Hemisphere Mixing

A Fully Data-Driven Model
Of QCD Multijet Backgrounds
For LHC Searches

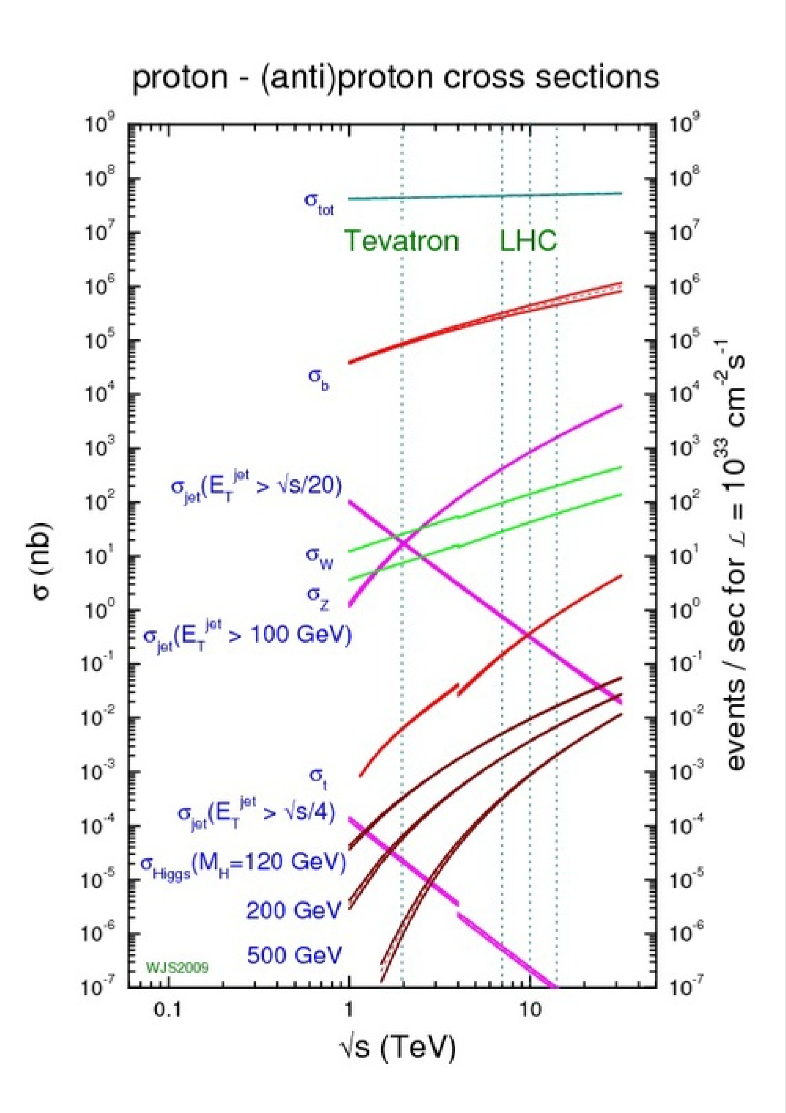
T.Dorigo, INFN – Padova

This Report is part of a project that has received funding from the **European Union's Horizon 2020 research and innovation programme under grant agreement N°675440**

Istituto Nazionale di Fisica Nucleare

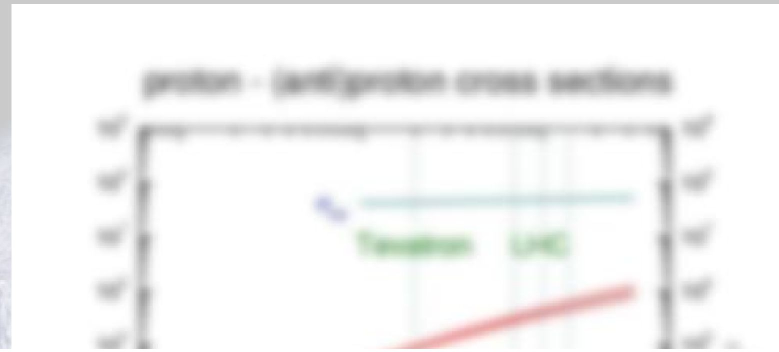
The Problem

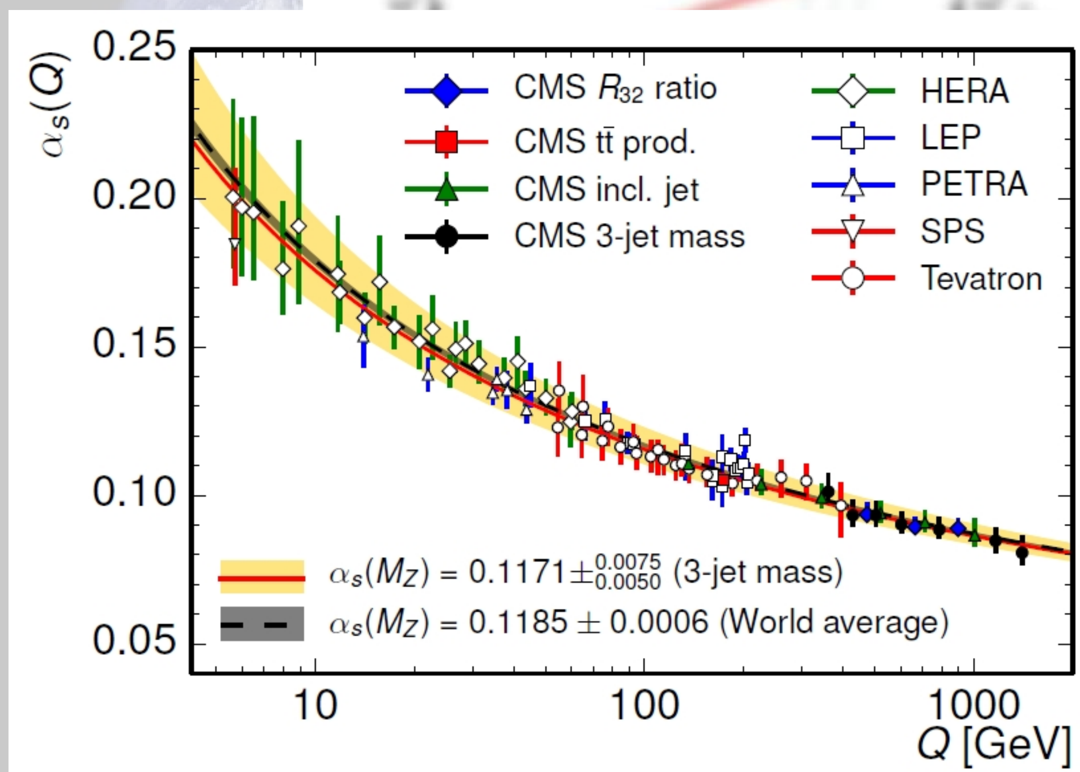
- At the LHC, we search for rare phenomena amidst a huge production of quark-gluon interactions



The Problem

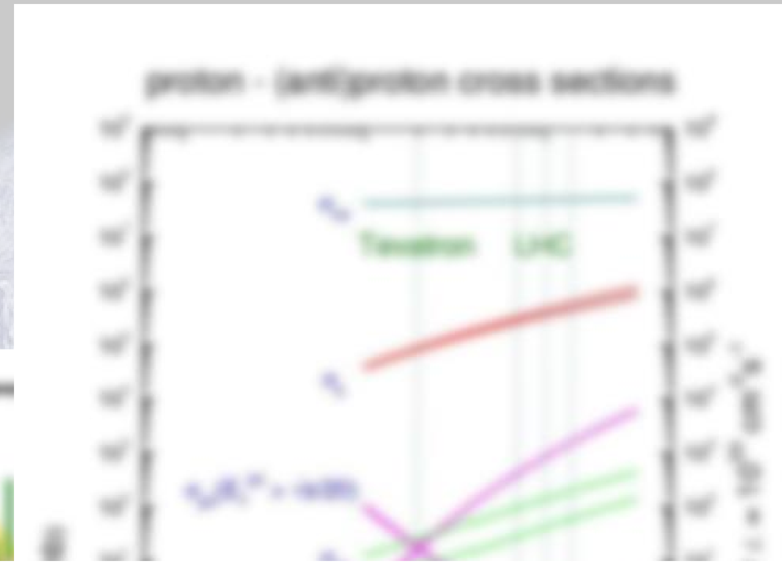
- At the LHC, we search for rare phenomena amidst a huge production of quark-gluon interactions
- We know how QCD works, but we cannot calculate it at low Q^2

