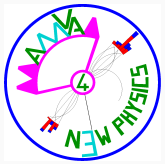
# Background modelling by hemisphere mixing

## Statistical Methods for Physics Analysis in the XXI Century
## CONF12 - Thessaloniki

Pablo de Castro [1] Alexandra Carvalho [1] Martino Dall'Osso [1] Tommaso Dorigo [1] Mia Tosi [2]

September 2, 2016





[1]Universita e INFN, Padova

[2]CERN

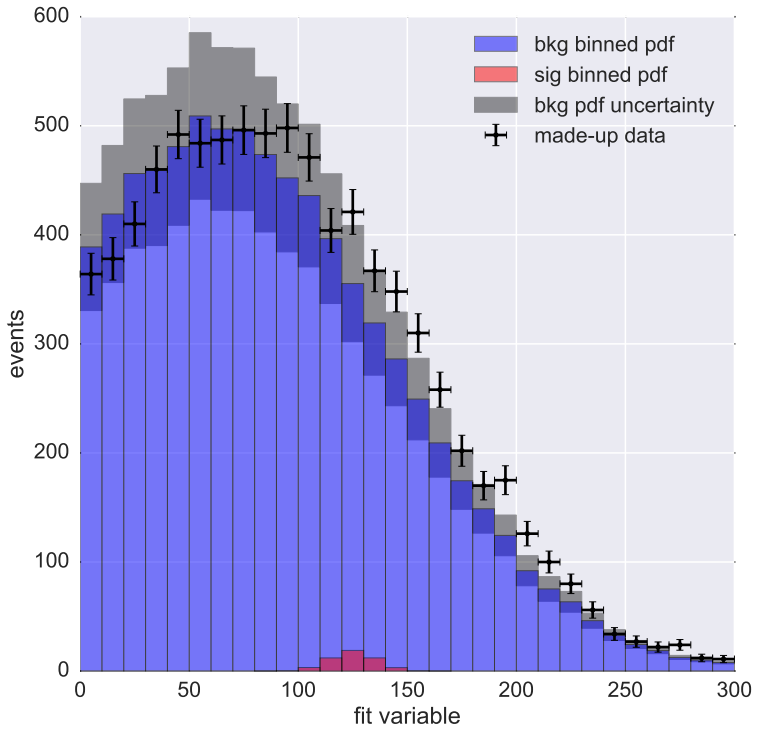
# Research problem description

Doing statistical inference on a LHC dataset, want to quantify or set limits on the the fraction of signal (physical process of interest which is studied) present in a certain amount of data.

## Focus

Analyses fitting a small peaking signal within large background (e.g. $HH \rightarrow b\bar{b}b\bar{b}$ and QCD multijet).

Then, a way to accurately model the bkg. is essential for carrying out a powerful analysis in this scenario.



# Background modelling

Precise knowledge of the background *pdf* as a function of variables of interest of fit is crucial for statistical inference.

## MC simulation driven

Estimate *pdf* from simulated backgrounds events.

Might need very high statistics; dominated by modelling uncertainties .

## Parametric

Use a functional parametric *pdf*, á la $H \rightarrow \gamma\gamma$.

