MC tuning experience from collider physics

Andy Buckley
University of Glasgow

NuTune 2016 workshop, Liverpool, 11 July 2016

Introduction

- **MC tuning is a necessary evil** – data description is needed, and models aren't as predictive as we'd like.
- Data and models also aren't as perfect/complete as we'd like: we need pragmatic best-fits and estimates of systematics.
- *Hope that somewhere along the way we also gain better physical understanding…*

In this talk I'll focus on:

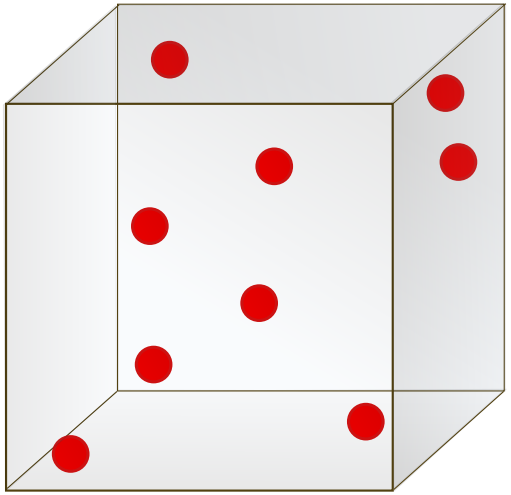
- The **Professor** numerical machinery developed for LHC MC tuning, and now widely used (including beyond tuning)
- Several rounds of LHC tuning, including ATLAS' tuning programme
- Fit statistics and systematics
- Some general lessons learned

Disclaimer: I don't know anything about neutrino MC! Please set me straight…

Professor

The Professor method

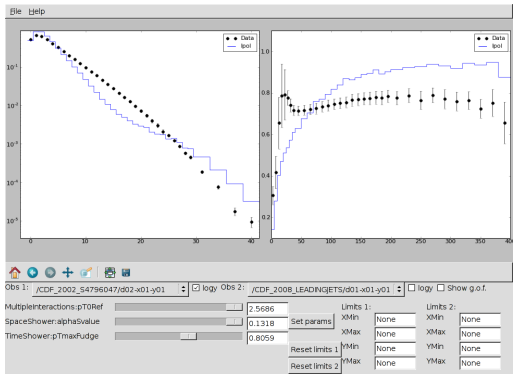
- *No magic, just a numerical assistant to take the grind out of tuning.*
  Physics insight is crucial to pick base config, best params, and keep it physical.
- **MC is slow: $\sim$ 1 day per run** $\Rightarrow$ can't use in serial optimisation e.g. MCMC.
- Trivially parallelise MC runs through "reasonable" subset of param space, and use sampled points to parametrise each bin's behaviour (for $\mathcal{O}(15)$ params).
- Use SVD polynomial fits for generality – values must vary in a polynomial fashion *or be transformed to do so*. Experimental "machine learning" fits in Prof v2.
- Minimisation or other applications of analytic interpolations are easy. Much strength is in the "system" features around the core machinery.



Professors 1 vs. 2

Professor v1

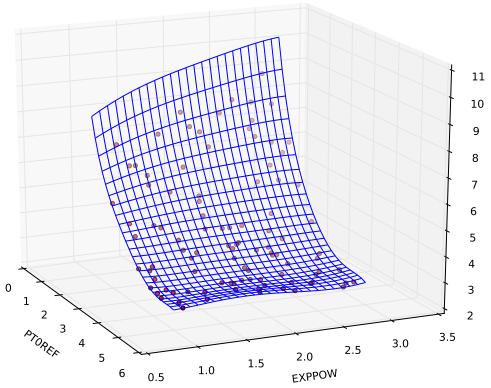
- Python, using scipy and pyminuit. Heavy/awkward code framework – historical baggage from early dev.
- $\mathcal{O}(20)$ scripts…not all well maintained! Lots of "advanced" features: correlations, sensitivities, GUI, etc.
- Still max functionality, but **hit "maximum entropy" point of development!**



Professors 1 vs. 2

Professor v2

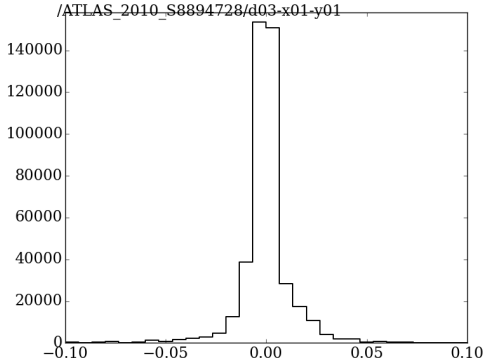
- Redesign from scratch: lean, mean & flexible again. **Thanks to Holger Schulz**
- Core code now in C++; independent of concepts like "bins". Very generic
- Wrapped into Python, and used as core of a library for data I/O ⇒ lessons learned from v1
- Now fewer scripts – and they are just UI wrappers around library functionality: use them as examples/templates

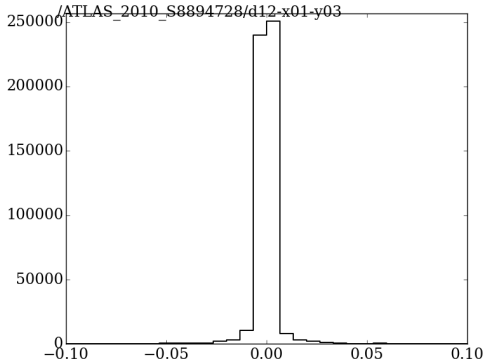


Testing the parameterisation

Can't just assume that parameterisation is working. . . but this is often done / inferred much later. New **`prof2-residuals`** tests explicitly.

- Loop over runs and histogram absolute & relative residuals between ipol and MC, e.g. $(f(p_i) - \mathrm{MC}_i)/\mathrm{MC}_i$
- Breakdown by observable, and value / error. Easily extended.
- For better testing, train interpolation on one run subset and test on the remainder



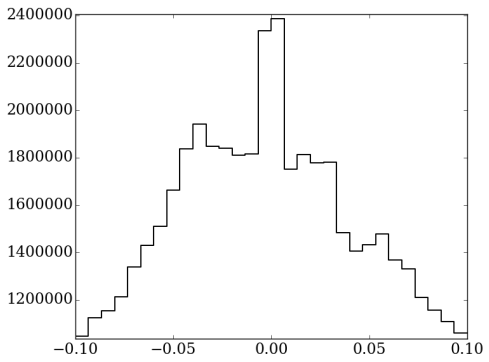


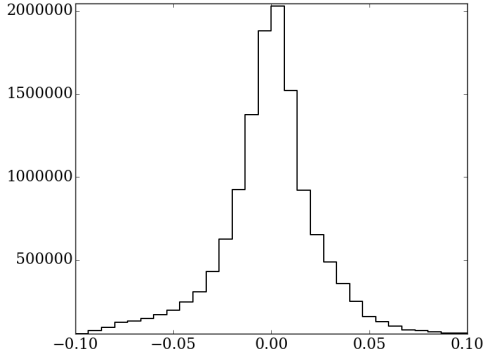
Measuring goodness-of-fit

- Historically used a simple pseudo-χ^2:

$$\chi^2(\vec{p}) = \sum_b w_b \frac{(f_b(\vec{p}) - \mathrm{ref}_b)^2}{\Delta\mathrm{ref}_b^2 + \Delta f_b^2 + \epsilon^2}$$

- Several limitations: no stat/syst separation, weight has $\sqrt{}$ of intuitive effect, $\Delta f_b^2 = \mathrm{median}(\Delta\mathrm{MC})$ i.e. const!
- Expt correlations available in Prof 1.4; coming soon in v2
- v2 uses linearised weights. Also more natural entry in correlations via cov_{ij}
- Prof 2 allows error parameterisation: greatly improves residuals.
  - Denom is of equal importance in χ^2! *Needs to be regularised in fit.*





Tuning systematics

Hard to get a handle on the uncertainties of an MC tune. Conflates several issues:

- Incompleteness of the model
- Discrete variations in models
- Param space sampling & parameterisation issues
- Statistical limitations and tensions in data

Sanity checks can be useful, e.g. assessing observable sensitivity to params: $\mathrm{d}\ln y_b / \mathrm{d}\ln p_i$

