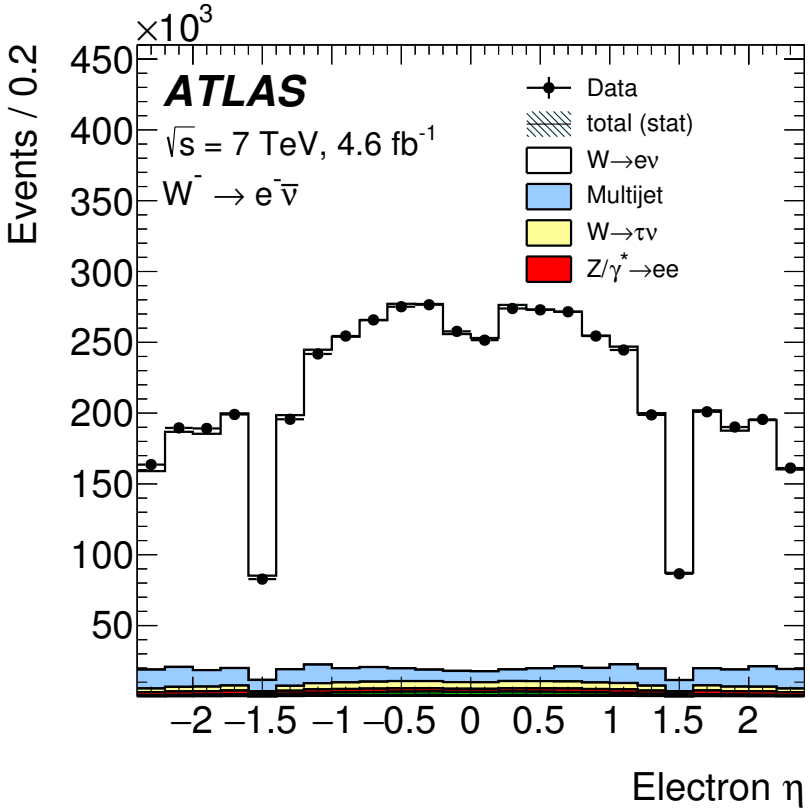
# Unfolding with ML using multiple reconstructed variables.

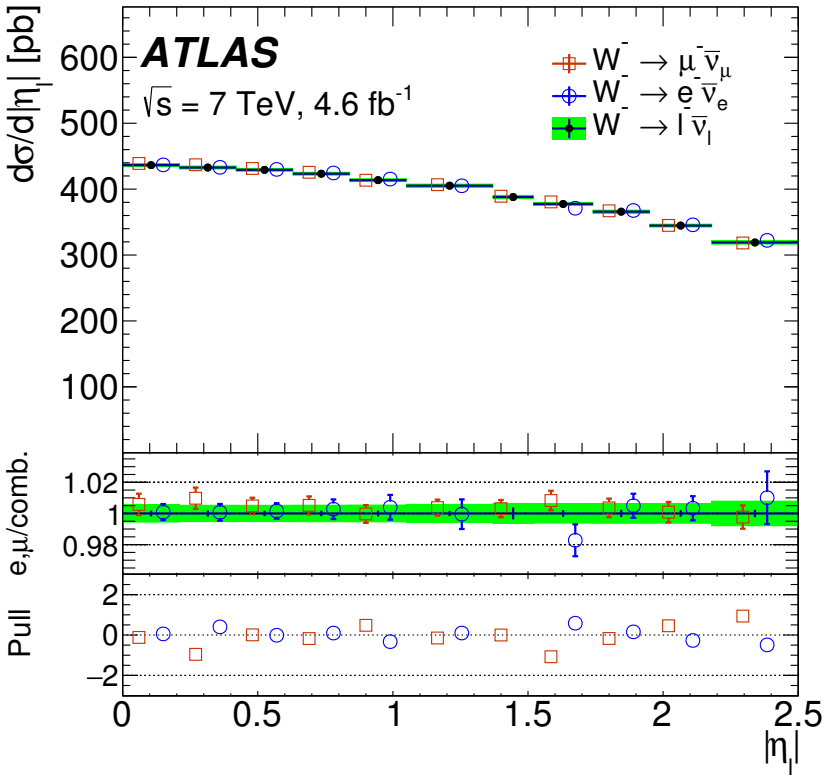
S. Glazov, SMP-J workshop 23/01/18.

