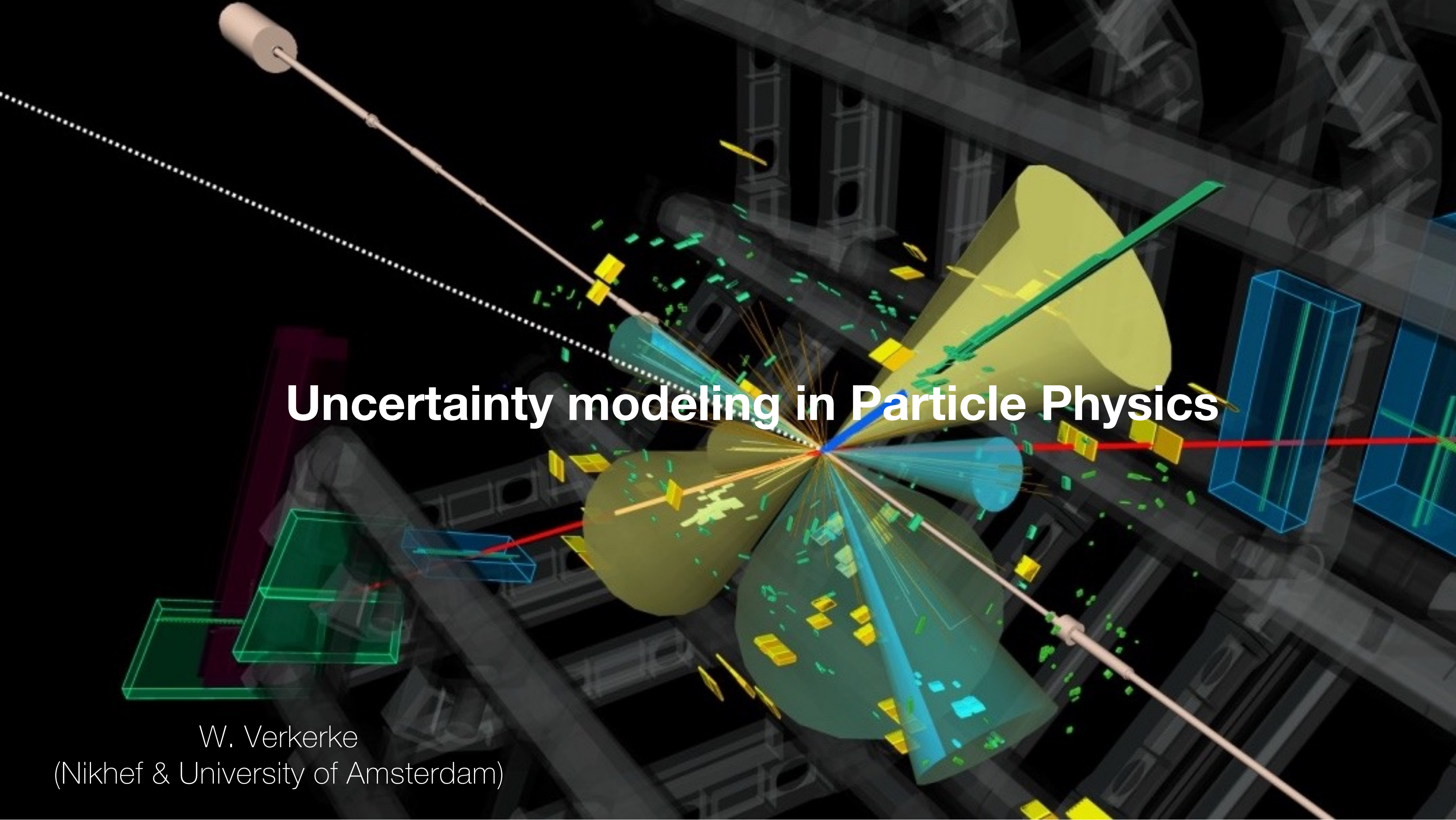
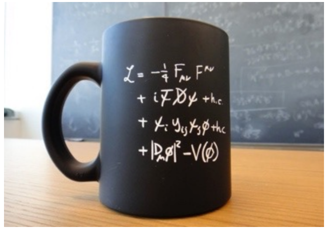
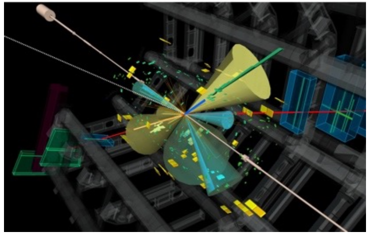


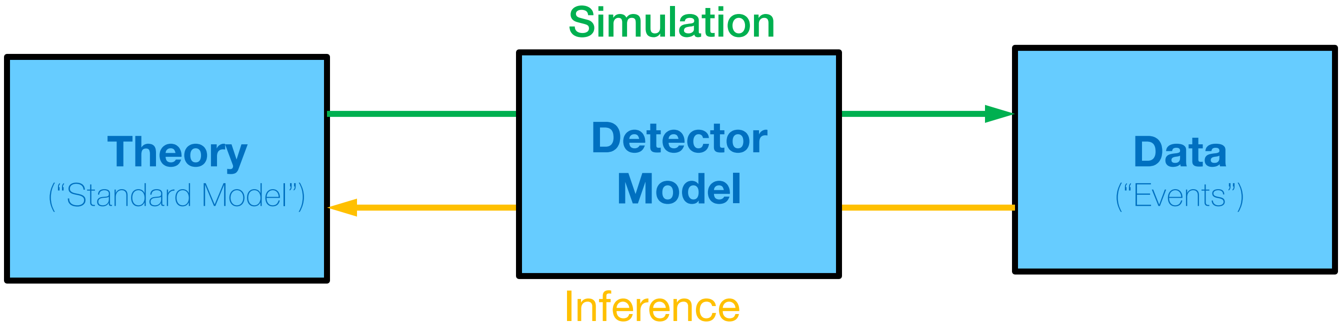
Uncertainty modeling in Particle Physics

W. Verkerke
(Nikhef & University of Amsterdam)

Particle physics data analysis in a nutshell

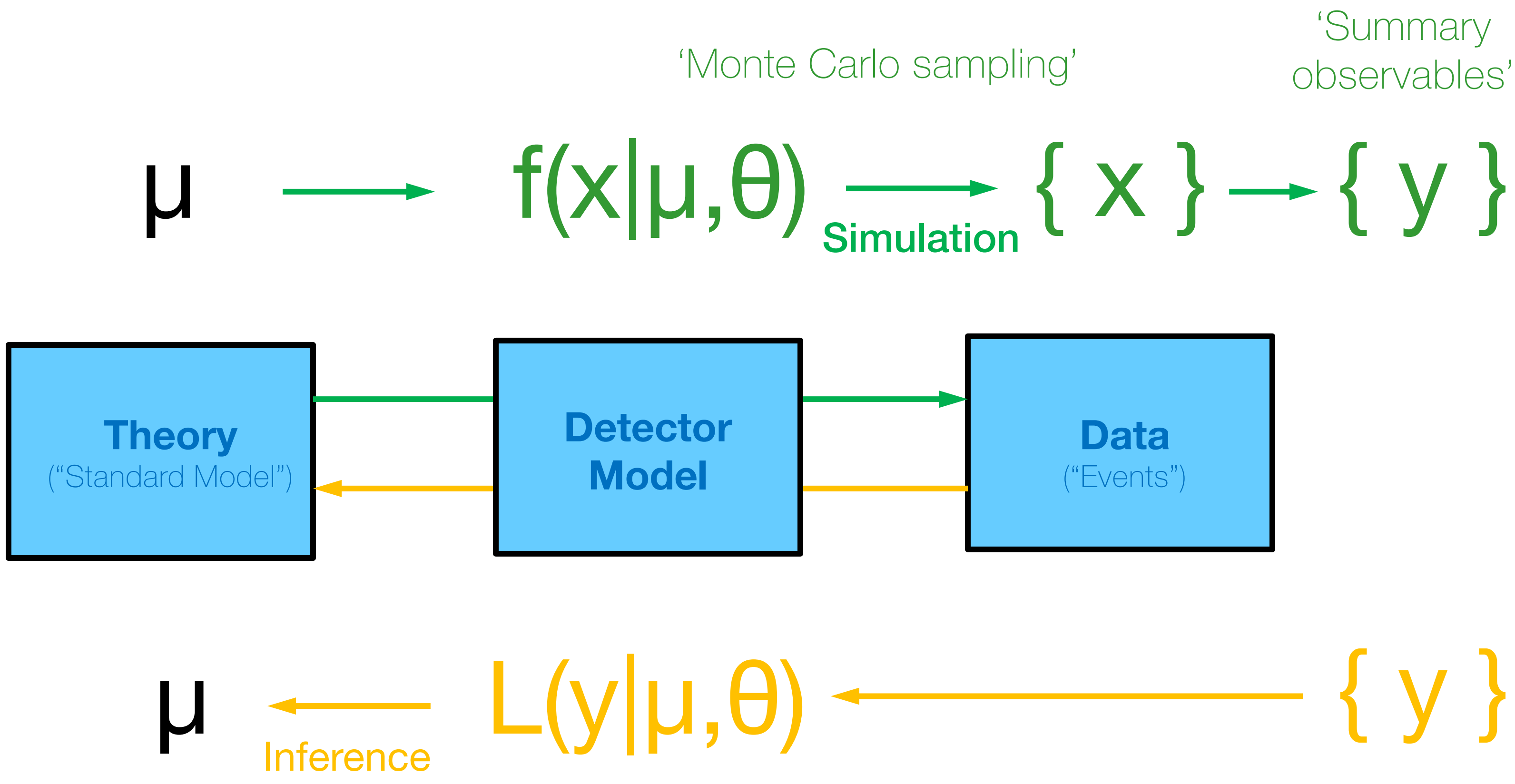






- Simulation is 'easy' (but imperfect)
- Inference is 'hard' as observable space is huge
- **Lots of opportunity here for AI/ML – but beware the imperfections of the simulation**

Particle physics data analysis in a nutshell

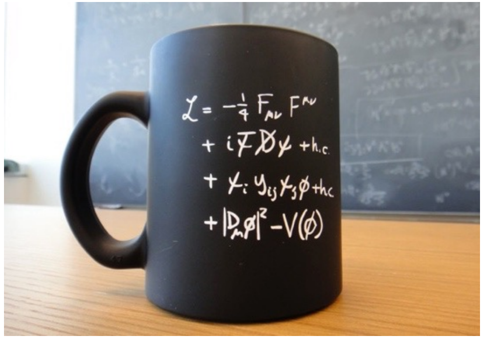


Overview

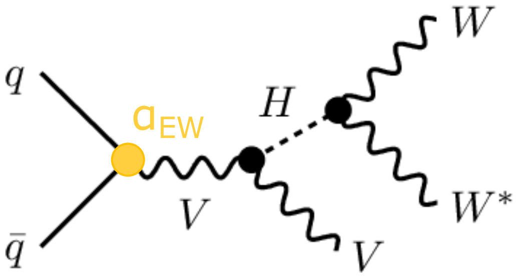
1. Source of uncertainty in the particle physics simulation chain
2. Anatomy of a typical LHC analysis – minimizing depedence on uncertainties
3. Statistical treatment of uncertainties - Frequentist concepts
4. Modeling of simulation uncertainties in the likelihood – general approach
5. Common issues with modeling of specific uncertainties
6. Summary & conclusion

Why are simulation predictions uncertain?

Standard Model some intrinsic uncertainty (through its 17 parameters) but these are almost always irrelevant in practice



However, ability to calculate SM prediction precisely varies very much depending on the regime evaluated

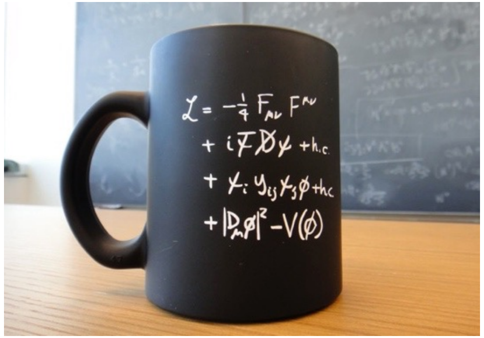


Each process calculable as ***infinite sum*** of amplitude contributions

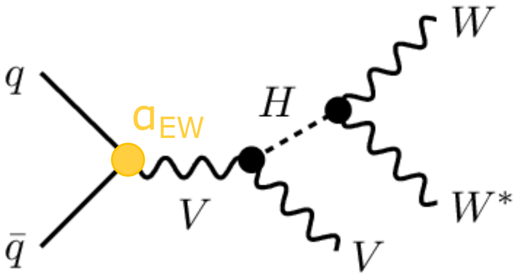
Tractable because contributions are a priori orderable, pertubation series in powers of α (α_{EW} = 1/137)

Why are simulation predictions uncertain?

