T^{gen}}{p_T^{reco}}\right)_{MC_{corr}}$$



For mean, the effect is ~ 0.1%

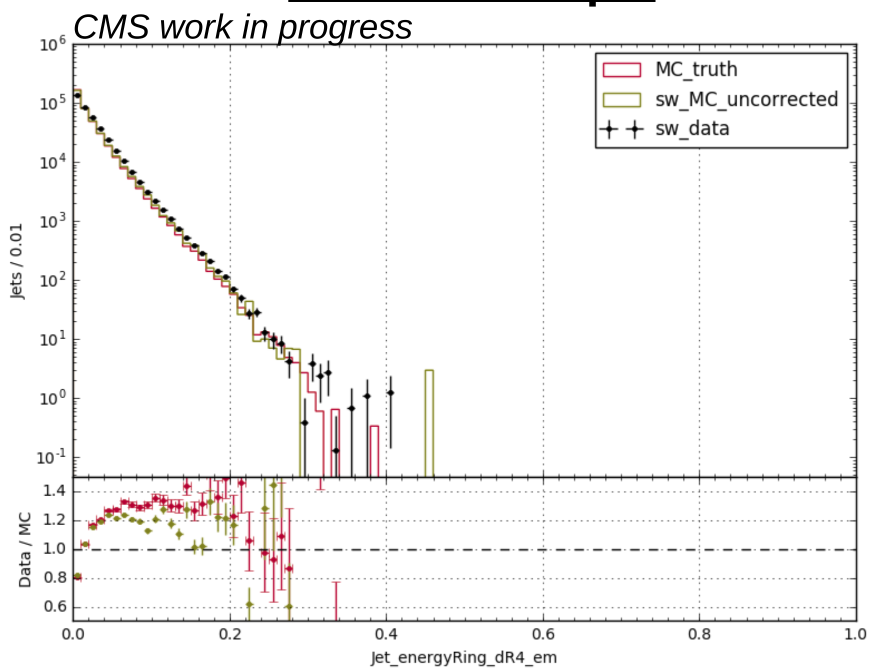
→ Regression does not introduce data/MC discrepancies on energy scale.

**Upto the current status of analysis.*

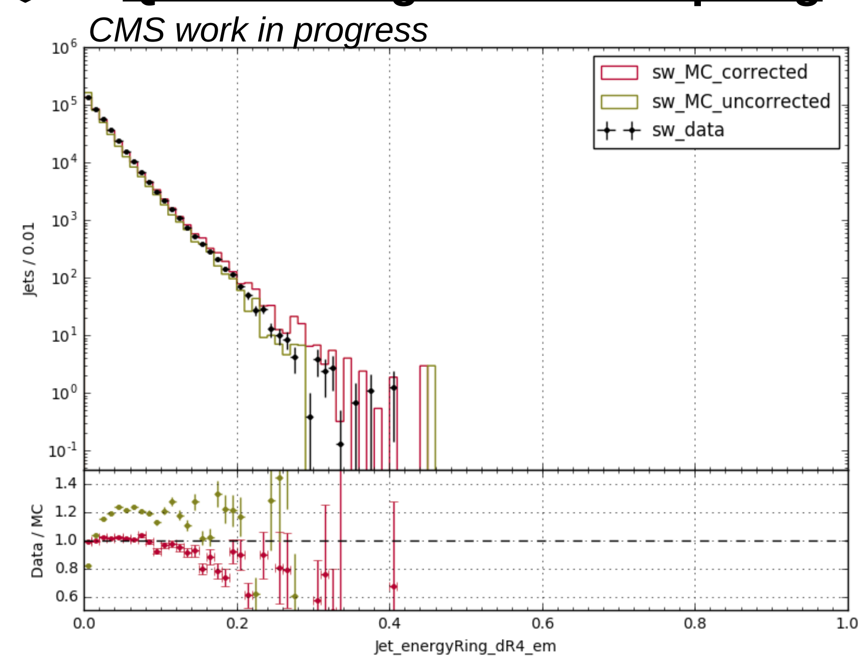
Summary

- Motivation: Match data/MC bjet distribution for better reconstruction of $X \rightarrow bb$.
- Differential correction: Quantile Regression Morphing and Stochastic Morphing
- bjet sample used: ttbar leptonic decay (with ISR background)
- Pure sample of bjet: sPlot technique
- The corrections derived are jet-by-jet and largely analysis independent !