# Choice of reconstructed variables for unfolding

- The choice of truth-level definition is connected to the variable used at reconstructed level.
- Usually they are chosen as close as possible to each other, which typically leads to reduced model dependence.
- However in some cases improved resolution may be more important than increased model dependence. Moreover, known explicit model dependence may be preferred if its uncertainty can be estimated in a well-defined way.
- To improve resolution, one may consider using several reco-level variables for unfolding.
- The combination of the information can be non-trivial due to in general non-Gaussian shape of reco. variables: use ML methods for that.

# Example 1: $\eta_\ell$ distribution

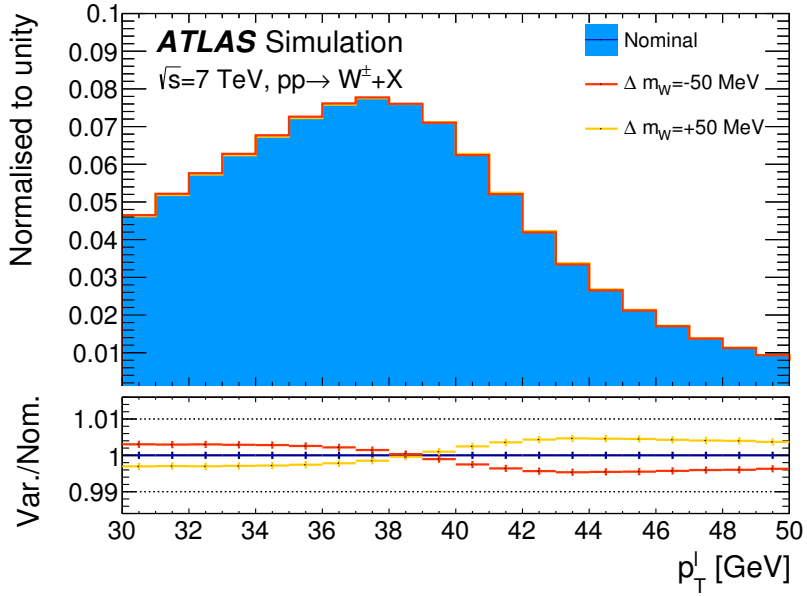


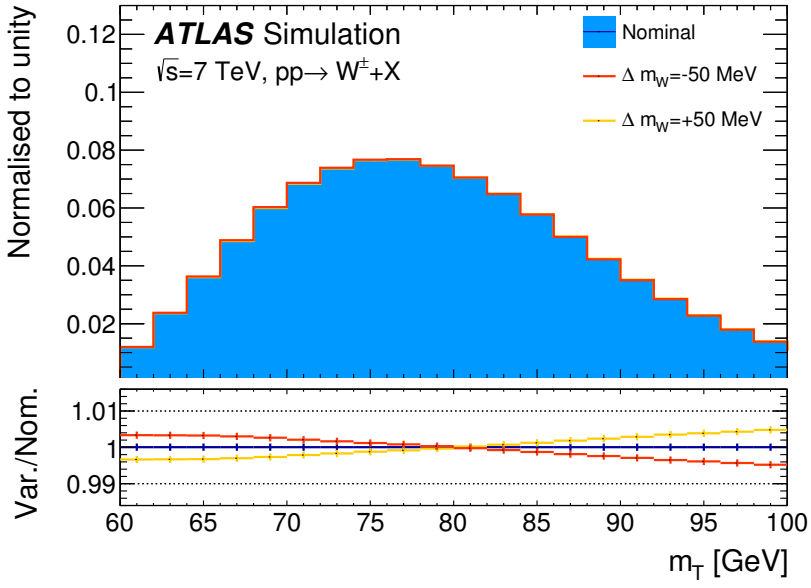




- Excellent resolution for $\eta_\ell$ with purity at 90% level for the bin sizes close to optimal in terms of underlying physics (e.g. variation of PDFs).
- → no need to think about extra variables.