Standard Model some intrinsic uncertainty (through its 17 parameters) but these are almost always irrelevant in practice



However, ability to calculate SM prediction precisely varies very much depending on the regime evaluated



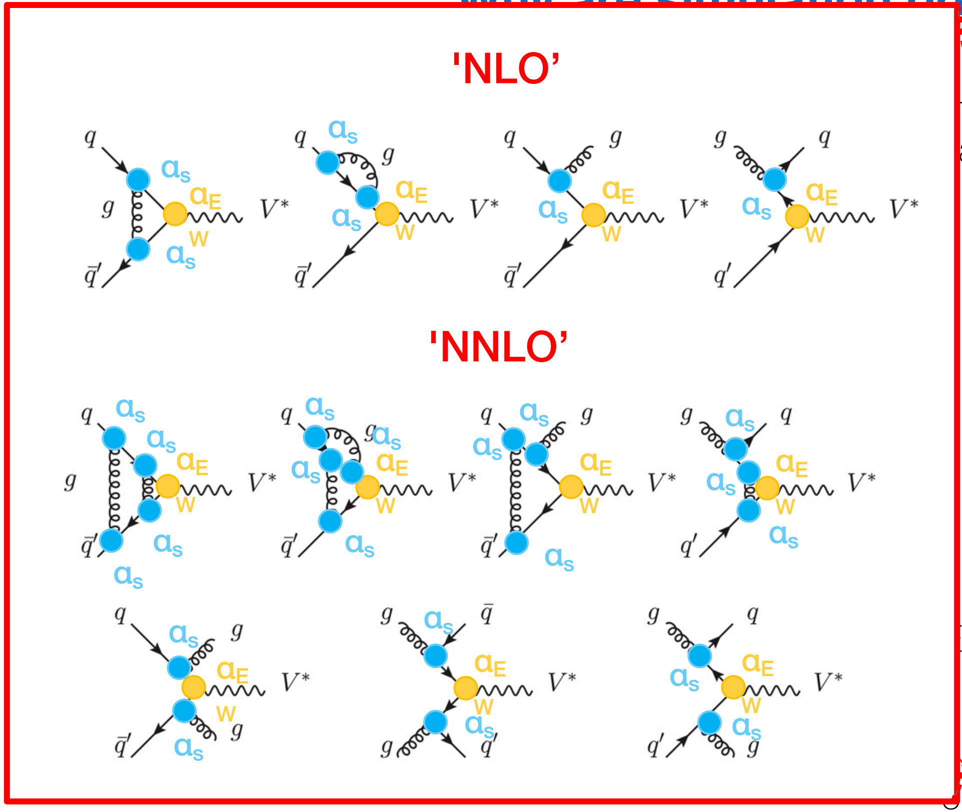
Each process calculable as ***infinite sum*** of amplitude contributions

Not tractable when α not small,
e.g. for strong interaction α_s depends on energy scale

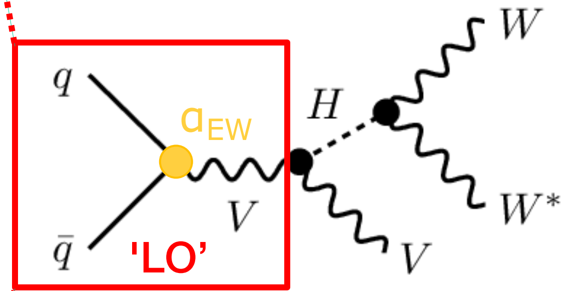
E = 91 GeV
$\alpha_s = 0.12$

E = 1 GeV
$\alpha_s = 0.5$

Why are simulation predictions uncertain?



owever, ability to calculate SM prediction precisely aries very much depending on the regime evaluated



ach process calculable as ***infinite sum*** of mplitude contributions

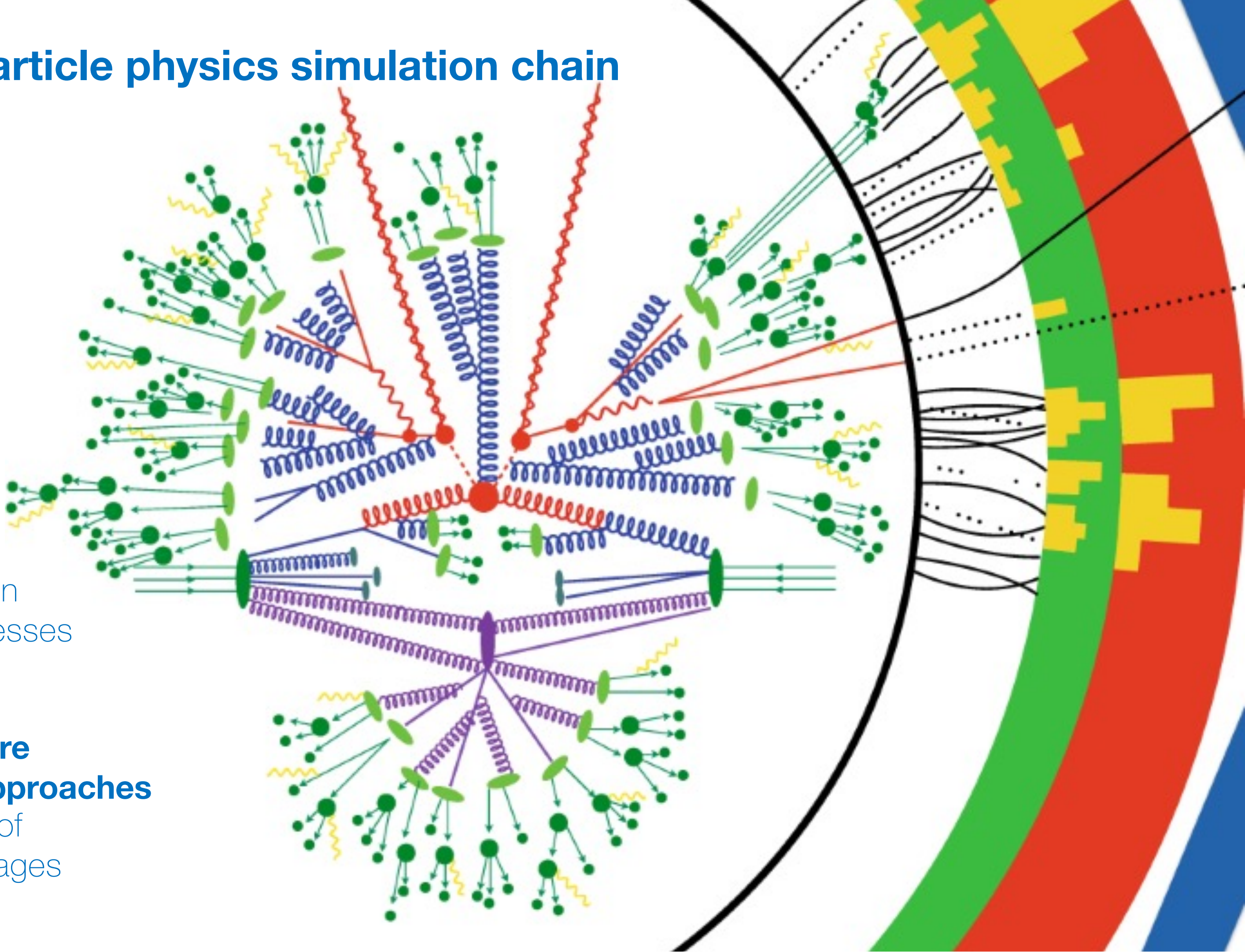
ot tractable when α not small,

.g. for strong interaction α_s depends on energy scale

E = 91 GeV ⟺ E = 1 GeV

$\alpha_s = 0.12$ ⟺ $\alpha_s = 0.5$

The (simplified) particle physics simulation chain



Simulation of a pp collision
at the LHC involves processes
at **many energy scales**

Different regimes require
separate calculation approaches
→ Impemented as chain of
separate simulation packages

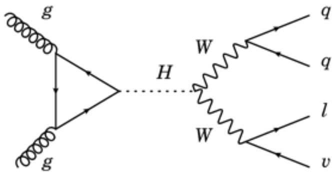
The (simplified) particle physics simulation chain

Proton structure

E($\lesssim$ 1 GeV)

Non-perturbative
('not calculable')

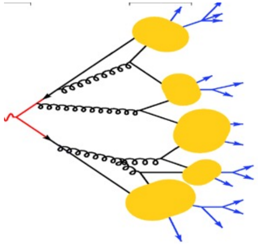
Matrix Element calculation



E(~1 TeV)

Perturbative
calculations
LO, NLO, NNLO,
sometimes N3LO

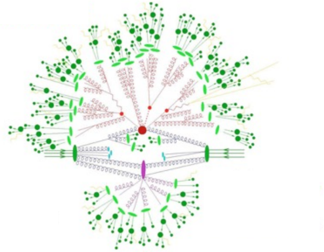
Parton showers



E(100 GeV ~1 TeV)

Perturbative
calculations,
with factorization
assumptions

Hadronization/Fragmentation



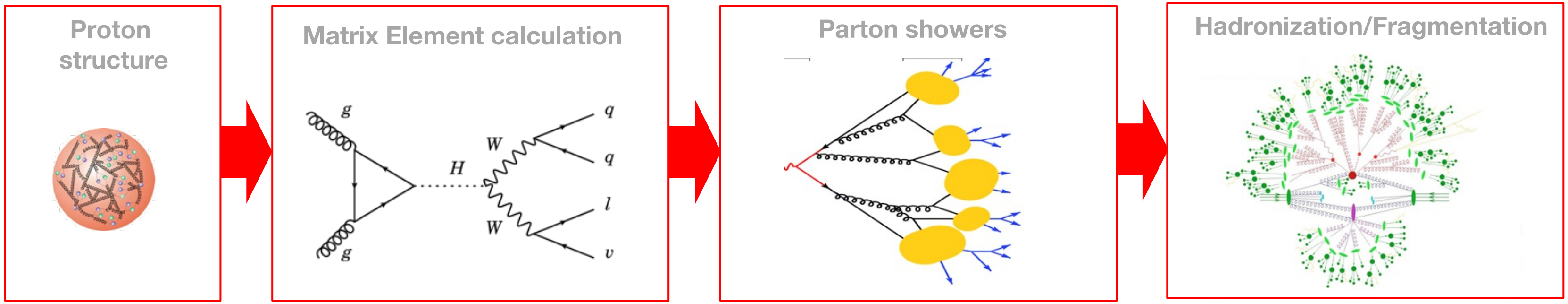
E($\lesssim$ 1 GeV)

Non-perturbative
('not calculable')

Also important, but not shown here: simulation of

- *Underlying Event (proton parts not involved in hard collision)*
- *Color Reconnection events*
- *Addition collisions in the same bunch crossing ("pile up")*

The (simplified) particle physics simulation chain



Estimation procedure

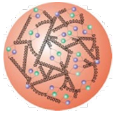
Uncalculable from theory.
All estimates based
on large-scale
fits to experimental data

Example uncertainties

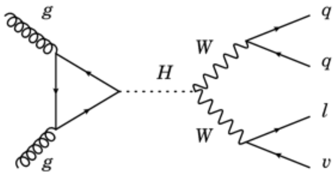
Fit method and
statistical uncertainties

The (simplified) particle physics simulation chain

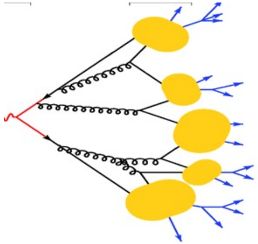
Proton structure



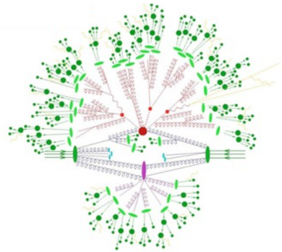
Matrix Element calculation



Parton showers



Hadronization/Fragmentation



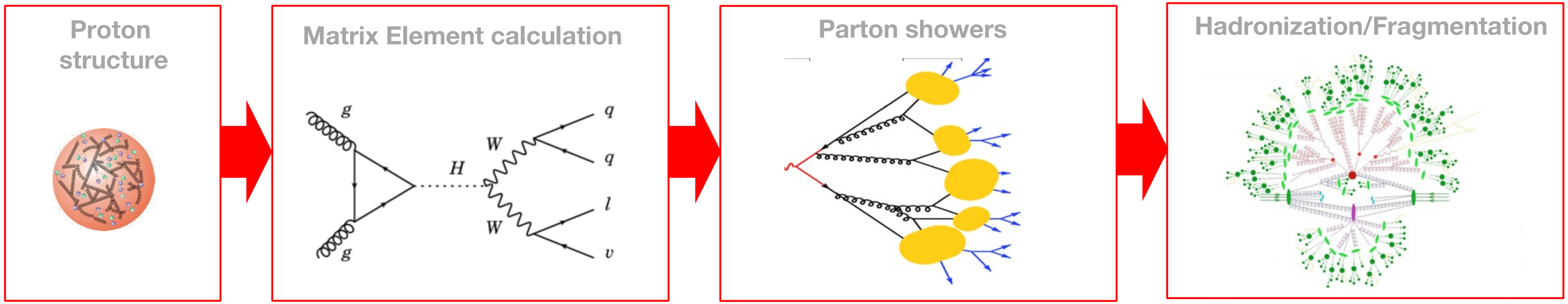
<u>Estimation procedure</u>

Theory calculations
(Monte Carlo simulation,
or fixed-order calculation)