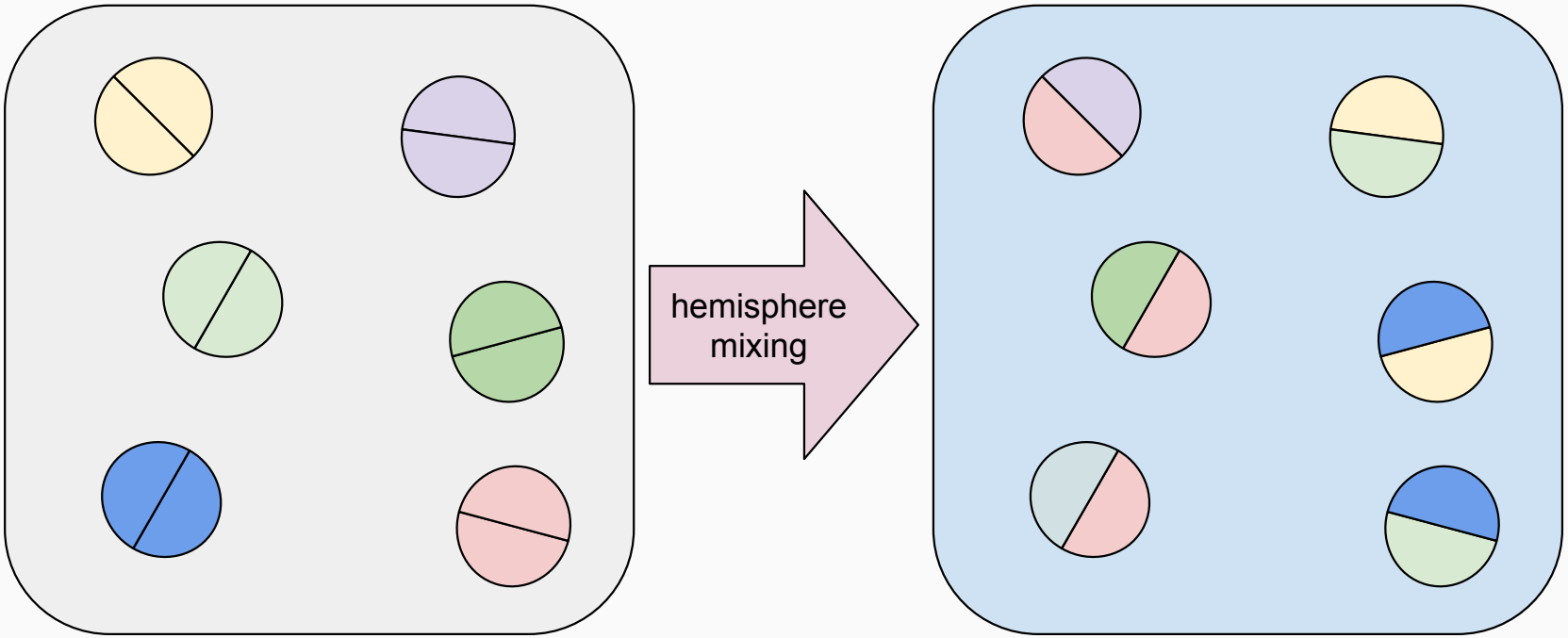
Very difficult when complexities present (e.g. comp. trigger or 2D fit).

## Non-parametric

Extract *pdf* from the data itself, without assuming func. form.

Typically based on extrapolating from "signal free" control region. **Hemisphere mixing is another proposed alternative**.

# Hemisphere mixing




**Original dataset**: bkgs and potentially a small signal fraction

**Mixed dataset**: new composed event that represent bkg-only

## Basic concept

Given a dataset, mix parts of different events to compose a new dataset which models non-peaking majority component.

