Peter Loch
June 16, 2009

Remarks concerning the Evaluation of Jet Reconstruction Performance

Jet performance should be evaluated with respect to the chosen calorimeter signal definition (towers and clusters) and the jet definition as suggested in the Les Houches 2007 proceedings (algorithm, algorithm parameters, recombination scheme). Before embarking on details for performance evaluation strategies, I would like to lay out my own view on what needs to be supported in the preliminary remarks discussed below.

Preliminaries

The large number of final states and the related large number of physics questions which can be addressed at LHC may actually require supporting high precision calibration for at least two different jet algorithms for the lifetime of ATLAS. For example, while in most cases recursive recombination algorithms like kT or Anti-kT perform better and are safer than even seedless cone algorithms, mainly due to the absence of the somewhat arbitrary split/merge parameter, some decays like the hadronic Z' may be better reconstructed with a fixed cone algorithms like SISCone, see Figure 1, taken from Gavin Salam's presentation at the last hadronic calibration workshop[1] in 2008.

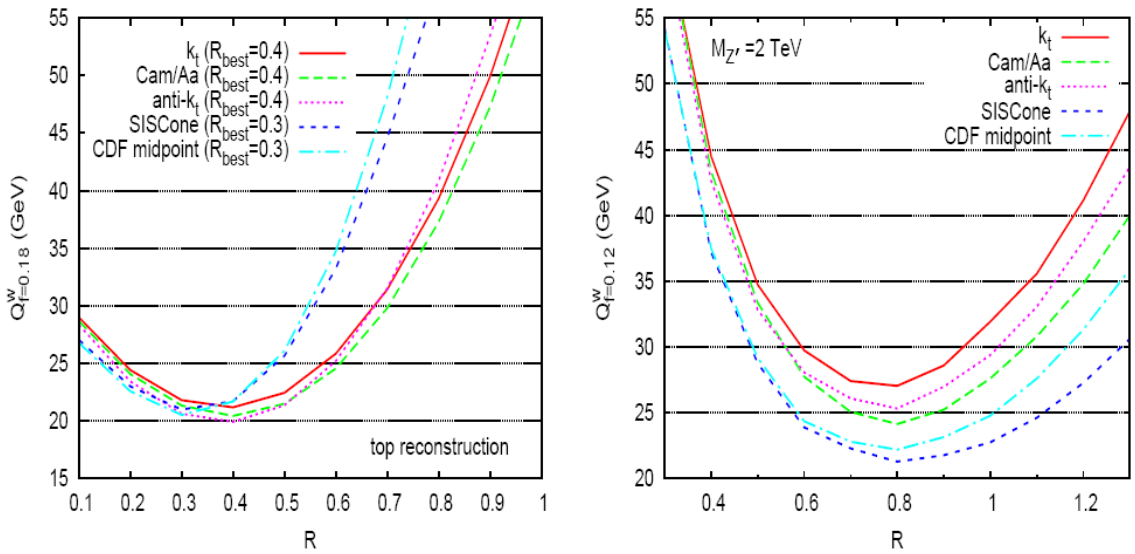

Figure 1: Minimum width of the mass distribution in top (left) and Z' (right) decay reconstruction containing a fixed fraction of all events, for various jet finders. From G. Salam et al., presented at the ATLAS Hadronic Calibration Workshop at Tucson in March 2008.

[1] See http://indico.cern.ch/getFile.py/access?contribId=5&sessionId=1&resId=0&materialId=slides&confId=26943

In addition, it is also quite clear that even for a given algorithm the jet size parameter needs to be adjusted for optimal resolution. Examples here are (inclusive) QCD jet cross-section measurements, which are often performed best if wider jets are used, or hadronic W mass spectroscopy in busy events like $t\bar{t}$ production, where narrow jets may perform better. These arguments may be a little to naïve, as narrow jet QCD analysis may actually be as efficient as using wider jets with less problems from pile-up, while heavily boosted top quarks may better be identified in full hadronic decays by analyzing the substructure of a wider jet merging the two light quark and the b quark jet, instead of trying to reconstruct each jet individually.