<u>Example uncertainties</u>

Missing higher orders

The (simplified) particle physics simulation chain



<u>Estimation procedure</u>

Perturbative parton shower calculations

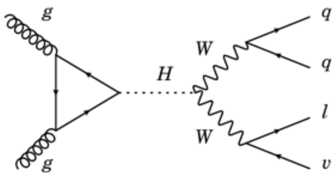
<u>Example uncertainties</u>

Matching of energy scale to that of Matrix Element calculations

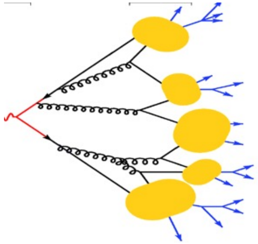
The (simplified) particle physics simulation chain

Proton structure

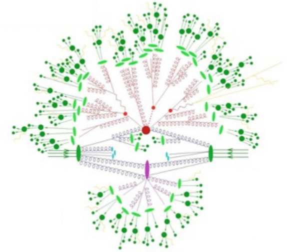
Matrix Element calculation



Parton showers



Hadronization/Fragmentation



Estimation procedure

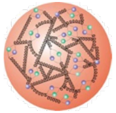
Monte Carlo simulation
based on mostly empirical models
(Multiple implementation, with varying degrees of tradeoff between concepts and tuning)

Example uncertainties

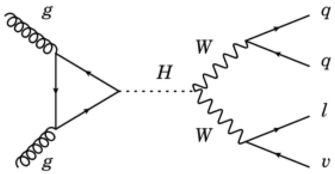
Tuneable parameters with poorly defined physical meaning
Disagreement between packages

The (simplified) particle physics simulation chain

Proton structure

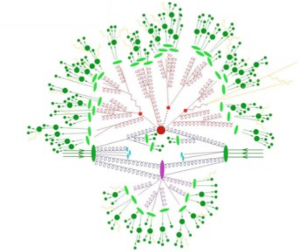


Matrix Element calculation

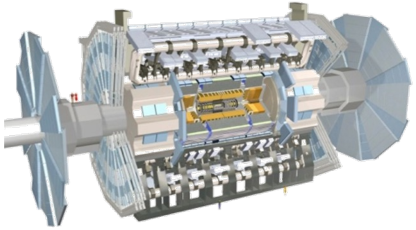


Parton showers

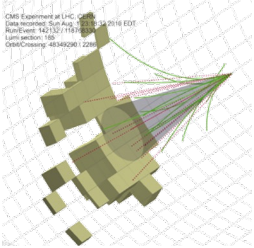
Hadronization/Fragmentation

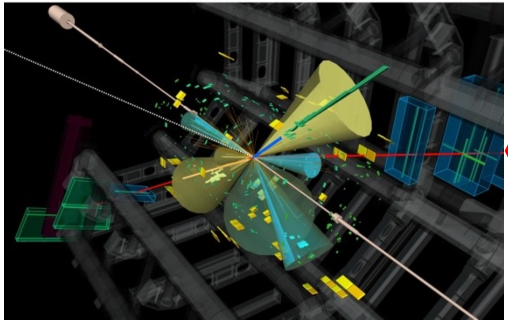


Detector simulation



Object Reco & ID

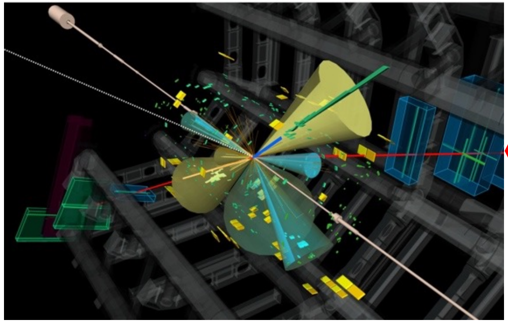


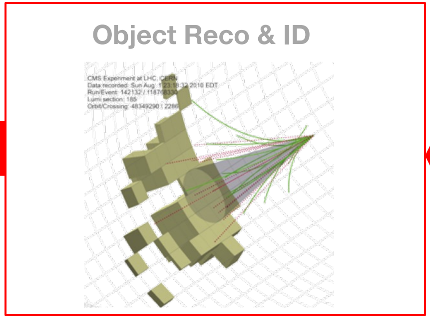


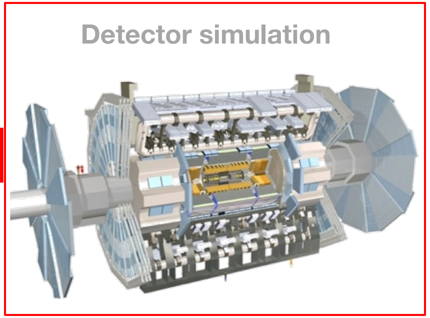
The (simplified) particle physics simulation chain

Estimation procedure
GEANT4 or “fast simulation”

Example uncertainties
Many tuneable parameters
in physics model of GEANT4
(notable hadronic showers),
parametrized model for digitatization
of detector response







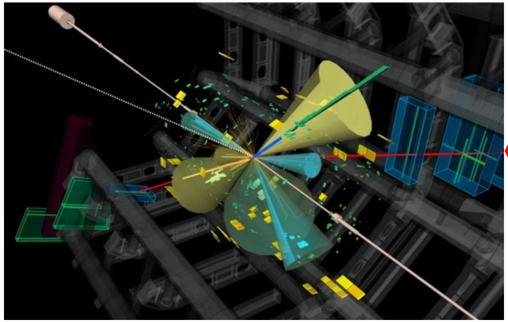
The (simplified) particle physics simulation chain

Estimation procedure

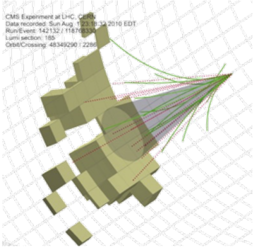
Separate custom-made procedures for each particle type (e,μ,τ, b/c/l-jets, E_t^{miss},…)

Example uncertainties

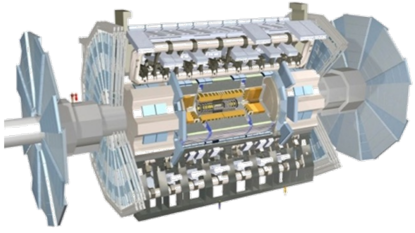
Wide ranging, including physics *simulation* uncertainties and *measurement* uncertainties from data-driven calibrations.



Object Reco & ID

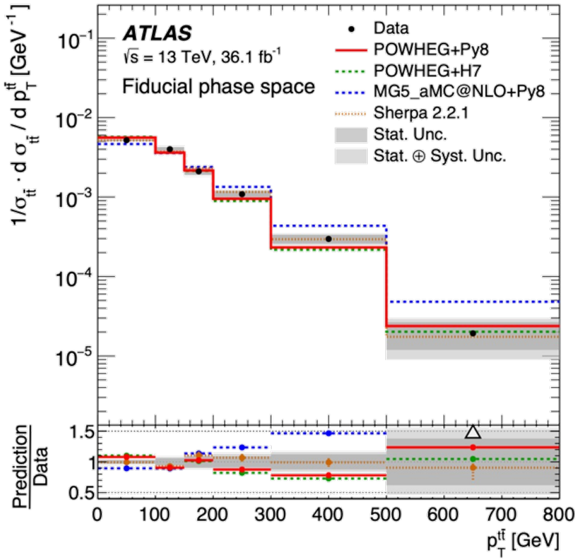


Detector simulation



Is the simulation generally accurate?

- Despite (sometimes) decades of work on simulation packages, and amazing precision in many measurements, **some specific processes and kinematic regimes that are often crucial appear really hard to be correctly modeled in simulation**
- A handful famous/notorious examples
  - QCD multijet production – the (by far) dominant process at the LHC – is almost impossible to simulate as background. (Multitude of physics and technical reasons for this)
  - Differential distributions of top-quark pair kinematics (a dominant background in many analysis) very difficult to get right in simulation →
  - Simulated inclusive cross-section of processes like V+HF production production rates are still off by O(40%) w.r.t observation despite many advances in calculations
  - Efficiency of most object-identification procedures (notably jet-related) are multiplied with data-driven phase-space dependent correction factors applied to simulation.



- **Validation of simulation is generally not exhaustive**
  - Mostly focused on O(1)-dim differential distributions of high-p_T physics objects

Overview

1. Source of uncertainty in the particle physics simulation chain
2. Anatomy of a typical LHC analysis – minimizing depedence on uncertainties
3. Statistical treatment of uncertainties - Frequentist concepts
4. Modeling of simulation uncertainties in the likelihood – general approach
5. Common issues with modeling of specific uncertainties
6. Summary & conclusion

Anatomy of a typical LHC analysis

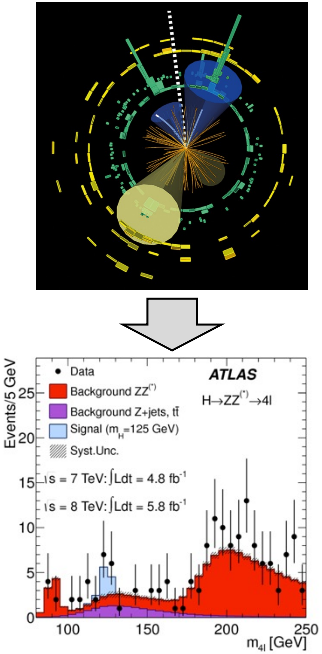
- Given the many caveat and approximations made in simulation, try to be careful not to rely too much on its details ('data driven analysis')
- Typically – HEP analysis is a two-step process

1. Data reduction

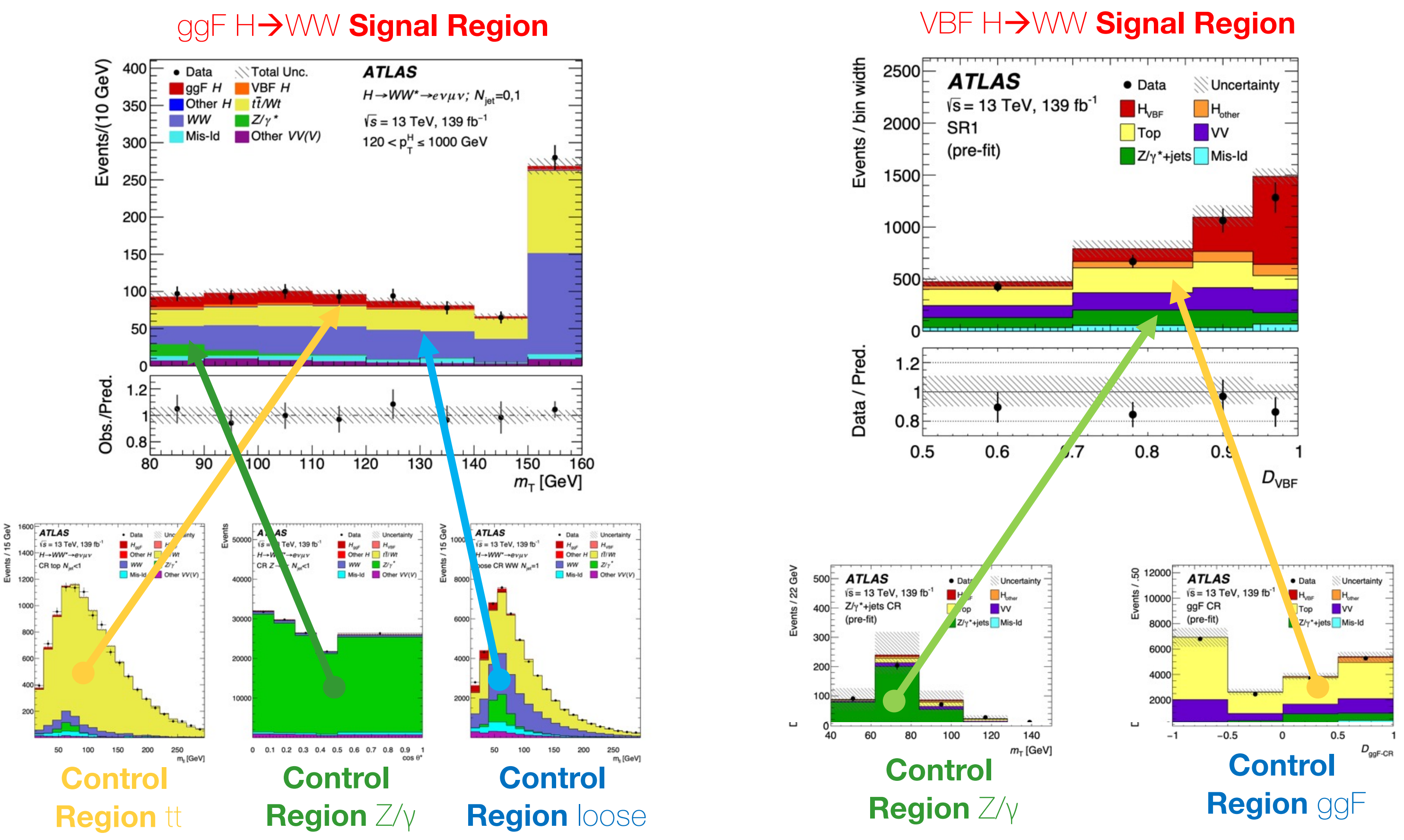
- **Approximate modeling of simulation uncertainties generally acceptable**
- In case of mismodeling, selection could be suboptimal, but effects can be corrected for inference
- ML/AI abundantly used here (mostly BDTs traditionally)