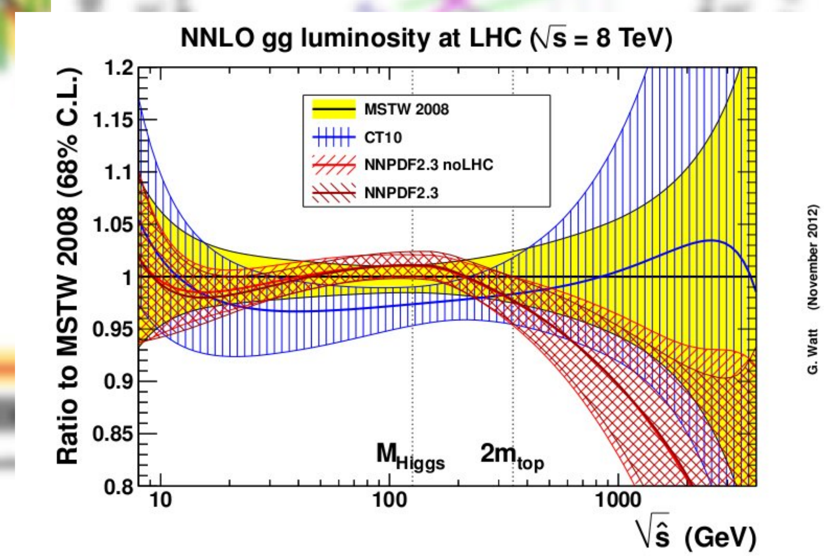




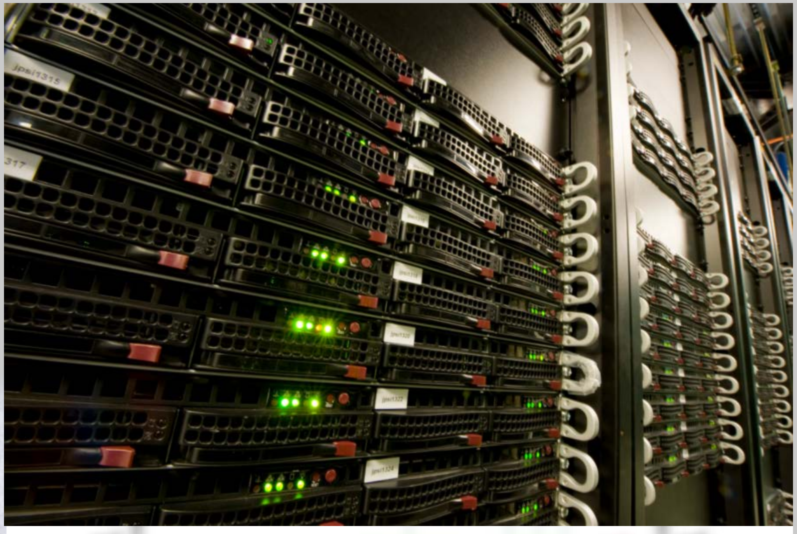
The Problem

- At the LHC, we search for rare phenomena amidst a huge production of quark-gluon interactions
- We know how QCD works, but we cannot calculate it at low Q^2
- We can still model the physics, but model uncertainties (PDF, UE tunes, hadronization) affect our predictions
  - The issue is especially relevant when we deal with multijet final states

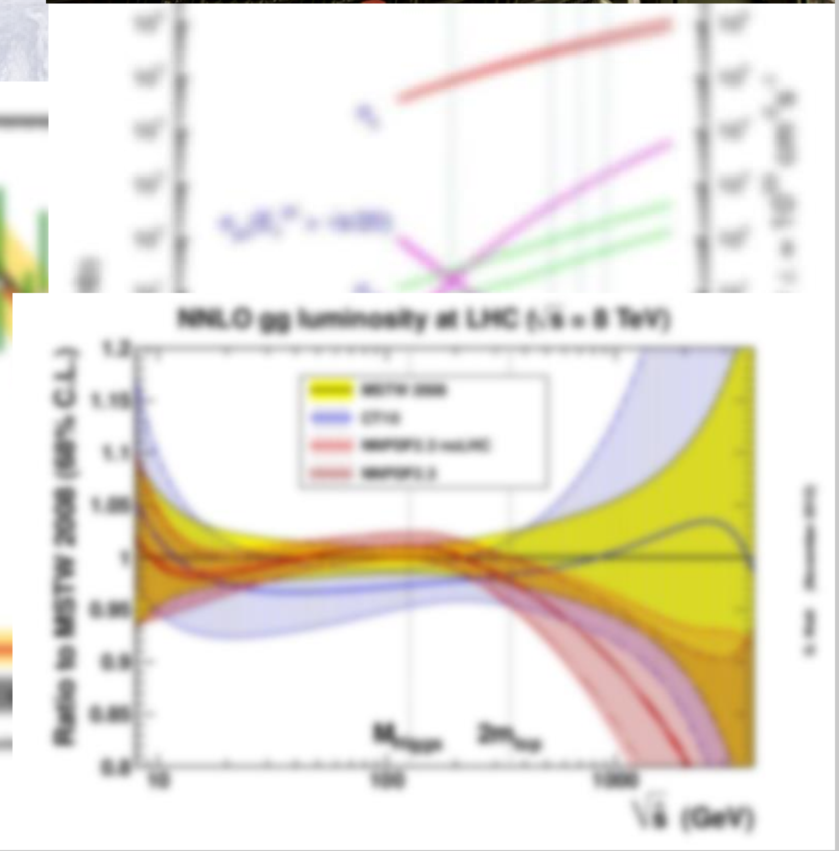




The Problem



- At the LHC, we search for rare phenomena amidst a huge production of quark-gluon interactions
- We know how QCD works, but we cannot calculate it at low Q^2
- We can still model the physics, but model uncertainties (PDF, UE tunes, hadronization) affect our predictions
  - The issue is especially relevant when we deal with multijet final states
- In addition, CPU is a limiting factor
  - Centrally provided QCD samples give effective luminosity much smaller than experimental data
  - **How can we reduce our systematics in our searches for new phenomena?**

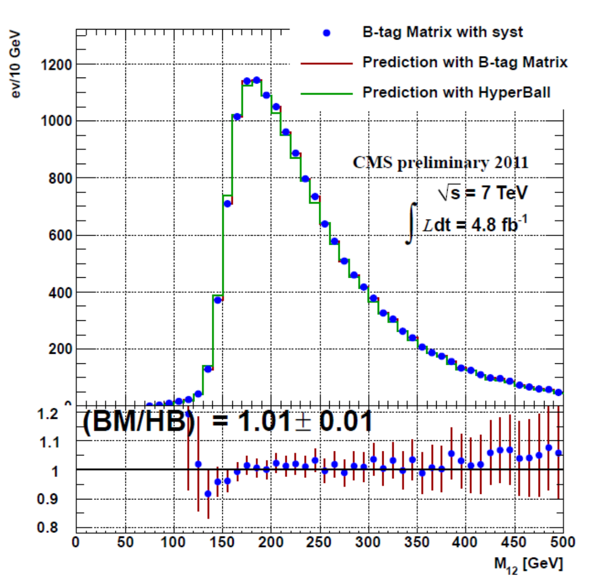


Data-Driven Modeling

In searches for NP or precision measurements at the LHC we usually either

1) **rely on common data-driven techniques** to predict relevant spectra:

- Sideband-based methods
- ABCD extrapolations → b-tag matrices → kNN
- Access to large-enough "control samples" often limits the accuracy of these predictions



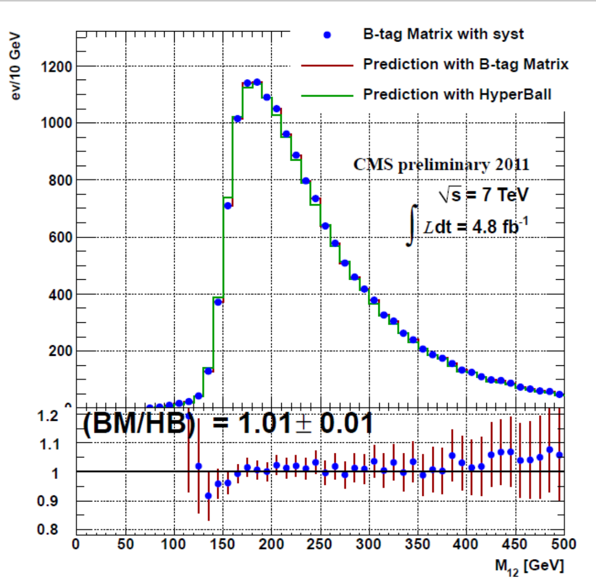
Top: B-tag and kNN-based dijet mass models in search for bbH→bbbb, CMS-HIG-12-027

Data-Driven Modeling

In searches for NP or precision measurements at the LHC we usually either

1) **rely on common data-driven techniques** to predict relevant spectra:

- Sideband-based methods
- ABCD extrapolations → b-tag matrices → kNN
- Access to large-enough "control samples" often limits the accuracy of these predictions

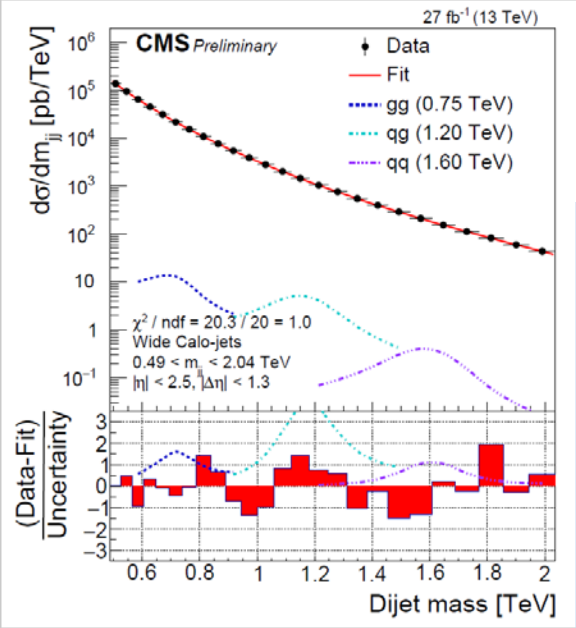


Top: B-tag and kNN-based dijet mass models in search for bbH→bbbb, CMS-HIG-12-027