Of course, providing precise calibration for two jet algorithms with two different jet size configurations for all expected final states is very challenging, if not impossible. But it may be possible to exploit the commonalities of the reconstructed jets for calibration. For example, it is very likely that the same jet reconstructed with two different algorithms or algorithm configurations has a very similar core with respect to the transverse momentum contribution and spatial distribution of the contributing particles or calorimeter signals – after all, both jets are reconstructed at the same direction and thus are based on the same signals or particles. Main differences are expected at the jet margins, or more general, in the low p_T constituent contribution. This means that the major correction in the jet calibration is universal with respect to algorithm choice, if the prominent constituents provide sensitive variables to parameterize those. It is quite clear, though, that achieving a (better than) 1% systematic uncertainty on the jet energy scale requires some hopefully small jet definition, and likely also topology dependent corrections. Studying universal jet features and calibrations, even though not necessarily very important for the initial data due to the use of more detailed and likely at first less understood signal features, should therefore be part of the roadmap to optimal jet reconstruction.

Kinematic variables for systematic evaluations

The kinematic variables helpful to evaluate for a full characterization of jet reconstruction and calibration performance can viewed in two categories. The first contains the observable variables accessible for any individual jet, while the second contains variables reconstructed from tow or more jets as well as overall event kinematics.

Kinematic jet variables are all energy, direction, and mass reconstruction related variables which can be calculated from the fully reconstructed $(E, \vec{p})$ for a given jet. Note that the basic jet measurement in calorimeter is direction and energy by combination of signals from energy deposited at a given location in the detector. This introduces different but correlated quality "scales" for energy, momentum, and mass reconstruction which should be examined. For example, the quality of direction reconstruction actually determines the transverse momentum resolution for high rapidity jets starting from $y \approx 4$, while at lower rapidity the quality of the energy measurement is the determining contribution. The reconstruction quality evaluated for an individual jet can still depend on the topology of the collision this particular jet has been created in. The sensitivity of all reconstructed jet variables to this environment should therefore be evaluated as well. In addition, specific jets like from heavy flavour quarks, or single jets reconstructed from boosted heavy particle decays, may have other quality requirements to be considered. For

example, jets from boosted heavy particles often use sub-jets to reconstruct the original particle mass, meaning that the energy scale and direction reconstruction uncertainties and fluctuations of these sub-jets are important for the performance of this measurement. This is probably a second order quality estimator, but should be considered, as the sub-jet mass scale, if resolvable at all, can be important for the discovery of new particles.

Multi-jet and **event kinematics** add other handles to evaluate the jet reconstruction quality. Of particular interest are variables accessible in "quiet" topologies, like di-jet mass and the corresponding rapidity gap between jets in QCD, and variables like the transverse momentum balance, expressed as a ratio or in bi-sectional variables if appropriate, in di-jet, Z/γ+jet(s), and multi-jet systems. Care has to be taken to avoid strong model dependences[2] in the evaluation of jet reconstruction performances, or at least those need to be understood and highlighted within the context of the comparison. W mass spectroscopy in hadronic decays is attractive because of the in-situ truth reference available in data. The jet topology variables mentioned earlier may actually not have a stable jet by jet truth references. In some cases event by event truth references may be available, especially in MC, but in others statistical comparisons between data and MC may be sufficient and safer with respect to error evaluation.

Some variables indicating differences in jet reconstruction quality, which maybe not the highest priority for the first data, are related to jet areas and transverse momentum densities. Strategies to measure these are likely different for different calorimeter signals, jet algorithms, and jet algorithm configurations. They are also more complex with respect to the truth reference they are evaluated against, as in case of perfectly matched truth jets the areas at particle and detector level may be considerably different. This should be studied at some point. Note that jet area reconstruction is sensitive to pileup, and this sensitivity can be evaluated in the context of a jet reconstruction performance comparison. Also, there are likely topology dependencies of this measure, e.g. jets from color singlets or color octets, jets close to each other or close to other reconstructed particles, etc.

Significance of variables and timeline for comparisons

The ability to safely compare different jet reconstruction schemes and calibration approaches is strongly related to the progress in understanding the detector acceptance and signal features, together with the need to provide a reasonable jet response for physics analysis as early as possible. Unfortunately, the related process can easily lead to a non-optimal choice when initial lack of understanding (and knowledge) of detector signal features masks performance problems for the selected strategy. To avoid this, a staggered or decision is preferable. This strategy must be documented and communicated extremely well, especially to physics groups, and a best but conservative systematic scale uncertainty must be provided for each jet calibration evaluation. And one more time, it is strongly suggested to start debugging of tools available for even the most detailed analysis already

[2] Note that rapidity gaps and angular separation in azimuth in QCD di-(multi-)jet events are interesting for MC generator evaluation (hard scatter models as well as minbias), especially when looking beyond the two hardest jets, see e.g. S. Mrenna's studies for second and third jet direction separation at Tevatron/CDF.

in the early stage, to avoid losing too much time with tool debugging and event selections when higher quality data becomes available.