sPlot Technique



Quantile Regression Morphing



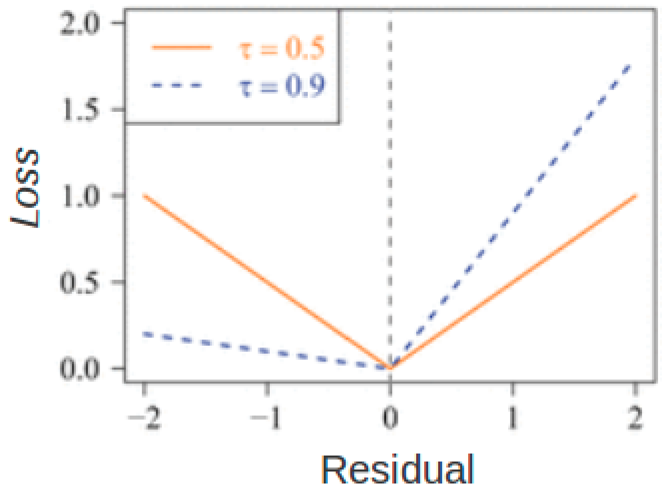
Jet electromagnetic energy fraction in ring ΔR = 0.4-0.3

Back-up

Training Stage:

Train a grad-boosted decision tree to minimize the "*quantile loss*" function

$$
\begin{aligned}
Loss &= \tau * |y_i - q_\tau(\bar{x})| & \text{if} \quad y_i - q_\tau(\bar{x}) > 0 \\
&= (1-\tau) * |y_i - q_\tau(\bar{x})| & \text{if} \quad y_i - q_\tau(\bar{x}) < 0
\end{aligned}
$$



to estimate conditional quantile function $q_\tau(x_i)$

$$\hat{q}_\tau(\bar{x}) = \underset{(q_\tau(\bar{x}))\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{N} \text{Loss}(y_i, q_\tau(\bar{x}))$$

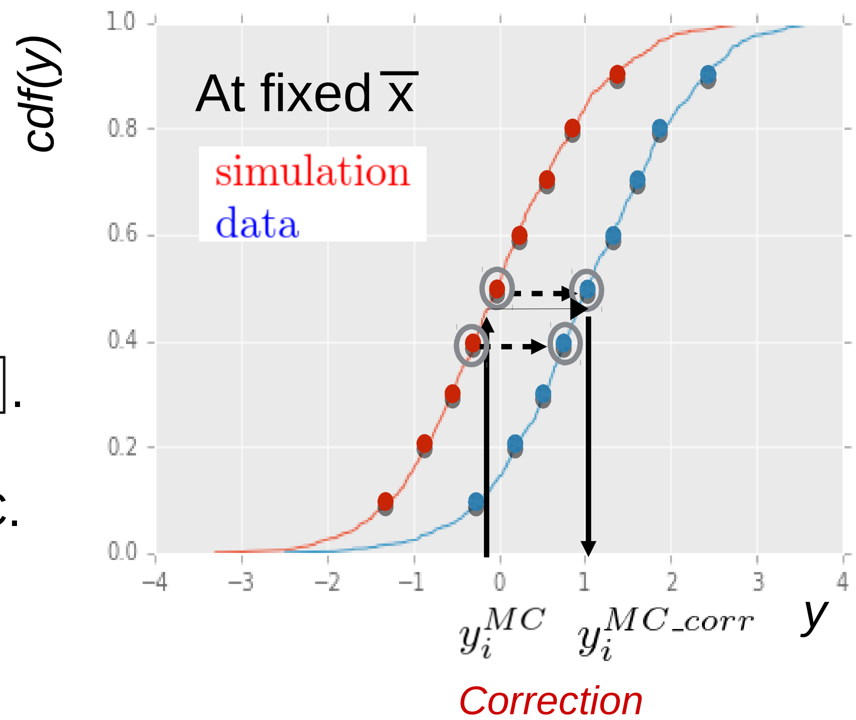
for data and MC for each quantile value τ .

From sklearn.ensemble import GradientBoostingRegressor

Application Stage: Differential corrections

For each MC variable $\mathbf{y}_i$ to be corrected

1. Compute $[q_\tau^{data}(x_i)]$ and $[q_\tau^{MC}(x_i)]$.
2. Run binary search of $\mathbf{y}_i$ on $[q_\tau^{data}(x_i)]$ and $[q_\tau^{MC}(x_i)]$.
3. Use linear interpolation to estimate cdf of data and MC.
4. Match corresponding quantiles on both cdfs to get corrections.