2. Inference

- **Accurate modeling of simulation uncertainties crucial**
- Mismodelling may bias result and/or underestimate uncertainties on final result
- Extensive strategies to minimize influence of systematics, e.g. large number of **control and validation regions** common, express results fiducial regions, perform calculations using ratios
- **Extensive explicit modeling of simulation systematic uncertainties**.
- ML/AI use increasing



Anatomy of a typical LHC analysis



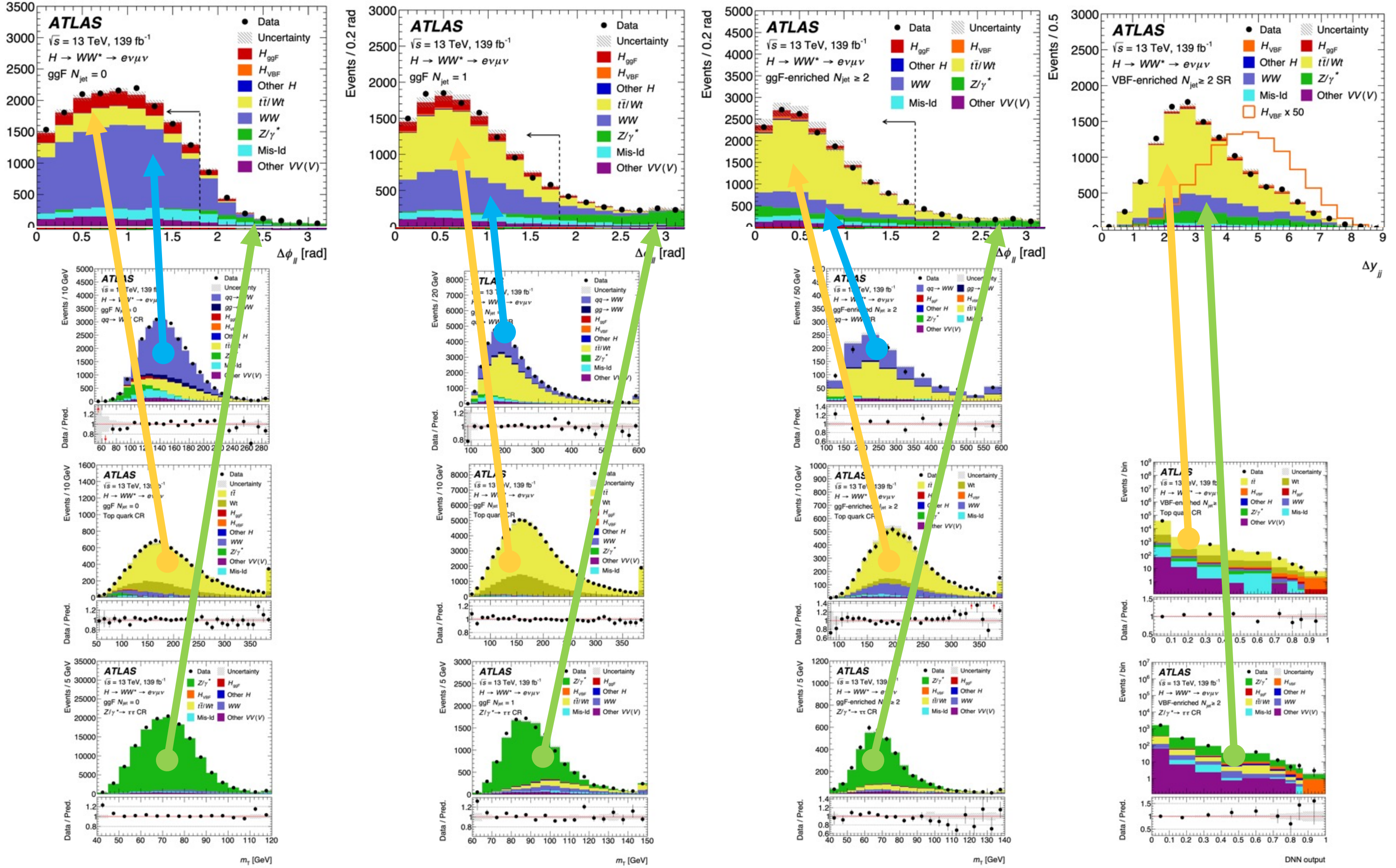
Anatomy of a typical LHC analysis

ggF H→WW
Signal Regions (0j,1j,2j-ggF,2j-VBF)

SM WW
Control Regions

top quark
Control Regions

Z/γ
Control Region



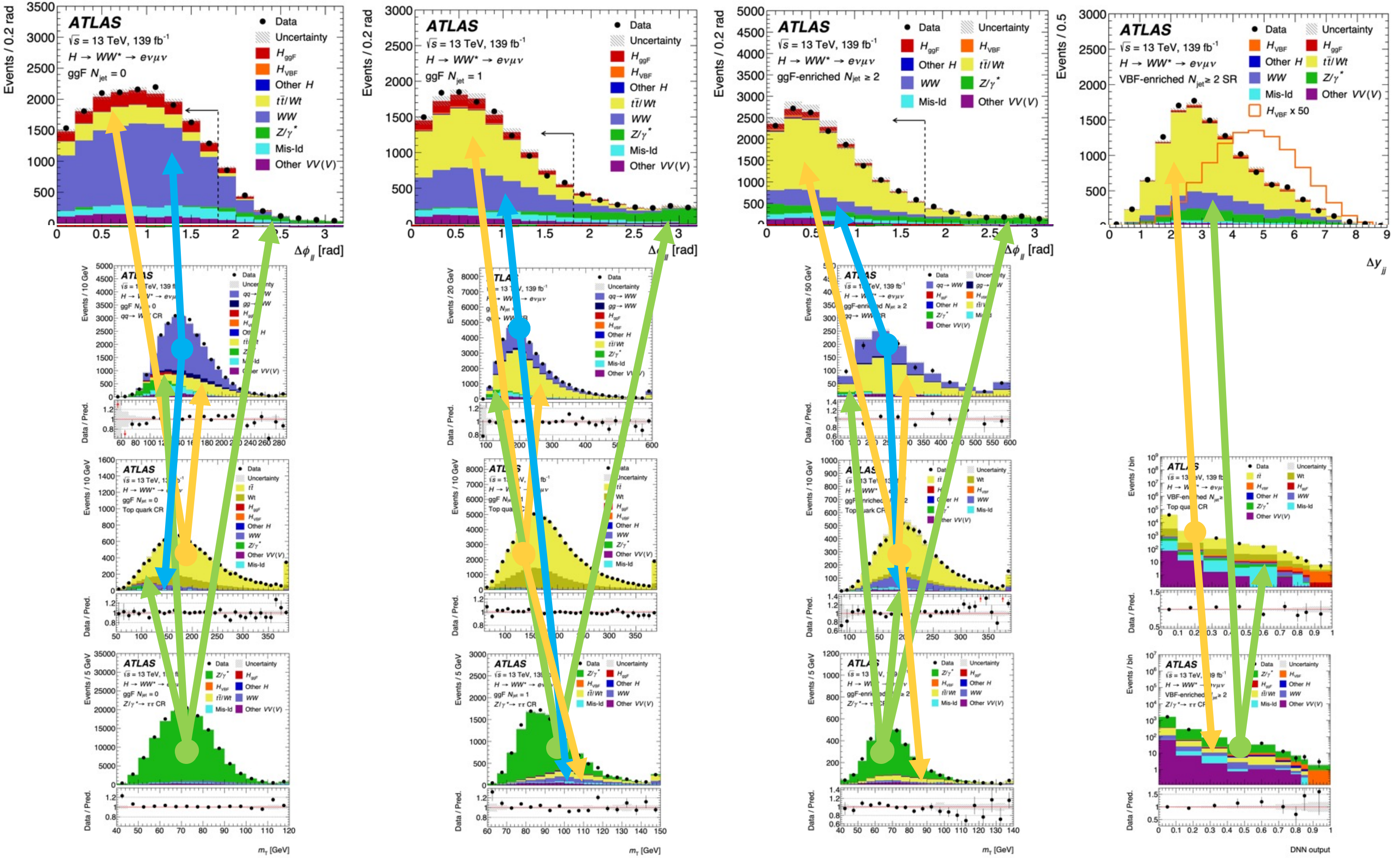
Anatomy of a typical LHC analysis

ggF H→WW
Signal Regions
(0j,1j,2j-ggF,2j-VBF)

SM WW
Control Regions

top quark
Control Regions

Z/γ
Control Region



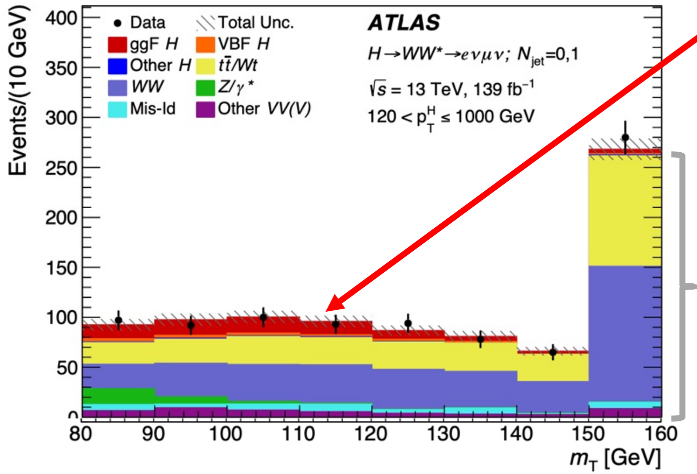
Anatomy of a typical LHC analysis

Yield Distribution

$$L(N_k|\mu,\theta) = \prod_r \text{Poisson}\left(N_{k,r}\,\middle|\, s_k(\mu,\theta)\cdot f_s^{k,r}(\theta) + b_{k,r}(\theta)\right),$$

r = bin index

k = region index



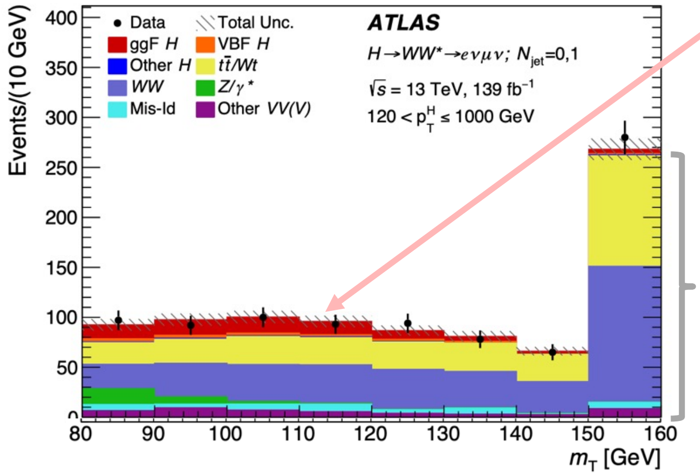
Anatomy of a typical LHC analysis

Yield Distribution

$$L(N_k|\mu,\theta) = \prod_r \text{Poisson}\left(N_{k,r}|s_k(\mu,\theta)\cdot f_s^{k,r}(\theta) + b_{k,r}(\theta)\right),$$

r = bin index

k = region index



Relates yield to cross-section

(summation over k' as multiple regions may contribute to k)

$$s_k(\sigma,\theta) = \mathcal{L}\times\sum_{k'}\sigma^{k'}\times\epsilon_k^{k'}(\theta),$$