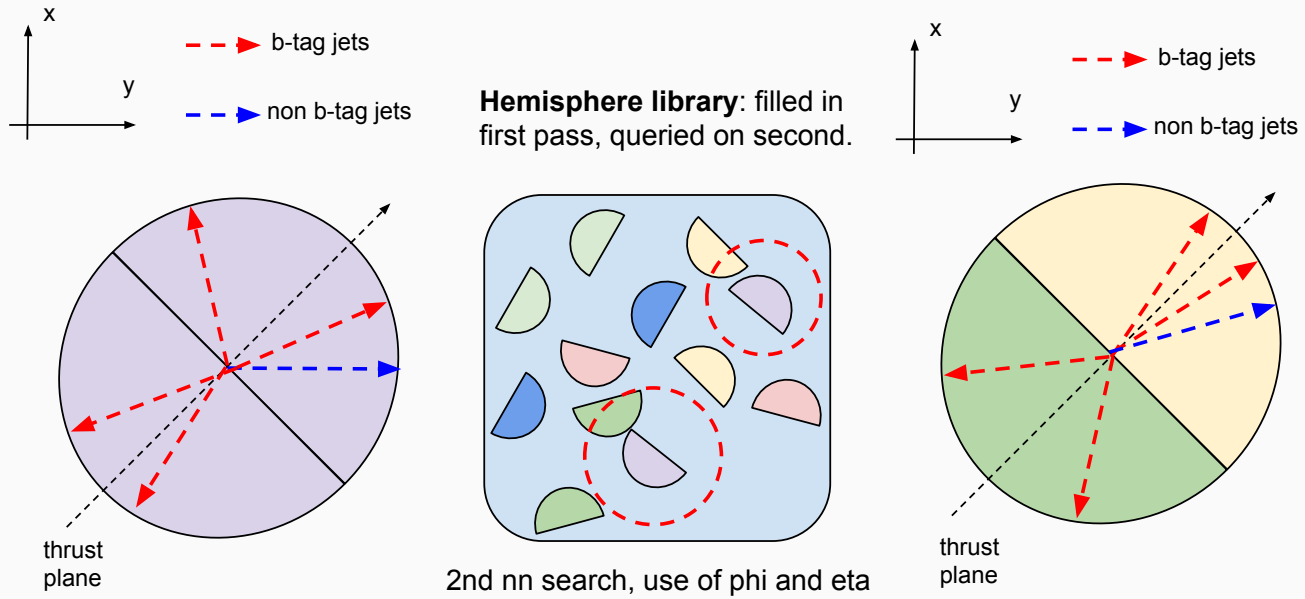
# Hemisphere mixing recipe

## create hemisphere library

divide event in 2 hemispheres, $\perp$ to transverse thrust plane, which is the one that maximizes $\sum p_T \cos\theta$

## modelling events

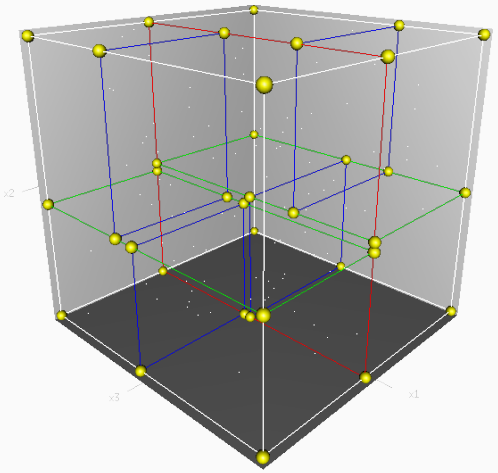
also divide event in hemispheres, but substitute both with best matching from the library $\rightarrow$ Frankenstein event



# $2^{nd}$ nearest neighbor search

**best matching hemisphere?**

search in library for the $2^{nd}$ nearest neighbor ($1^{st}$ is the same hemisphere) according to a certain distance metric.



k-d tree for $0(\log n)$ average search performance

**Require same:**

- num. of jets or $\geq 4$
- num. of b-tag jets or $\geq 4$

**n-dimensional $L^2$ distance**, combining hemisphere vars:

- $\sum^{\text{all hem. jets}} p_T \cos\theta$
- $\sum^{\text{all hem. jets}} p_T \sin\theta$
- $\sum^{\text{all hem. jets}} p_z$
- hemisphere inv. mass

Note: each variable is divided by $\sqrt{\text{Var}(X)}$.

# Use case - CMS non-resonant hh $\rightarrow b\bar{b}b\bar{b}$ with 2015 data

This technique was developed for modelling QCD multijet background within CMS hh $\rightarrow b\bar{b}b\bar{b}$ analysis at 13 TeV with 2015 data which is now public at CDS as CMS-PAS-HIG-16-016.