arXiv:1612.03016

# Example 2: $m_W$ unfolding

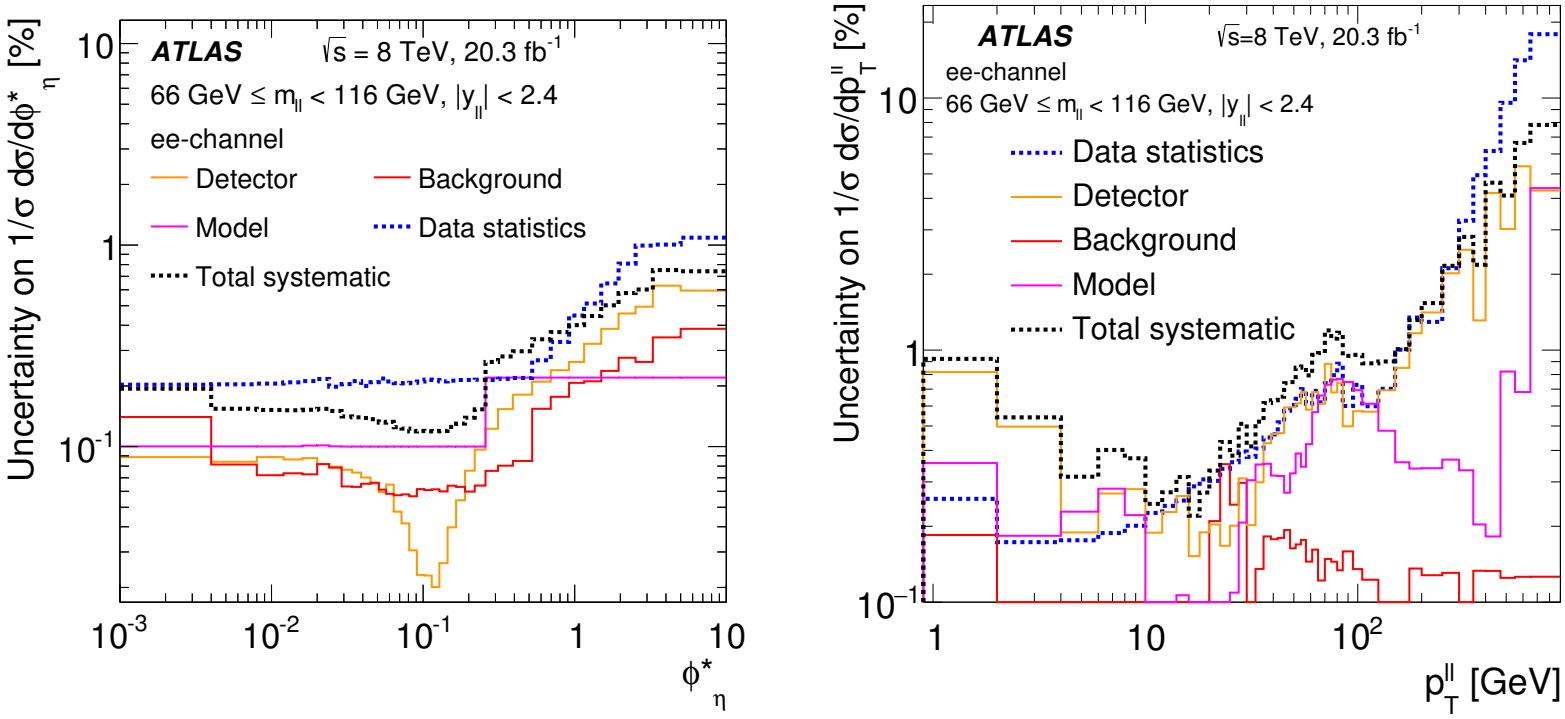



- $m_W$ "unfolded" using reco level fits to $p_T^\ell$ and $M_T$ distributions.
- Reco and unfolded variables are pretty far from each other, this can not be avoided.
- The combination is performed after unfolding, providing extra consistency check.
- Modelling uncertainties have the largest contribution for the result, however they are studied in a systematic way.