Approaching a decision for a certain jet reconstruction and calibration can introduce a timeline given below (assuming physics collisions at 10 TeV; all collected data sample sizes to be confirmed). Note that the real time estimates are likely optimistic.

Initial stage: commissioning for collision physics – first run and/or 1-3(?) months of collisions

Samples/triggers.....minimum bias triggers, jet triggers, photon triggers
Detector reconstruction..... calorimeter cells/towers/clusters, tracks
Physics reconstruction .....jets, missing Et
MC .....selective, mostly minbias
Data/quality.....few pb^{-1}/initial to good detector signals
Beam energy (center of mass).....900 – 10,000 GeV

This initial stage started with the first recorded particle events, and lasts until a certain level of the calorimeter signals has been achieved. Important studies during this phase, besides the ones characterizing the basic cell level detector signal, are single isolated track response in minimum bias, and a first look at jet response with at least one previously derived (MC based) calibration. At this stage the focus should be on finding unexpected response problems (e.g., dead or noisy regions in the calorimeters) by looking at the jet and/or cluster and tower response (balancing not needed!) as function of direction for different high statistics trigger samples. First use and real data commissioning of tools like MPF, jet-jet, and jet-photon balance. At the end a basic understanding of the hadronic signal efficiency in the active regions of the calorimeters, i.e. the raw signal at cluster and tower level, should have developed, including some time dependent changes. Early detailed MC can be useful for the track response in minimum bias. *This stage prepares for calibration comparisons, it should not be part of them - especially it should not be part of only one specific approach!*

Early stage: jet calibration for first physics and initial comparisons – first 6(?) months of collisions after commissioning

Samples/triggers..... minimum bias triggers, di-jets, photon+jets
Detector reconstruction..... calorimeter cells/towers/clusters, tracks
Physics reconstruction .....jets, missing Et
MC .....minbias, QCD photon+jets and di-jets
Data/quality..... up to a few 100 pb^{-1}/good detector signals
Beam energy (center of mass).....10,000 GeV

Once the detector signal quality has reached a first level of good quality, jet reconstruction can be evaluated in more detail. Studies related to the underlying event activity and pile-up (if present) can start, and pT balance techniques can be explored. The data col-

lected in this period should be of sufficient quality to allow a data driven calibration[3], and first comparisons of data and MC. The focus should be to deliver a first calibration with a flat jet response in transverse momentum and direction for initial physics studies (e.g., Anti-kT 0.4, with topological clusters), while at the same time also start to evaluate the performance of different other configurations appropriate for Standard Model physics analysis. For this evaluation, the same events used in data driven calibration (photon-jet and di-jets) should be used. In addition, energy flow pattern from minimum bias and underlying events should be explored with as much reliable experimental input as possible (e.g., # secondary vertices).

The important observables at this stage are the response as function of jet p_T and η, with the appropriate tools (jet calibration task force wiki). The relative jet energy resolution can be measured with di-jets for the considered calibration schemes, and first comparisons of the least MC dependent jet observables can be performed for the various approaches. These can include comparisons of event features like di-jet invariant mass, underlying event analysis a la R. Field, and missing Et distribution shapes. Also included should be evaluations of jet reconstruction efficiencies and fake rates. In any case, the data quality should be good (no technical defects, corrected or removed problematic cells, sectors, regions, large central detector acceptance, ...).

An additional task at this stage is the evaluation of the calorimeter signal definition with respect to its stability against (pile-up and electronic) noise, and its usefulness for pile-up baseline suppression and missing Et resolution and scale. The aim is to make a strong recommendation for one particular calorimeter signal. This may require selective explorations into performance beyond signal linearity and resolution, e.g, with respect to jet shapes.