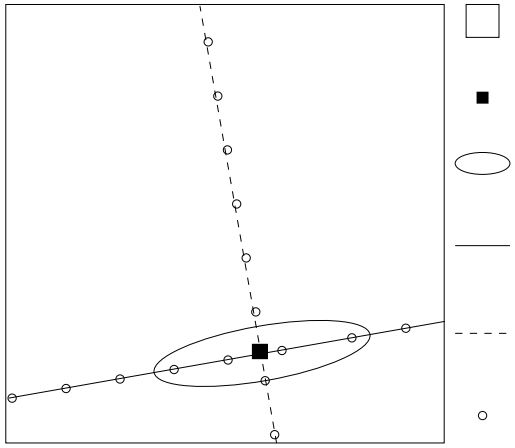
Stability of fits can be assessed by making many replica optimisations on parameterisations made from random subsets of sampled param points

Variations should be derivable from the shape of the goodness-of-fit around the optimum, cf. PDF fits. *Right?*

Eigentunes and coverage

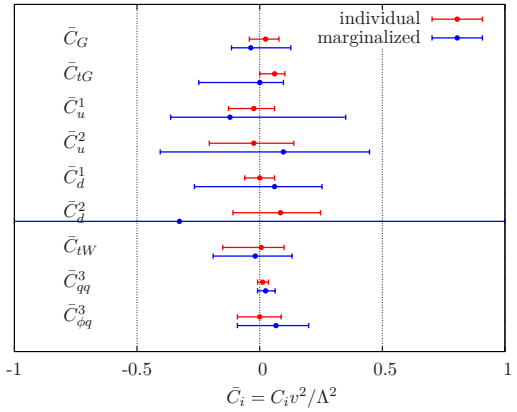
Eigentunes have gained acceptance as a "robust" way to create systematic variation tunes.

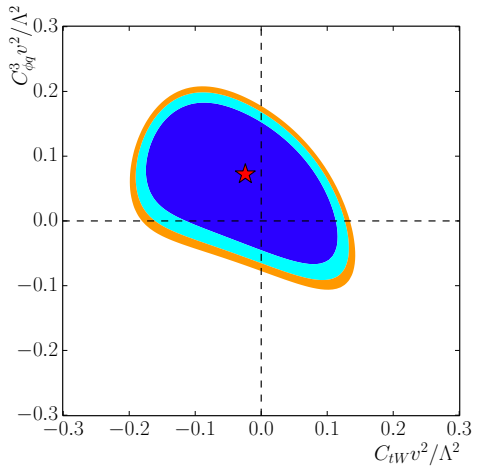
- Motivation cf. PDF Hessian errors.
- *Directions* are robust: physics in the components of principle directions
- But distance along vectors not well-defined. If true χ^2, expect $\Delta \sim$ num params; actually more like num bins for coverage
- Effect of large correlated systematics? Not experimental…but in model??
- Can we define a statistically robust $\Delta\chi^2$ for tunes? Perhaps instead aim for iterated minimal data coverage.
- More robust dimensional reduction wanted / needed? cf. ATLAS A14 procedure – *see later*



Aside: Prof4BSM – life beyond tuning

- Much recent development/use hasn't been for tuning at all…
- Fast parameterisation also finds use in BSM physics, e.g. **arXiv:1506.08845**, arXiv:1511.05170, **arXiv:1512.03360**
- Use parameterisation of observables in Wilson coefficient space to build confidence limit contours
- Speed important for marginalising limits in many dimensions for projections
- Now being worked on via GPGPU hyper-parallel polynomial evaluation





Pre-LHC tuning

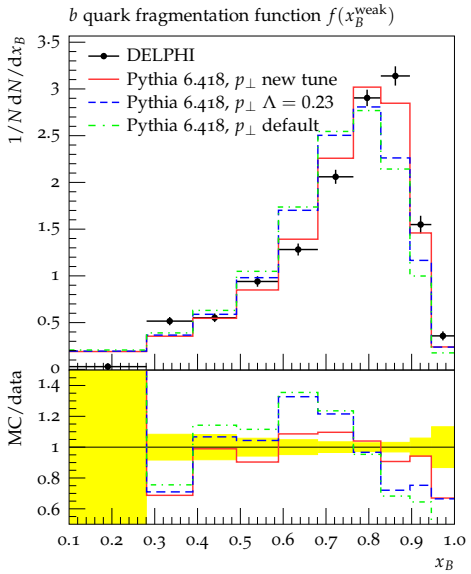
Early hadron collider tunings

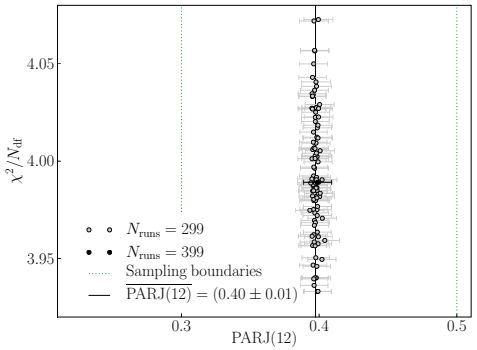
- Circa 2006, only established tuning activity was from Rick Field, on PYTHIA6 vs. Tevatron data.
- Like playing whack-a-mole:
  - Tune A describes $p\bar{p}$ *underlying event* well...
  - ...but destroys Z $p_\perp \Rightarrow$ Tune AW; ...
  - ...which in turn fails on *dijet decorrelation*! $\Rightarrow$ Tune DW. Etc.
- Full observable coverage essential
  - and certainly $N_{\text{df}} \sim N_{\text{bins}} \gg N_{\text{params}}$
- We also get itchy feet about messing too much with perturbative physics: there's meant to be predictivity in there

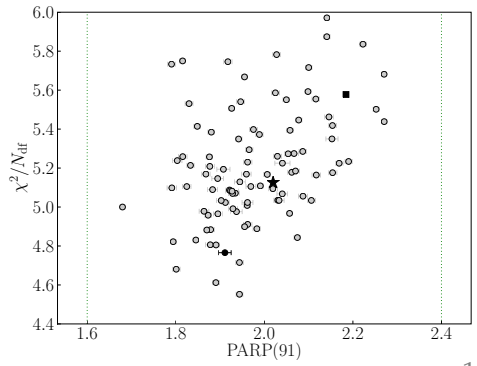
Original Professor tunings

Professor development started also with PYTHIA6:

- **Divide and conquer strategy:** tune to LEP first (*hadronisation, FSR*), then hadron colliders (*ISR, MPI*)
  - Often a good approach. Isolate parameters as far as possible, *assume* hadronisation universality. Here allows $\mathcal{O}(30)$ params to be used in 3–4 ∼orthogonal stages.
  - All input analysis done using **Rivet** (developed in parallel) – still main tuning data source
- More philosophy: do as much as possible by perturbative means, only tune to fill in gaps
- Resulting paper and PY6 Prof-Q^2/pT tunes: **arXiv:0907.2973**







