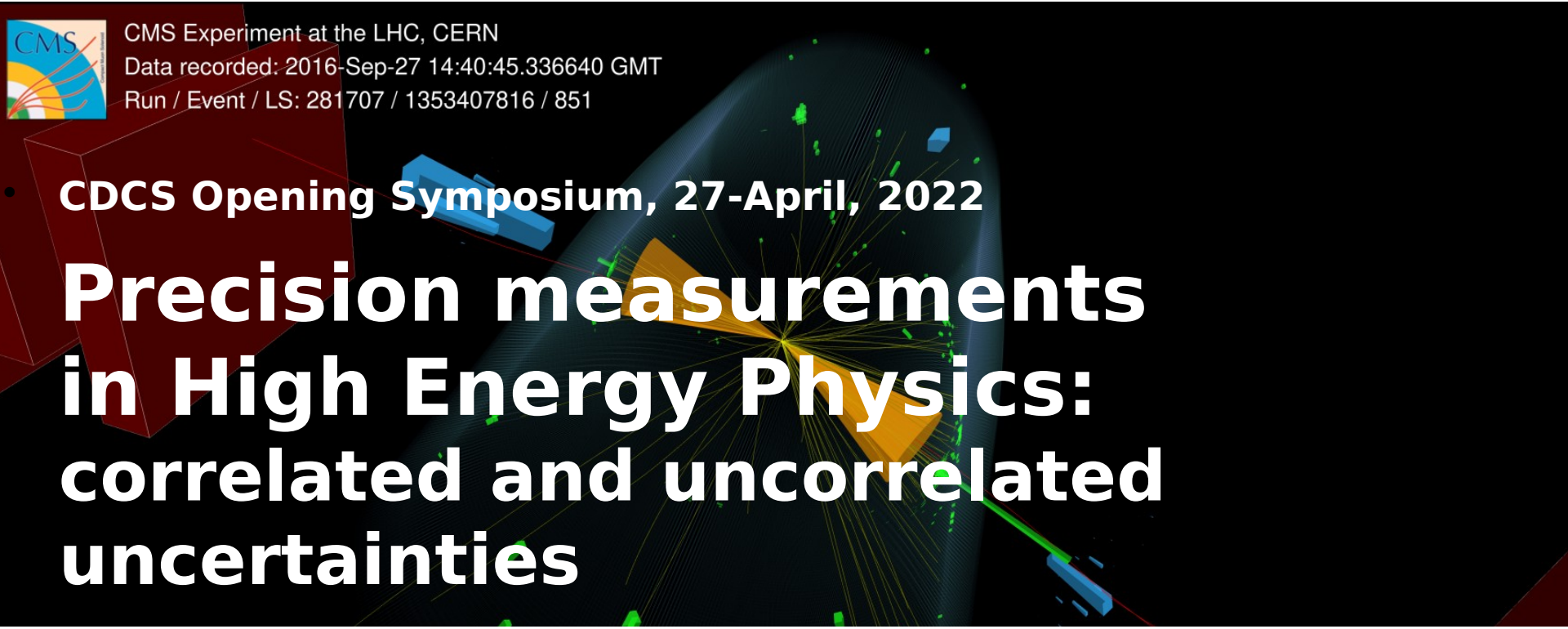
Patrick L.S. Connor (IEXP & CDCS)

CDCS
CENTER FOR DATA AND COMPUTING
IN NATURAL SCIENCES



- **CDCS Opening Symposium, 27-April, 2022**

Precision measurements in High Energy Physics: correlated and uncorrelated uncertainties

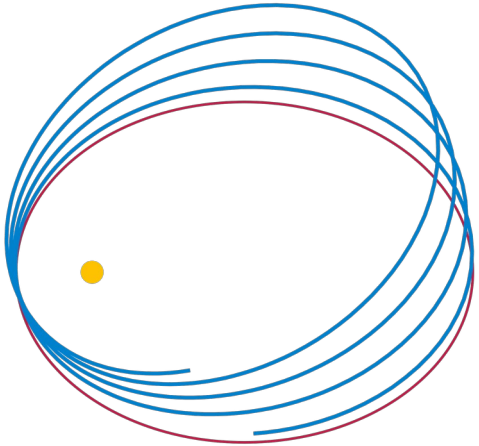
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DESY.