2) or **throw our hands up**:

- Find a "reasonable" functional form, fit it to data, look for local deviations as possible hints of new particles

Statistical precision of Run 2 datasets challenges methods based on "QCD inspired" parametric forms



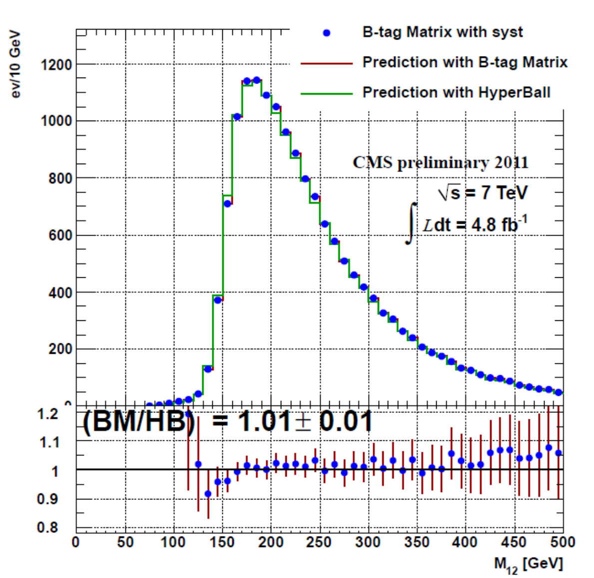
Left: five-parameter fit to dijet mass shape in CMS-EXO-16-056; Bottom: residuals from fit

Data-Driven Modeling

In searches for NP or precision measurements at the LHC we usually either

1) **rely on common data-driven techniques** to predict relevant spectra:

- Sideband-based methods
- ABCD extrapolations → b-tag matrices → kNN
- Access to large-enough "control samples" often limits the accuracy of these predictions

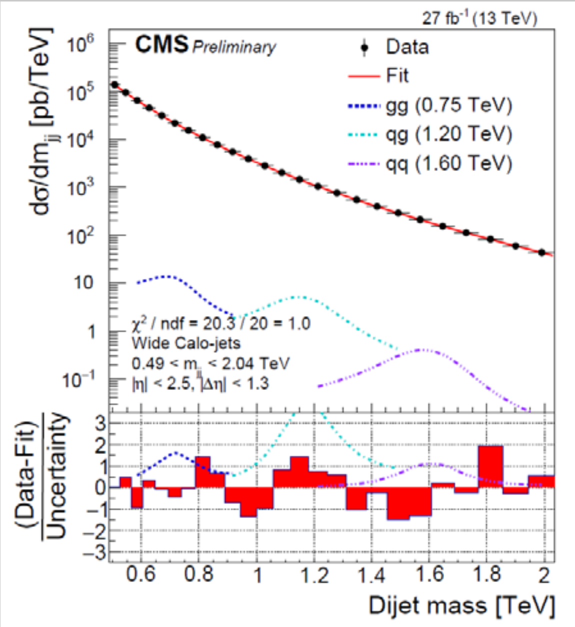


Top: B-tag and kNN-based dijet mass models in search for bbH→bbbb, CMS-HIG-12-027

2) or **throw our hands up**:

- Find a "reasonable" functional form, fit it to data, look for local deviations as possible hints of new particles

Statistical precision of Run 2 datasets challenges methods based on "QCD inspired" parametric forms



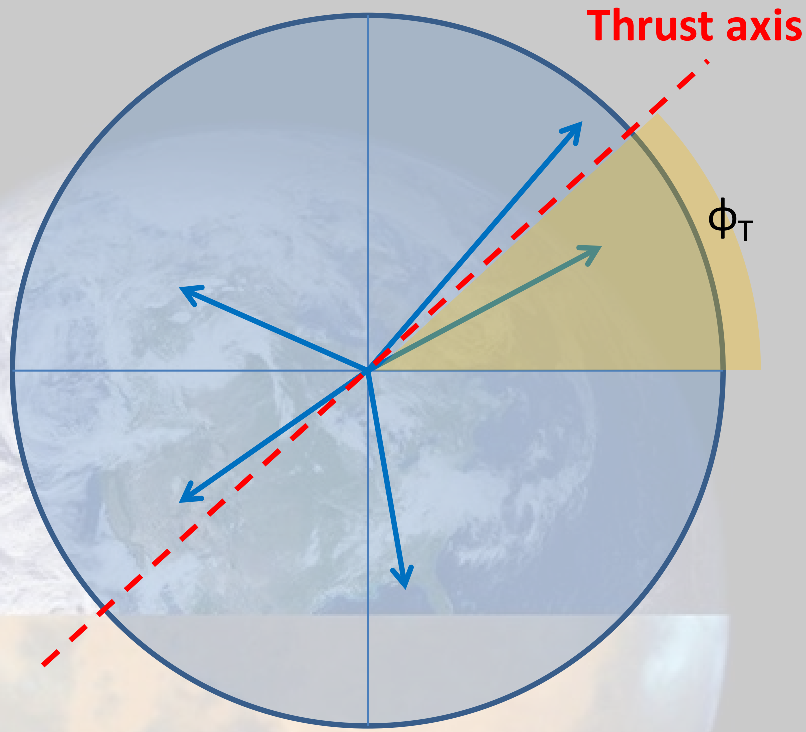
Left: five-parameter fit to dijet mass shape in CMS-EXO-16-056; Bottom: residuals from fit

The modeling problem is made harder by the booming of statistical learning methods: one does not content oneself to model just a 1D PDF, but wants a model of the **full multi-D space**

QCD events laid bare

- High-energy QCD events come from a complicated matrix element, but in essence they originate from a 2→2 process when the final state is enriched in complexity by ISR, FSR, MPS, PU…

- In the days of e^+e^- machines one studied hadronic events by defining a thrust variable to interpret the event
  - Thrust axis = axis that maximizes **T = $\Sigma p_T * |\cos\phi|$** with ϕ = angle particle-axis (or jet-axis)



The axis is supposed to coincide with the direction of the two final-state partons – at least at LO in e^+e^- collisions

QCD events laid bare

In hadron collisions one has a boost along z which **breaks the axis into two semiaxes**, back-to-back in azimuth but not in R-z

- Never mind – we can use the T axis *in the transverse plane*
- What do we do with it ?

→ Define hemispheres (or hemi-cylinders):

$\cos\phi > 0 \, / \, \cos\phi < 0$

QCD events laid bare

In hadron collisions one has a boost along z which **breaks the axis into two semiaxes**, back-to-back in azimuth but not in R-z

- Never mind – we can use the T axis *in the transverse plane*
- What do we do with it ?

→ Define hemispheres (or hemi-cylinders):

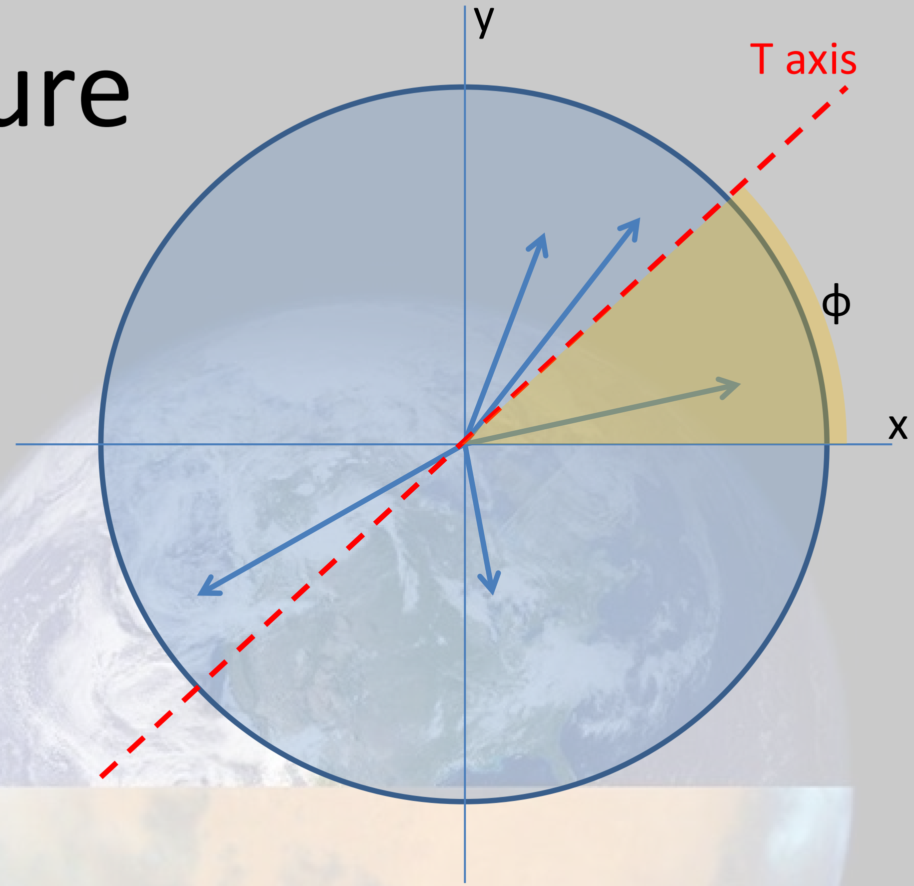
cosϕ > 0 / cosϕ < 0

Working assumption: In large T events, all the physics arising from ISR, FSR, MPS, PU is "second order" in defining the topology of the produced jets; and each of the two leading order partons does not influence the physics on the other hemisphere