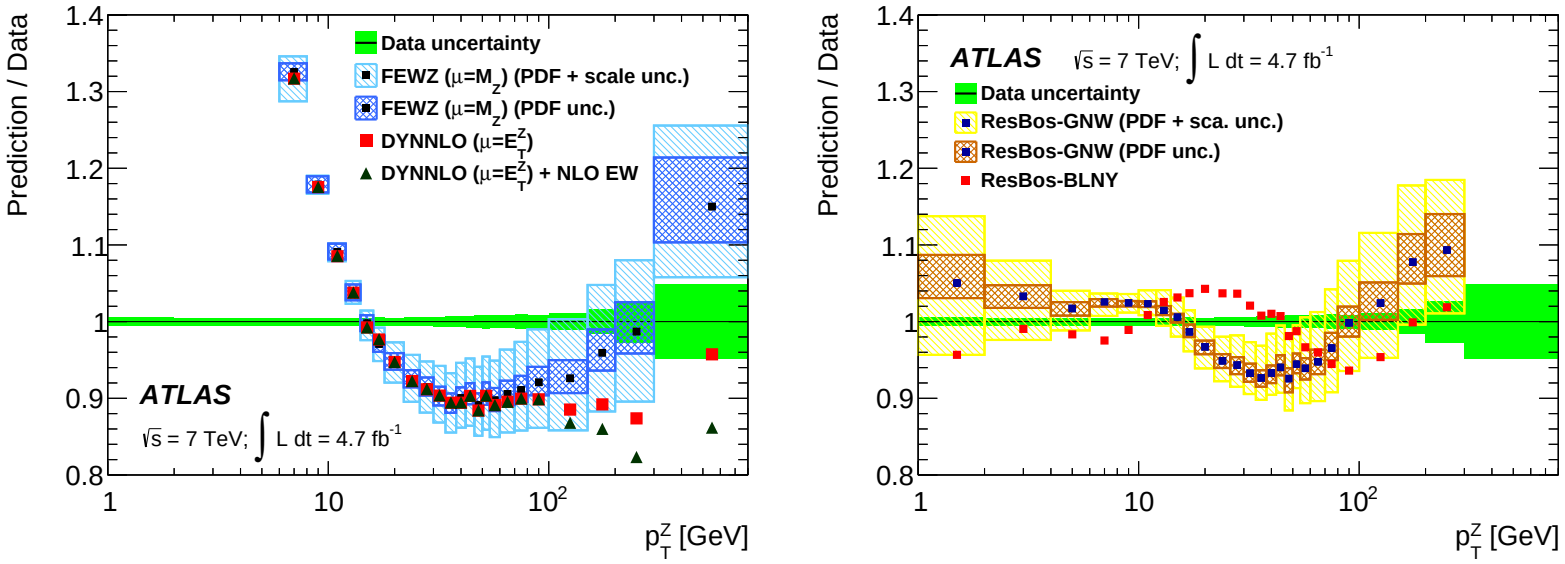
Current ATLAS tunes

ATLAS AZ/AZNLO EW boson production tunes

ATL-PHYS-PUB-2013-017

New 7 TeV Z boson transverse momentum measurement [arXiv:1406.3660] used to develop a tune for W mass measurement: use modelling for unfolding, systematics reduction, etc.

Finite Z (and W) p_{T} wholly generated by recoil against QCD emissions and intrinsic parton p_{T} in the incoming protons.



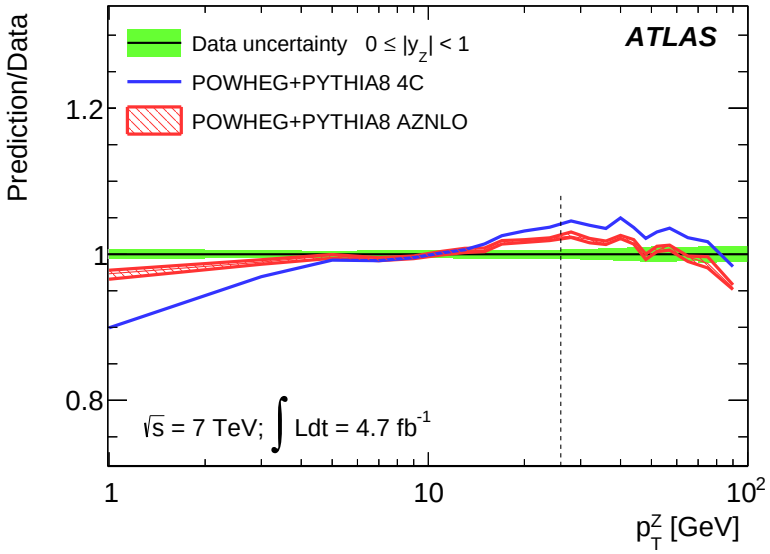
Significant disagreements w.r.t. fixed-order FEWZ, DYNNLO.

ATLAS AZ/AZNLO EW boson production tunes

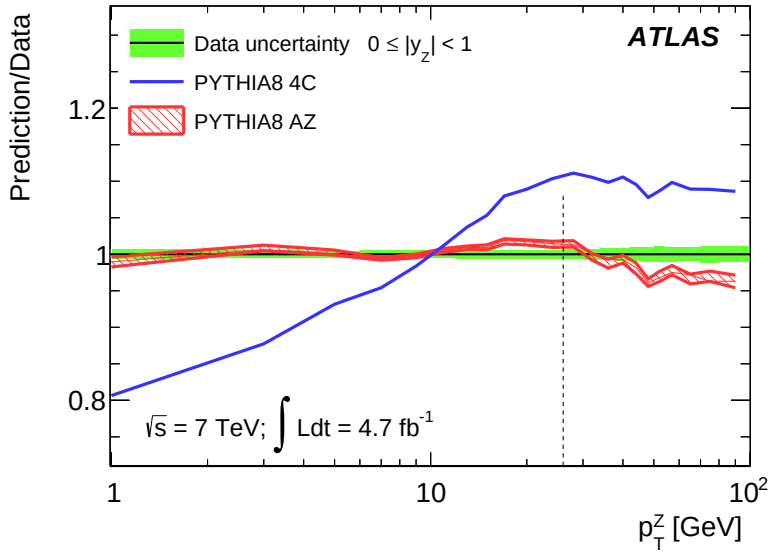
ATL-PHYS-PUB-2013-017

ATLAS (POWHEG+)Pythia8 AZ tunes give few-% description:

POWHEG+Pythia8



Pythia8



Tuning of primordial k_{T}, ISR α_{s} and ISR cutoff cutoff variables.

Less freedom at NLO: good!! But perturbative setup details important. Similar studies now performed for $t\bar{t}/ttH$, etc.

Cautionary tale: operationally useful, but minimal physics content.

ATLAS A14 global tunes

- **General purpose tuning:** need for close data modelling to be used in unfolding & where control regions can't be defined.
  "Data-driven" often just means normalization.
- **A14** is a set of Pythia 8 combined ISR, FSR, and MPI tunes: 500 points of 3M events in *ten* parameter hypercube; interpolate; iterate…
  Start from 'Monash' tune.
- Equivalent tunes to four LO PDFs: CTEQ6L1, MSTW2008LO, NNPDF23LO and HERAPDF15LO. Consistency check / optimisation
- Aimed at covering LO BSM samples at high-$p_\perp$– but has since become ATLAS standard tune for ∼everything! Hmm. Replacement for 2017?

A14 tune observables: all ATLAS Run 1 data

Observable	Weight
Z p_T: ATLAS_2011_S9131140	
Z p_T (< 50 GeV) (20 distributions)	10
Track jet properties: ATLAS_2011_I919017	
Charged jet multiplicity, z, p_T^{rel}, $\rho_{ch}(r)$ (200 distributions)	10
Jet shapes: ATLAS_2011_S8924791	
Jet shape ρ in jet events (59 distributions)	10
Dijet decorr: ATLAS_2011_S8971293	
Decorrelation $\Delta\phi$ (> 0.75) (9 distributions)	20
Multijets: ATLAS_2011_S9128077	
3-to-2 jet ratios (8 distributions)	100
Substructure: ATLAS_2012_I1094564	
Jet mass, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, τ_{21}, τ_{23} (36 distributions)	5
Track-jet UE: ATLAS_2012_I1125575	
Transverse region N_{ch}, mean p_T profiles (8 distributions)	10
Jet UE: ATLAS_2014_I1298811	
Transverse, trans-max, trans-min sum incl. profiles (8 distributions)	20
Transverse sum E_T/sum p_T ratio profiles (2 distributions)	5
Transverse mean p_T incl. profiles (2 distributions)	10
Transverse, trans-max, trans-min incl. distributions (30 distributions)	1
$t\bar{t}$ gap: ATLAS_2012_I1094568	
Gap fraction vs Q_0, different $\lvert y \rvert$ (4 distributions)	100, 80, 40, 10
$t\bar{t}$ jet shapes: ATLAS_2013_I1243871	
Jet shapes $\rho(r)$, $\psi(r)$ (20 distributions)	5

A14 tunes *NB. just to give a feel for procedure…*

Sampled param ranges and fitted values:

Param	Min	Max	CTEQ	MSTW	NNPDF	HERA
SigmaProcess:alphaSvalue	0.12	0.15	0.144	0.140	0.140	0.141
BeamRemnants:primordialKThard	1.5	2.0	1.72	1.82	1.88	1.83
SpaceShower:pT0Ref	0.75	2.5	1.30	1.62	1.56	1.61
SpaceShower:pTmaxFudge	0.5	1.5	0.95	0.92	0.91	0.95
SpaceShower:pTdampFudge	1.0	1.5	1.21	1.14	1.05	1.10
SpaceShower:alphaSvalue	0.10	0.15	0.125	0.129	0.127	0.128
TimeShower:alphaSvalue	0.10	0.15	0.126	0.129	0.127	0.130
MultipartonInteractions:pT0Ref	1.5	3.0	1.98	2.22	2.09	2.14
MultipartonInteractions:alphaSvalue	0.10	0.15	0.118	0.127	0.126	0.123
BeamRemnants:reconnectRange	1.0	10.0	2.08	1.87	1.71	1.78

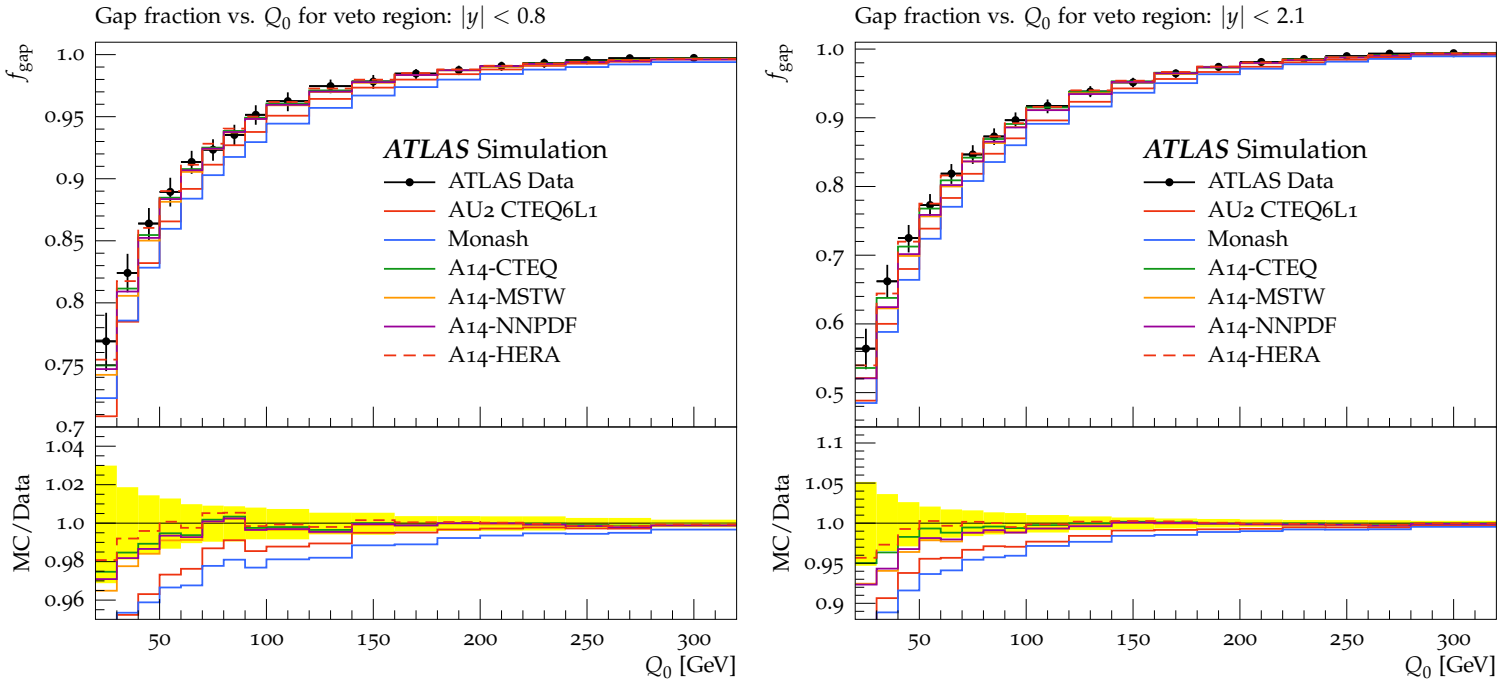
Remarkable consistency, particularly between ISR and FSR α_s! Also indicates tune stability. Shower $\alpha_s \to$ quite low; hard process $\alpha_s \to$ quite high.

Colour reconnection also quite robust: similar to Monash = 1.80, a little bigger than 4C = 1.5. Important for e.g. $t\bar{t}$/exotic colour flow studies.

Comparisons

$t\bar{t}$ gap fractions

Fraction of $t\bar{t}$ events with a $|y|$ gap *not* disrupted by a jet of $p_T \geq Q_0$

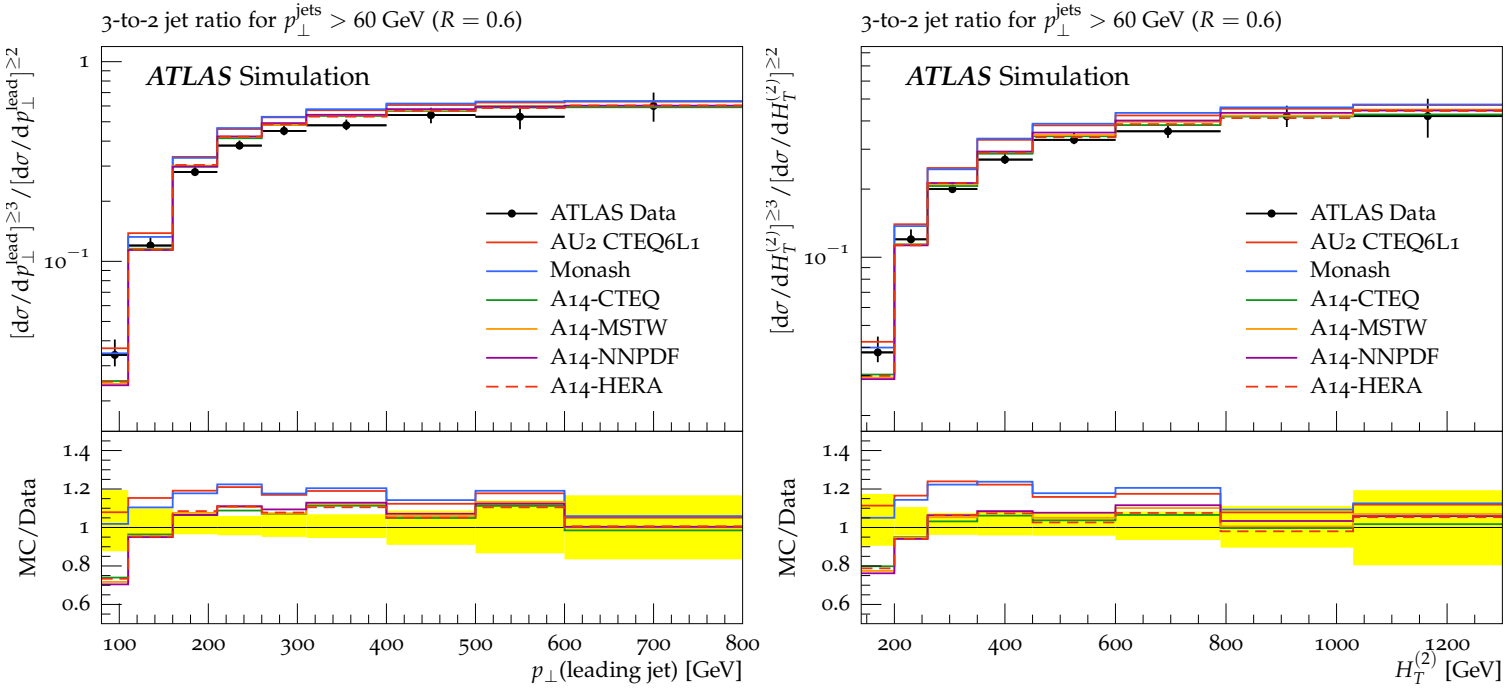


"Damped shower key to improvement, otherwise in horrible tension with other ISR observables incl. Z p_T" – **physics insight!**

Comparisons

3/2 jet ratios (3rd jet from shower)

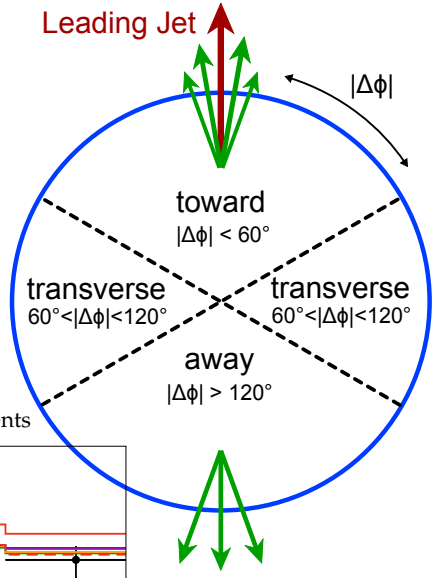
New tunes improve 3rd jet data description except at low scale.