The goal after this phase is to provide a first jet calibration with realistic systematic errors for physics analysis. It is also the first phase of the comparison between different jet algorithms. A realistic precision goal is probably ~5% absolute scale error in the central and endcap region, and ~10% in the forward region. *Note that at this stage a first decision is taken by providing a first jet calibration – but this should really be considered as temporary! Future improvements should not be limited to improvements of the initial strategy but include alternatives promising better precision, maybe at the price of more detailed validations with MC etc.*

Final stage: conclusion of comparisons after about 1-2(?) years of collisions

Samples/triggers	minimum bias, di-jets, $\gamma / Z +$ jets, $W \to jj$ in $t\bar{t}$
Detector reconstruction	calorimeter cells/towers/clusters, tracks
Physics reconstruction	jets, missing Et, muons
MC	all processes
Data/quality	up to 10 fb^{-1}/high quality detector signals
Beam energy (center of mass)	10,000 GeV

[3] Even if the photon reconstruction is not ultimately precise, the absolute jet energy normalization should be derived from these events with the appropriate systematic error. The jet energy scale error will be reduced as the understanding of the photon signal improves.

The final stage includes very detailed evaluations of all jet related observables with different algorithms and calibration schemes, now including first topology dependences and more calibration signals like the W mass peak. At this stage detailed comparisons to MC should be possible, and the effect of calibrations on jet shape variables should be understood. A focus should be to provide incorporate the advantages of all calibration models into a common approach, and evaluate new concepts especially for universal calibration strategies, where jet observables most independent of the jet algorithm or its configuration are used to correct the jet energy scale. In general in this phase the absolute precision goal should be <3% central and ~5% forward. Also, at this time the jet performance group should be able to provide well understood and tested tools useful for refined calibration in a physics analysis, including jet shape dependent corrections to improve resolutions, or the inclusion of other signal objects like reconstructed tracks for the same reason. Jet classification with respect to the associated vertex should be understood and be standard part of jet reconstruction at this stage.

Refined jet calibration (rest of the lifetime of the experiment?)

The jet performance group should continue to explore refinements of the jet energy scale calibration, especially focusing on the 1% absolute precision goal. This likely includes exploration of the calibration in the context of a given event topology[4], i.e. including the most sensitive measures of isolation with respect to other jets, reconstructed particles, and unused but significant calorimeter signals. More sophisticated calibration schemes based on the experimental ability (or lack thereof) to separate gluon from quark jets should be explored, with the highest considerations for the stability of the jet signal. The goal is order 1% calibration at least for selected topologies and pile-up scenarios.

The evaluation of the performance even at the earliest stages should include a full statistical analysis of the observed response functions, as discussed in the next section. The software tools for this can be made available in the `JetPerformance` package.

Statistical observables to compare jet reconstruction performances

The usual way of comparing the performance of various jet algorithms and configurations using calorimeter signal clusters, noise suppressed or inclusive (all cells) tower signals is presently a bit restrictive in that only the most obvious observables are taken into account, meaning the average response, resolution, efficiency and fake jet rate in different regions of the (p_T, η) phase space. This may not be sufficient for a full evaluation, as especially at LHC deviations from the assumed Gaussian shape of response functions[5] can have quite serious consequences for discovery and even Standard Model physics analysis. For example, tails in these resolution functions can be enhanced by calibration or correction functions while at the same time the fluctuations around the mean or most probable response are reduced. This may introduce hard to control biases not only in the kinematic

[4] Not so much fully reconstructed final states, as the best calibrated jets are part of that.
[5] Usually defined by the distribution of the ratio, the relative ratio, or the difference between a fully reconstructed kinematic variable and the corresponding expected or "true" value.

reconstruction of jets but especially also in the missing transverse momentum reconstruction. The goal of any jet calibration must therefore be to establish a relation between the reconstructed energy and the chosen truth reference that is "as Gaussian as possible". Note that really only then the golden rule of sampling calorimetry applies, meaning that the most likely true energy E_{true} for a given particle or jet with an observed and calibrated signal E_{rec} is most likely *and* on average E_{true}, and that the probability for this to be true is given by the width of a normal distribution around E_{rec}. It is therefore essential to understand and minimize the tails of response functions and to focus on calibration methods which best restore the Gaussian shape of this function.

