TECHNISCHE UNIVERSITÄT DRESDEN
Fakultät Physik Institut für Kern- und Teilchenphysik

Monte Carlo modelling of signals and backgrounds

Frank Siegert

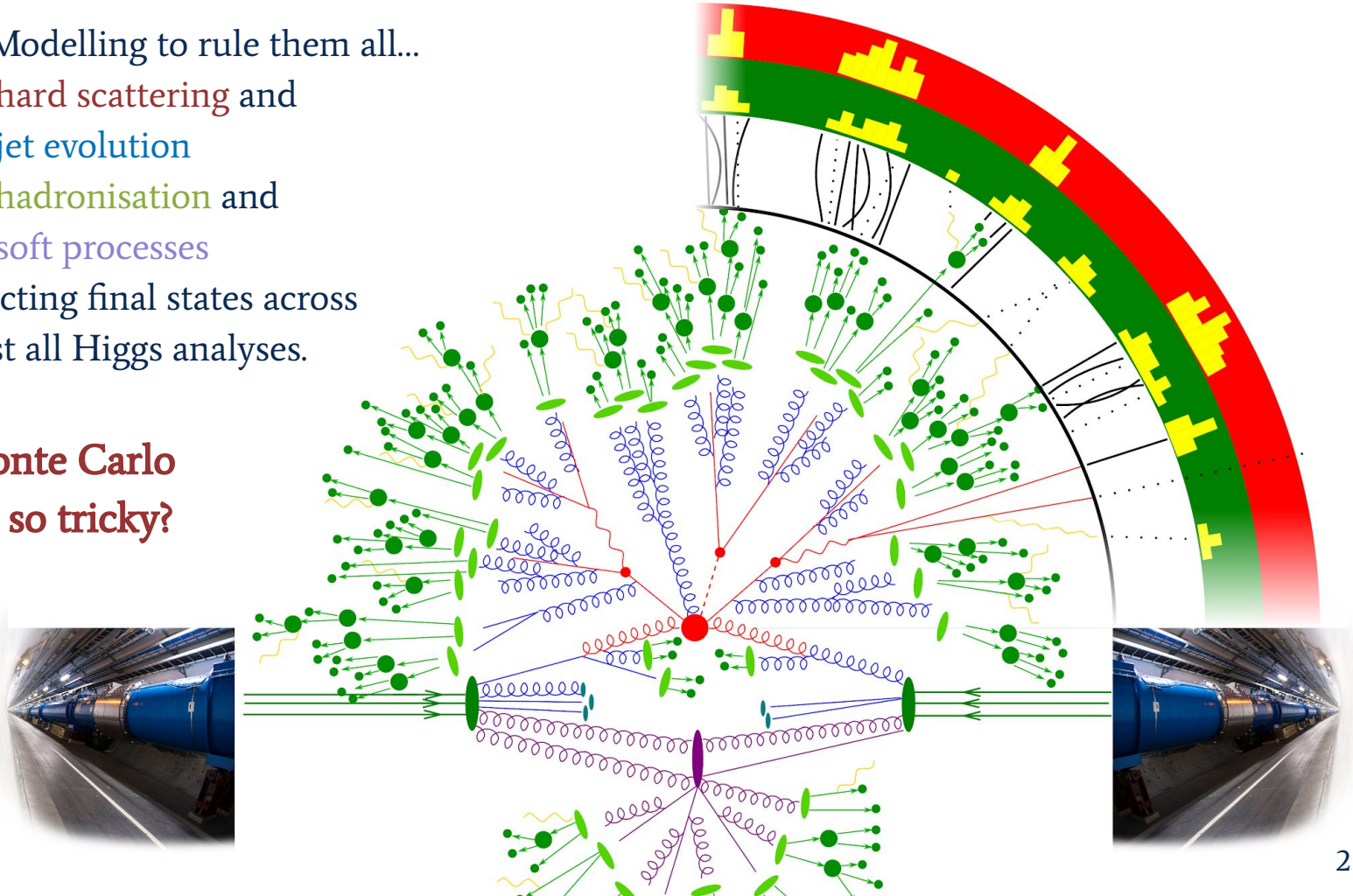
Higgs2021
18-22 October 2021

One Modelling to rule them all...

- hard scattering and jet evolution
- hadronisation and soft processes

... affecting final states across almost all Higgs analyses.

Why is Monte Carlo modelling so tricky?



The art (or ambiguities) of constructing a PS MC

- A parton-shower Monte Carlo is not a fixed-order prediction
  - It is much more powerful!
  - And at the same time much more ambiguous!

Ambiguities = Uncertainties

(and in addition there can be bugs of course)

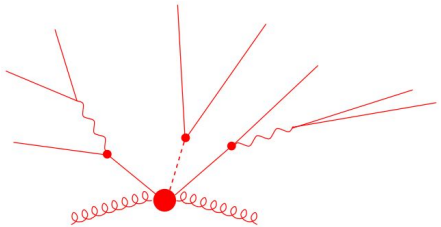
Let's review them briefly ...

The art (or ambiguities) of constructing a PS MC

- A parton-shower Monte Carlo is not a fixed-order prediction
  - It is much more powerful!
  - And at the same time much more ambiguous!
- Typical sources of ~~trouble~~ ambiguities:

The art (or ambiguities) of constructing a PS MC

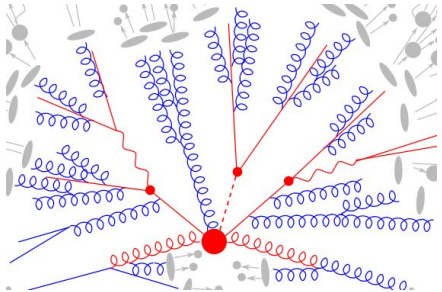
- A parton-shower Monte Carlo is not a fixed-order prediction
  - **It is much more powerful!**
  - **And at the same time much more ambiguous!**
- Typical sources of ~~trouble~~ **ambiguities**:
  - Hard scattering
    - » Limited perturbative accuracy
      ⇒ **ambiguity in scale and PDF choices**
    - » Factorised decays and narrow-width approximation
      - Spin correlations between production and decay MEs in the chain
        ⇒ **ambiguity in polarisation treatment**
        - particularly tricky for tau decays, as they can be hadronic!
      - Diagram overlap (e.g. tt and tWb)
        ⇒ **ambiguity in overlap removal**
    - » NLO EW corrections
      ⇒ **ambiguity in combination between NLO QCD and NLO EW**
    - » Multi-leg merging of ME & PS
      ⇒ **ambiguity in transition**



The art (or ambiguities) of constructing a PS MC

- A parton-shower Monte Carlo is not a fixed-order prediction
  - **It is much more powerful!**
  - **And at the same time much more ambiguous!**
- Typical sources of ~~trouble~~ **ambiguities**:
  - Parton shower: QCD corrections with three major ambiguities
    - » Functional form of splitting kernels
      (approximation of real-emission MEs)
      ⇒ **ambiguity which (finite) pieces to keep**
    - » Kinematics recoil
      (how to construct 1→2 splittings with m=0 away from collinear limit)
      ⇒ **ambiguity where to distribute recoil for momentum conservation**
    - » Evolution variable
      (direction in which logs are resummed)
      ⇒ **ambiguity what "from hard to soft" means exactly**

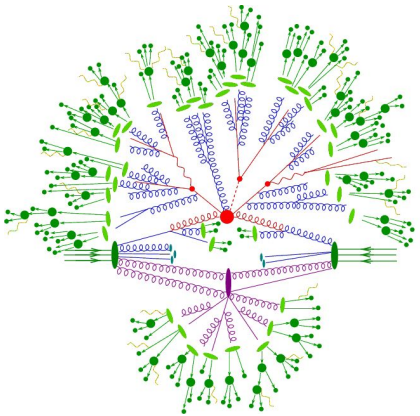
Simon's talk



Additionally many ambiguities for treatment of quark masses in the above!