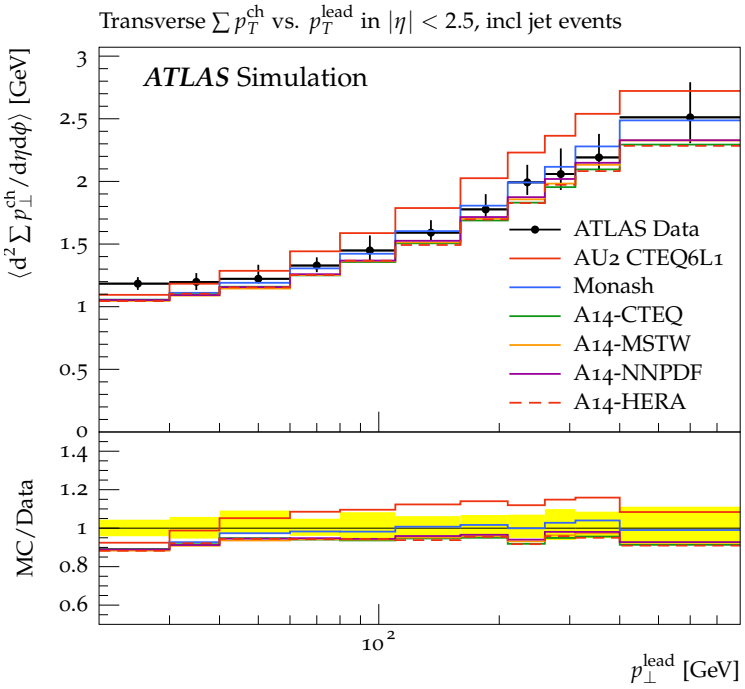
Focus of these tunes is where no NLO/multileg simulation available, e.g. BSM processes at high scales

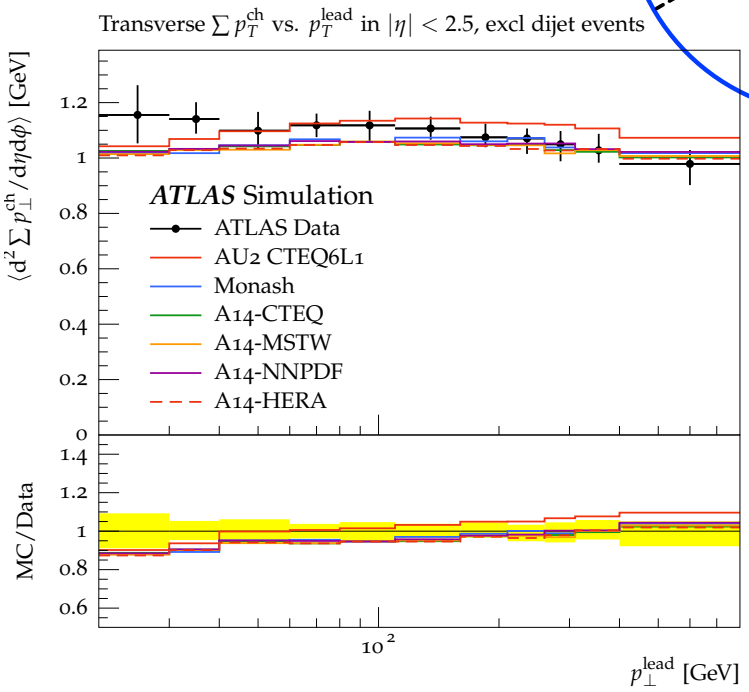
Comparisons

Jet underlying event



Standard UE w.r.t leading jet: new result incl. exclusive dijet selection

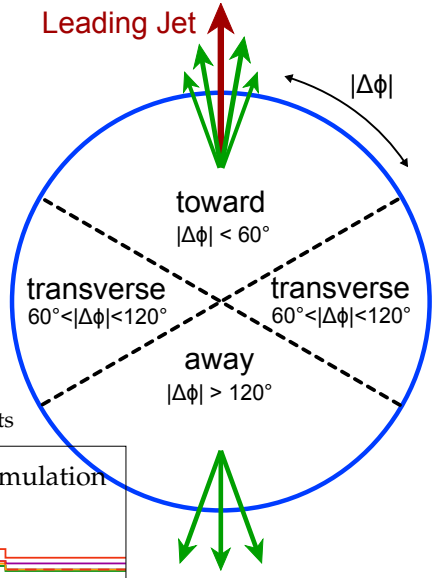




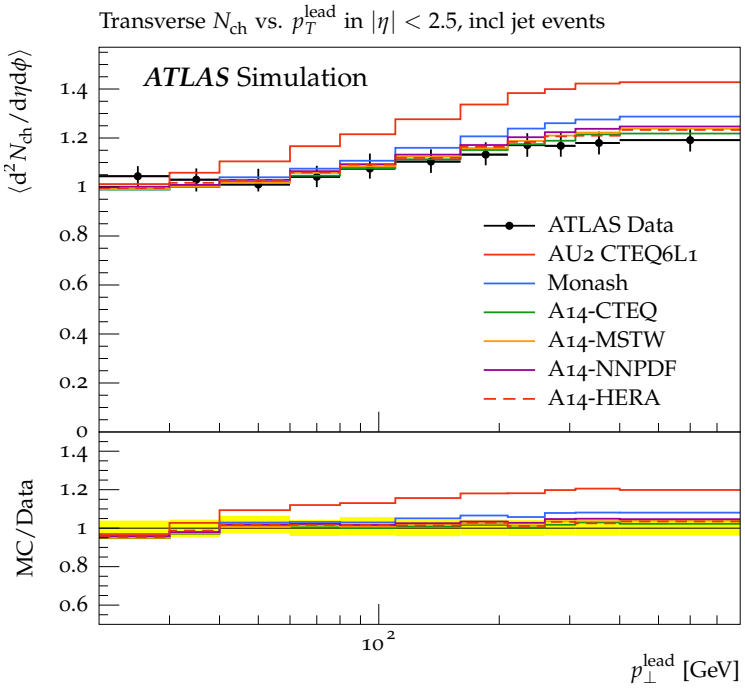
Undershoots in lowest-p_T^{lead} bins – not the intended application area.
Note difficulty of tuning to observables in which model components are highly coupled