Several variables to evaluate this shape need to include stable measures for the tails and general asymmetry. Truncated Gaussian fits, i.e. fitting only a chosen range of the distribution, are popular but likely not the most stable approach, as the result may actually depend on the accumulated statistics. Details of the shape around the most probable response are also often washed out by these fits as well. Less statistics model dependent (no fit) distribution shape measures should at least be included. The following estimators should be consistently evaluated and be part of the comparison process:

$\langle R \rangle$	average response in a given kinematic bin
R_{mop}	most probable response in a given bin
R_{med}	median response in a given kinematic bin
$\sigma = \sqrt{\langle R^2 \rangle - \langle R \rangle^2}$	standard deviation (RMS) of the response function
$\frac{\sigma}{\langle R \rangle}$	fractional resolution
$\gamma_3 = \frac{\sum_{i=1}^{N} \left(R_i - \langle R \rangle\right)^3}{N\sigma^3}$	skewness of the response function
$\gamma_4 = \frac{\sum_{i=1}^{N} \left(R_i - \langle R \rangle\right)^4}{N\sigma^4} - 3$	kurtosis of response function
$Q^w_{f=50\%}$	narrowest value range in the distribution containing a fixed fraction (here 50%) of events
f^-_{tails}, f^+_{tails}	fraction of events in low/high tails of distribution

This list is a starting point, there may be more statistical estimators useful to determine the quality of jet reconstruction. Note that the response can be defined many ways - basically all quantities expected to be Gaussian distributed can be used (ratio of signal and true energy in MC, in-situ p_T balance, reconstructed mass distributions in hadronic decays after subtracting background shapes, etc.). The following discusses the features of the less used quantities in some detail. Note that in general bin independent definitions of all variables are preferred!

$\boldsymbol{R_{mop}}$, the most probable value of the response function, is not easy to measure, as it often requires a model describing the peak of the distribution. This measure should therefore be

used with care, especially concerning its error – probably not the most useful and discriminant statistical quantity for calibration quality evaluations and comparisons.

R_{med} is the median of the response function. It can be safely calculated from binned or un-binned data using e.g. ROOT, and has a well defined error.

$\langle R \rangle = R_{mop} = R_{med}$ holds for a symmetric distribution. Of course this is insufficient to determine the Gaussian character. Here variables sensitive to the shape are needed.

(Normalized) moments around the mean: σ, μ_3, μ_4 are sensitive to the shape of the distribution. They can be used to estimate the Gaussian character, because especially in the case of the higher order moments expected values are $\mu_3 = \mu_4 = 0$. Any deviation from this indicates asymmetry ($\mu_3 \neq 0$, with $\operatorname{sgn} \mu_3$ indicating the direction of the skew) or symmetric deformation ($\mu_4 > 0$). The standard deviation σ for a Gaussian is of course defined such that the range $\langle R \rangle \pm \sigma$ contains $\approx 68\%$ of all events. Note that a bin-independent and thus preferred calculation of the higher order moments requires a priori knowledge of the mean, i.e. the data points analyzed should be cached.

Quality estimator $Q^{w}_{f=50\%}$**:** this measure is also very attractive as it does not depend on any assumption of the shape of the distribution. It is the narrowest range of data containing a given fraction f of events (e.g., 50%). While it is not directly a resolution measure, it can be used in the context of comparisons because a smaller range clearly indicates better resolution at a fixed number of events, without any fit to a function model. Also, all statistical variables can be recalculated within the indicated range (truncation without fit). And the range delimiters can be used to determine the fractional amount of events in low and high end tails, again without analytical model. It seems that this estimator is rather stable for all peaked resolution functions. Figure 2 shows an example for the quality estimator for a di-quark mass distribution.

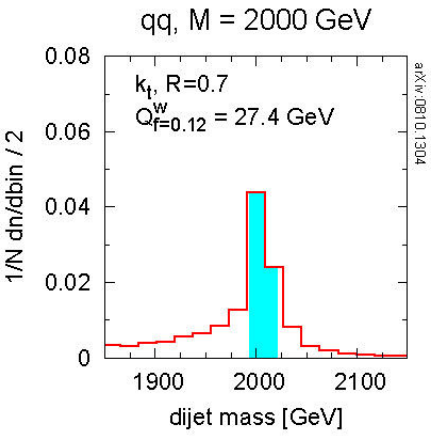

Figure 2: Example for the quality estimator in a di-quark mass distribution (shaded area), from Cacciari, Salam, Soyez (http://quality.fastjet.fr)

Resolution function tails f^{-}_{tails}, f^{+}_{tails} measure the fractional number of events found in the tails of the resolution function (low and high tails). These are important as the aim of jet calibration must be to reduce these potentially very problematic tails which can introduce not only fake selection rates in the event filters but also disturb the offline event selection, the jet energy resolution in an often non-quantified way, and the scale and resolution of the missing transverse energy reconstruction. Safe estimators of tails are not easy to get. The model-dependent approach is fitting e.g. a Gaussian on the full or truncated distribution and calculating