TUHH
Technische Universität Hamburg-Harburg

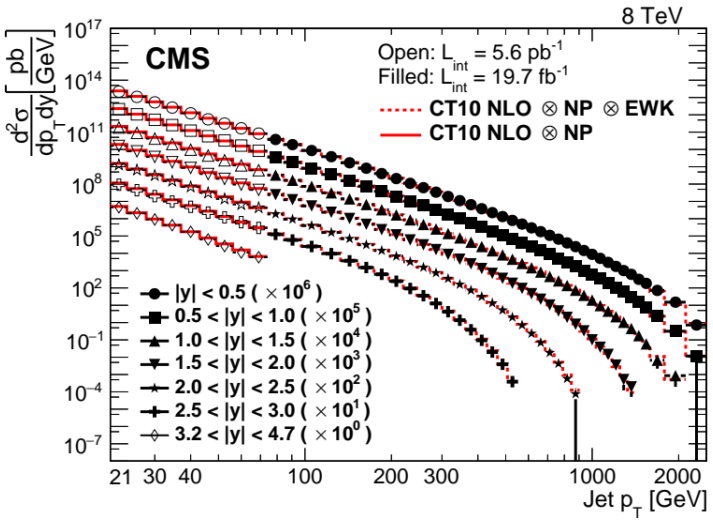
Motivation (1)

- **Kepler & Brahe**: planets‘ orbits are ellipses
- **Newton**: perihelion precession
- **Le Verrier**: residual discrepancy of Mercury‘s orbit of 42.980±0.001″ per century
- **Einstein**: tiny (but significant!) gap between experimental observations and prediction explained by General Relativity



Uncertainties tell you how much you can trust your data

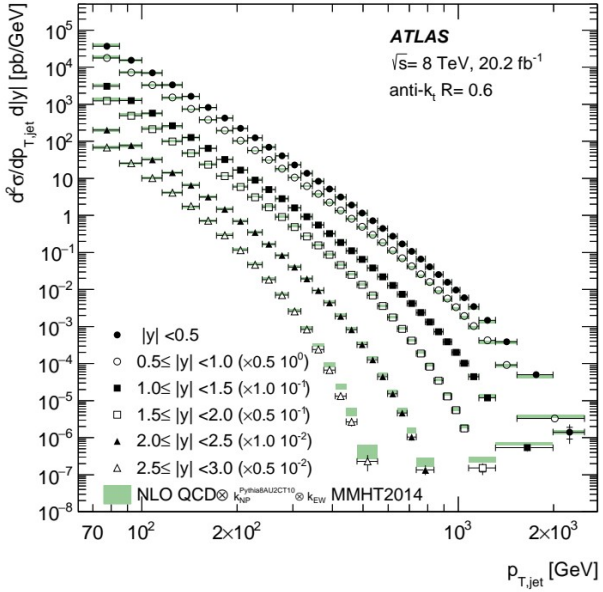
Motivation (2)



Jet differential distributions

→ e.g. inclusive jet

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_\mathrm{T}\,\mathrm{d}y} = \frac{1}{\mathcal{L}}\frac{N_\mathrm{jets}^\mathrm{eff}}{\Delta p_\mathrm{T}\,\Delta y}$$



Uncertainties of such spectra require a sophisticated description.

- Bin-to-bin fully correlated, partially correlated, and fully decorrelated from bin to bin.
- Typically O(10) different sources of systematic uncertainties.
- Direct impact on extraction of fundamental physics quantities such as the α_s.