## Aim

Setting upper limit on $\sigma_{hh}^{\text{SM-like}} \times \text{BR}(hh \rightarrow b\bar{b}b\bar{b})$

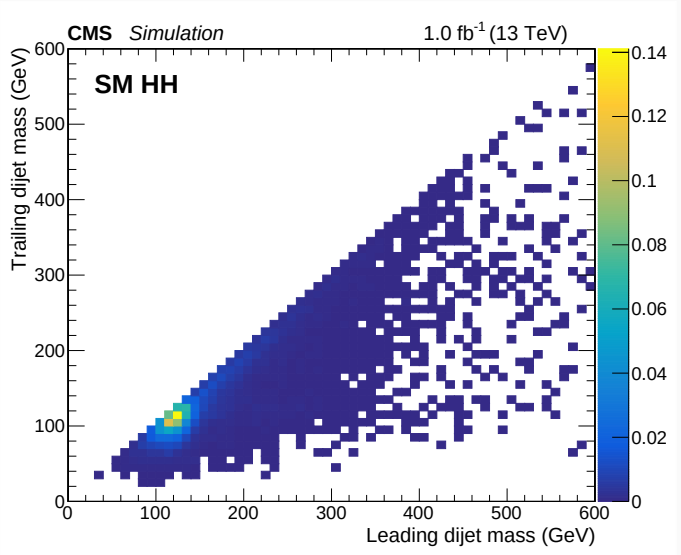
## Trigger

Kin. jet req. and online b-tagging.

## Jet selection

jet $p_T \geq 30$ GeV and $|\eta| \leq 2.5$
min 4 b-tagged jets (CSVv2M)

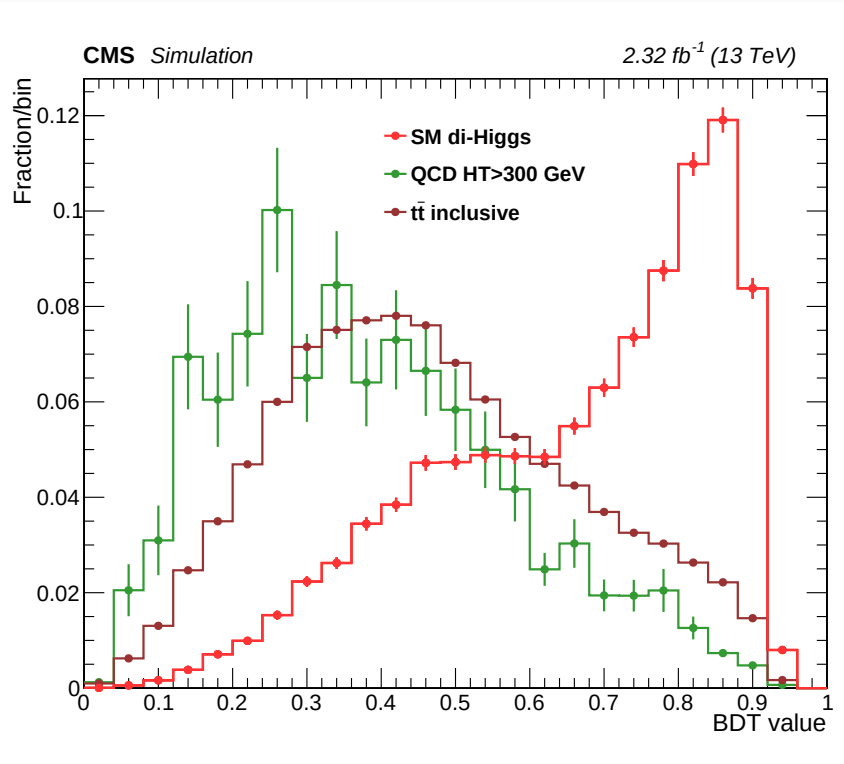
2D $m_{jj}^{\text{leading}}$ : $m_{jj}^{\text{trailing}}$ binned likelihood fit used for inference



# Mixing - multivariate classifier - jet pairing interplay

## Dijet pairing (done again after mixing)

From all selected jets, two pairs have to be chosen as dijet candidates. Take 4 jets with the highest b-tagging discriminator and find the combination that minimizes $\Delta m$.