Compute ‘sWeights’ for both data and MC

$$ {}_s\mathcal{P}_{sig}(bdt_i) = \frac{V_{ss} f_s(bdt_i) + V_{sb} f_b(bdt_i)}{N_s f_s(bdt_i) + N_b f_b(bdt_i)} $$

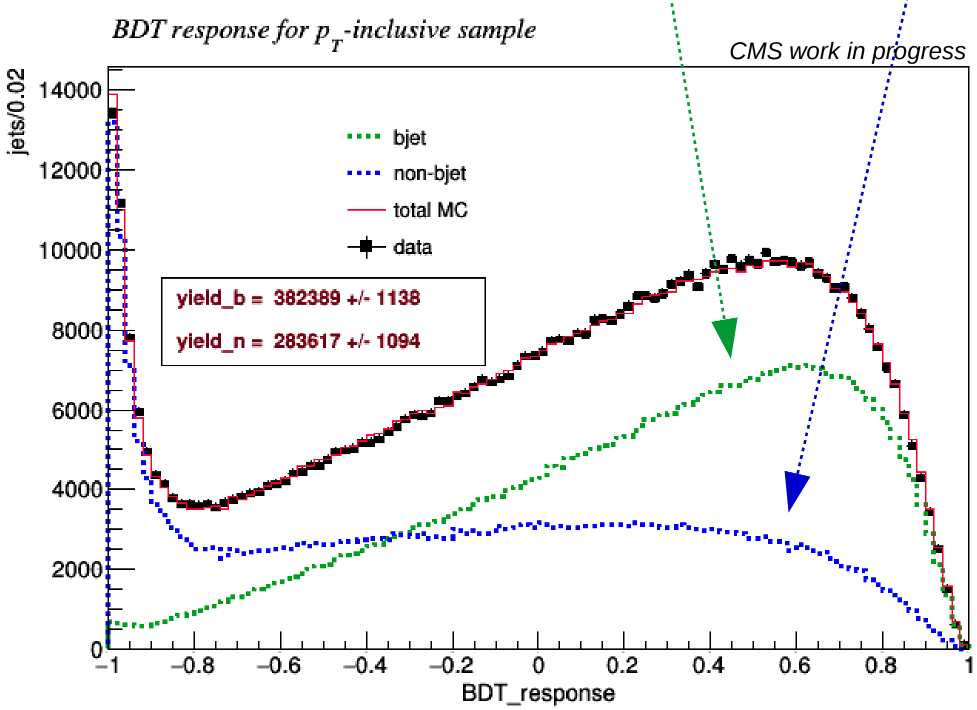
$$ V^{-1}_{nj} = \frac{\partial^2 \mathcal{L}}{\partial N_n \partial N_j} = \sum_{i=1}^{N} \frac{f_n(bdt_i) f_j(bdt_i)}{\left(\sum_{k=1}^{N_s} N_k f_k(bdt_i)\right)^2} $$

Obtained through likelihood fit

$$ \mathcal{L} = \sum_{i=1}^{N} log\Big(N_s f_s(bdt_i) + N_b f_b(bdt_i)\Big) - N $$

$f_s(bdt_i)$	Signal pdf
$f_b(bdt_i)$	Background pdf

MC Truth info



Fit over data

In practice: RooStats::Splot Class

→ Likelihood fit
→ sweights

We compute sWeights for MC as well so as to study the bias produced by the sPlot technique

Limitations of sPlot technique:
BDT should be uncorrelated with variables to be unfolded!

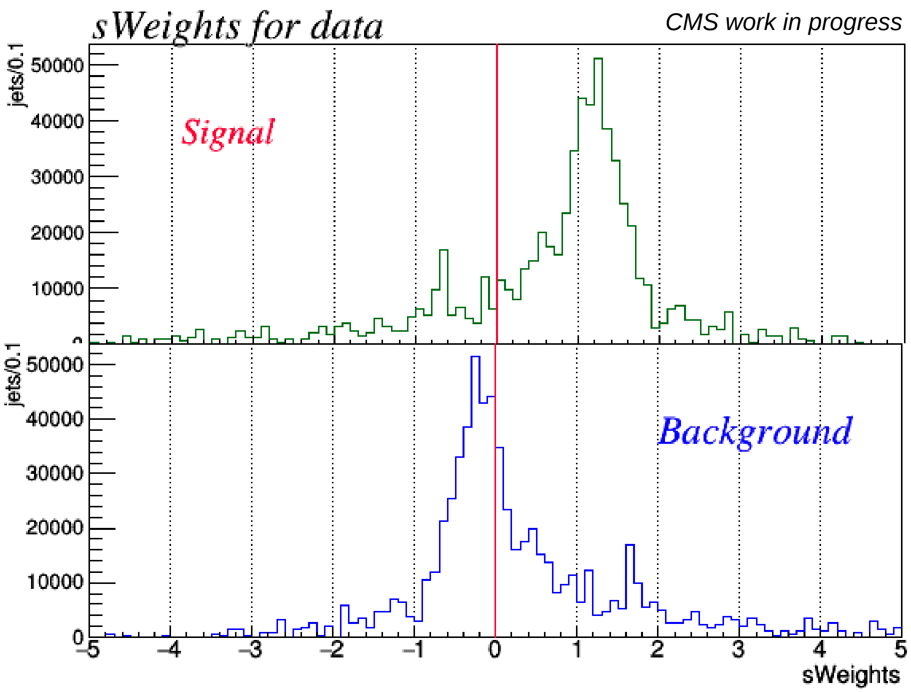
Since the discriminatory variable BDT depends on kinematic variables especially p_T