The art (or ambiguities) of constructing a PS MC

- A parton-shower Monte Carlo is not a fixed-order prediction
  - **It is much more powerful!**
  - **And at the same time much more ambiguous!**
- Typical sources of ~~trouble~~ **ambiguities**:
  - Hadronisation: Soft QCD modelling without “first principles”
    - » B-hadron production from partons
      ⇒ **ambiguity of flavours formed** (e.g. meson or baryon, B* or B, ...)
  - Hadron decays: Effective field theories for heavy-flavour decays
    - » B-hadron decays
      ⇒ **ambiguity of decay matrix elements** (form factor models)



The art (or ambiguities) of constructing a PS MC

- A parton-shower Monte Carlo is not a fixed-order prediction
  - **It is much more powerful!**
  - **And at the same time much more ambiguous!**
- Typical sources of ~~trouble~~ **ambiguities**:

Let's look at recent developments for some of the modelling issues from Adinda+Nicolas ...

- Background modelling
- Signal modelling
- Statistics and practicalness

$t\bar{t}$ modelling
- Good modelling of bulk of phase space by th after tuning
  - Though sizable discrepancies remain
- Difficulty: uncertainties in tails / corners of ph
  - Not easy to get enough MC statistics:
    - filtering / slicing strategies

W/Z+$b\bar{b}$ largest bkgs in VHbb search
Difficulty: generate enough MC events in relevant phase space (high pT(V)), filtered for W/Z+hf

ATLAS: Select function, and estimate maxir bias 'spurious signal'
- Requires vast amounts of MC events
- Limitation for high luminosity

Statistics and practicalness:

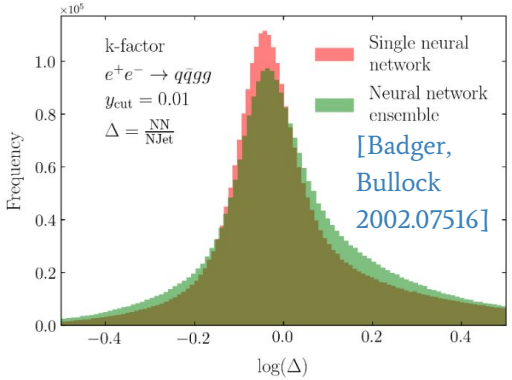
data
Even if it does - would need very large samples to avoid large MC statistical uncertainties
Hybrid solution: Embedding

Countless CPU hours required for MC generation
Filters (in)efficiency, spread of MC weights

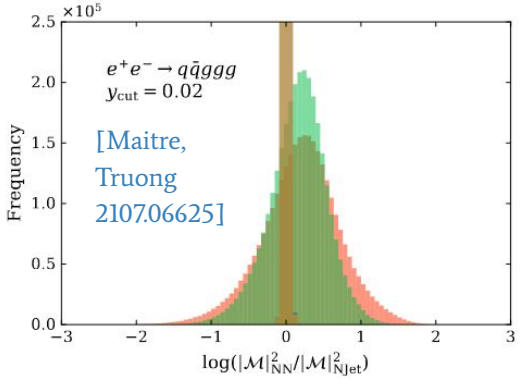
MC stat noise in uncertainty evaluation smoothed by use of ML techniques for n-dim reweighting

Fast: Machine learning for matrix elements

- Boom in ML techniques has also met the Monte Carlo landscape [review]
- Most relevant in the context of efficiency: Matrix elements!
  - Many surrogate models on the market → fast, but how accurate?

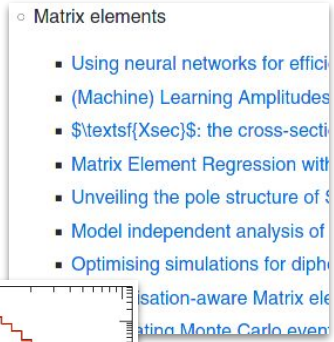


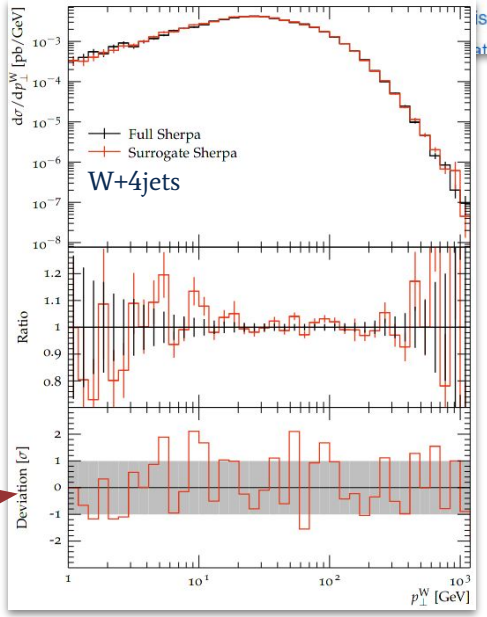
[Badger, Bullock 2002.07516]



[Maitre, Truong 2107.06625]

- Main question: How to embed them faithfully into Monte Carlo event generators?
  - Novel unweighting based on surrogates: faithful!
    [Danziger, Janßen, Schumann, FS 2109.11964]



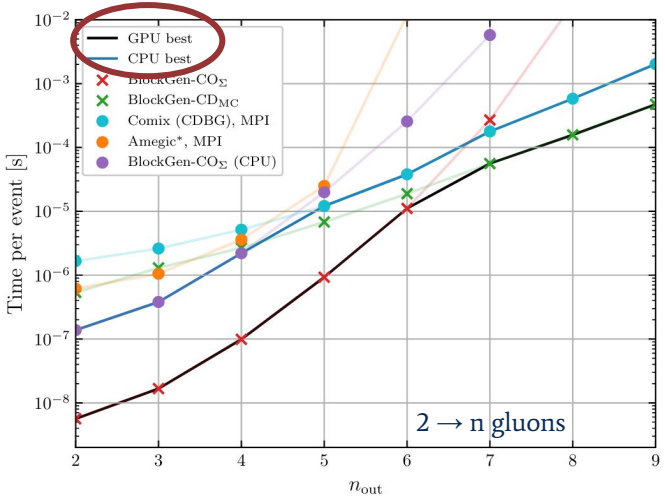


Fast: GPU computing

- June 2021: The month of matrix elements on GPUs:

Sherpa/BlockGen [Bothmann et al 2106.06507]

- Automated ME construction with Berends-Giele recursion
- Cross-platform for GPU/CPU (Kokkos)



MadGraph5_aMC

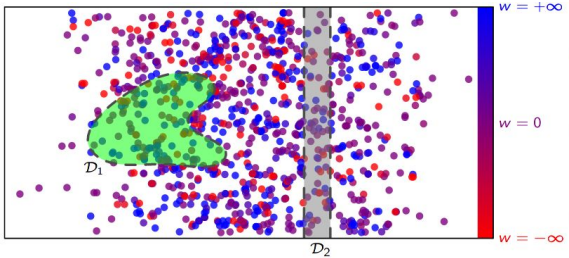
- MadFlow [Carrazza et al 2106.10279]
  - Cross-platform (TensorFlow) framework for GPU MEs
- MG4GPU [Valassi et al 2106.12631]
  - Converts process code from Fortran to GPU, aiming for automation
- Phase space with Rambo@GPU

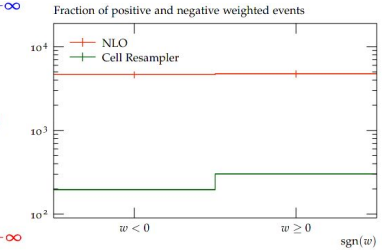
Process	MadFlow CPU	MadFlow GPU	MG5_aMC
$gg \to t\bar{t}$	9.86 μs	1.56 μs	20.21 μs
$pp \to t\bar{t}$	14.99 μs	2.20 μs	45.74 μs
$pp \to t\bar{t}g$	57.84 μs	7.54 μs	93.23 μs
$pp \to t\bar{t}gg$	559.67 μs	121.05 μs	793.92 μs

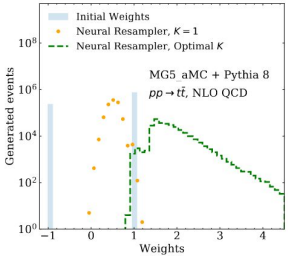
Particularly important for utilisation of HPC resources!

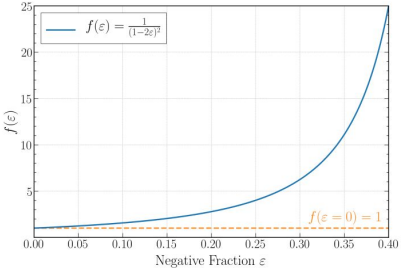
Favourable: Negative weight reduction

- Fraction ε of negative weights reduces sample size by factor $(1-2\varepsilon)^2$
- Two recent directions of improvements:
  - resampling methods [Andersen et al 2005.09375, 2109.07851], [Nachman et al 2007.11586]
    - » a posteriori combination of "close" events

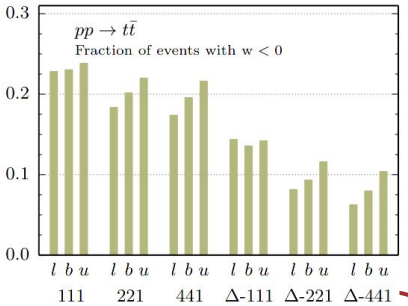


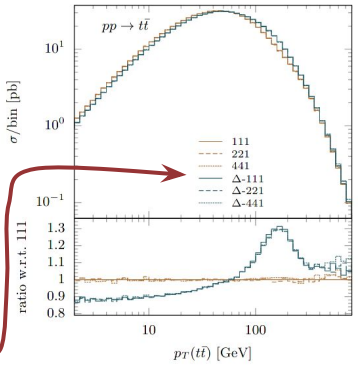


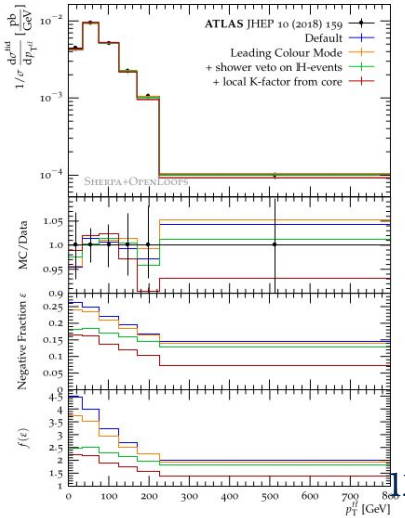




  - event generator improvements [Frederix et al 2002.12716], [Danziger et al 2110.xxxxx]
    - » a priori reduction of negative weights during event generation
    - » modification of NLO+PS matching or multi-leg merging
    - » sometimes also affect physical distributions



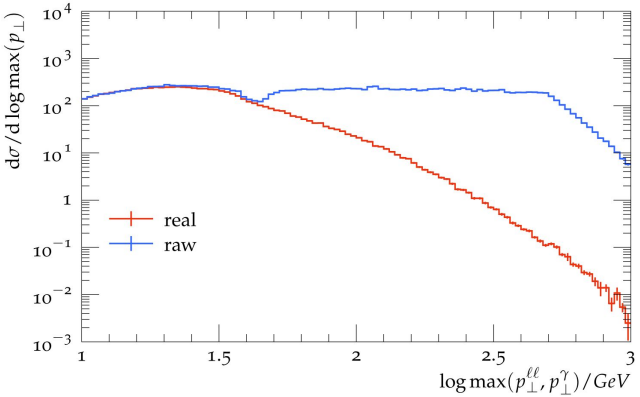




Filtered: Phase space biasing

- Phase space biasing
  - Experimental analyses often use sliced event samples to populate rare phase space
  - Slicing not always practical:
    - » Selection at hadron level inefficient and slow
    - » Migrations from low-p_T to high-p_T due to shower/UE/had
      → spikes due to low luminosity in low-p_T slices
    - » Non-continuous stat. unc. at slice boundaries
  - Alternative starting to be used/explored:
    Continuos phase-space biasing
    - » Effectively modify Monte Carlo integrand
    - » Correct event weight for real distribution after unweighting!

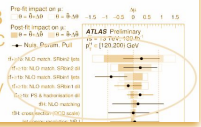
Example: biased llγ with Sherpa



- Heavy flavour filtering remains challenging
  - LCG workflows adapted to produce filtered evgen samples slightly more efficiently
    - » e.g. V+b, V+c, V+light from same evgen stream
  - HF Fusing (→ later) could mitigate this (at least for b) due to separate “direct” component
    - » “fragmentation” component still needs filtering/enhancement

Textbook example: $t\bar{t}b\bar{b}$, for ttHbb

- $t\bar{t}b\bar{b}$ dominant bkg and low S/B
  - Complex process to model by MC
- Very large theory uncertainty
  - Cross-section well constrained by profiling, measured ~1.3x



VHbb: W/Z+hf backgrounds

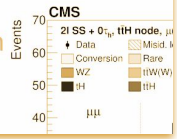
- W/Z+$b\bar{b}$ largest bkgs in VHbb search

Background modelling:

- Heavy Flavour
- tt+V
- Embedding

ttH in multilepton final states: ttW/ttZ

- ttH ML: complex final states with many bkgs
- ttW/ttZ leading ones



Resonant backgrounds - embedding

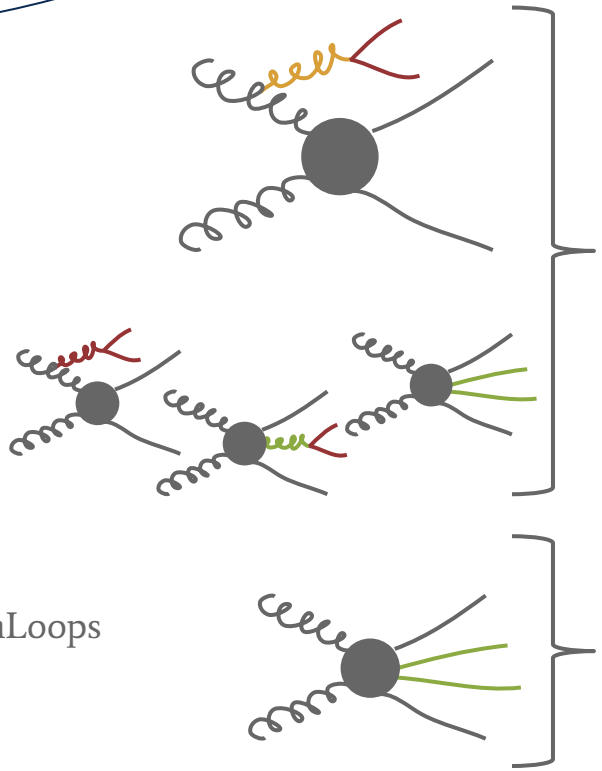
- E.g. Z boson decays in fermionic channels

Heavy-flavour associated final states

Traditional approaches for X+b(b) MC predictions:

X = Z, W, tt, ...

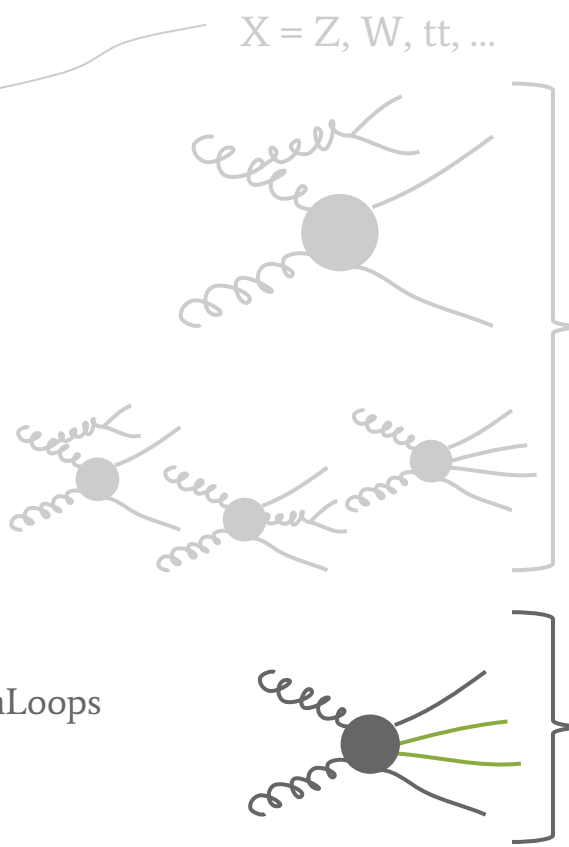
- "Inclusive" NLO+PS sample with HF production from **parton shower g→bb**
  - e.g. {Powheg,aMC@NLO}+{Pythia,Herwig}
- Multi-leg merged sample with HF from **higher-order MEs** (hard b's) or **parton shower g→bb** (soft/collinear b's)
  - e.g. MG5_aMC+Pythia, Sherpa+OpenLoops
- NLO+PS Xbb using **matrix elements** with **massive** b-quarks
  - e.g. Powheg+OpenLoops+Pythia8, Sherpa+OpenLoops



Heavy-flavour associated final states

Traditional approaches for X+b(b) MC predictions:

- "Inclusive" NLO+PS sample with HF production from **parton shower g→bb**
  - e.g. {Powheg,aMC@NLO}+{Pythia,Herwig}
- Multi-leg merged sample with HF from **higher-order MEs** (hard b's) or **parton shower g→bb** (soft/collinear b's)
  - e.g. MG5_aMC+Pythia, Sherpa+OpenLoops
- NLO+PS Xbb using **matrix elements** with **massive** b-quarks
  - e.g. Powheg+OpenLoops+Pythia8, Sherpa+OpenLoops



Anatomy of 4FS NLO+PS for ttbb

- 2→4 NLO QCD matrix elements with massive b-quarks

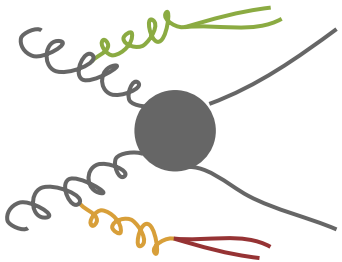
Final state g→bb **dominant**

- massive b’s → no (jet) cuts!
- collinear g→bb produced in ME

No initial state b in 4FS PDFs

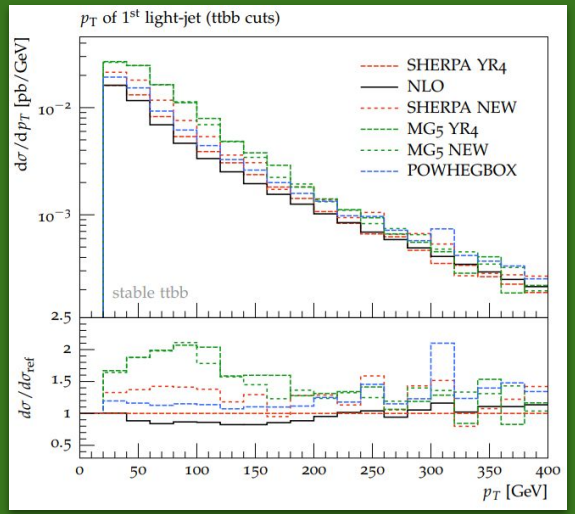
- IS g→bb in ME

- Matched to parton shower for additional emissions
  - **“double-splitting”** contribution becomes relevant!



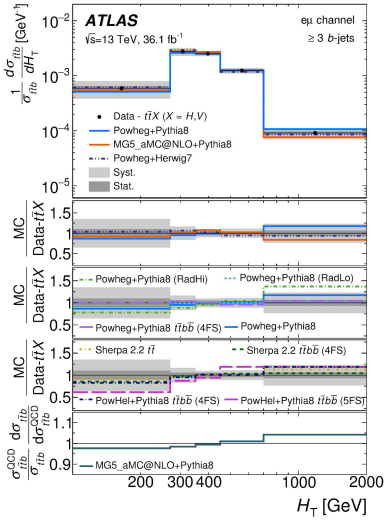
Anatomy of 4FS NLO+PS for ttbb

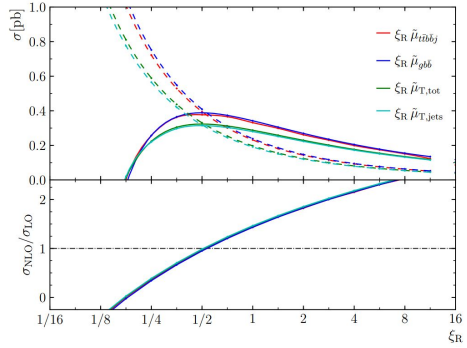
- History:
  Large discrepancies
  in NLO+PS programs!
  - Improve or accept as unc's?
- Arguably one of the most challenging processes for NLO+PS matching
  - Strong interest to understand unc's as prototype for other processes!