Goal & Outline

Discuss tools for more or less broad application:

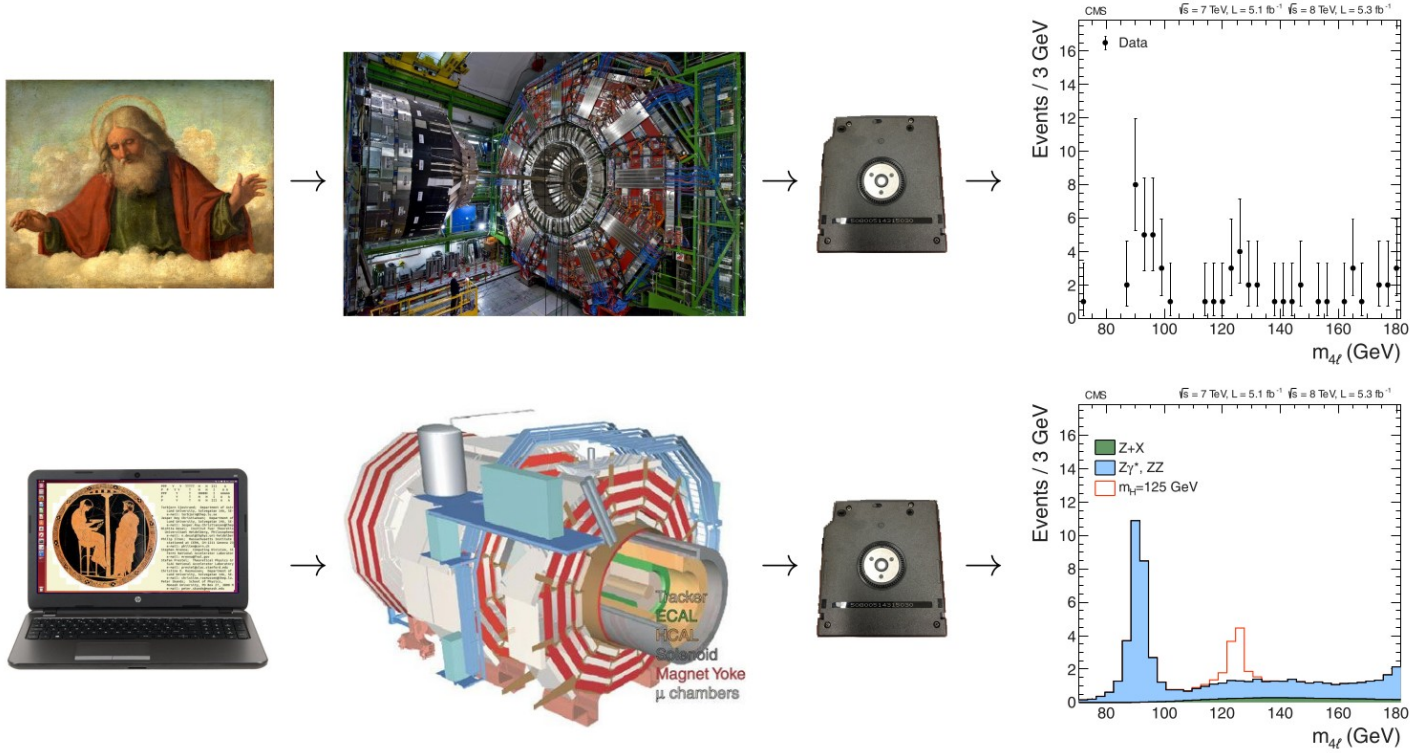
1) **DAS**: Das Analysis System
   - Jet analysis in context of the CMS Collaboration

2) **RAN**: Refinement Adversarial Networks
   - By-pass computing-demanding simulation of detector with Geant4

3) **Step**: Smoothness Tests based on Expansion of Polynomials
   - Investigate the quality of the statistical description of a differential measurement

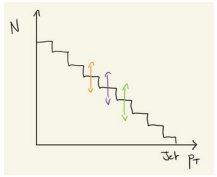
4) **Teddy**: Treat Entries ouf of the Diagonal DecentlY
   - Extract high-level distributions from histograms with complex uncertainty schemes