Physics POI
(process cross-section in some particle-level kinematic region k')

Acceptance x Eff.
(relates reco-level region k to particle-level region k')

Anatomy of a typical LHC analysis

Yield Distribution

$$L(N_k|\mu,\theta) = \prod_r \text{Poisson}\left(N_{k,r} \,|\, s_k(\mu,\theta) \cdot f_s^{k,r}(\theta) + b_{k,r}(\theta)\right),$$

r = bin index

k = region index

Relat

Sensitivity to simulation modeling uncertainties

(summation over k' as multiple regions may contribute to k)

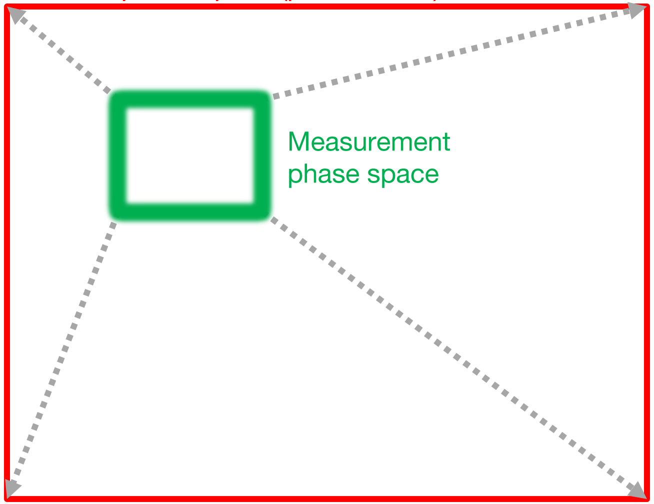
$$s_k(\sigma,\theta) = \mathcal{L} \times \sum_{k'} \sigma^{k'} \times \epsilon_k^{k'}(\theta),$$

Physics POI
(process cross-section in some particle-level kinematic region k')

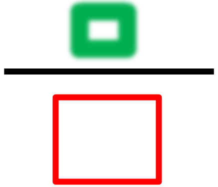
Acceptance x Eff.
(relates reco-level region k to particle-level region k')

Anatomy of a typical LHC analysis - acceptance

Inclusive phase space (particle-evel)



Acceptance =



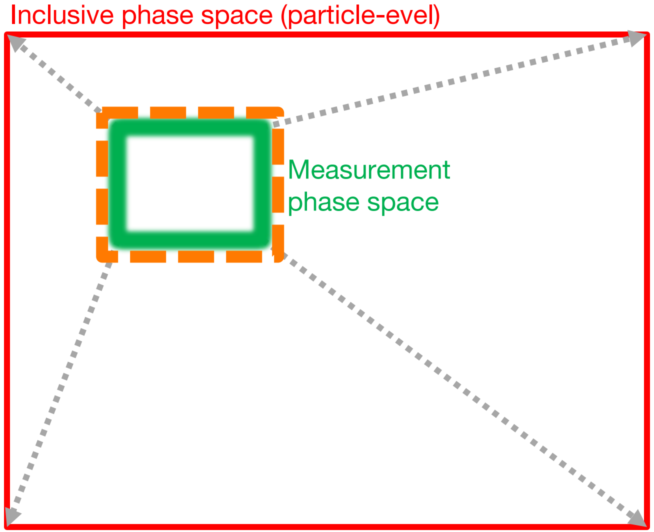
Large extrapolation → Large sensitivity to modeling uncertainties

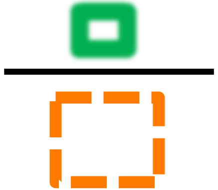
$$s_k(\sigma,\theta) = \mathcal{L} \times \sum_{k'} \sigma^{k'} \times \epsilon_k^{k'}(\theta),$$

Physics POI
(process cross-section
in some particle-level
kinematic region k')

Acceptance x Eff.
(relates reco-level
region k to particle-level
region k')

Anatomy of a typical LHC analysis - acceptance



Acceptance = 

Small extrapolation → Smaller sensitivity to modeling uncertainties

$$s_k(\sigma,\theta) = \mathcal{L} \times \sum_{k'} \sigma^{k'} \times \epsilon_k^{k'}(\theta),$$

Physics POI
(process cross-section in some particle-level kinematic region k')

Acceptance x Eff.
(relates reco-level region k to particle-level region k')

Anatomy of a typical LHC analysis – cross-sections

SM (calculated to high order)

BSM (calculated to lower order)

$$\sigma^{k'} = \underbrace{\sigma_{\mathrm{SM}}}_{\text{SM}} + \underbrace{\sigma_{\mathrm{int}} + \sigma_{\mathrm{BSM}}}_{\text{BSM}}$$

Apply ratio corrections (assumes factorization)

$$\sigma^{k'} = \sigma_{\mathrm{SM}}^{((\mathrm{N})\mathrm{N})\mathrm{NLO}} \times \left(1 + \frac{\sigma_{\mathrm{int}}^{(\mathrm{N})\mathrm{LO}}}{\sigma_{\mathrm{SM}}^{(\mathrm{N})\mathrm{LO}}} + \frac{\sigma_{\mathrm{BSM}}^{(\mathrm{N})\mathrm{LO}}}{\sigma_{\mathrm{SM}}^{(\mathrm{N})\mathrm{LO}}}\right)$$

$$= \mathcal{L} \times \sum_{k'} \sigma^{k'} \times \epsilon_k^{k'}(\boldsymbol{\theta}),$$

Physics POI (process cross-section in some particle-level kinematic region k')

Acceptance x Eff. (relates reco-level region k to particle-level region k')

Overview

1. Source of uncertainty in the particle physics simulation chain
2. Anatomy of a typical LHC analysis – minimizing depedence on uncertainties
3. Statistical treatment of uncertainties - Frequentist concepts
4. Modeling of simulation uncertainties in the likelihood – general approach
5. Common issues with modeling of specific uncertainties
6. Summary & conclusion

Uncertainties in particle physics

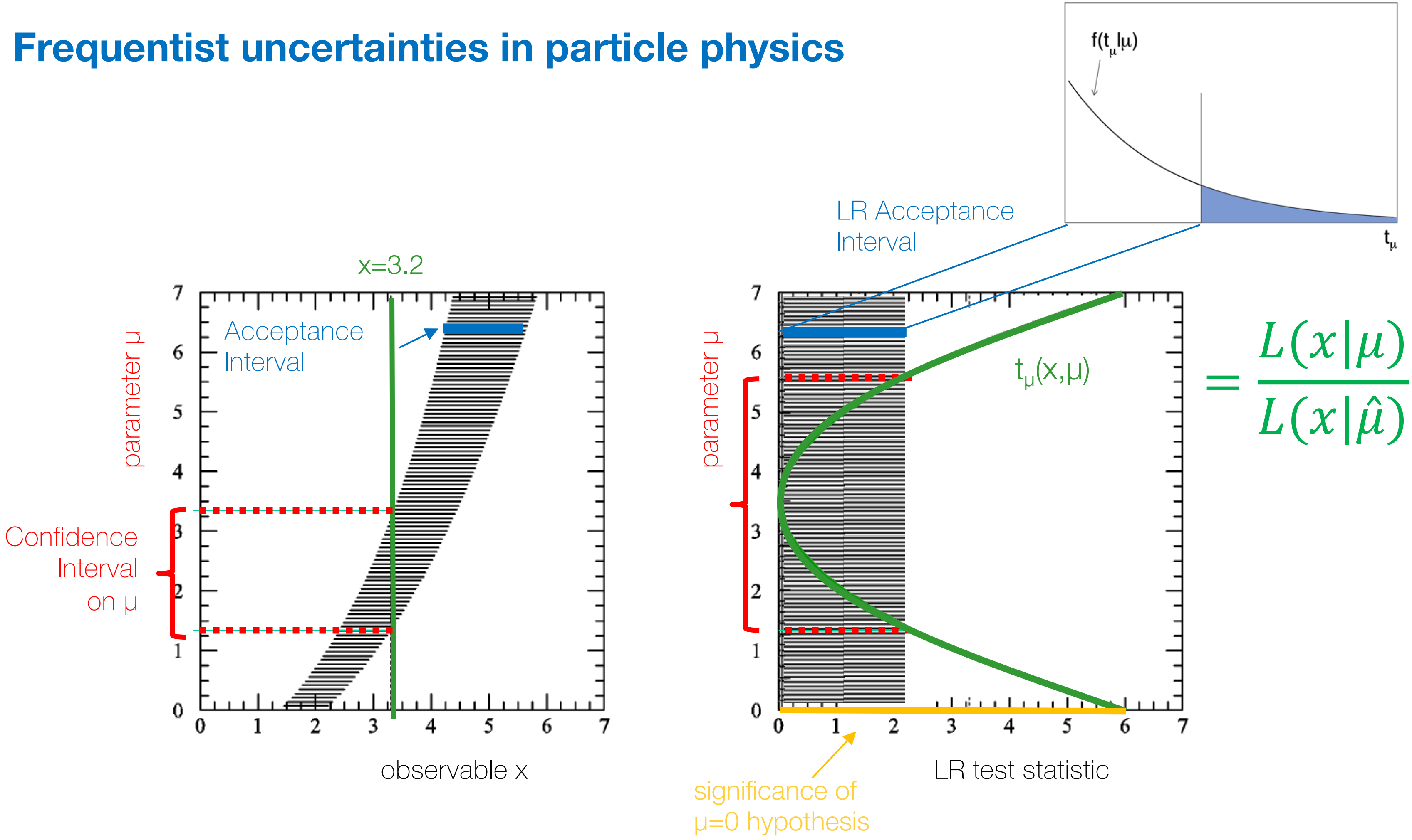
Inference

$$\mu \longleftarrow L(y|\mu,\theta) \longleftarrow \{\, y \,\}$$

- Statistical methodology in particle physics is (very) predominantly **frequentist**
- Notion of **coverage** is central in definition of uncertainties (68%, 95%)
- Computational procedures for frequentist methodology quite different from those for Bayesian influences practical aspects of how systematics uncertainties are modelled.

- **A 30-second nutshell reminder of Frequentist approach**
  - Observations **{ y }** are summarized with a test statistic **T(y)**, in practice a **likelihood ratio** testing for compatibility of the data with a certain hypothesis μ
  - With knowledge of the distribution **T_μ(y)** under given hypothesis μ can define an **acceptance interval** that captures 68% of the observed outcomes
  - A confidence belt maps the acceptance interval for each value of μ, and allows to construct a **confidence interval** in μ for a given observed value of **T_μ(y)**