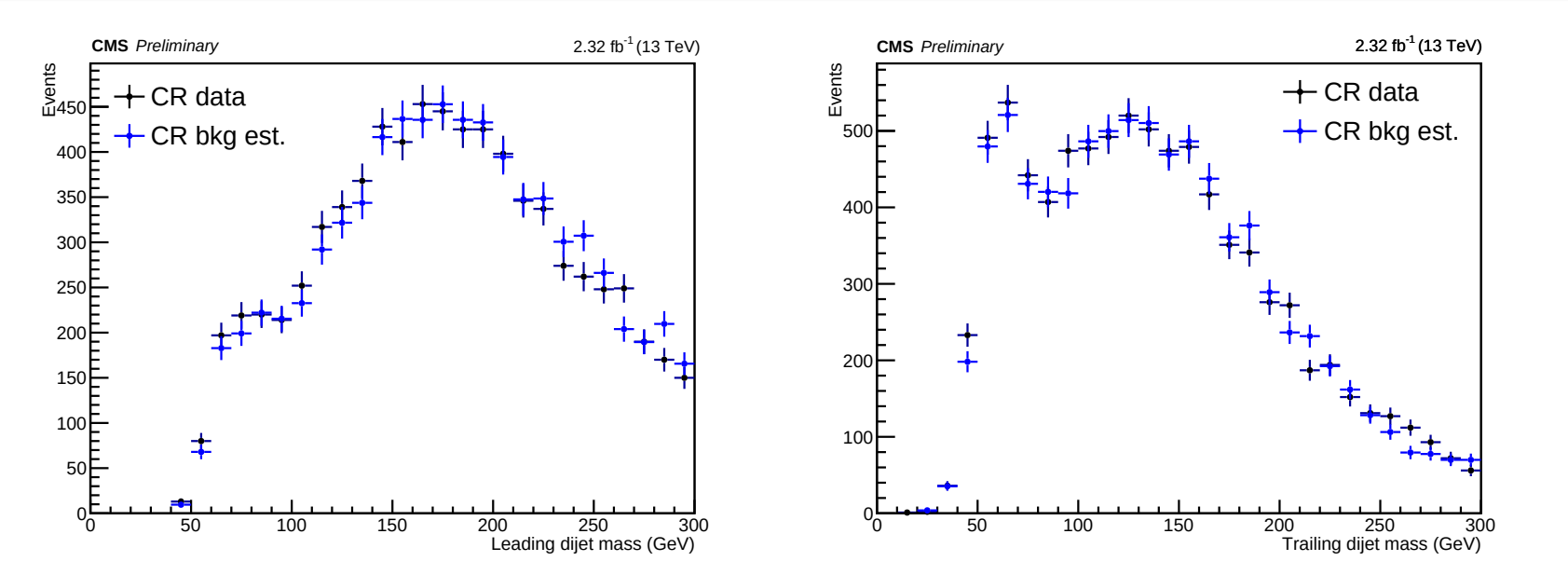

## Multivariate classifier cut

Analysis included a cut on a supervised ML classifier trained with MC samples.

Makes bkg modelling more challenging, mixed data has to behave like the bkg for all features used in the MVA.

# Hemisphere mixing checks

This technique was initially developed and tested using an unweighted toy dataset (only 0.5 $fb^{-1}$) made with events from CMS simulation samples of HT-binned QCD, $t\bar{t}$ and $hh \rightarrow b\bar{b}b\bar{b}$.