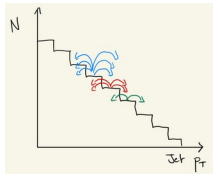
Typical analysis strategy

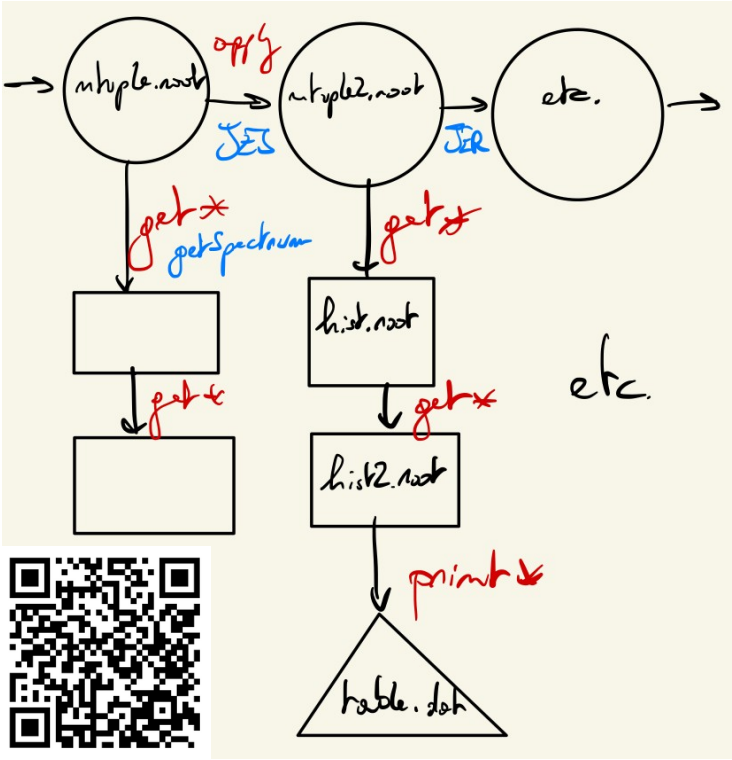


DAS (with CMS colleagues)

- Optimised for **debugging**
  - **Event loop** based
  - **Factorising** all steps
- Accounts natively for **systematic variations** at event level
  - **Vectors** of weights and factors
- Currently dedicated to **CMS jet analysis** exclusively
  - Modular design makes it easily **extendable…**
  - Essentially depends on demand and person power

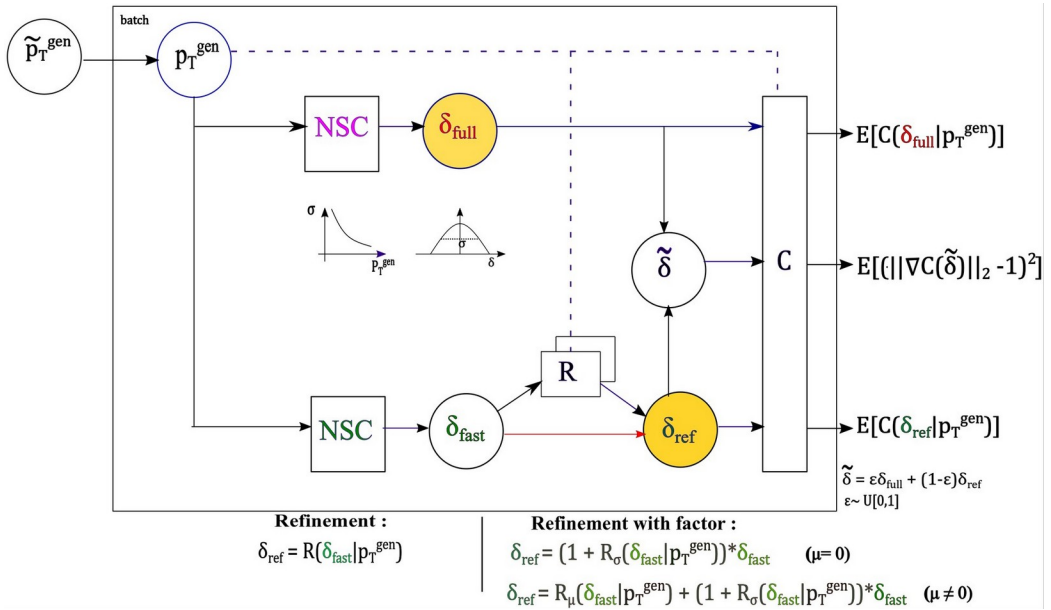






RAN (with TUHH and IEXP colleagues)

- Replace full simulation based on Geant4 by fast simulation based on higher-level **phenomenological** models (e.g. Delphes)
- **Refine** obtained simulation with adversarial networks
- **Interface** with DAS...?