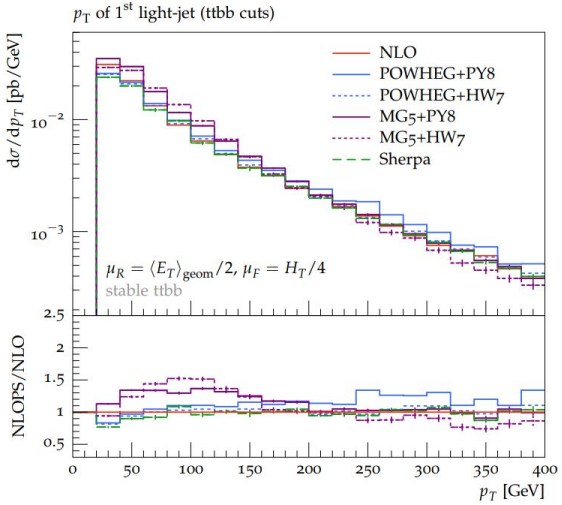
4FS NLO+PS for ttbb

- New inputs:
  - Experimental data?
    Not yet precise enough to discriminate
  - Fixed-order studies of ttbbj@NLO with OpenLoops2+Sherpa
    [Buccioni, Kallweit, Pozzorini, Zoller 1907.13624]
    - » Reduced μ_R stabilises K-factor





- Application of reduced scale to tuned NLO+PS comparisons
  [Garzelli, Jezo, Kardos, Pozzorini, Reuschle, FS, Zaro, ...]
  - Improved agreement between NLO+PS tools
  - Still sizable O(40%) differences in N_{2b} region → origin?



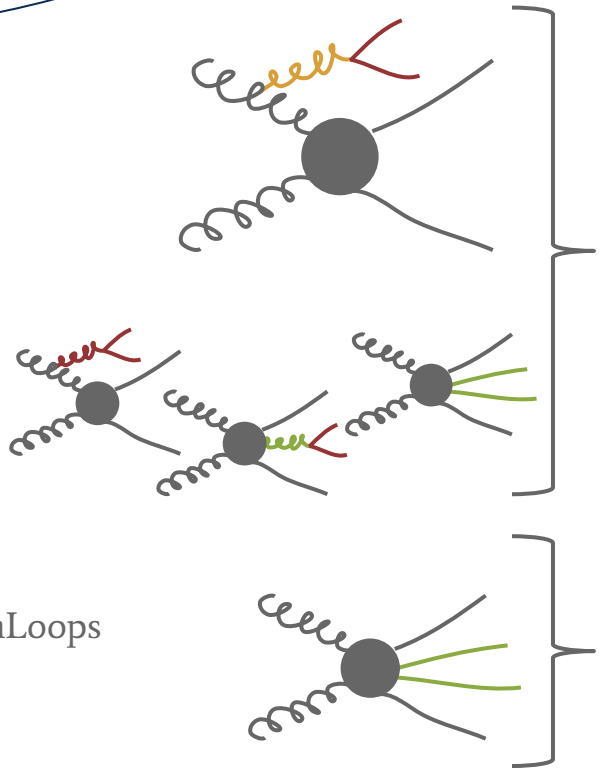
- Aim within ttH/tH subgroup:
  - Final recommendation at LHCHWG General Meeting (Dec)
  - WG note in preparation

Heavy-flavour associated final states

Traditional approaches for X+b(b) MC predictions:

X = Z, W, tt, ...

- "Inclusive" NLO+PS sample with HF production from **parton shower g→bb**
  - e.g. {Powheg,aMC@NLO}+{Pythia,Herwig}
- Multi-leg merged sample with HF from **higher-order MEs** (hard b's) or **parton shower g→bb** (soft/collinear b's)
  - e.g. MG5_aMC+Pythia, Sherpa+OpenLoops
- NLO+PS Xbb using **matrix elements** with **massive** b-quarks
  - e.g. Powheg+OpenLoops+Pythia8, Sherpa+OpenLoops



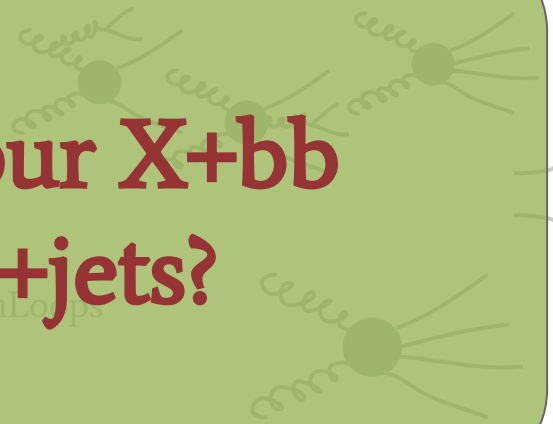
Heavy-flavour associated final states

Traditional approaches for X+b(b) MC predictions:

- "Inclusive" NLO+PS sample with HF production from **parton shower g→bb**
  - e.g. {Powheg,aMC@NLO}+{Pythia,Herwig}
- Multi-leg merged sample with HF from **higher-order MEs** (hard b's) or **parton shower g→bb** (soft/collinear b's)
  - e.g. MG5_aMC+Pythia, Sherpa+OpenLoops
- NLO+PS X**bb** using **matrix elements** with **massive** b-quarks
  - e.g. Powheg+OpenLoops+Pythia8, Sherpa+OpenLoops

X = Z, W, tt, ...

"5-flavour" schemes

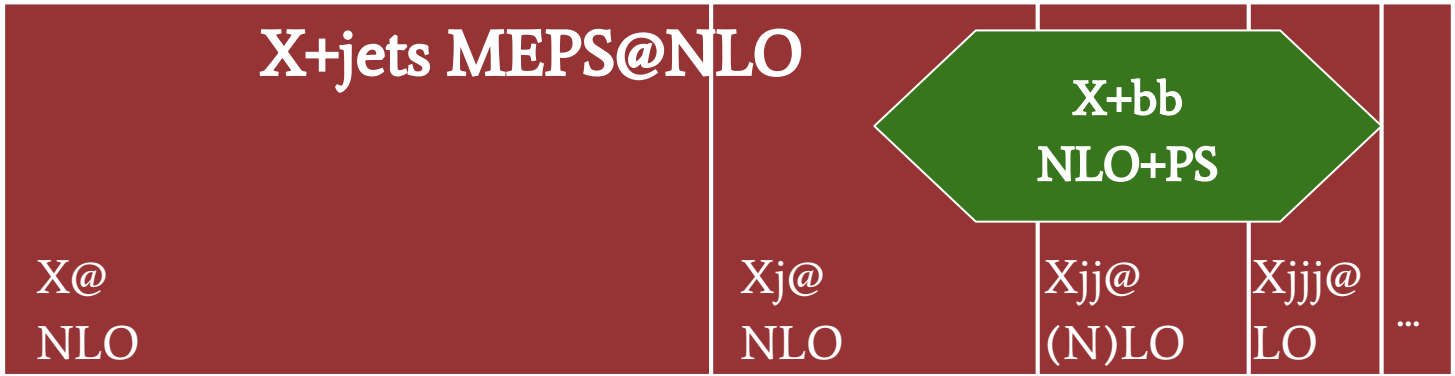


"4-flavour" schemes

Combining 4-flavour X+bb and 5-flavour X+jets?