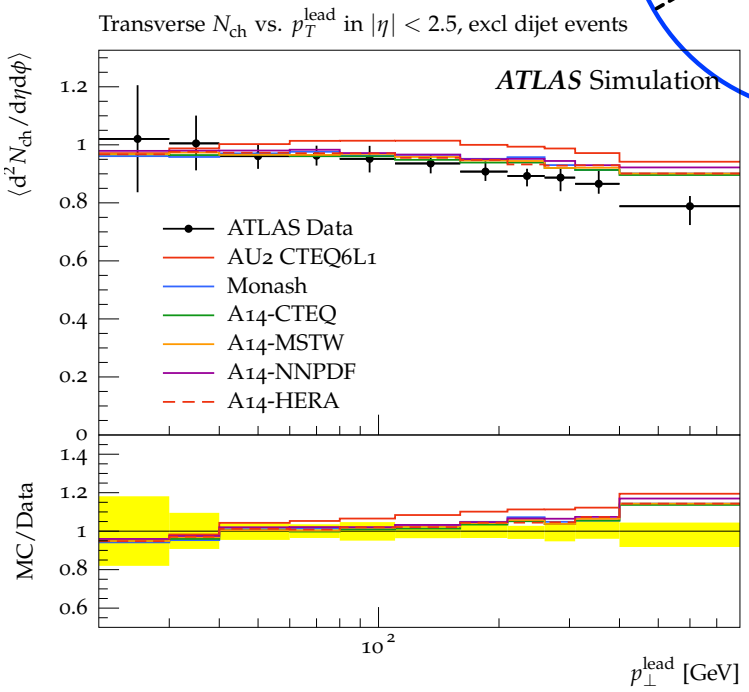
Comparisons

Jet underlying event



Standard UE w.r.t leading jet: new result incl. exclusive dijet selection

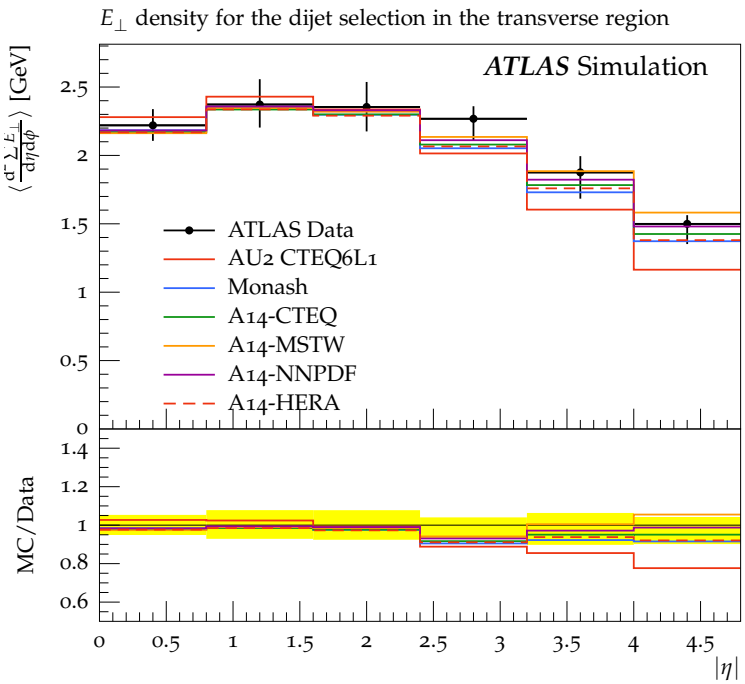


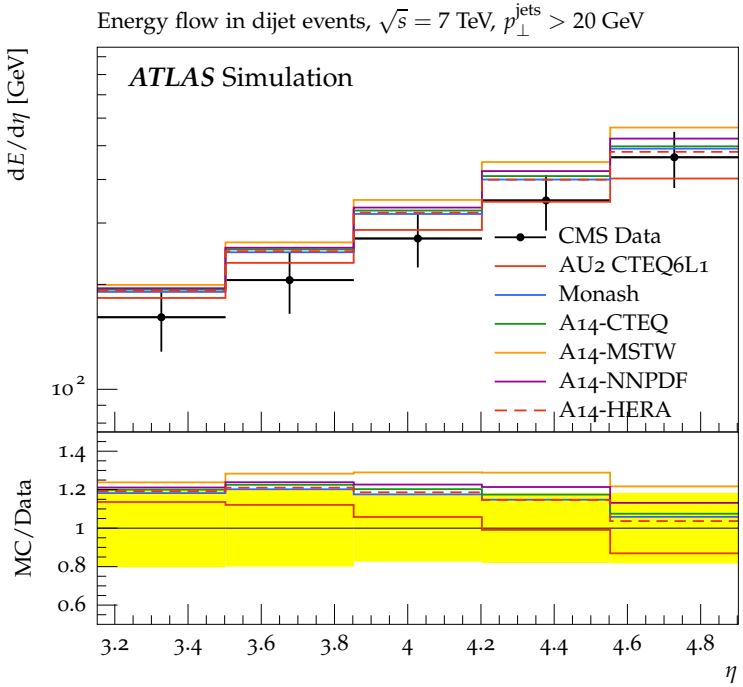


Undershoots in lowest-p_T^{lead} bins – not the intended application area.
Note difficulty of tuning to observables in which model components are highly coupled

Comparisons

Forward energy flow (ATLAS vs CMS)





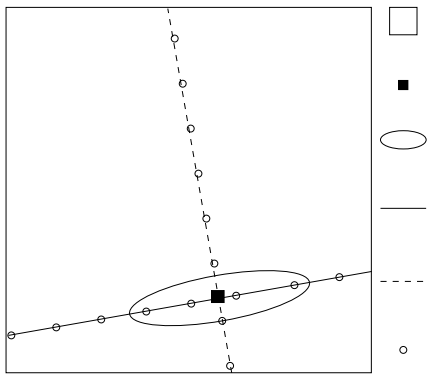
Slight ATLAS/CMS tension (just $\sim 1\sigma$ systematic level)

Eigentune constructions for A14

If tuning for model-based unfolding or extrapolation ☠ also need an estimate of model/tune uncertainty

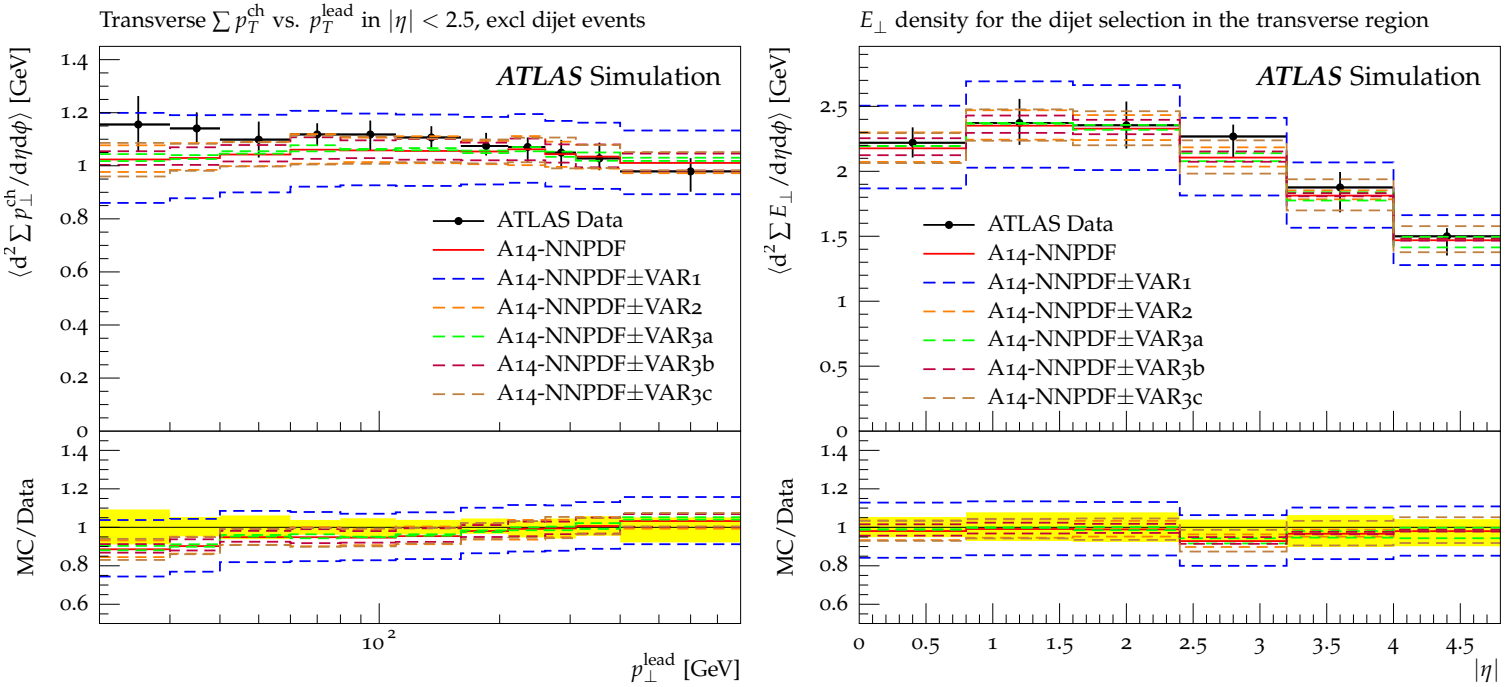
- **Eigentunes:** *systematic variations on the parameters of the A14-NNPDF tune, using the principle directions of the covariance matrix around the tune minimum and making excursions to find a fixed* $\Delta\chi^2$.
- Tune $\chi^2 = 3.44 \times 10^5$ for $N_{\mathrm{df}} = 80315 \Rightarrow \chi^2/N_{\mathrm{df}} = 4.29$. NB. Systematics correlations not included. Used $\Delta\chi^2 = N_{\mathrm{df}}/2 \Rightarrow$ obtain *correlated* $\sim 1\,\sigma$ variation bands. **Ad hoc!**
- Combined FSR, ISR + MPI tune has ten parameters $\Rightarrow$ 20 eigentunes

  No-one would use them!
- Manually reduced to a tractable covering set of 5×2 variations covering mainly UE, jet structure, and $3\times$ extra jet effects. Can be reduced depending on analysis.