Promising approach to improve the model dependence of the data reduction!

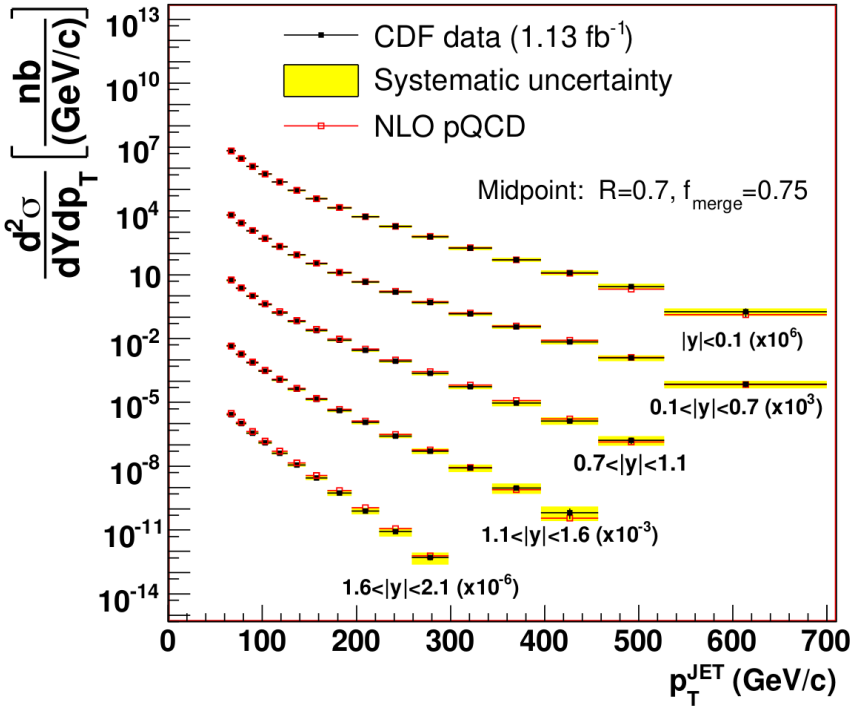
STEP (1) (with Radek Zlebcik)

- Typical jet measurements span over several orders of magnitude.

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_\mathrm{T}\,\mathrm{d}y} = \frac{1}{\mathcal{L}}\frac{N_\mathrm{jets}^\mathrm{eff}}{\Delta p_\mathrm{T}\,\Delta y}$$

- Even *counting* jets is not trivial, while we typically target %-level precision.
- Residual artifacts in the spectrum may render the data difficult to fit.
- Issues may often be spotted only after the data have been published.