Frequentist uncertainties in particle physics

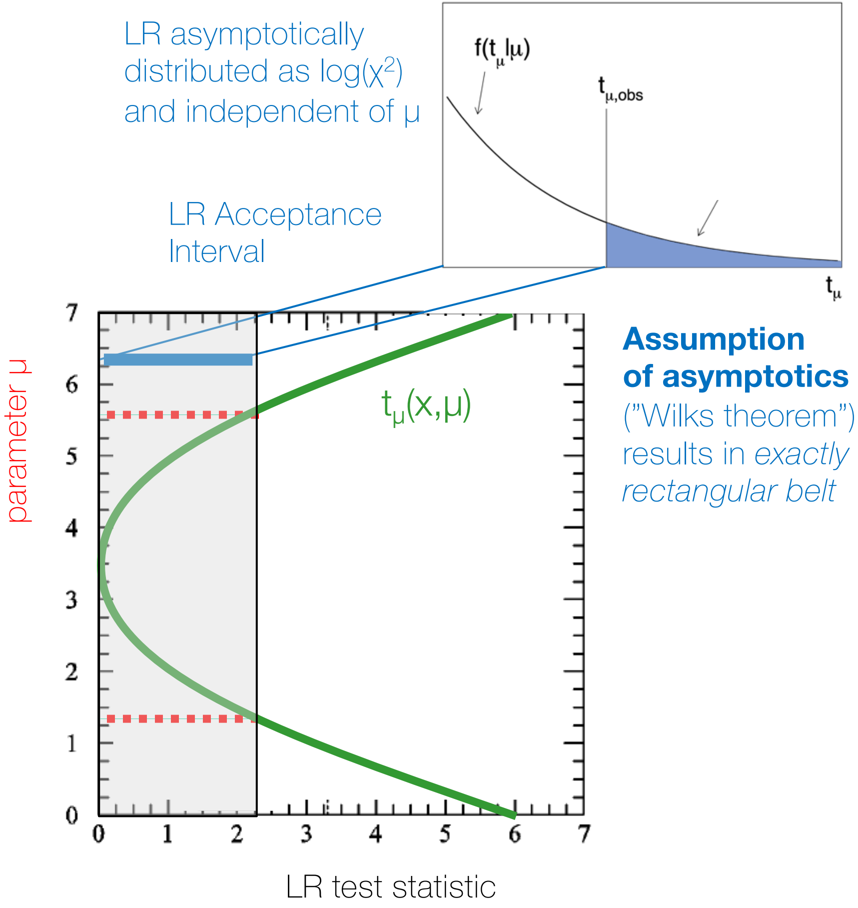


$$= \frac{L(x|\mu)}{L(x|\hat{\mu})}$$

Frequentist approach – asymptotic approximation

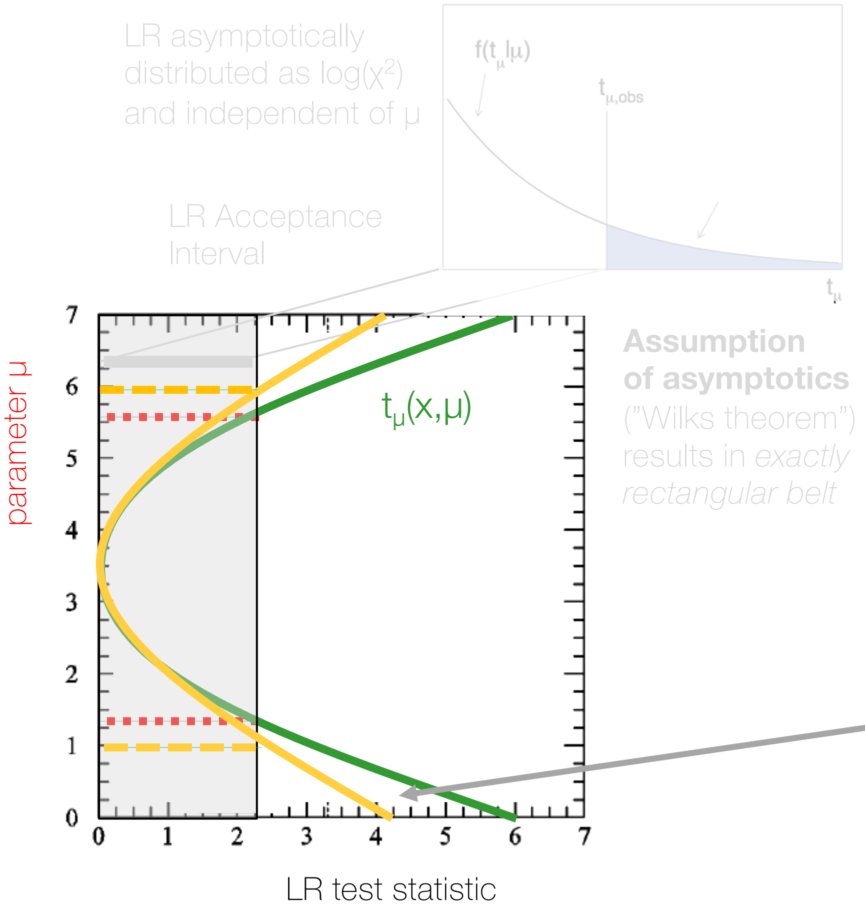
LR asymptotically distributed as log(χ^2) and independent of μ

LR Acceptance Interval



Assumption of asymptotics ("Wilks theorem") results in *exactly rectangular belt*

Frequentist approach – with nuisance parameters



Nuisance parameters (to incorporate modeling uncertainties) included in **profile likelihood ratio**

$$t_\mu = \frac{L(x|\mu)}{L(x|\hat{\mu})}$$

$$t_\mu = \frac{L(x|\mu, \hat{\hat{\theta}}(\mu))}{L(x|\hat{\mu}, \hat{\theta})}$$

Frequentist approach – asymptotics & the profile likelihood ratio

$$t_\mu = \frac{L(x|\mu,\hat{\hat{\theta}}(\mu))}{L(x|\hat{\mu},\hat{\theta})}$$

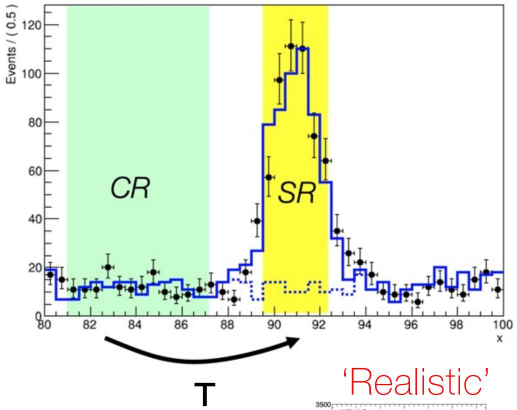
- Note 1: that $\mathbf{t_\mu}$ in profile likelihood can *in principle* depend on values of θ in hypothesis
  - Practical approach at LHC → always assume values values $\hat{\theta}$
- Note 2: computation of $\mathbf{t_\mu}$ is relatively cheap even if even if dimension of θ is large
  - **No practical penalty on introducing many nuisance parameters**.
  - Many LHC analyses often have hundreds nuisance parameters, and often enough more than 1000
- Note 3: notion of coverage should also extend to knowledge on nuisance parameters,
  - **Often difficult due to imprecise or incomplete definitions of nuisance parameters**
  - In practice only an issues if they result in large uncertainties in μ, but that happens often enough

Overview

1. Source of uncertainty in the particle physics simulation chain
2. Anatomy of a typical LHC analysis – minimizing depedence on uncertainties
3. Statistical treatment of uncertainties - Frequentist concepts
4. **Modeling of simulation uncertainties in the likelihood – general approach**
5. Common issues with modeling of specific uncertainties
6. Summary & conclusion

Modeling (simulation) uncertainties in the likelihood

'Simple'



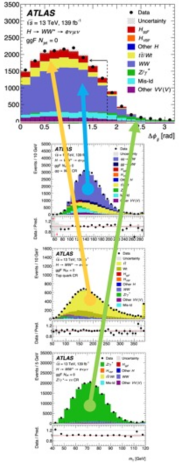
- Simple data-driven

$$L(N_{SR}|N_{CR}) = \text{Poisson}(N_{SR}\,|\,s + b) \cdot \text{Poisson}(N_{CR}\,|\,\tau \cdot b)$$

- Fully simulation-based

$$L(N_{SR}) = \text{Poisson}(N_{SR}|\,s + b) \cdot \mathbf{Gauss(\,b_{sim}\,|\,b,\,\sigma_{b,sim})}$$

'Realistic'



- Realistic data-driven

$$L(N_{SR}) = \text{Poisson}(N_{SR}|\,s + b) \cdot \text{Poisson}(N_{CR}\,|\,\tau \cdot b) \cdot \mathbf{Gauss(\,\tau_{sim}\,|\,\tau,\,\sigma_{\tau,sim})}$$

Modeling (simulation) uncertainties in the likelihood

- Generalization of modeling approach

$$L(N_{SR}) = \text{Poisson}(N_{SR} | s + b) \cdot \mathbf{Gauss}(\, \mathbf{b_{sim}} \,|\, \mathbf{b}, \boldsymbol{\sigma_{b,sim}})$$

$$L(N_{SR}) = \text{Poisson}(N_{SR} | s + \mathbf{b(\alpha)}) \cdot \mathbf{Gauss}(\, \mathbf{0} \,|\, \boldsymbol{\alpha}, \mathbf{1})$$

"Response function"

Can be non-linear

"Subsidiary measurement"

For additive systematics, can always be reduced to a unit Gaussian

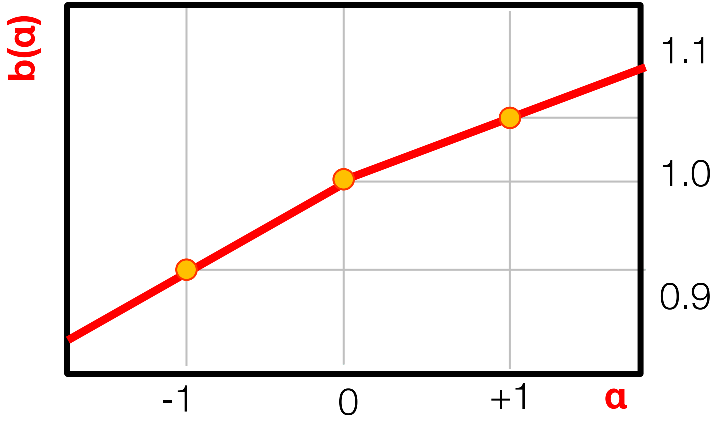
Alternatively:
Poisson - For systematic effects of a statistical nature
LogNormal – For multiplicative systematics where a positive-definite NP is required

Modeling (simulation) uncertainties in the likelihood

- Generalization of modeling approach

$$L(N_{SR}) = \text{Poisson}(N_{SR} | s + b) \cdot \mathbf{Gauss}(\, b_{sim} \,|\, b, \sigma_{b,sim})$$

$$L(N_{SR}) = \text{Poisson}(N_{SR} | s + \mathbf{b(\alpha)}) \cdot \mathbf{Gauss}(\, 0 \,|\, \alpha, 1)$$



Empirical approximation of true response

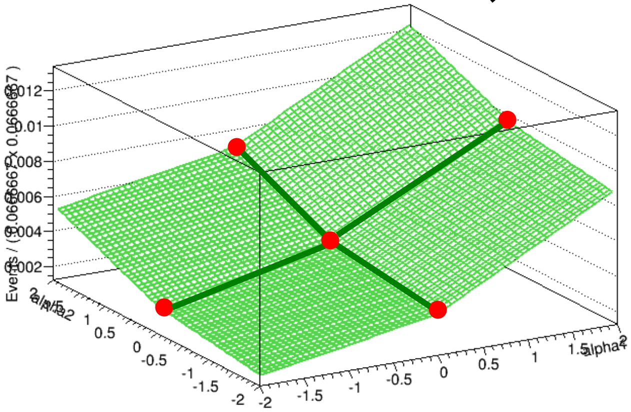
- **Sample simulation** response at α=-1,0,+1
- Apply piece-wise linear **interpolation** (or higher-order smooth functions if needed)

Modeling (simulation) uncertainties in the likelihood

- Generalization of modeling approach

$$L(N_{SR}) = \mathrm{Poisson}(N_{SR} | s + b) \cdot \mathbf{Gauss}(\, b_{sim} \,|\, b, \sigma_{b,sim})$$

$$L(N_{SR}) = \mathrm{Poisson}(N_{SR} | s + \mathbf{b}(\alpha_1, \alpha_2)) \cdot \mathbf{Gauss}(\, 0 \,|\, \alpha_1, 1) \cdot \mathbf{Gauss}(\, 0 \,|\, \alpha_2, 1)$$

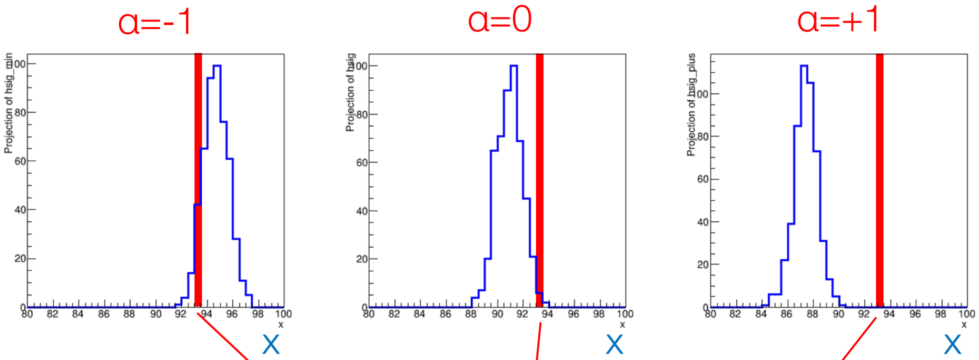


Interpolated response function generalizes easily to multiple nuisance parameters

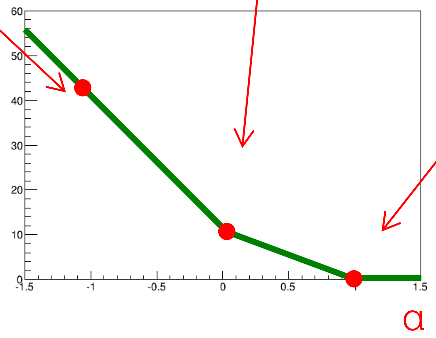
Typically only 'star topology' sampled, i.e. no correlation effects in response function of a single bin