Fusing X+bb and X+jets in the Sherpa MC

aka “Multi-jet merging in a variable flavour number scheme”



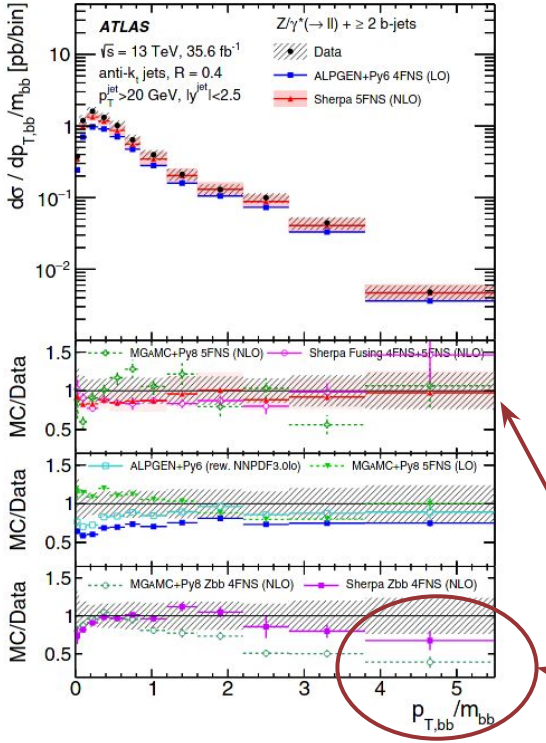
Three main ingredients:

1. Interpreting X+bb as merged contribution
2. Overlap removal
3. Matching 4F/5F in PDFs and α_s

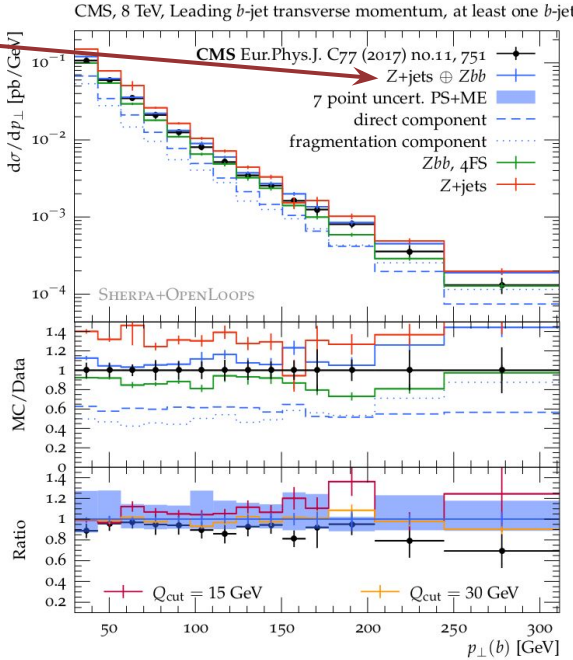
Can be applied for LO and NLO merging!

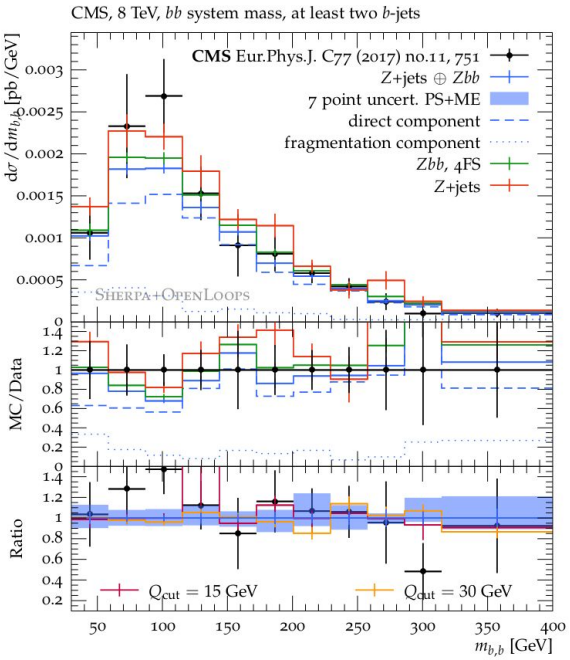
Validation for Z+HF production

- Implementation in Sherpa & validation in Z+HF vs. CMS data [Höche, Krause, FS 1904.09382]



[2003.11960]

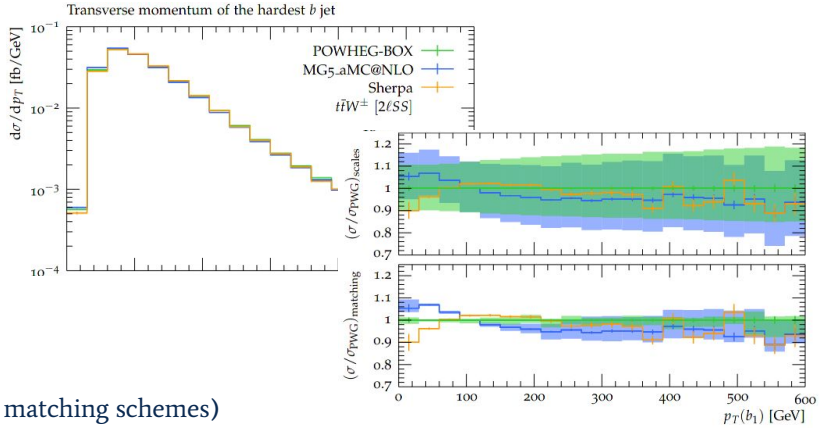




- Recently used in ATLAS for 13 TeV analysis
  - particularly interesting: “inverse” hierarchies

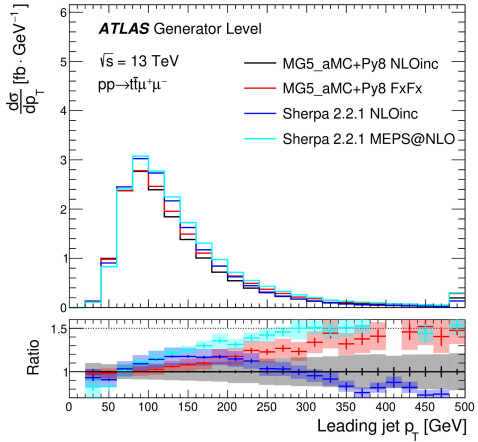
tt+V generator comparison

- First ttV implementation in PowhegBox
  + comparison to MG5_aMC@NLO and Sherpa
  [Febres Cordero, Kraus, Reina 2101.11808]
  - NLO+PS with factorised decays
    and LO spin correlations
  - Generally good agreement
    within perturbative/matching uncertainties
  - Differences mainly at low p_T (as expected from different matching schemes)