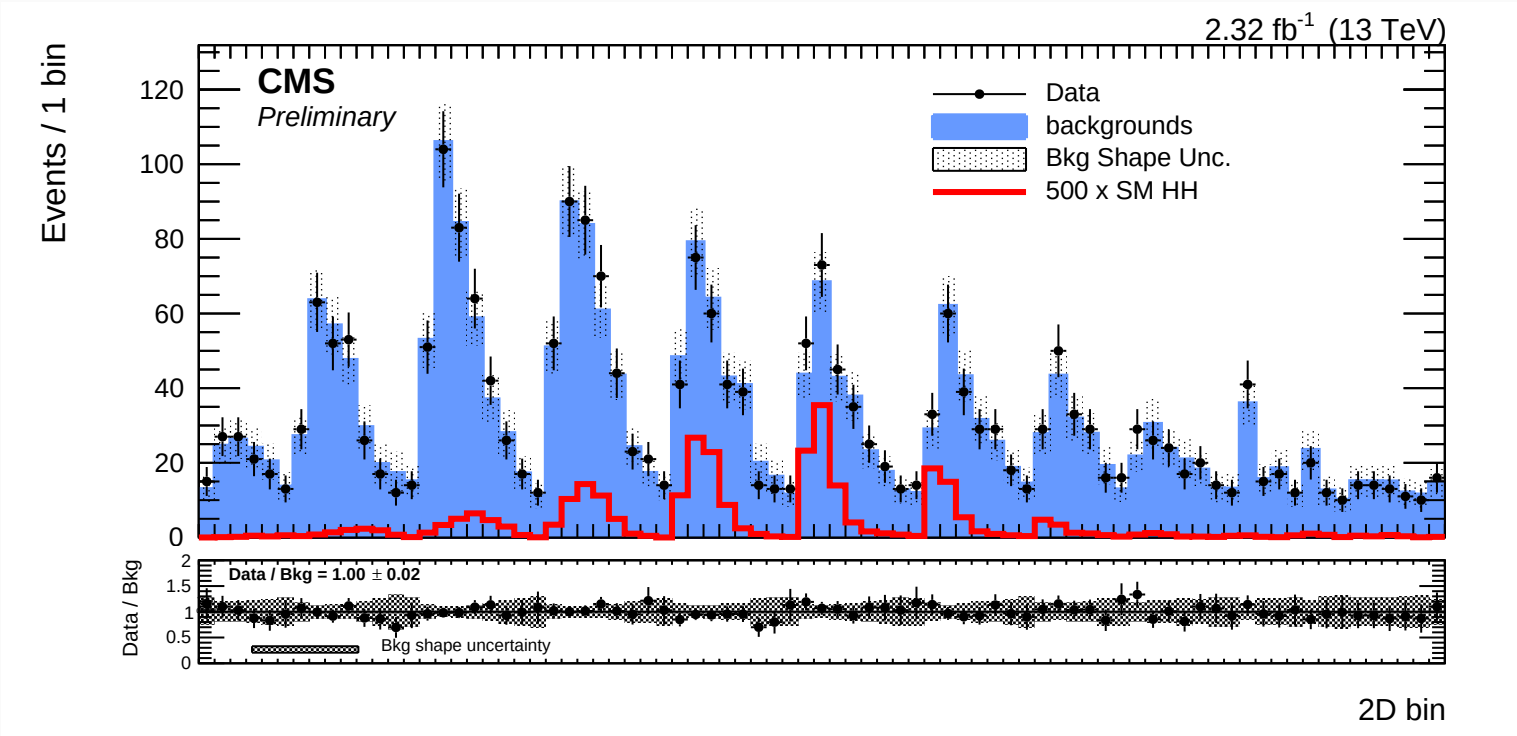



In addition, a data-based closure check was carried out using a BDT control region, obtaining statistically compatible results.

# Background modelling uncertainty assesment

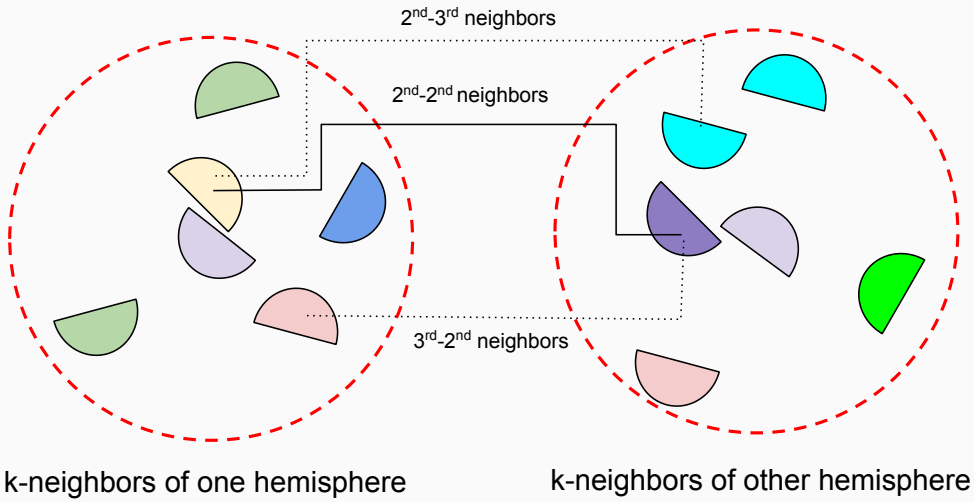
Background *pdf* has a statistical uncertainty of the same order than the data (normalization is free), so a per bin nuisance has to be added to the likehood (MC stats like).