If that were true, we would have a simple recipe for generating large samples of QCD events from smaller samples:

Mix and match hemispheres that correspond to outgoing partons of "similar" kinematics

The mixing procedure



1) **For each event** in the original sample:
   - Find transverse thrust axis

i.e., determine angle ϕ such that

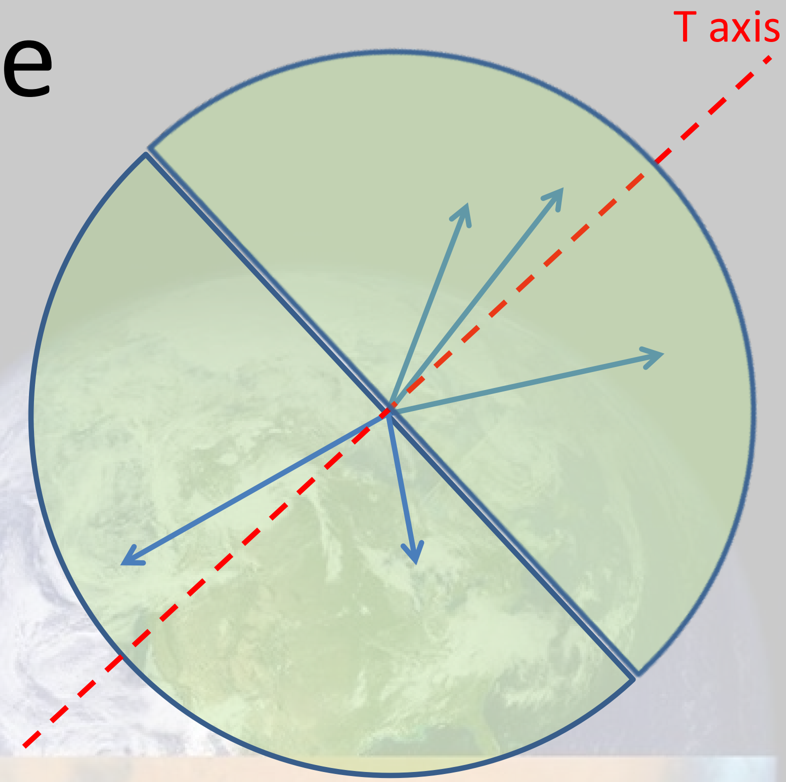
$T = \sum p_T{}^{jet} cos(\varphi_\mathrm{T} - \varphi_\mathrm{jet})$

is maximized

The mixing procedure

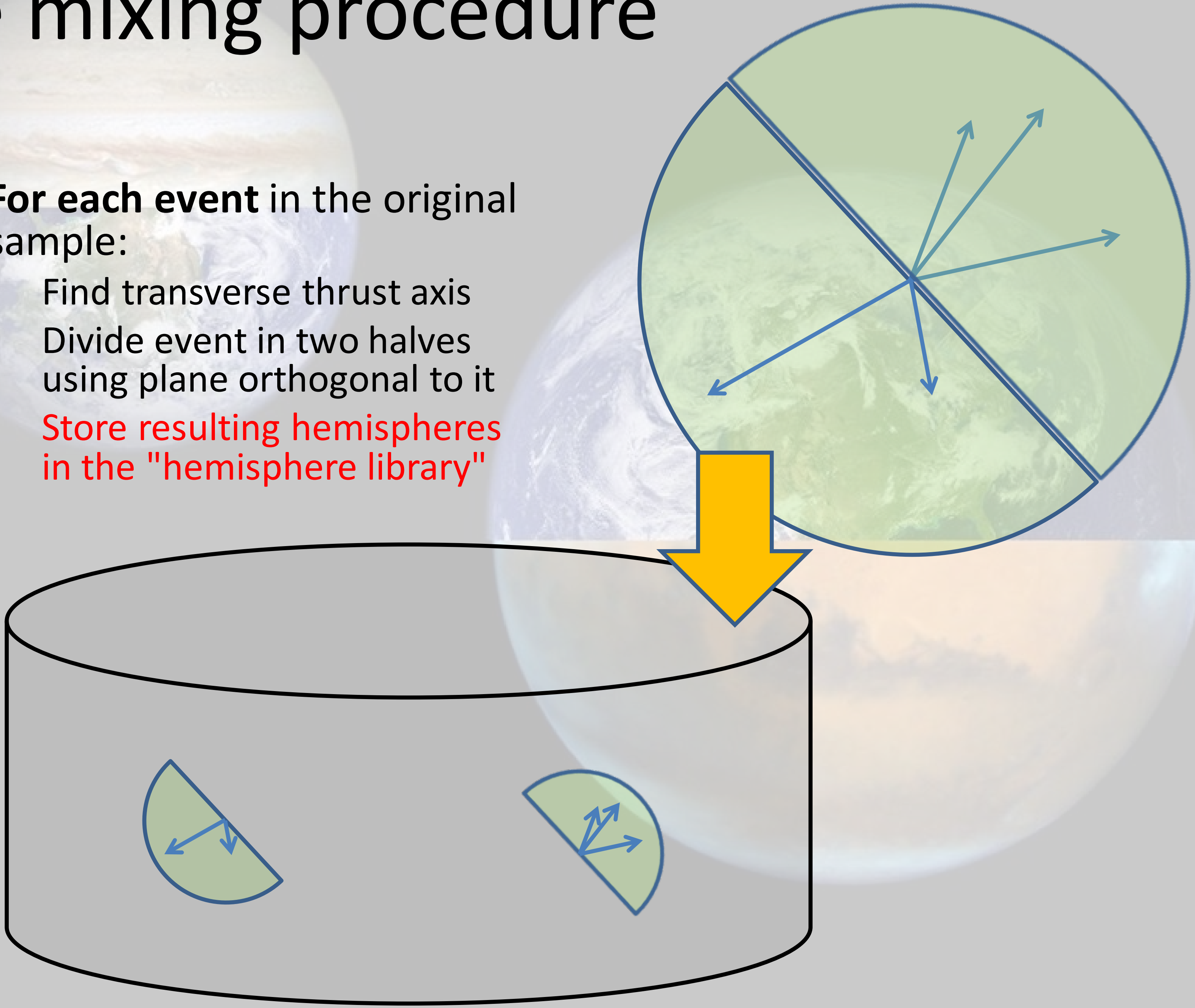
1) **For each event** in the original sample:
   - Find transverse thrust axis
   - Divide event in two halves using plane orthogonal to it

   This defines *two* jet collections for each event (hemispheres)



The mixing procedure

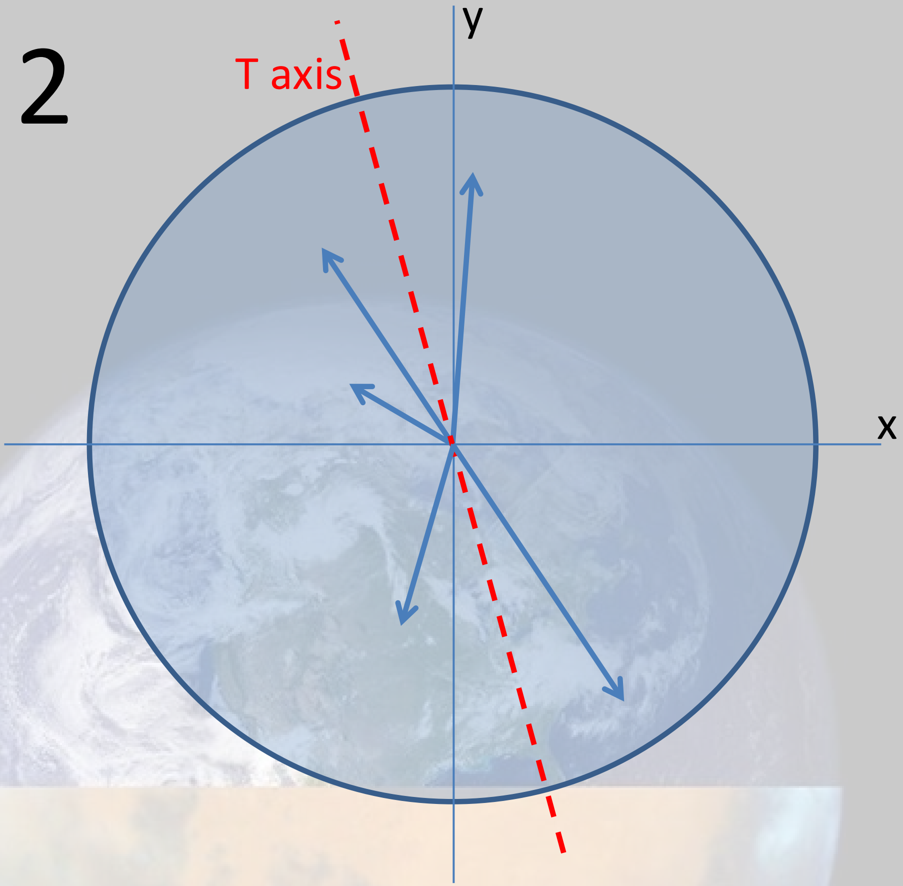
1) **For each event** in the original sample:
   - Find transverse thrust axis
   - Divide event in two halves using plane orthogonal to it
   - Store resulting hemispheres in the "hemisphere library"