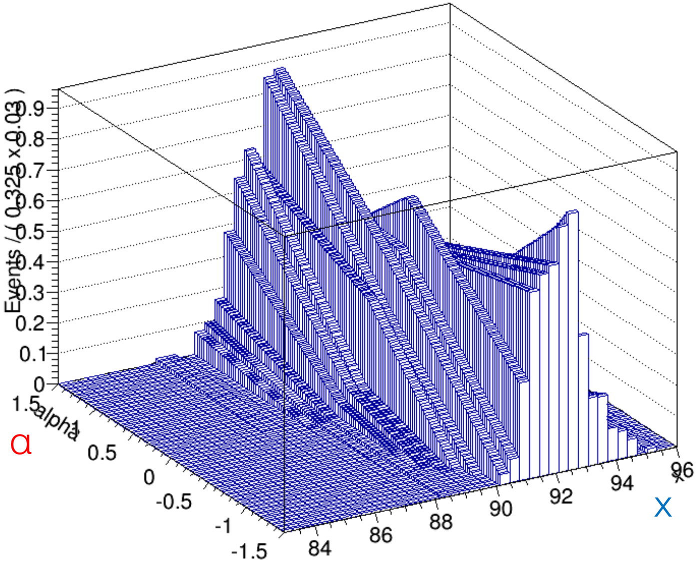
Modeling (simulation) uncertainties in the likelihood

- Generalization of modeling approach to distributions



Piecewise linear interpolation response model for a one bin





Bin-by-bin piece-wise interpolation robust enough for small-to-moderate distortions typically introduced by systematic variations

Modeling (simulation) uncertainties in the likelihood

- **Modeling uncertainties across regions** – choice of correlated or uncorrelated

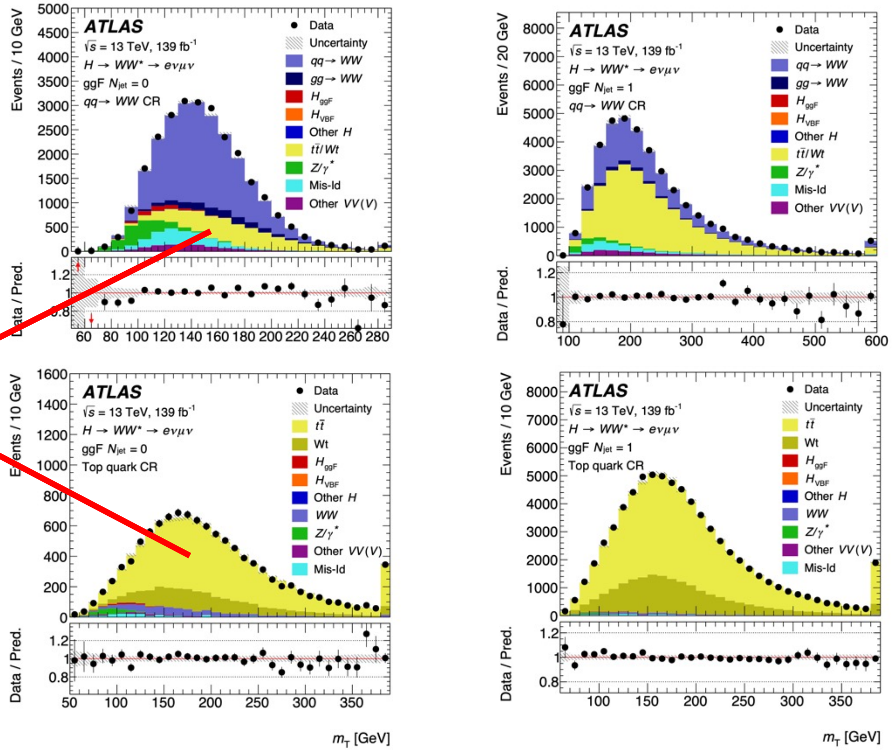
0-jet

1-jet

WW CR

common
normalization
parameter
$\alpha_{top\text{-}0jet}$

top CR



Wouter Verkerke, NIKHEF

Modeling (simulation) uncertainties in the likelihood

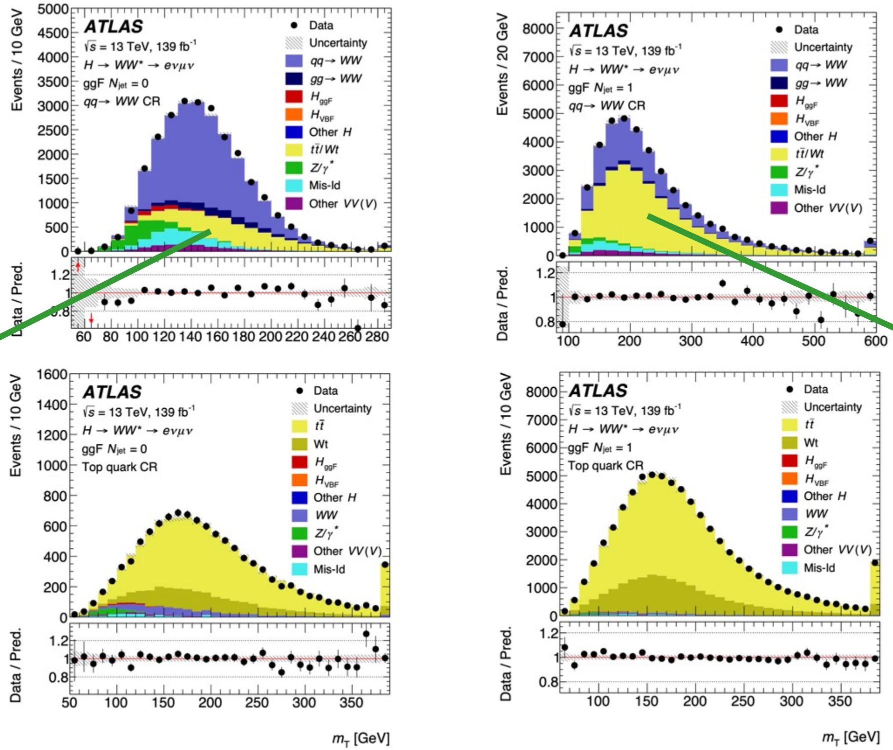
- **Modeling uncertainties across regions** – choice of correlated or uncorrelated

0-jet

1-jet

WW CR

top CR

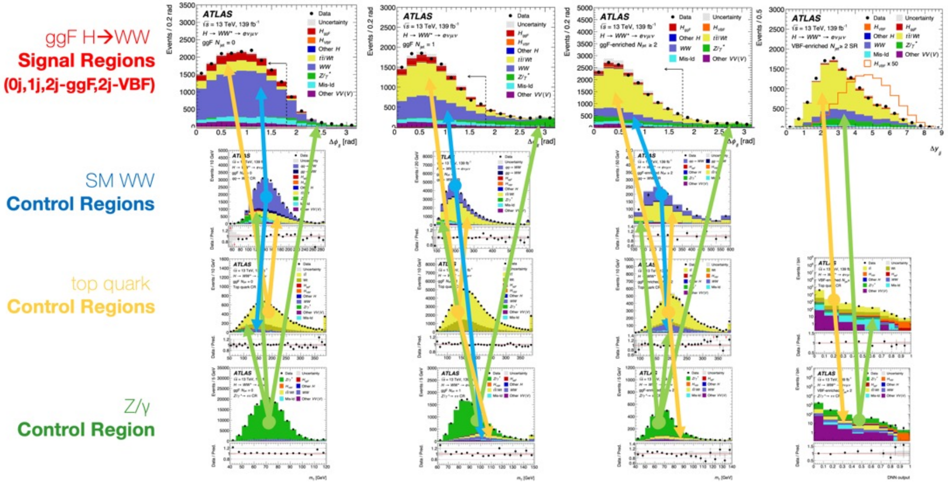


separate
normalization
parameter
$\alpha_{top\text{-}0jet}$

separate
normalization
parameter
$\alpha_{top\text{-}1jet}$

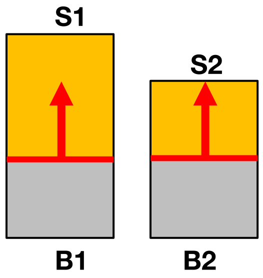
Modeling (simulation) uncertainties in the likelihood

- In a complete analysis there will be **many nuisance parameters**, with typical number ranging from **100-1000**
- Number driven by approach to break down uncertainties into *individual sources* that map to known concepts in theory or detector
- **NP correlation scheme is always a major point of attention**, as for many modeling systematics it is *not always clear if source uncertainties are correlated or uncorrelated* across kinematic regions
- *Partial correlations in individual sources/NPs uncommon*.
In NP groups that collectively describe a systematic uncertainty source, partial correlations are modeling through mix of correlated and uncorrelated components

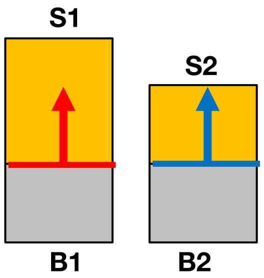


Implementing "appropriately conservative" uncertainties

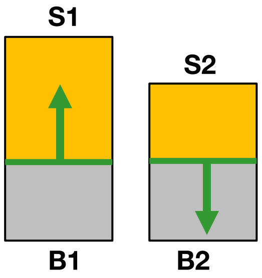
- **Correlation model** of NPs can present host of thorny issues if there is no clear **guidance from systematic source**
- Illustration with '2-bin' analysis



NP: 10% bkg uncertainty
correlated modeling



NP1: 10% bkg unc. – bin 1
NP2: 10% bkg unc. – bin 2



NP: 10% bkg uncertainty
anti-correlated model

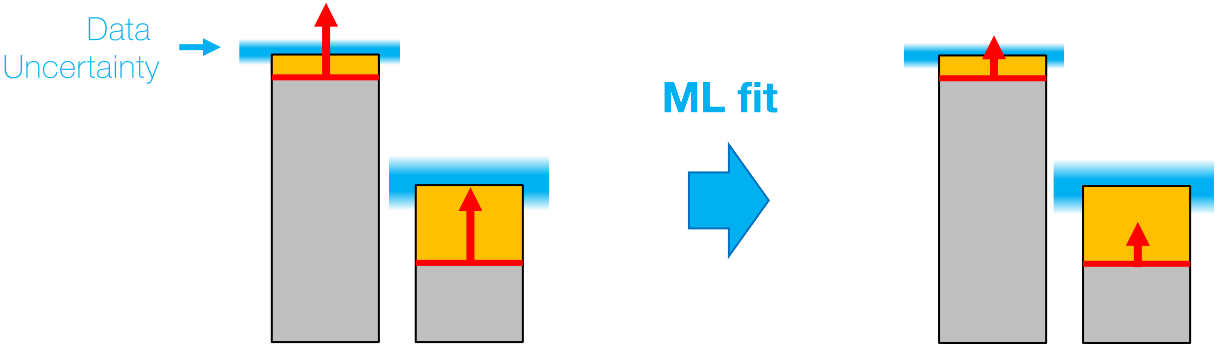
POI ∝ S1+S2	Conservative	Optimistic?	Very Optimistic
POI ∝ S1/S2	Very Optimistic	Optimistic?	Conservative

Implementing "appropriately conservative" uncertainties

- Beware propagation of **constraining effects of high-statistic measurements** through correlation modeling assumptions

'pre fit'

'post fit'



1 NP representing 10% bkg uncertainty
correlated effect in both bins

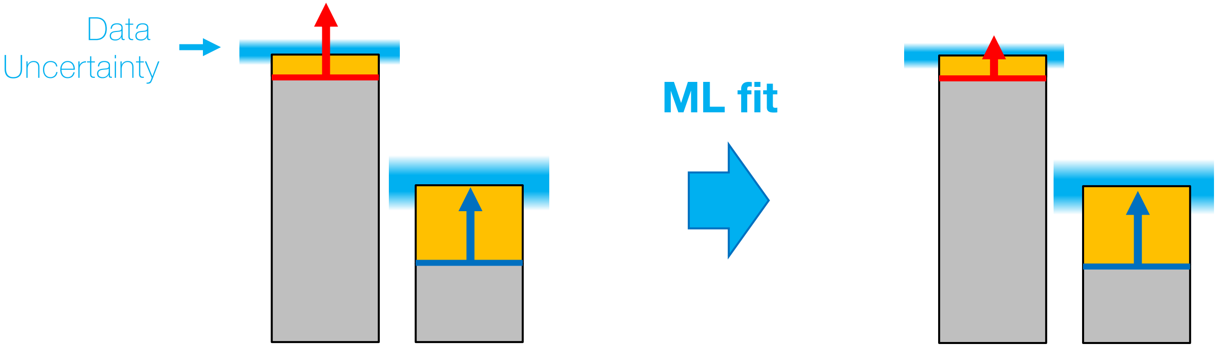
1 NP representing 5% bkg uncertainty
correlated effect in both bins

Uncertainty reduction in <u>both</u> bins through contraining power of bin 1

Implementing “appropriately conservative” uncertainties

- Beware propagation of **constraining effects of high-statistic measurements** through correlation modeling assumptions