$f_{tails}^{-}, f_{tails}^{+}$ using the fitted width and the number of events outside of some range defined by it. This scenario introduces problems with the interpretation of tails. First, for any fitted model the probability densities can differ between the chosen function and the actual distribution, even within the fitted range. This can then affect the probability interpretation in general. Second, the errors on the tails can be hard to estimate (bi-nominal?). Using the quality estimator $Q_{f=50\%}^{w}$ with an unbiased definition of sample probability may be one solution to estimate $f_{tails}^{-}, f_{tails}^{+}$ more safely and with a cleaner bi-nominal error estimate.

Rules of engagement

Naturally, meaningful comparisons of the performance of different jet algorithms, algorithm configurations, and calorimeter signal definitions requires agreed-upon event samples and object selections (cuts). Any deviation from these selections or samples must be clearly stated and applied to competing calibration and reconstruction schemes as well. In general the jet response cannot (and should not) be optimized by changing the calibration event selection – the immediate goal must be to reconstruct all jets in the accessible phase space and above a detector imposed threshold with the highest possible quality. This does not mean that the quality achieved can not be different in different regions of the phase space, of course. For example, even though corrections for jet response in problematic detector regions should be included in all calibration approaches, comparisons of these can generate different results after applying fiducial volume cuts to avoid these regions. In particular, the error in these comparisons can be more meaningful in this case. Nevertheless, a full evaluation is best to be done for the full detector coverage, including regions of reduced acceptance and/or response. Decisions for one or another scheme then include scenarios where one strategy performs better than another in high acceptance regions, while in problematic regions the finding may be opposite. These findings may make the evaluation inconclusive at first, and may require inclusion of more collision physics final state oriented aspects (e.g., topology dependences of reconstruction performance), possibly resulting in accepting both approaches in the end.

The range of jet p_T for performance comparisons should be driven by realistic expectations, especially concerning the lower edge. Here the event topology may play a larger role than presently reflected in reconstruction performance evaluations. There is no real reason not to believe that if a certain low p_T jet reconstruction performs not at optimum in QCD di-jet events, the same reconstruction strategy can work in e.g. W+n jet(s) topologies – and vice versa. This needs to be carefully explored. In general a large number of agreed upon final states should be included in the full evaluation, each with a clear definition of the calibration normalization. All calibrations should be cross-examined with the full list of accessible final states, with available data handles as well as data-MC comparisons. The samples need to include all relevant Standard Model QCD and electroweak (Higgs, W, Z production associated with jets like VBF, jets from recoiling hadronic systems, or including jets from decays) final states with jet activity. Additional samples with very high activity from models beyond the Standard Model should be included, especial-

ly if they occupy otherwise less covered regions of the phase space, like some SUSY final states with a large number of jets and leptons, or very high energy jets.

Pileup may be an issue at LHC quite early on, and should be considered in the evaluation of jet reconstruction performance anyway. While event categorization using secondary vertices seems to be a good data driven strategy, it requires not only the understanding of the vertex reconstruction efficiency and quality, but also the jet vertex association. These are likely not available early on with sufficient precision, but should be useful after some initial running.

In a side remark, it is probably useful to think about (calorimeter) resolution as the quantifier for the likelihood to reconstruct a true quantity for a given signal. In case of perfectly Gaussian fluctuations, and with the already mentioned principle of calorimetric energy reconstruction in mind, the distribution of the true energy for a (narrow) bin of fully calibrated and corrected signals must be Gaussian, with an average corresponding to the bin average[6] and a width indicating the error on this mean. Of course, the true energy distribution can be replaced by a response distribution, thus measuring the fractional (relative) resolution.

Software tools

`JetPerformance` for relevant plots, uses combinations of **`JetAnalysisTools`** tools, one specific evaluation/tool

`JetAnalysisTools` for jet (and others!) filters, selectors, object matching, shape evaluations including prominent constituents analysis tools (re-calculation of jet mass and shapes from restricted set of constituents…), distribution moment calculations, quality estimators for distributions, probe and reference variable caching, simple histogram descriptions and booking from python scripts,

[6] Assuming the probability density function for the true variable is flat, i.e. cross-section dependencies, trigger efficiency, etc. are unfolded or irrelevant within the narrow signal bin.