In addition, a worst-case scenario percentage bias was estimated from BDT CR and added in quadrature with variance.



# Use of several nearest hemispheres combinations

One idea to reduce the statistical uncertainties on the background estimation is to take not only the $2^{nd}$ nearest neighbors but also the $3^{rd}$, $4^{th}$, ..., $n^{th}$ for each hemisphere.




A total of $(n-1)^2$ combinations of those can be considered so creating several mixed events per original event and therefore reducing statistical uncertainties if assumed independent.

# New DELPHES open dataset for benchmarking

However, as more data is used (e.g. 2016 adquired CMS data), background stat. uncertainties from this method will be reduced so possible method biases will dominate.

A larger toy simulated dataset is required to further study this technique and compare with possible alternatives. So datasets (MADGRAPH5_AMC@NLO+PYTHIA8+DELPHES) were produced within AMVA4NewPhysics ITN for this and other purposes:

| sample | # events | $\sigma$ (pb) | equiv. lumi. ($fb^{-1}$) |
|---|---|---|---|
| $pp \to jjjj$ | 10M | 1.45E+07 (LO) | 6.91E-04 |
| $pp \to b\bar{b}jj$ | 10M | 4.25E+05 (LO) | 2.35E-02 |
| $pp \to b\bar{b}b\bar{b}$ | 10M | 1.75E+03 (LO) | 5.72 |
| $pp \to t\bar{t} \to b\bar{b}jjjj$ | 10M | 2.05E+02 (LO) | 48.9 |
| $pp \to hh \to b\bar{b}b\bar{b}$ | 10M | 1.13E-02 (NLO+NLL) | $\infty$ |

**Available at root://eospublic.cern.ch//experiment/amva4np**

# Creating toy datasets and planned checks

## Unweighted toy datasets

By sampling without replacement up to the expected number of events for a certain lumi (e.g. $5\text{fb}^{-1}$).

For the time being, bkg is purely composed of $pp \to b\bar{b}b\bar{b}$ (stats too low for the rest of QCD, $t\bar{t}$ to be added).

Several datasets with different signal fraction / times $\sigma_{hh}^{SM}$ (signal free case is the first thing to be studied).

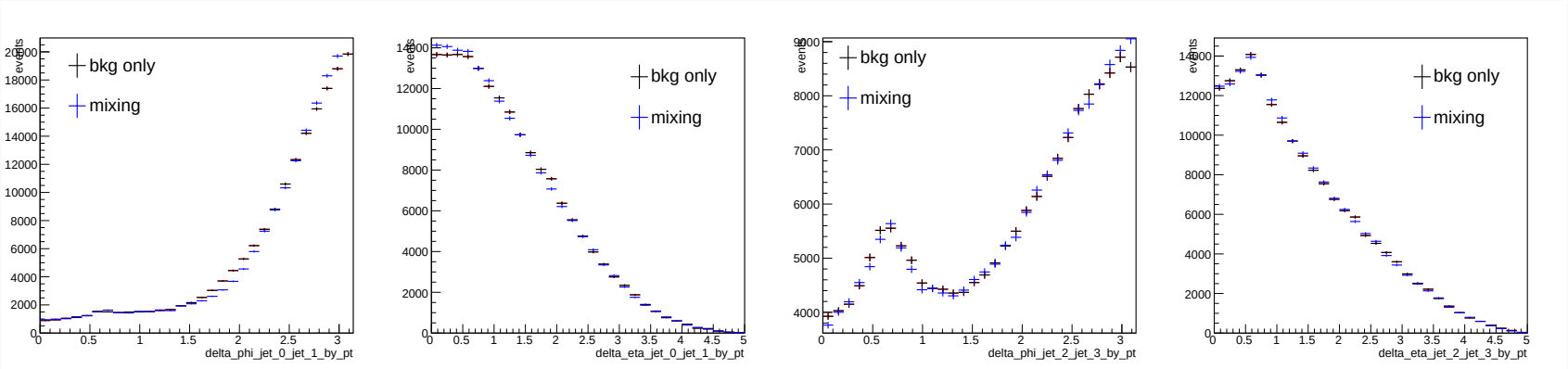
## Open questions:

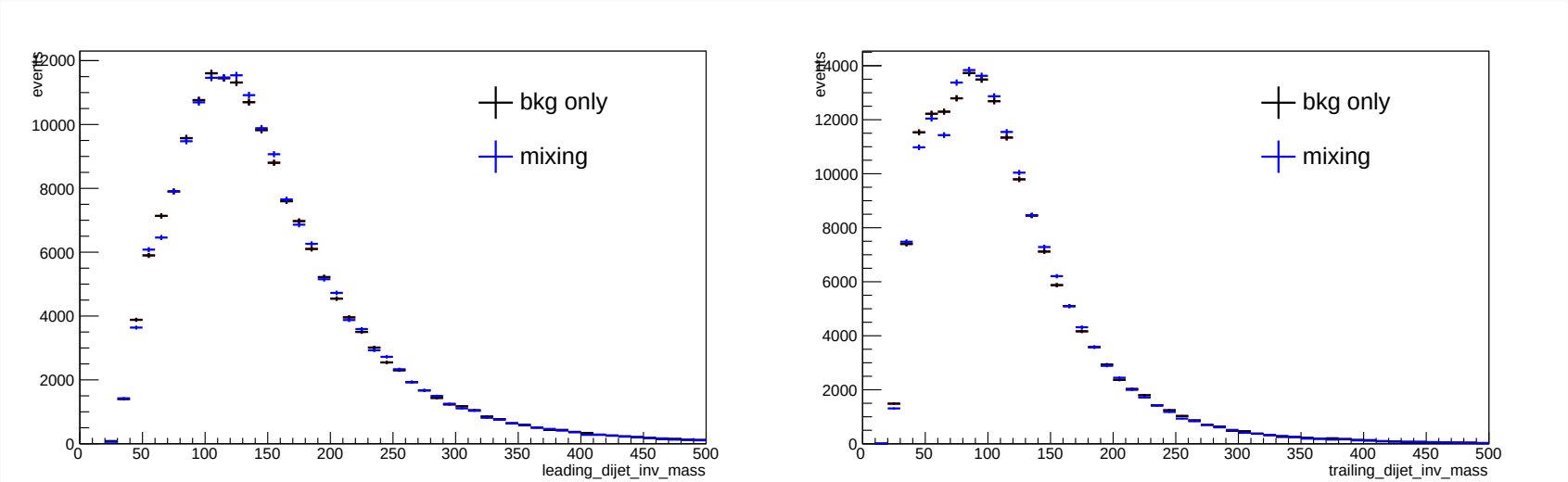
The technique was proven to provide stat. compatible modelling for low stats, will biases appear for larger statistics?

What is a good figure of merit to optimize technique hyperparameters (e.g. distance metric)?

Using several hemisphere combinations can lead to reduced bkg. statistical uncertainties without paying a large bias?

# Fresh from the oven - Mixing with $pp \to b\bar{b}b\bar{b}$ QCD only ($5fb^{-1}$)




Some statistically significant discrepancies can be see in the first preliminary studies carried out, currently in the process of understanding their source and impact.

# Conclusions

A non-parametric technique to estimate the *pdf* of QCD multijet background component in analysis with small relative signal fraction, by mixing hemispheres from the original data events, has been presented.

Its applicability to the non-resonant hh $\rightarrow$ $b\bar{b}b\bar{b}$ has been described, for which has been proven to given good background modelling for the data statistics available in 2015.

Further checks using a larger toy dataset are on the way, which will provide insighful information about the goodness of this technique for higher statistics datasets ($\geq 5 fb^{-1}$) where biases might become important.