Mixing procedure - 2

2) Take again original sample: for each event

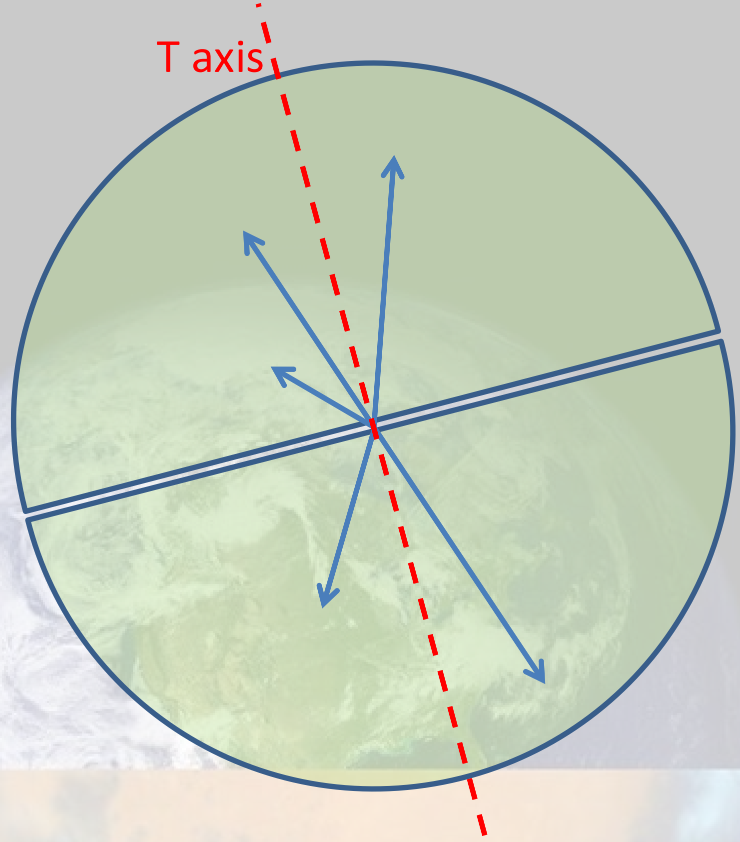
- Find transverse thrust axis,



Mixing procedure - 2

2) Take again original sample: for each event

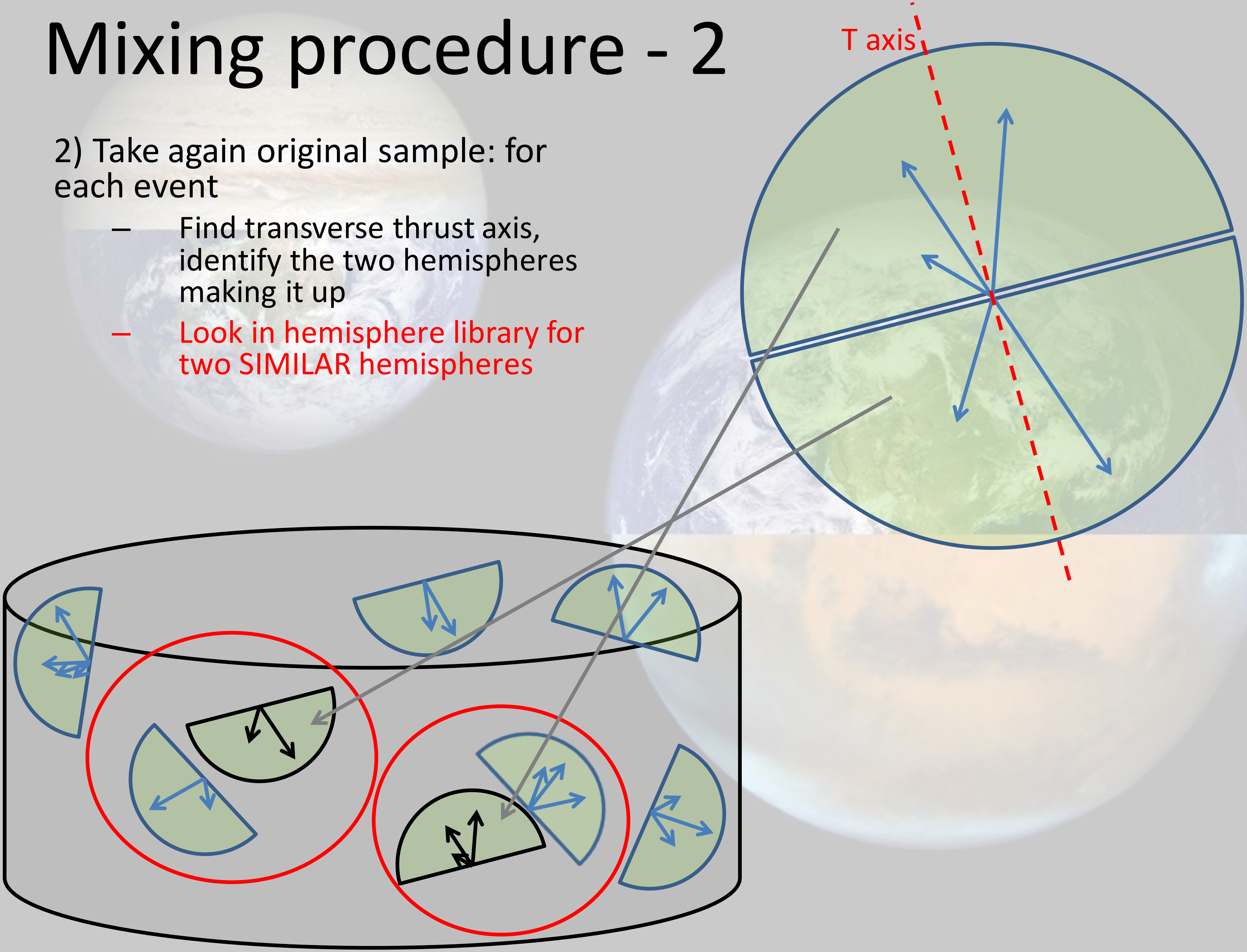
- Find transverse thrust axis, identify the two hemispheres making it up



Mixing procedure - 2

2) Take again original sample: for each event

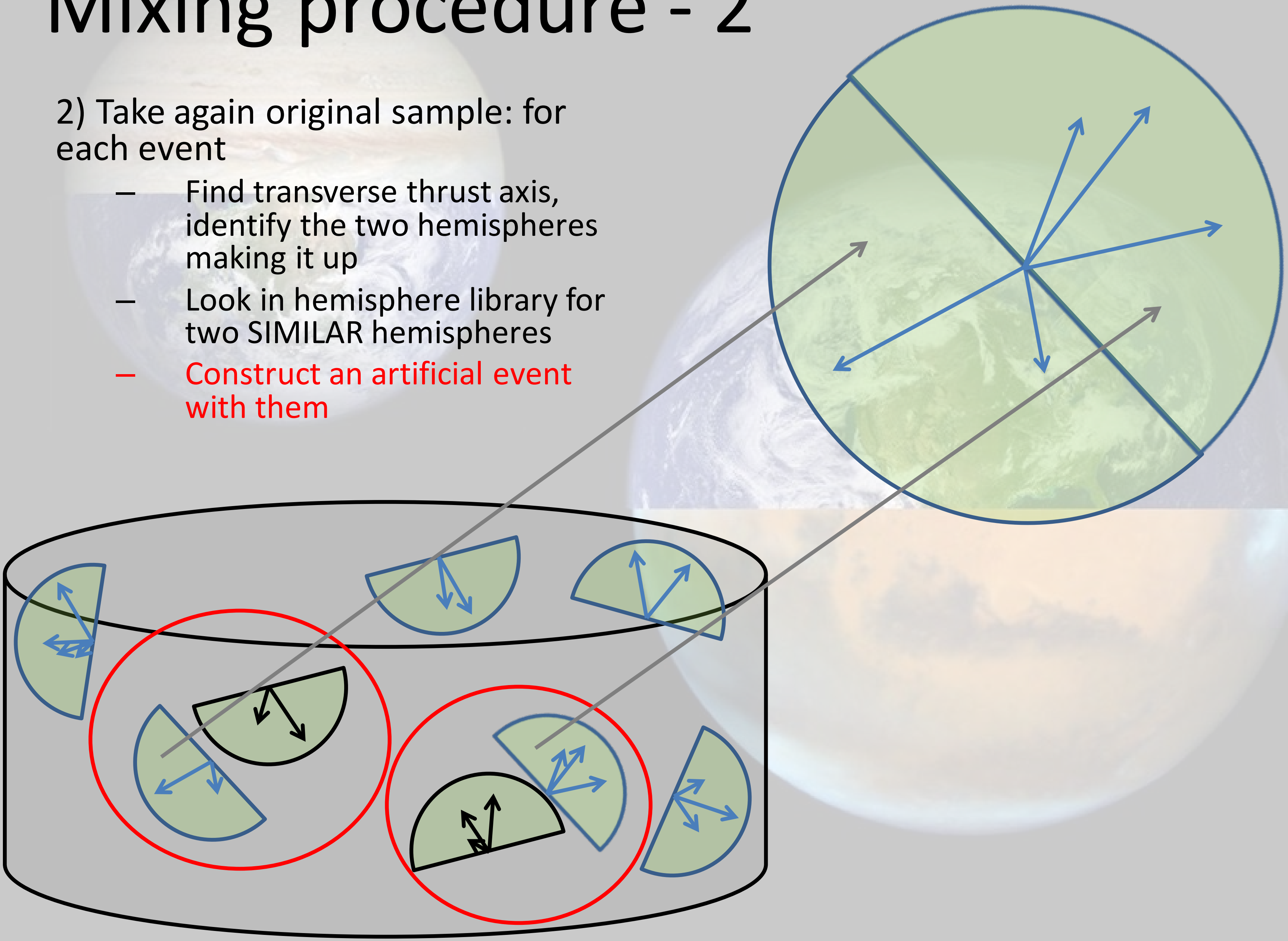
- Find transverse thrust axis, identify the two hemispheres making it up
- Look in hemisphere library for two SIMILAR hemispheres