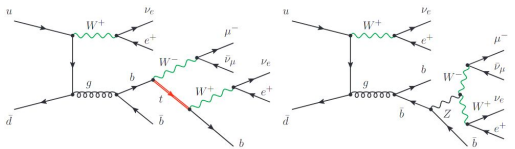


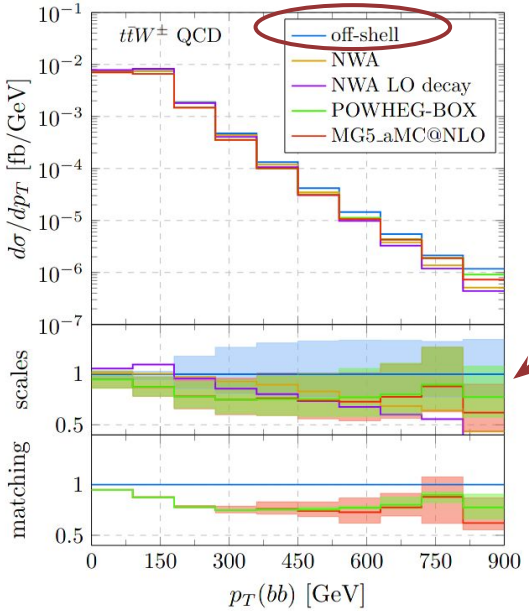
ttW/ttZ leading ones
- Description by MC complex
- Significant differences between generators

- NLO+PS limited in predictivity
  [ATL-PHYS-PUB-2020-024]
  - High scales in ttW → additional hard jet production
  - NLO multi-jet-merged calculations predict significantly harder spectra than NLO+PS



tt+V: non-resonant & EW corrections

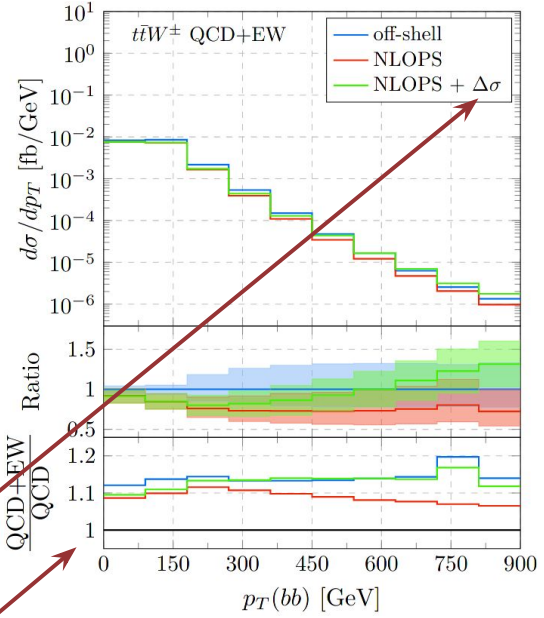


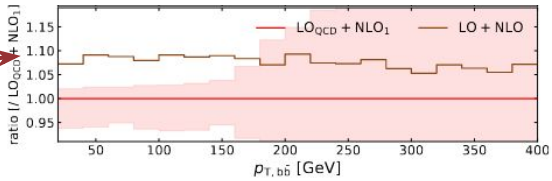


- Offshell contributions in realistic multi-lepton signatures
  [Denner, Pelliccioli 2102.03246]
  [Bevilacqua et al 2109.15181]
  - Include single- or non-resonant diagrams
  - Effects visible in tails of distributions for 3-lepton signature
  - No exclusive NLO+PS generator available → Proposal for additive combination at distribution level:

$$\frac{d\sigma^{\text{th}}}{dX} = \frac{d\sigma^{\text{NLO+PS}}}{dX} + \frac{d\Delta\sigma_{\text{off-shell}}}{dX}, \quad \text{with} \quad \frac{d\Delta\sigma_{\text{off-shell}}}{dX} = \frac{d\sigma^{\text{NLO}}_{\text{off-shell}}}{dX} - \frac{d\sigma^{\text{NLO}}_{\text{NWA}}}{dX}$$

- Also relevant for full prediction: subleading/NLO EW at ~10% level
  - Partially available (for onshell ttV) in MCs through EWvirt approximation



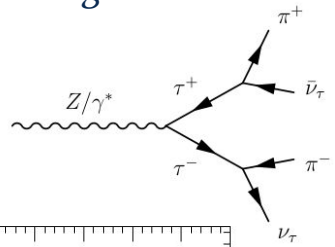


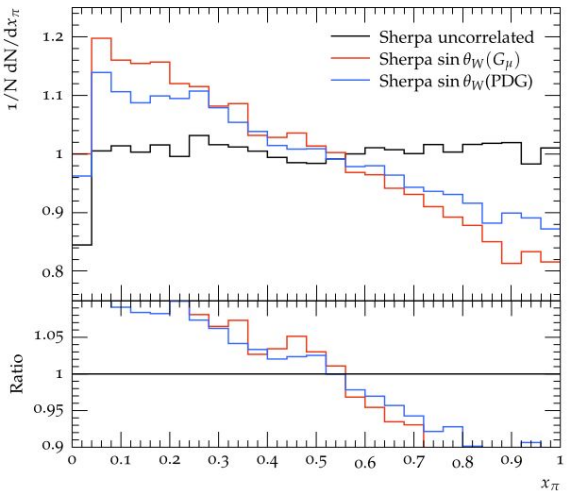
Thoughts on embedding

- Normally not an event-generator modelling topic!
  - e.g. Z(→bb)+jets as bkg for VBF H→bb:
    embedding simulated b-jets in Z(→μμ)+jets data events ~independent from modelling
- Potentially more tricky: Embedding taus into Z(→μμ)+jets for Z(→ττ)
  - Problem: Taus decay and their spins are correlated through production from Z!
  - Correcting for spin correlations after embedding?
    E.g. reweighting based on 1D(?) double-ratio from MC?

$$w(x_\pi) = \frac{\mathrm{d}\sigma^{\text{correlated}}}{\mathrm{d}x_\pi} \Big/ \frac{\mathrm{d}\sigma^{\text{uncorrelated}}}{\mathrm{d}x_\pi}$$

    - » What about other variables, e.g. $m_{\pi\pi}$?
  - Careful: Depends on interplay with NLO EW corrections and choice of EW scheme!
    - » New $\sin^2\theta_{\text{eff}}$ scheme [Chiesa, Piccinini, Vicini 1906.11569] recently available in Powheg, Sherpa+OpenLoops
      → NLO EW corrections **and** PDG value of $\sin^2\theta$