→ ~~find~~ build a function with same shape

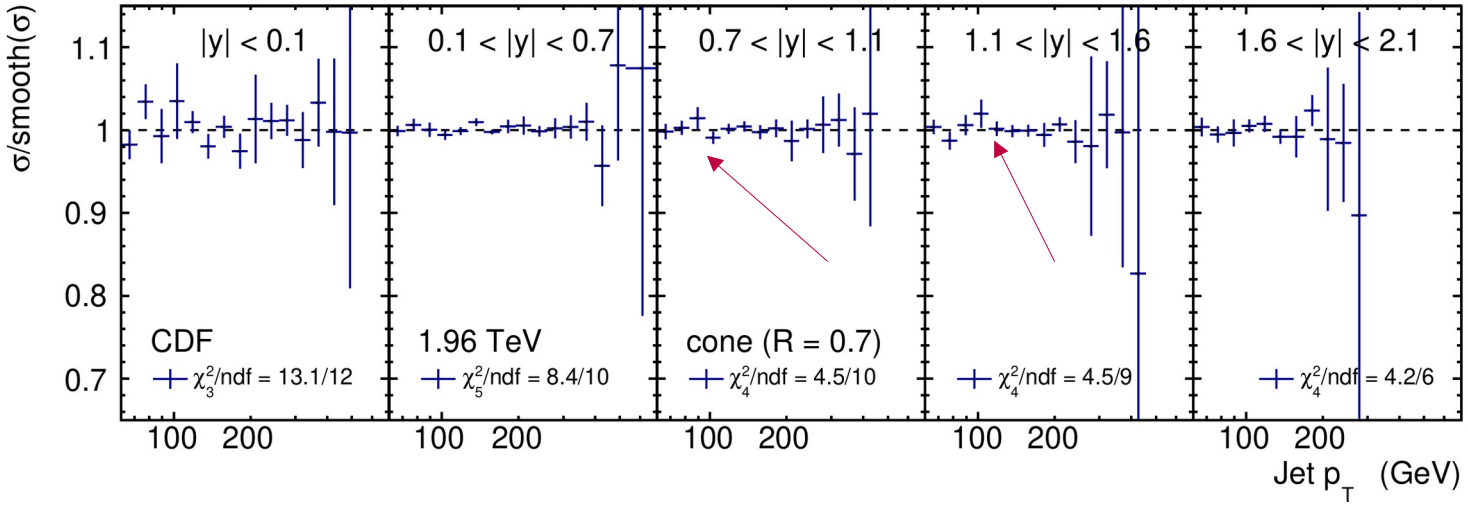


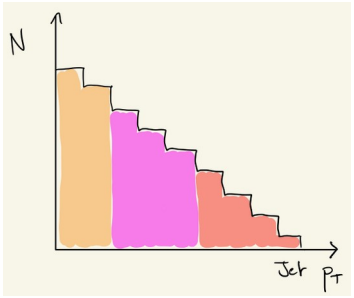
STEP (2) (with Radek Zlebcik)

- Fit with and divide by an ad hoc, (nearly) agnostic, smooth function

$$f_n(p_T) = \exp\left(\sum_{i=0}^{n} b_i T_i\left(2\frac{\log p_T/\log p_T^{\min}}{\log p_T^{\max}/\log p_T^{\min}} - 1\right)\right) \qquad \chi_n^2 = \min_{b_{i\le n}}\left[\left(\mathbf{x}-\mathbf{y}_{b_{i\le n}}\right)^{\mathsf{T}}\mathbf{V}^{-1}\left(\mathbf{x}-\mathbf{y}_{b_{i\le n}}\right)\right]$$

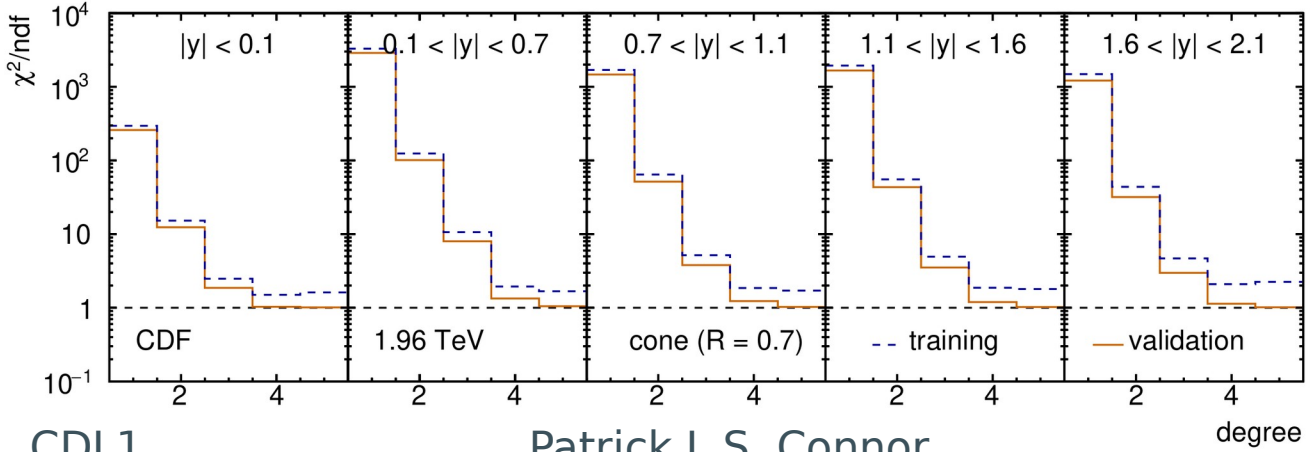
- Track down possible artifacts introduced in the data reduction (e.g. trigger)





STEP (3) (with Radek Zlebcik)

- Iterative fit procedure with early stopping criterion based on cross validation.
- Two sets of 10k replicas (fake data generated from statistical properties of the original spectrum) are produced:
  - training replicas are used to run the fit;
  - validation replicas are used to determine the level of overfitting.
- Not more than 10% of the validation replicas should have better χ^2 than the training replicas.