Mixing procedure - 2

2) Take again original sample: for each event

- Find transverse thrust axis, identify the two hemispheres making it up
- Look in hemisphere library for two SIMILAR hemispheres
- Construct an artificial event with them

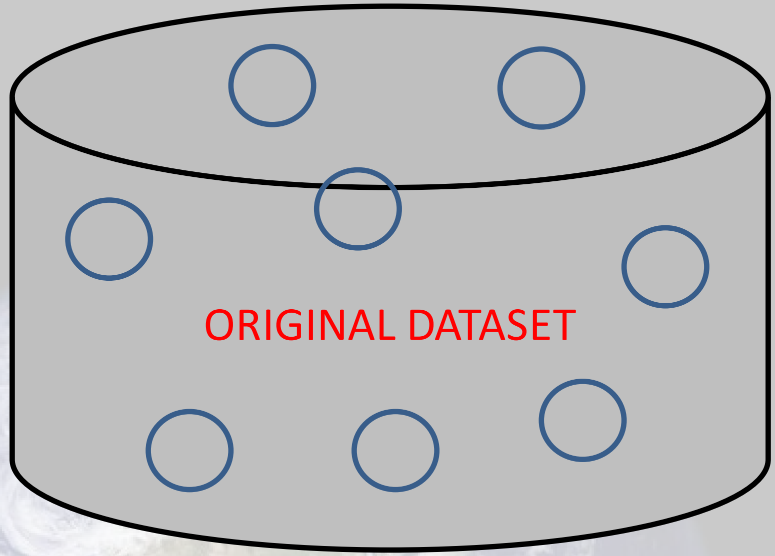


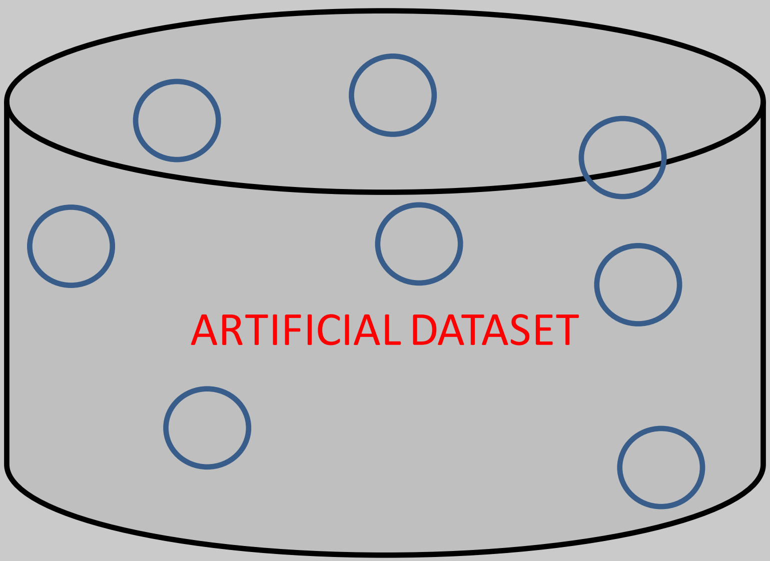
Mixing procedure - 2

2) Take again original sample: for each event

- Find transverse thrust axis, identify the two hemispheres making it up
- Look in hemisphere library for two SIMILAR hemispheres
- Construct an artificial event with them

The procedure creates an artificial dataset which can be used for modeling purposes





Hemisphere similarity criteria :

- Number of jets (req. equal)
- Number of b-tags (req. equal)
- Thrust
- Thrust minor
- Hemisphere mass
- Sum of jets p_z components

The 4 continuous variables are used to define a **kNN distance** which yields the similarity measure:

$$D(1p)^2 = \frac{(T(h_1)-T(h_p))^2}{V_T} + \frac{(M(h_1)-M(h_p))^2}{V_M} + \frac{(|P_z(h_1)|-|P_z(h_p)|)^2}{V_{P_z}} + \frac{(T_a(h_1)-T_a(h_p))^2}{V_{T_a}}$$

Test setup: HH→bbbb search

- As a test of the procedure we take fast-simulated LHC pp→multijet events
  - Events are selected to contain >=4 p_T>30 GeV jets, $|\eta|<2.5$, b-tagged with medium requirements (ε=0.6, a=0.01), mimicking a 2016 CMS analysis
  - Leading b-tagged jets are paired by minimum ΔM_{jj} criterion to compute M_{12}, M_{34} combinations
  - A total of 40 kinematic variables are used for tests

Test setup: HH→bbbb search

- As a test of the procedure we take fast-simulated LHC pp→multijet events
  - Events are selected to contain >=4 p_T>30 GeV jets, |η|<2.5, b-tagged with medium requirements (ε=0.6, a=0.01), mimicking a 2016 CMS analysis
  - Leading b-tagged jets are paired by minimum ΔM_{jj} criterion to compute M_{12}, M_{34} combinations
  - A total of 40 kinematic variables are used for tests
- **Data** is constituted by QCD multijet production (80%) and top pair-production (20%)
  - To study the effect of a contamination from non-resonant HH pair production and decay to two b-quark pairs we may add that process to the sample mixture
  - SM predicts HH fraction to be <0.01% at this level of selection

Test setup: HH→bbbb search

- As a test of the procedure we take fast-simulated LHC pp→multijet events
  - Events are selected to contain >=4 p_T>30 GeV jets, |η|<2.5, b-tagged with medium requirements (ε=0.6, a=0.01), mimicking a 2016 CMS analysis
  - Leading b-tagged jets are paired by minimum ΔM_{jj} criterion to compute M_{12}, M_{34} combinations
  - A total of 40 kinematic variables are used for tests
- **Data** is constituted by QCD multijet production (80%) and top pair-production (20%)
  - To study the effect of a contamination from non-resonant HH pair production and decay to two b-quark pairs we may add that process to the sample mixture
  - SM predicts HH fraction to be <0.01% at this level of selection

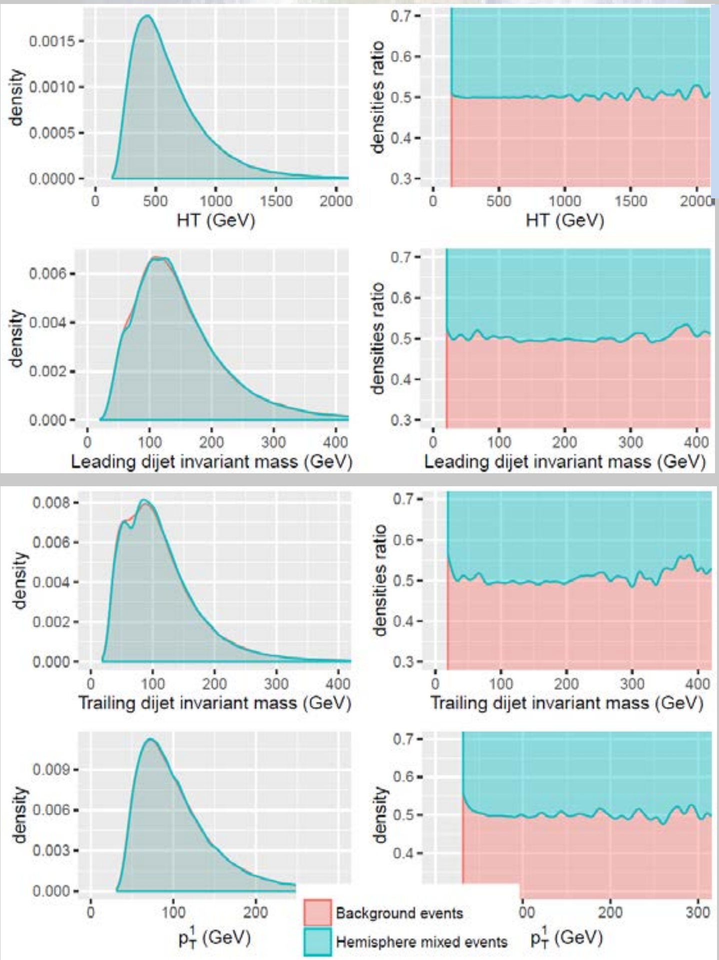
Then we do our magic:

1) The selected data constitutes the "original sample"
2) A hemisphere library is constructed with them
3) Event mixing is then applied, obtaining an artificial sample