We have ***p_T binned the sample*** to make BDT conditionally independent of jet p_T

30-40, 40-50, 50-60, 60-70, 70-80, 80-100, 100-120, 120+ GeV

Further, in our analysis, events with only **positive sWeights** are taken.

For *minimizing the loss function* quantile regression



Back-Up: Migrating events from peak to tail

Let the events at peak be labelled as Class 0 and the ones in the tail as Class 1.

Case 1: Suppose a particular event is at peak in MC. If for the same input, if the probability of data to be in tail > probability of MC to be in tail i.e.

$$\text{If} \quad P_1^{data}(x_i) > P_1^{MC}(x_i)$$

The probability with which that event has to be moved from peak to tail is:

$$w * P_0^{MC}(x_i) + P_1^{MC}(x_i) = P_1^{data}(x_i) \implies w = \frac{P_1^{data}(x_i) - P_1^{MC}(x_i)}{P_0^{MC}(x_i)}$$

i.e. for a random number $z \in [0, 1]$, if $z < w$, we move the event from peak to tail, else not.

In case of moving an event from peak to tail, we move it to tail according to the pdf/cdf of the tail. The cdf of tail is obtained by linearly interpolating the quantiles estimated as obtained previously.

Back-up: Migrating events from tail to peak

Case 2: Now, suppose a particular event is in tail of MC. If for the same input (x), if the probability of data to be at peak > probability of MC to be at peak i.e.

$$\text{If} \quad P_0^{data}(x_i) > P_0^{MC}(x_i)$$

The probability with which that event has to be moved from tail to peak is:

$$w * P_1^{MC}(x_i) + P_0^{MC}(x_i) = P_0^{data}(x_i) \implies w = \frac{P_0^{data}(x_i) - P_0^{MC}(x_i)}{P_1^{MC}(x_i)} \tag{6.12}$$

i.e. for a random number $z \in [0,1]$, if $z < w$, we move the event from tail to peak, else not.

Once this is done, we morph the events in tail of MC to data.

Back-up: Correlations while migrating events from peak to tail

In case of correlated variable set like:
The secondary vertex attributes of the jet

$$m^{vtx},\ p_T^{vtx},\ d_{3D}^{vtx},\ s_{3D}^{vtx}$$

d→ distance b/w PV and SV
s → significance

We train quantiles for

Remove the one which has to be corrected!

$$x = [p_T,\ \eta,\ \phi,\ \rho]$$

Use it to get an estimate of cdf for morphing

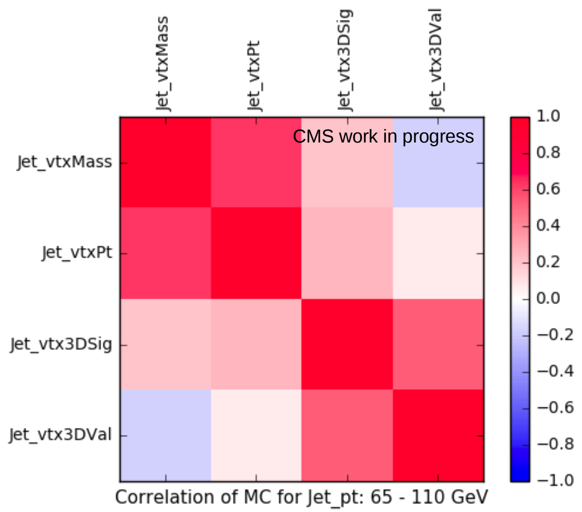
$$x = [p_T,\ \eta,\ \phi,\ \rho,\ m^{vtx},\ p_T^{vtx},\ d_{3D}^{vtx},\ s_{3D}^{vtx}]$$

Move an event as per the (pdf) cdf estimated by these quantiles in case of migrating an event from peak to tail.