2111.09968v2

TEDDY (1) (with Radek Zlebcik)

- Apply **transformation** on distributions with complex correlations
  - e.g. ratio of inclusive 2- and 3-jet cross sections
    - Many systematic effects cancel → stronger sensitivity to α_s

$$\frac{\mathrm{d}^2\sigma_3}{\mathrm{d}p_\mathrm{T}\,\mathrm{d}y} \Big/ \frac{\mathrm{d}^2\sigma_2}{\mathrm{d}p_\mathrm{T}\,\mathrm{d}y}$$

- **Normalise** 2D distribution in bins
  - e.g. jet substructure variable λ in bins of p_T
    - Factour out irrelevant physics effects

$$\frac{1}{N(p_\mathrm{T})}\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_\mathrm{T}\,\mathrm{d}\lambda} \quad \text{where} \quad N(p_\mathrm{T}) = \frac{\mathrm{d}\sigma}{\mathrm{d}p_\mathrm{T}}$$

- Extract a **fraction**, etc. etc.

→ generic approach based on MC techniques

TEDDY (2) (with Radek Zlebcik)

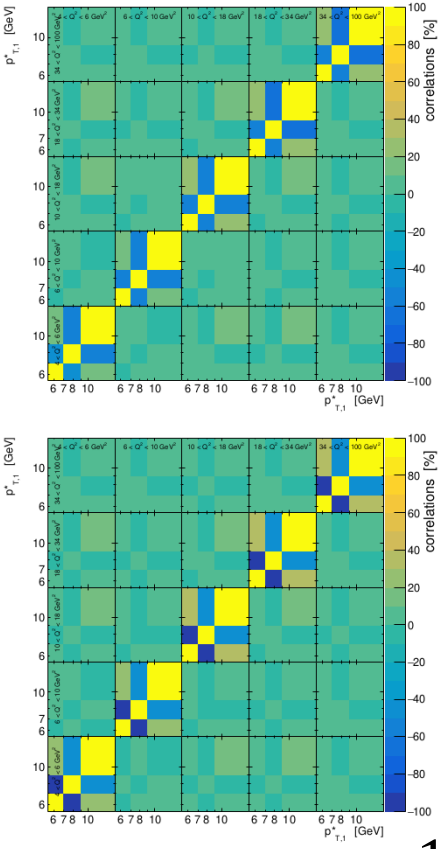
- We want to apply y = f(x)
  - ⇝ x has covariance matrix V

$$\boldsymbol{\theta}_n = \mathbf{f}\left(\mathbf{x} - \mathbf{R}^{-1}\boldsymbol{\delta}'_n\right) \quad \text{with} \quad \delta'_{n,i} \sim \mathcal{N}\left(0, \sqrt{\max(0, k_i)}\right)$$

- The covariance matrix of y is simply obtained by MC integration:

$$\mathbf{W} = \left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\theta}_n \cdot \boldsymbol{\theta}_n^{\intercal}\right) - \frac{1}{N^2}\left(\sum_{n=1}^{N}\boldsymbol{\theta}_n\right) \cdot \left(\sum_{n=1}^{N}\boldsymbol{\theta}_n\right)^{\intercal}$$

→ illustrated with partial normalisation of H1 dijet measurement



Summary & Conclusions

- **Statistical and systematic uncertainties** are crucial in measurements.
- When designing the analysis software, they should be accounted for from the **very beginning** → DAS
- Ongoing project with adversarial networks to help produce very large data sets and better cover **model uncertainties**.
- Generic tools to deal with bin-to-bin **partial correlations** → STEP & TEDDY.

Thanks for your attention!