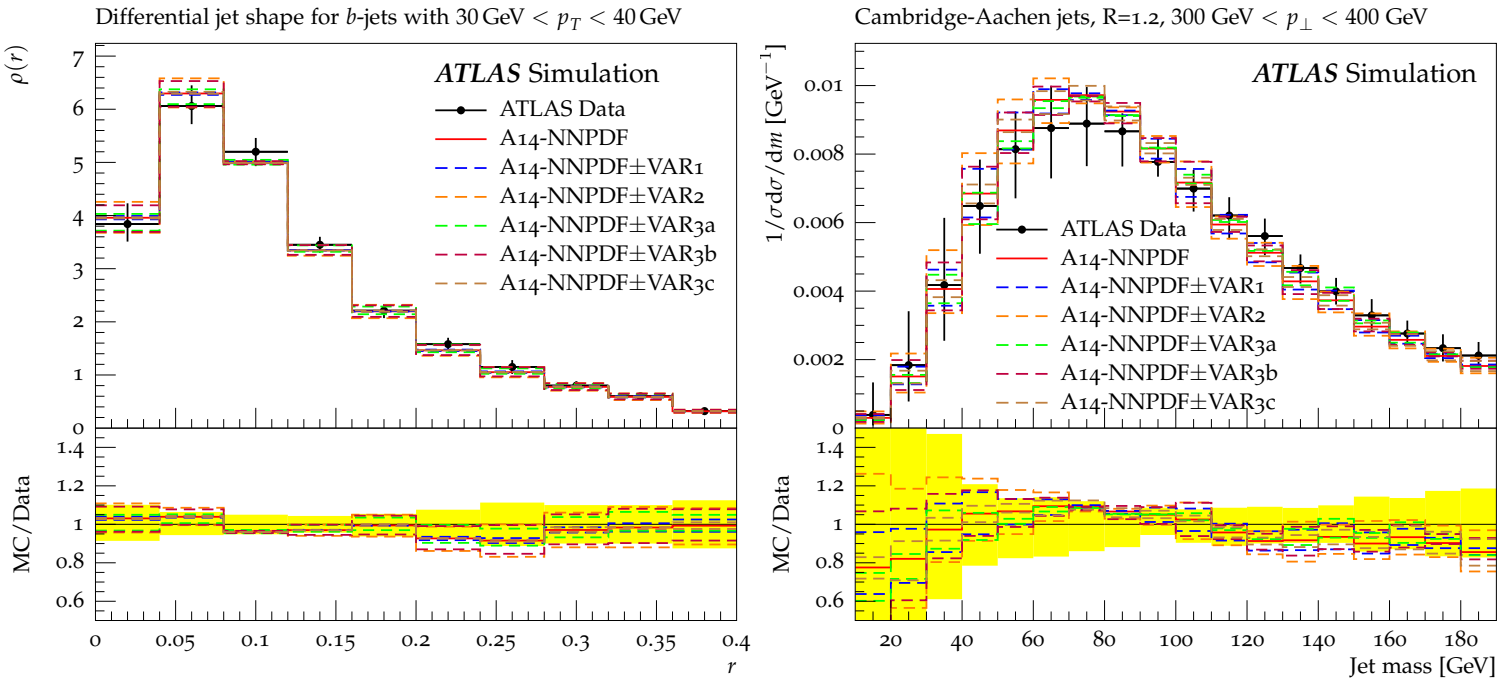
Eigentune bands

Underlying event



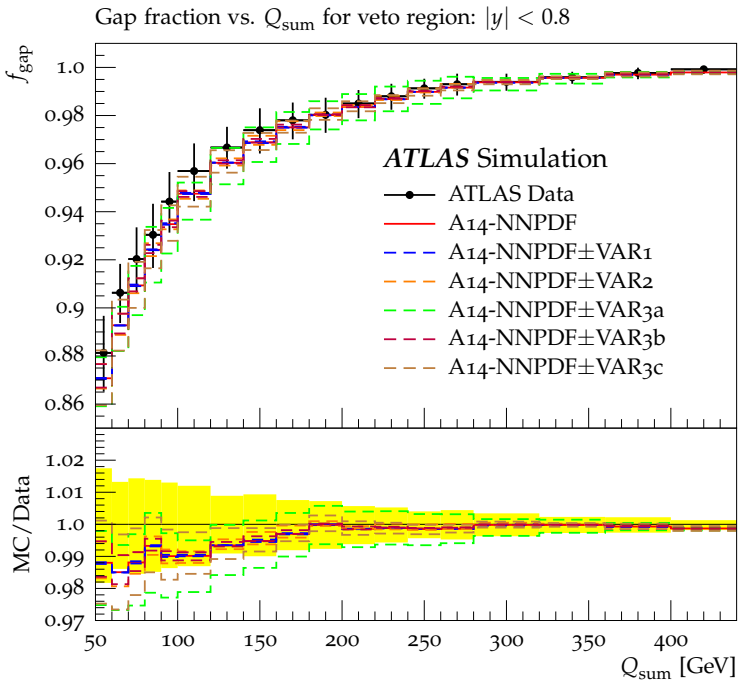
Eigentune bands

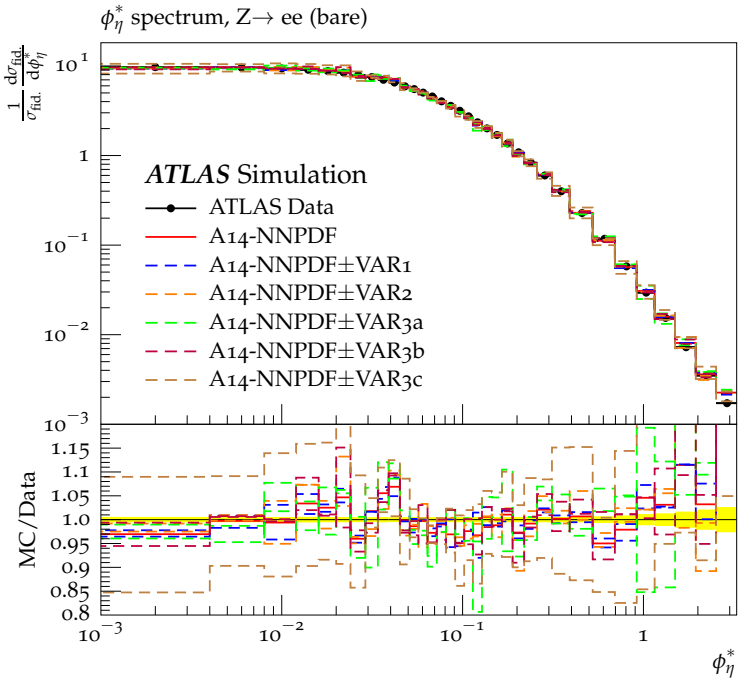
Jet structure



Eigentune bands

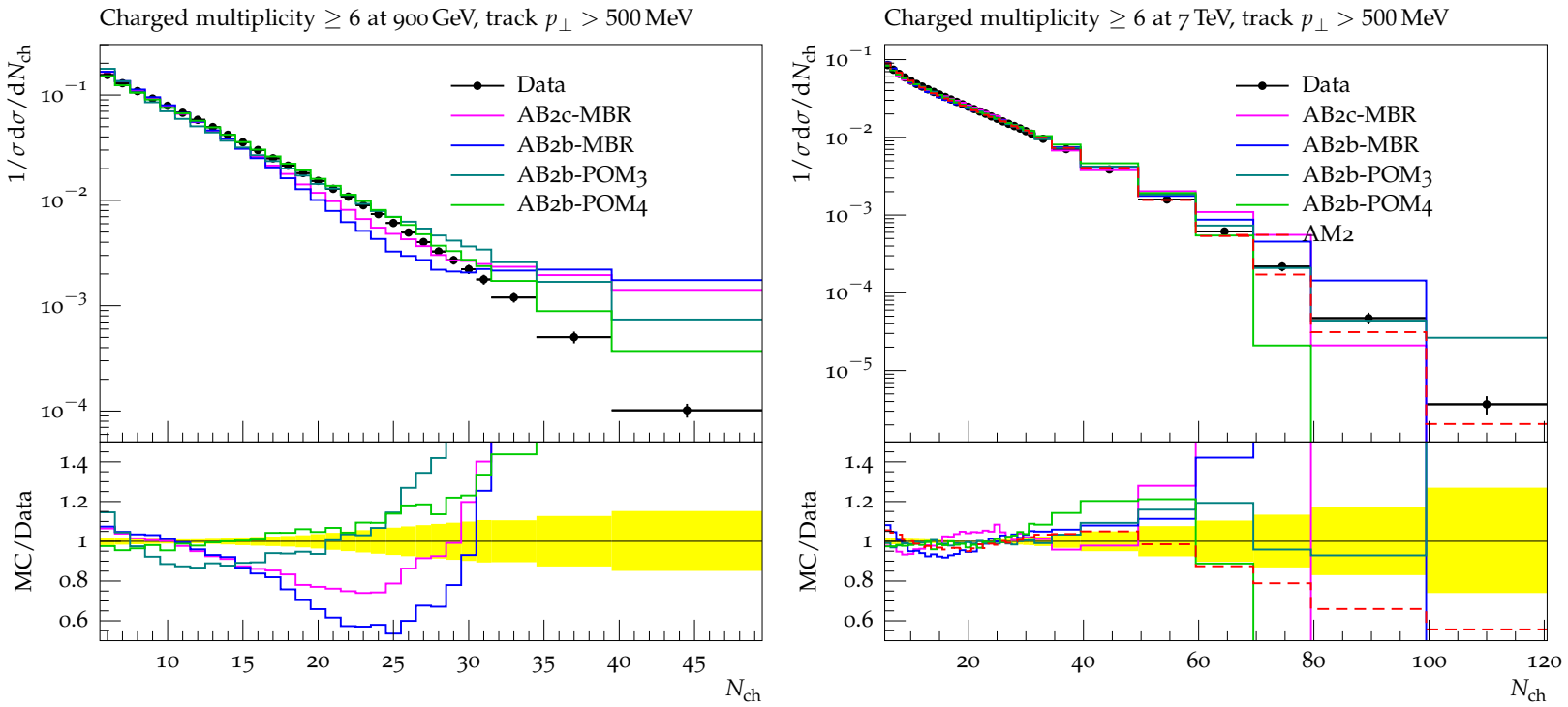
Extra jets and recoils





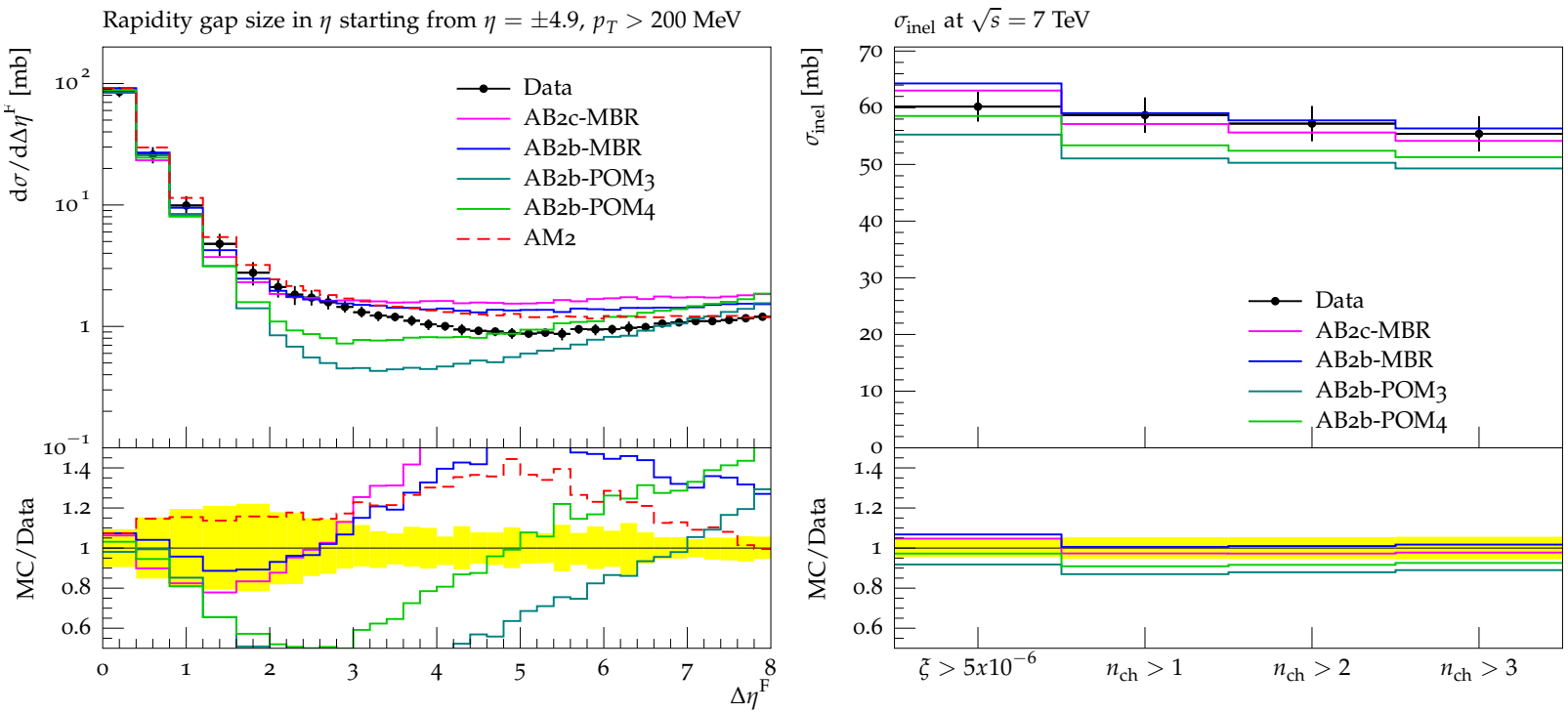
Minimum bias / diffraction tuning

Charged multiplicities at different $\sqrt{s}$



Sometimes there's no good model available. Tuning can be the only way to find that out for sure

Diffraction and total $\sigma_{\text{inelastic}}$



Sometimes there's no good model available. Tuning can be the only way to find that out for sure

Still need to get something imperfect but usable! e.g. total σ not that important for LHC operational purposes (i.e. pile-up MC)

Conclusions

- **Obviously this is incomplete!** Hope I gave some idea of the (still developing) methodology, areas of application, and issues/concerns
- Many good tunes, e.g. Perugia, Monash, without Professor. But they required more effort, and don't cover so systematically: machinery is not all, but it helps. Brains still useful.
- **Statistically hard to define exactly what's wanted:**
  - Some distributions more important (per bin). Maybe the tails are what's important...
  - But *how much more important?* Weights $\Rightarrow$ subjectivity
  - Importance of correlations, given subjectivity & heuristics elsewhere?
  - And what about uncertainty coverage when the "truth" does not lie in the model space?
  - The point is usually not statistical fairness or model testing, but finding a imperfect-but-useful operating point
- **And what about uncertainty coverage *when the "truth" does not lie in the model space?***
  - **The point is usually not statistical fairness or model testing, but finding a imperfect-but-useful operating point**