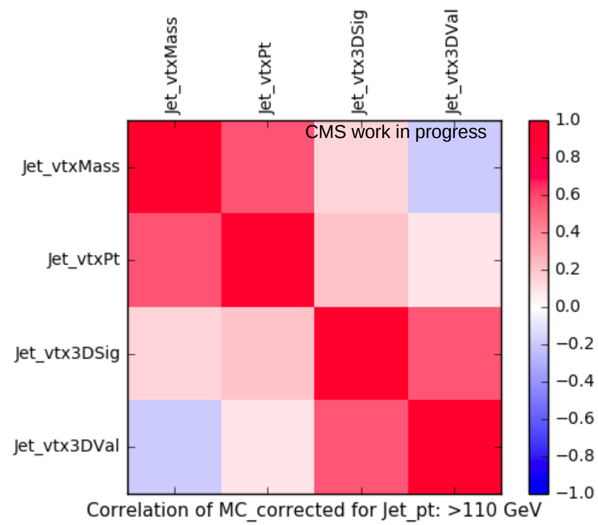
By doing this correlation with other correlated is taken into account.

Back-up: Correlation plots for secondary vertex variables

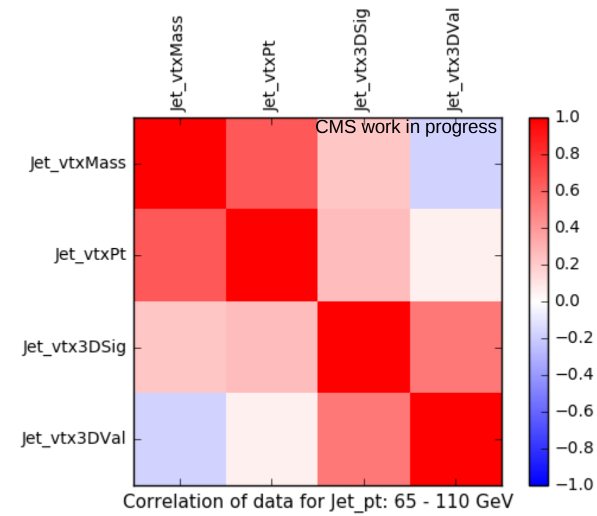
MC



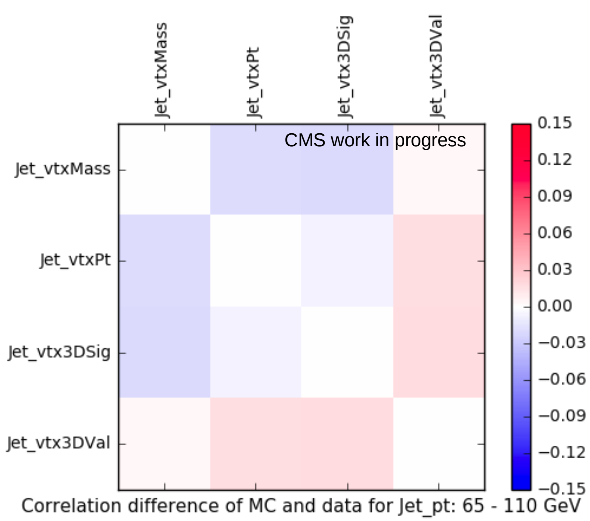
Corrected MC



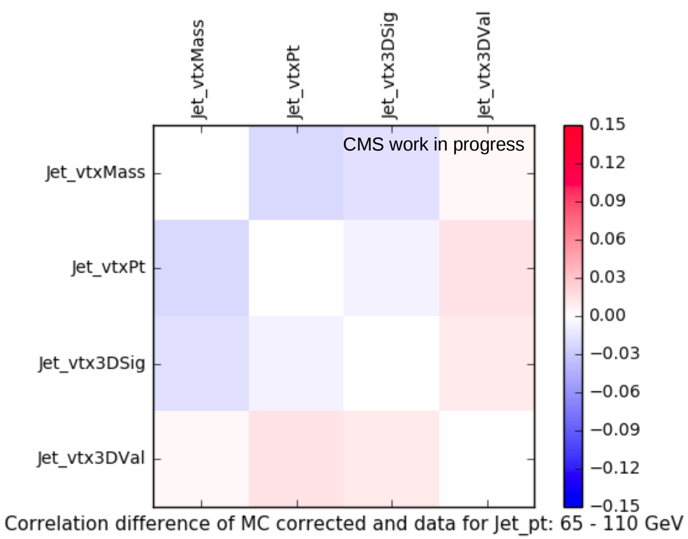
data



MC - data



Corrected MC - data



Jet p_T: 65-110 GeV