- Consolidating the estimation of these effects would be beneficial

Underlying event & parton shov

- This uncertainty is particularly large for VBF

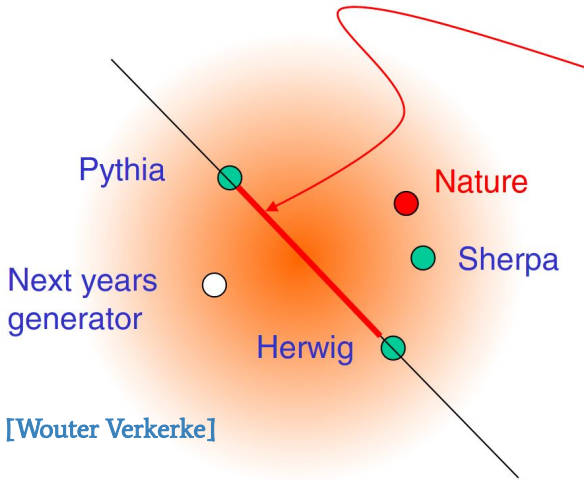
Signal modelling:

- Non-perturbative uncertainties
- Tuning
- ~~VBF-like configurations~~
  → Simon's talk

Underlying event & parton shower

Uncertainty source
Underlying Event and Parton Shower (UEPS)

Non-perturbative uncertainties



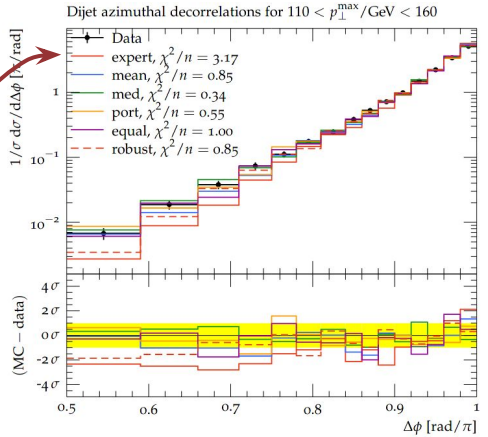
The problem child: 2-point uncertainties

- Particularly problematic in likelihood fits with constraints
- Traditionally even worse: “Pythia” vs. “Herwig”
  - not just one, but many 2-point unc’s in one
- Now (slightly) more controlled/separated
  - parton shower → better use parametric variations [Simon’s talk]
  - hadronisation variations still 2-point, but factorised: Cluster model vs. String model within same generator
  - matching/merging variations separately [e.g. ATL-PHYS-PUB-2020-023]
- New tools can help improve 2-point variations:
  - Tuning
  - ML parametrisation of variations

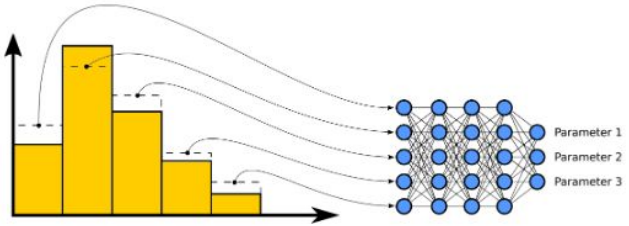
New tools for Monte Carlo tuning

Several recent Monte Carlo tuning developments

- Classical interpolation and minimization:
  "Apprentice" as successor of "Professor" [2103.05748]
- Extension: Automatic observable weights
  "BROOD" [2103.05751]
  - Less time, less subjective bias
  - Beats expert hand-tuning!



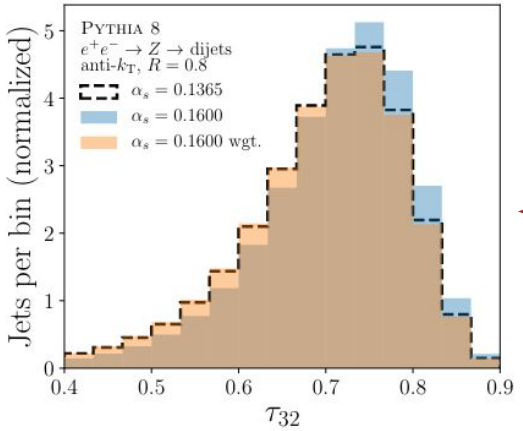
- Replace polynomial interpolation with NN regression:
  "MCNNTUNES" [2010.02213]
  - Adds option of direct (inverse) learning of parameters instead of interpolation
  - Not very robust/usable yet, but interesting idea



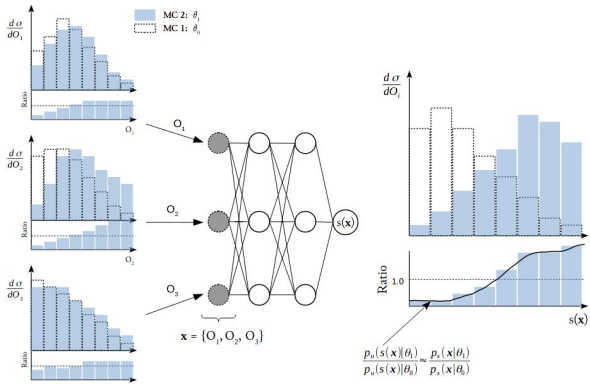
- Huge amount of new precise data for jet physics
  → Open question: How well does tuning work if models do not cover those data?

ML techniques for NP uncertainties

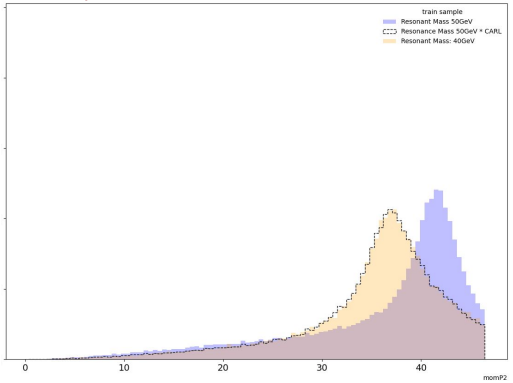
- On-the-fly weights are the standard for perturbative uncertainties (scales, PDFs) now!
  Can we achieve something similar for NP uncertainties?
  - Avoid duplicating simulated MC datasets → CPU and disk saving
  - Statistical fluctuations better under control
- ML techniques to learn differences between two MC samples



- Multi-dimensional mapping
  - » ATLAS VHbb analysis [→ last talk]
  - » CARL [1506.02169]
  - » Reweighting with BDTs [1608.05806]
  - » DCTR [1907.08209]
- Alternative approach:
  Optimal transport [2008.08604]
  to map kinematics onto each other



[Stephen Jiggins]



Conclusions

- Monte Carlo modelling remains key aspect in many analyses
- Modelling improvements from higher perturbative accuracy often come with reduced practicalness
  - Do we need proxy models based on high-precision event generators?
    - » Multi-jet merged LO samples tuned to N(N)LO?
    - » Machine Learning?

    → Uncertainties?
- Ramp-up of developments addressing limited Monte Carlo statistics
  - HSF Event Generator WG as forum [2004.13687]
  - Includes also many less spectacular but important improvements not mentioned here!

Thanks for your interest!