arXiv:1701.07240

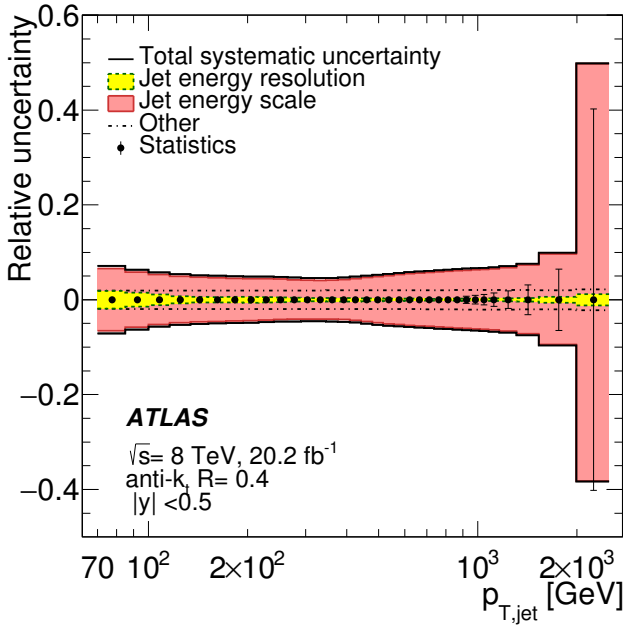
# Example 3: $Z_{p_T}$ and $Z_{\phi^*}$

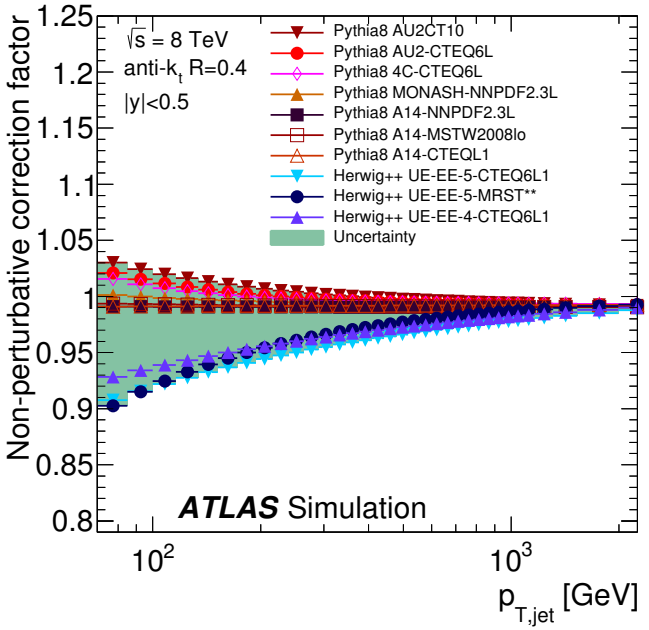



- $\phi^*_\eta$ and $p_{\mathrm{T}}^{\ell\ell}$ variables probe very similar physics, with $\phi^*_\eta$ having much better detector resolution and $p_{\mathrm{T}}^{\ell\ell}$ closer to the underlying physics.
- Additional interest is to probe low $p_{\mathrm{T}}$ region to study location of the peak: resolution is essential.
- → use both variables at reco level to unfold to single $p_{\mathrm{T}}$. Maybe add $M_{\ell\ell}$ information to the mix, to improve the resolution.

arXiv:1512.02192

# Example 4: jets



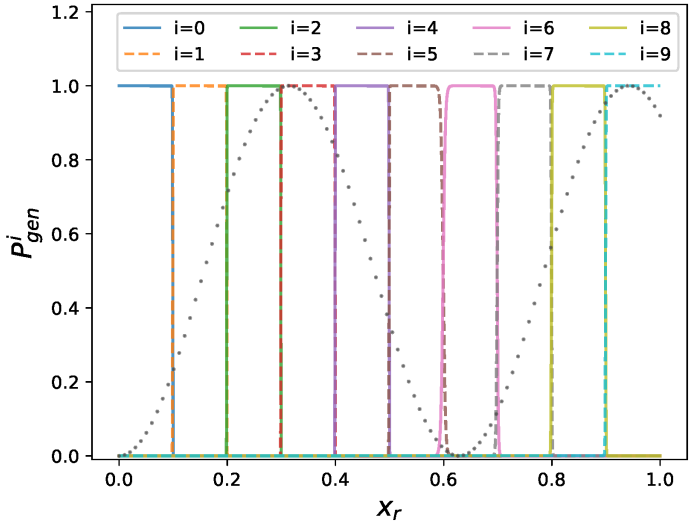




- Systematic uncertainties dominate over stats. for the bulk phase space
- Correlation model, systematic uncertainty PDF (additive vs multiplicative) play crutial role for interpretations
- Interplay between calibration and NP corrections

→ additional event-level information (HF tags, missing energy) may be important to reduce systematic uncertainties.