The kinematics of original and artificial data can be compared

A look at 1D kinematic distributions

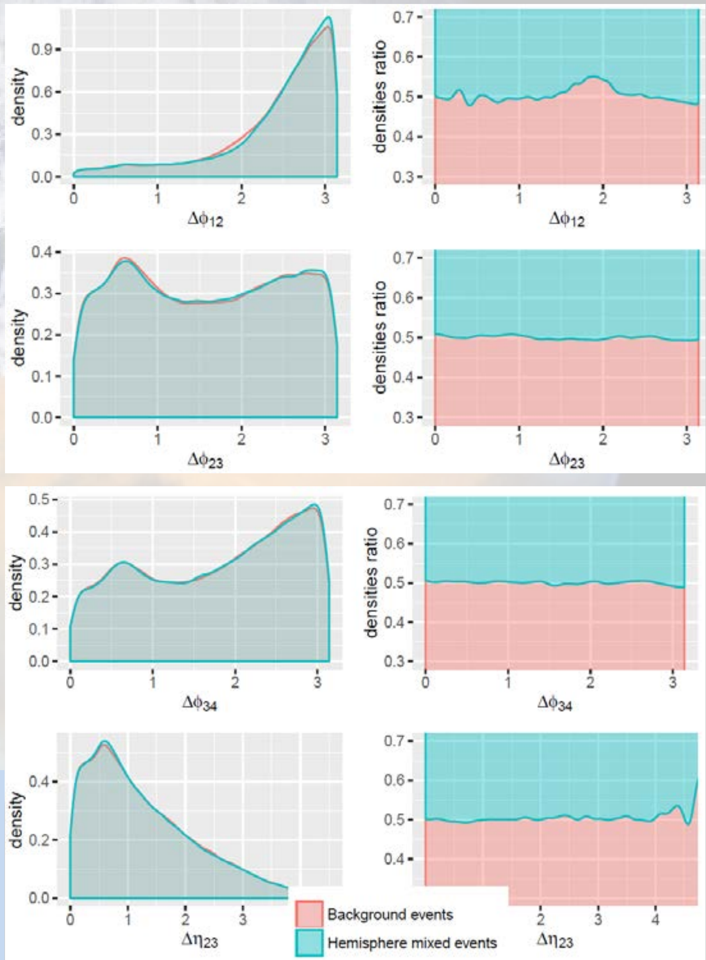
- The modeling of 1D marginals can be checked by comparing QCD+TT versus its artificial replica
- **No discrepancies are observed in any of the tested distributions,** e.g. see ones below



Left, top to bottom: H_T, M_{12}, M_{34}, leading jet p_T

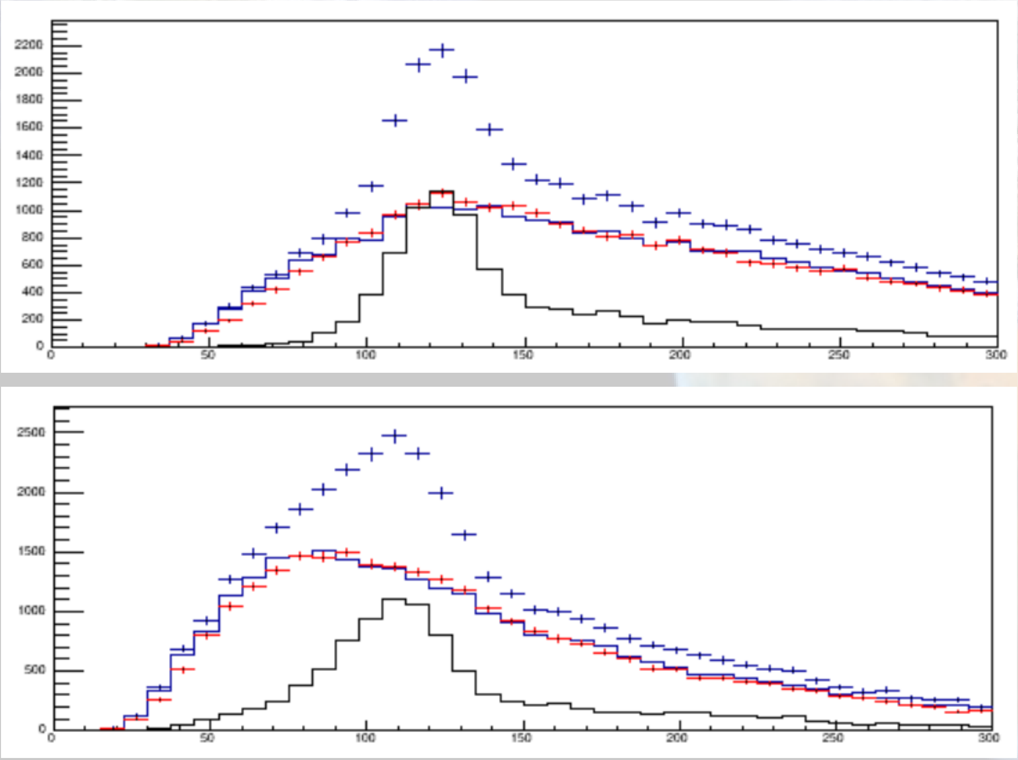
Distributions and ratio between original and artificial samples

Right, top to bottom: $\Delta\phi_{12}$, $\Delta\phi_{23}$, $\Delta\phi_{34}$, $\Delta\eta_{23}$



Signal injection tests

One may verify that the modeling ignores a **small** signal component by injecting it in the original sample before library creation, and comparing, e.g., dijet mass distributions (M_{12}, M_{34}) of original and artificial datasets



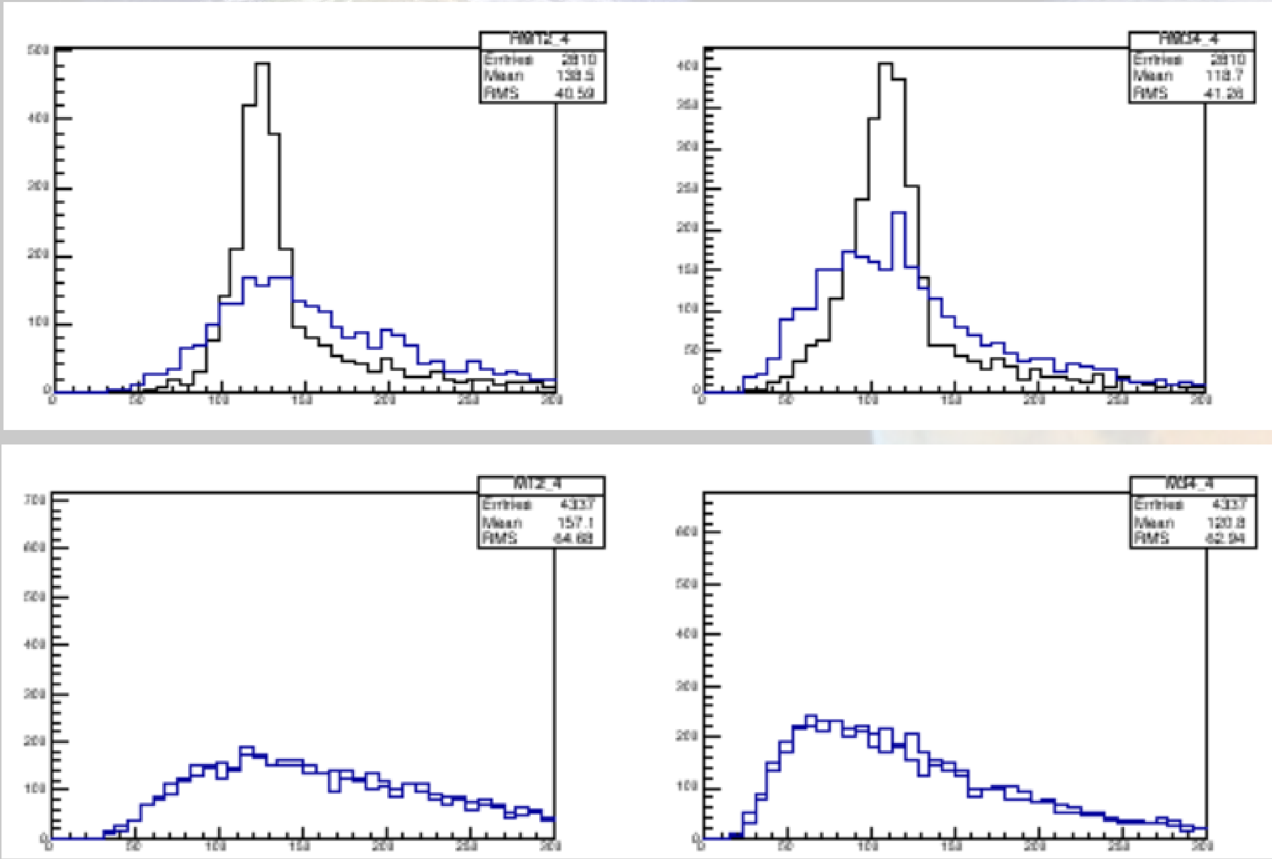
Top: M_{12} distribution for QCD+TT events with x10,000 HH contribution (blue points); artificial dataset (red points) rescaled to QCD+TT component alone (blue histogram); HH component (black histogram)

Bottom: same, for M_{34} distribution

[Fine print: above, to show the effect of a 0.5%-ish signal contamination we use a correspondingly populated hemisphere library. However a signal of that size would not be visible, so we apply the mixing to a sample with 100x larger signal contamination.]