arXiv:1706.03192

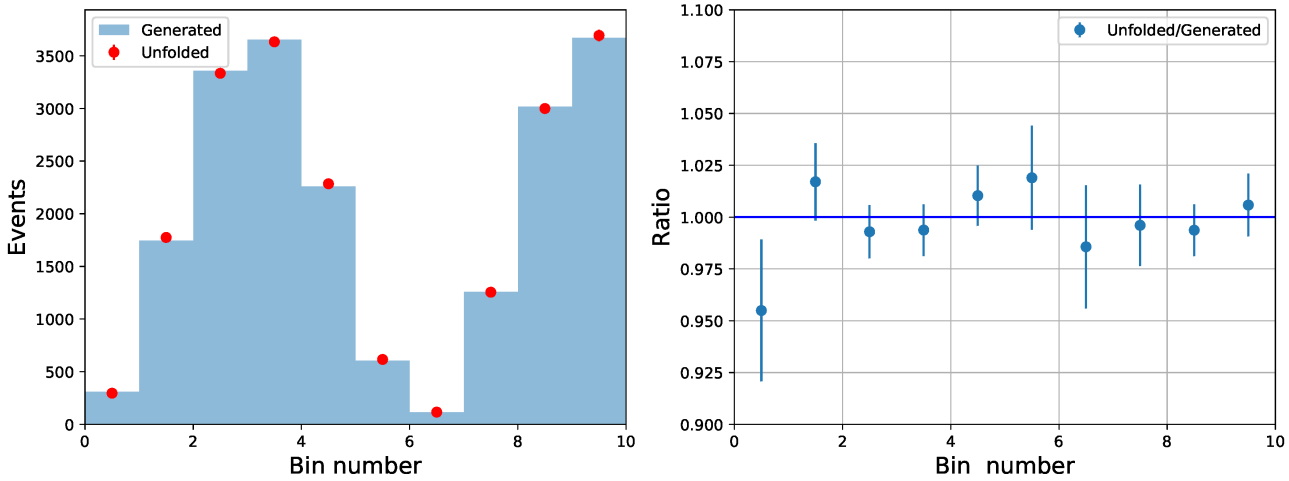
# A method for ML unfolding




- Determination of the truth-bin using reco-level variables is a classification problem, well suited for ML methods.
- Proof of principle test: use single reco-level variable, unfold $f(x) = \sin^2(5x)$ distribution for $0 < x < 1$ using 10 bins.
- Use 3-layer NN: input later, fully connected layer, and final "categorical" layer with softmax activation. Use categorical cross entropy as the loss function.
- Start with no-smearing: the net needs to find truth-bin boundaries.