Mapping of QCD and HH

In fact, one may check where signal and background events get mapped, by studying the dijet mass distributions of these events separately.



Distributions of M_{12} and M_{34} in signal events (top row) and background events (bottom row). Black: original data; blue: artificial (mixed) data

One sees that a small signal contamination acquires after mixing a background-like shape even in signal-distinctive distributions

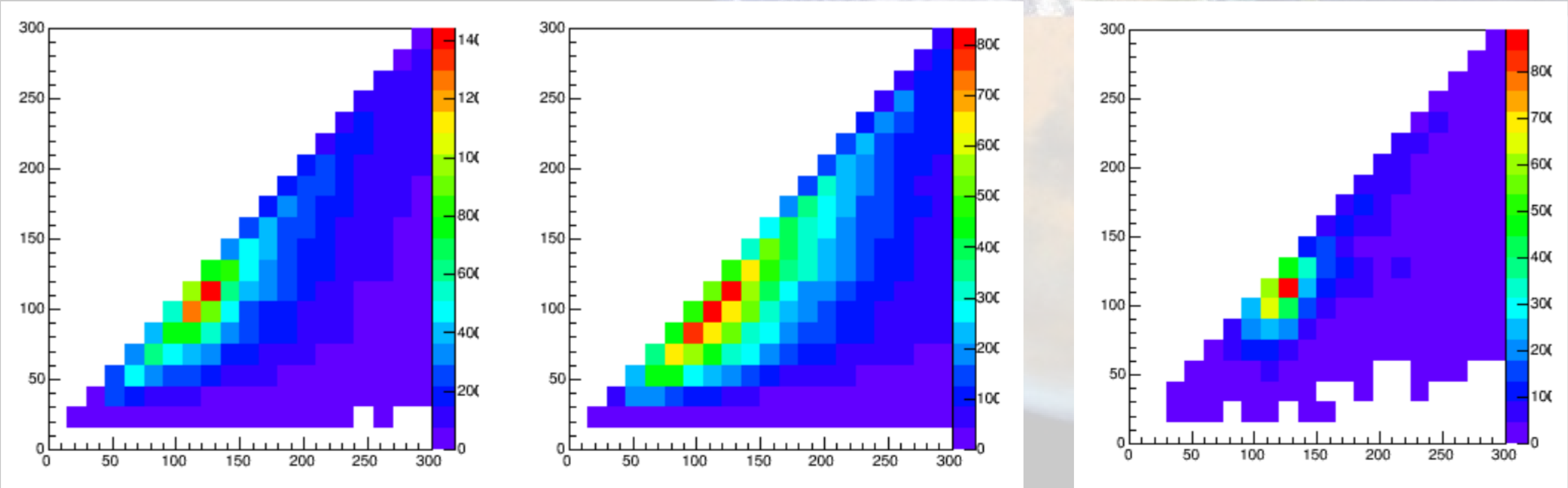
The majority component (QCD + TT) of the selection is instead mapped onto itself nicely, and remains insensitive of the signal contamination

Fits to the signal component

A more quantitative way to study the "dilution" of the minority component in the artificial dataset is to fit a discriminant variable in original data as the sum of signal+background, using the artificial data distribution as a model of the background

– E.g. we perform a 2-D fit to the M_{12}-M_{34} plane

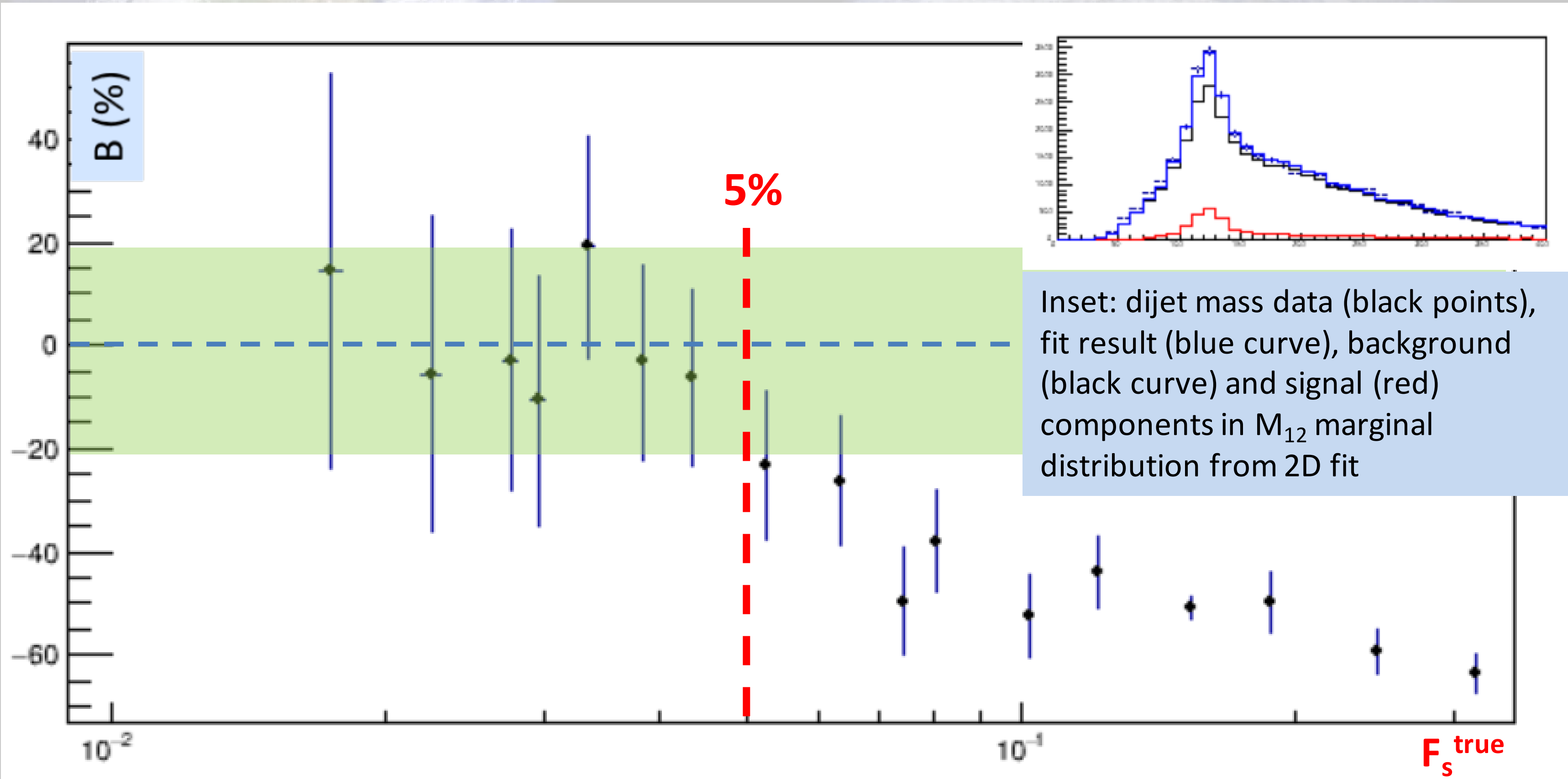
- If the background model provided by event mixing is sound, the bias on the extracted signal fraction should be small (<20% - the typical psychological threshold used in LHC searches)



2D mass distribution for original data (left), background model (center), and signal model (right)

Bias study

The bias to the signal fraction one may fit using artificial data as background model is compatible with zero for signal fraction of a few percents, and only becomes evident above 5%, highlighting that **the method is well suited to typical LHC searches**.



Inset: dijet mass data (black points), fit result (blue curve), background (black curve) and signal (red) components in M_{12} marginal distribution from 2D fit

Above: Bias (%) = $100*(F_s^{fit}-F_s^{true})/F_s^{true}$ as a function of the true signal fraction

Conclusions

- Contrarily to common wisdom, event mixing is a valid technique for high-p_T physics modeling at hadron colliders
  - The trick is to use the **transverse event characteristics** as a basis

- Multi-jet backgrounds can be **accurately modeled** for searches and measurements by creating and resampling hemisphere libraries
  - Particularly useful in small signal searches when QCD is dominant background
- The technique has already been used for a HH→bbbb search in 2015 LHC data (CMS-PAS-HIG-16-017), and is being extended to new searches

Conclusions

- Contrarily to common wisdom, event mixing is a valid technique for high-p_T physics modeling at hadron colliders
  - The trick is to use the **transverse event characteristics** as a basis
- Multi-jet backgrounds can be **accurately modeled** for searches and measurements by creating and resampling hemisphere libraries
  - Particularly useful in small signal searches when QCD is dominant background
- The technique has already been used for a HH→bbbb search in 2015 LHC data (CMS-PAS-HIG-16-017), and is being extended to new searches
- The modeling has been **shown to be valid in the full multi-D space**, enabling the use of artificial data as training sample for MVA classification tasks
- Mixing can also be used to **multiply the statistics** of the original sample, shrinking the statistical uncertainty of the model → very promising developments awaited soon
- A paper is in preparation
  - A public report (D4.1 of AMVA4NewPhysics) discussing multi-D hypothesis tests is already available at **https://tinyurl.com/yd2vfglt**

Thanks for your attention!