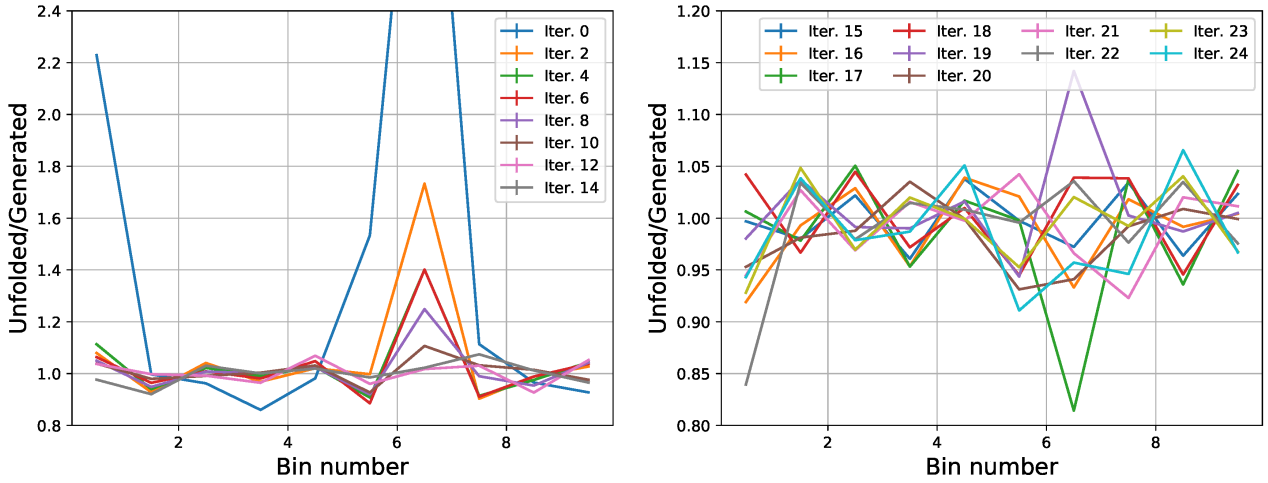
arXiv:1712.01814

# Proof of principle example



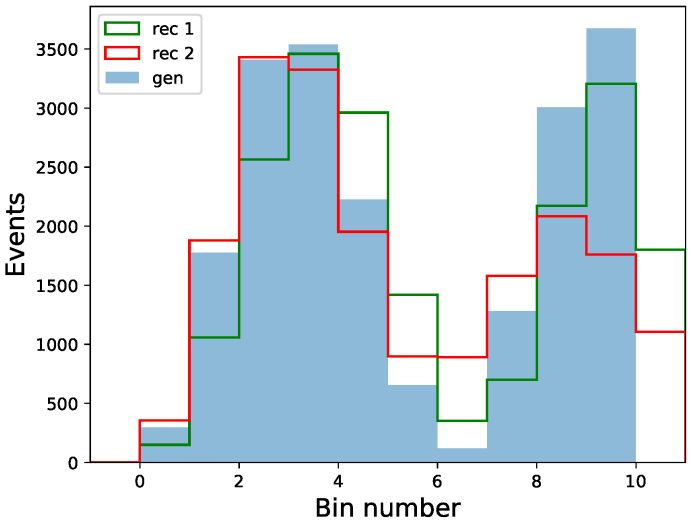

- Unfolding is done in an iterative procedure: updating the distribution of the truth variable based on previous iteration and re-training.
- Convergence is hard to prove mathematically: use numerical closure tests.
- Add smearing (0.5 bin Gaussian), first test uses truth-distribution as the prior. Use bootstrap method to estimate uncertainties.

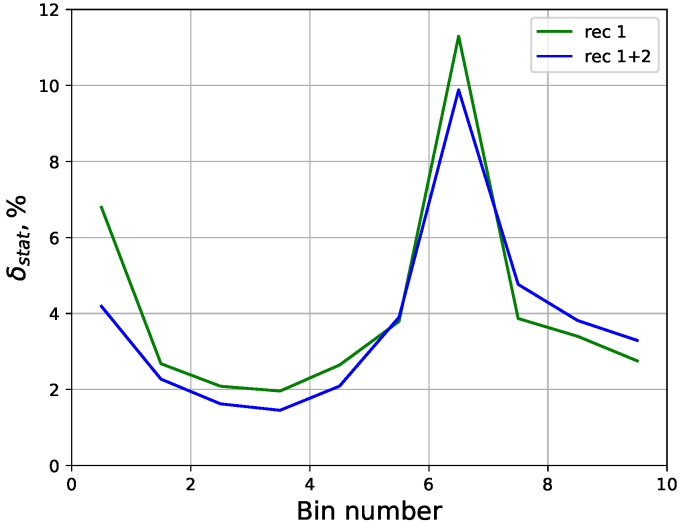
# Convergence of the method




- Start with the flat prior, observe convergence to the truth distribution after about ten iterations. After that the unfolding starts to oscillate around the minimum.
- More details can be found in `https://github.com/aglazov/MLUnfold.git` which contains a jupyter notebook, however tensorflow, and a server with a good GPU are needed to run it.

# Advantage of second variable




- Add second variable:
  - 1st with 0.5-bin Gaussian smearing and +0.5 bin shift;
  - 2nd: with 1.5 bin log-normal smearing and -0.5 bin shift.
- The second variable improves convergence (5 iteration instead of 10), leads to reduction of uncertainty for the first bins.

# Discussion

- Selection of the definition at the truth level can be complemented by selection of an optimal reco-level variable.
- The selection of the optimal reco-variable can be performed using ML methods, based on multiple input variables.
- This can be useful in various situations: when the connection with truth is not straightforward due to missing information (e.g. $m_W$, $p_{T_W}$, $y_W$); when the resolution is of critical importance (e.g. $p_{T_Z}$ at low $p_T$); when the full event information may bring additional calibration info (e.g. balance in $Z$+jet events for jet kinematics).