‘pre fit’ *‘post fit’*



1 NP representing 10% bkg uncertainty
correlated effect in both bins

NP1 representing 5% bkg uncertainty
NP2 representing 10% bkg uncertainty

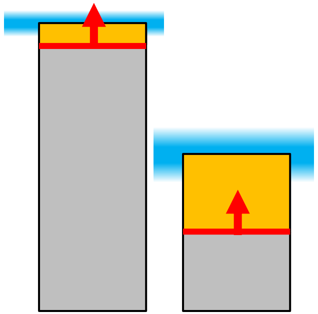
No uncertainty reduction in bin2 through constraining power of bin 1

Implementing "appropriately conservative" uncertainties

- Beware propagation of **constraining effects of high-statistic measurements** through correlation modeling assumptions

- If correlation assumption between regions well motivated
  → smart analysis strategy
- If no clear (physics) motivation behind correlation assumption then uncertainty reduction on POI may be spurious → attention needed!
- **Diagnostics** on constraining of NPs in data vital part of analysis

'post fit'



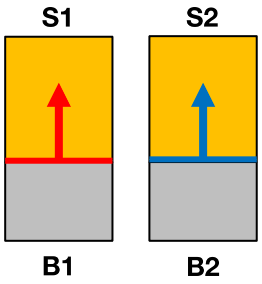
1 NP representing 5% bkg uncertainty
correlated effect in both bins

Uncertainty reduction in both bins through contraining power of bin 1

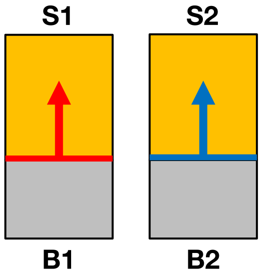
Implementing "appropriately conservative" uncertainties

- Beware propagation of **constraining effects of high-statistic measurements** through correlation modeling assumptions

- But beware that **decorrelating is not necessarily conservative**, effective √N reduction of 'sum POI's
- *Notably for many theory uncertainties nuisance parameters are 'proxies' with no proper connection to actual calculation*
- Notion of correlation model is ill-defined in many theory systematics,even discussion on what quantity uncertainty applies ('envelope' or 'integral')



NP: 10% bkg uncertainty
correlated modeling



NP1: 10% bkg unc. – bin 1
NP2: 10% bkg unc. – bin 2

Effective 7% bkg uncertainty
on POI ∝ S1+S2

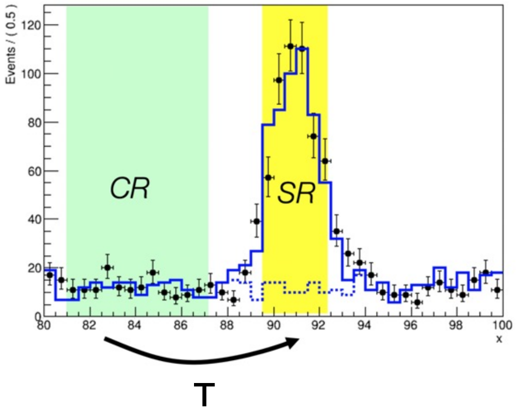
Overview

1. Source of uncertainty in the particle physics simulation chain
2. Anatomy of a typical LHC analysis – minimizing depedence on uncertainties
3. Statistical treatment of uncertainties - Frequentist concepts
4. Modeling of simulation uncertainties in the likelihood – general approach
5. Common issues with modeling of specific uncertainties
6. Summary & conclusion

Parameteric modeling of systematic uncertainties

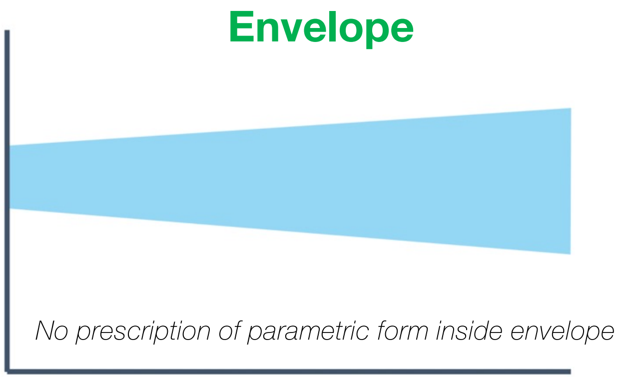
- Finding a parametric model for systematic uncertainties nuisance parameters that can over the 'true' distribution is the ultimate goal
  - But given that the true distribution is unknown, it is not a very practical goal
  - Instead aim to inventorize all known source of uncertainty, formulate parametric uncertainty model for them (response functions & subsidiary measurements) and implement them in the likelihood of a measurement

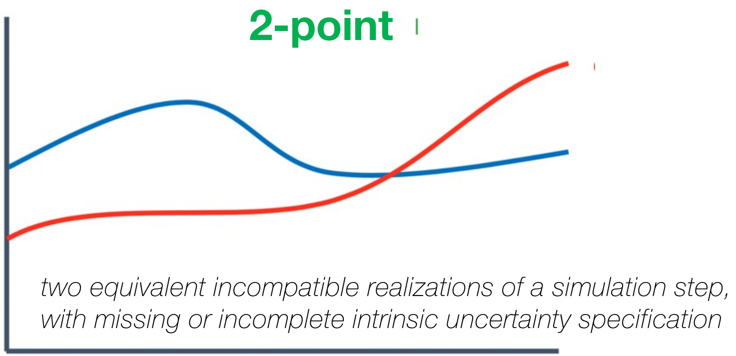
- **Easiest class of systematic uncertainties are those based on measurements**, but where the data are not part of the analysis dataset
  - Parametrization often physics- or detector motivated
  - Uncertainties on parameters have clear statistical interpretation
  - Main concern is any additional uncertainty on the 'transport factor' to the measurement space
  - **In HEP these are usually called 'good' systematics**



Parameteric modeling of systematic uncertainties

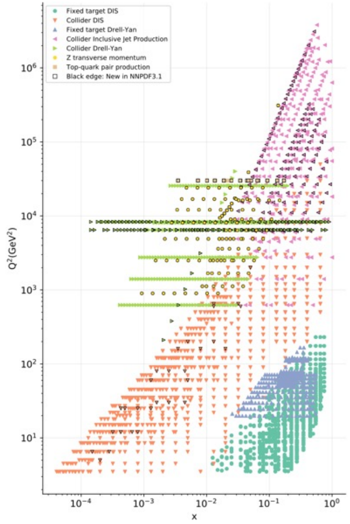
- **Difficult class of systematic uncertainties are those based on shortcomings of theory calculations**, with no relation to data
  - Only a general notion of the uncertainty is indicated, no meaningful parametric form of the uncertainty
  - No clear probabilistic intepretation of uncertainty prescription is provided
  - **In HEP these are usually called ‘ugly’ systematics**
- ‘Ugly’ systematic prescriptions generally come in one of two forms

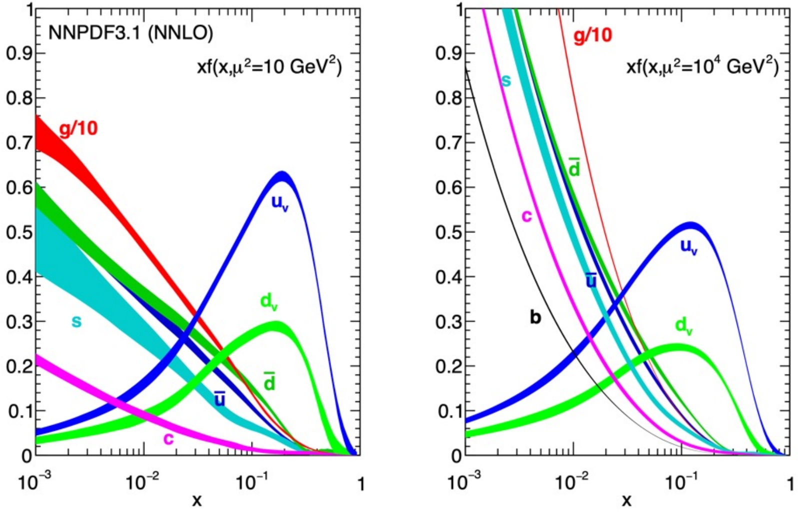


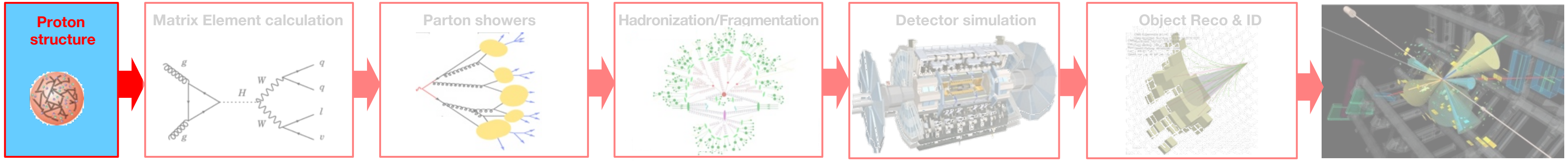


Proton structure – parton density models

- Proton density functions are *effectively an experimental measurement*
  → highly complex fits to large numbers of datasets



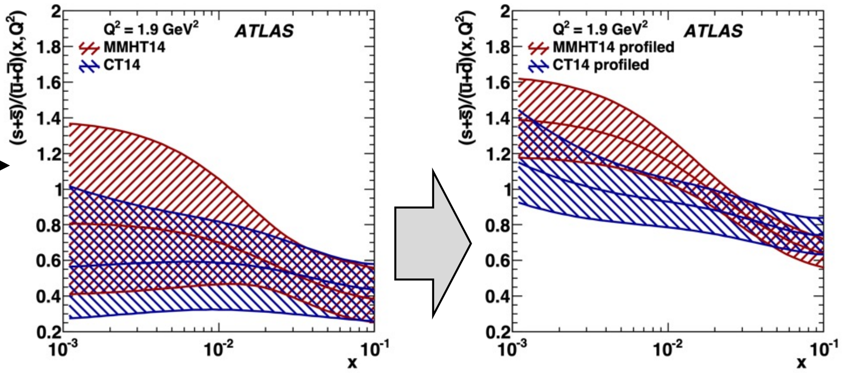


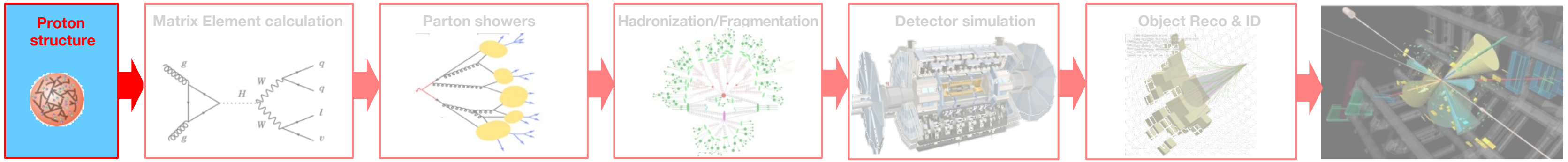


Proton structure – parton density models

- Proton density functions are *effectively an experimental measurement* → highly complex fits to large numbers of datasets
- **Detailed parametrization provided** (O(40) parameter Hessian – or replica sets, depending on PDF fitting collaboration
- Generally considered a 'good' systematic, parametric even used to constrain PDF uncertainties from fits to physics data
- But multiple PDF sets exist, that do not perfectly agree with each other

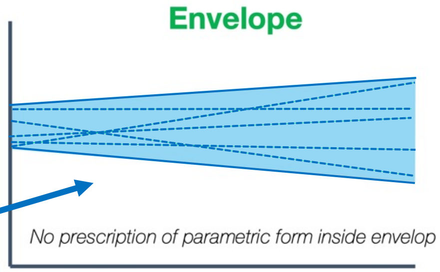
'prefit' 'postfit'

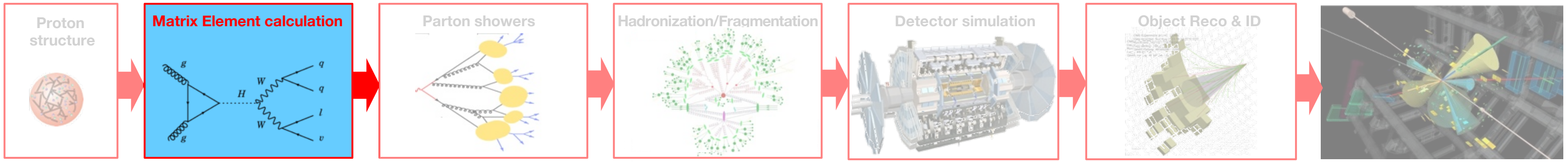




Hard Scatter – Missing Higher Orders

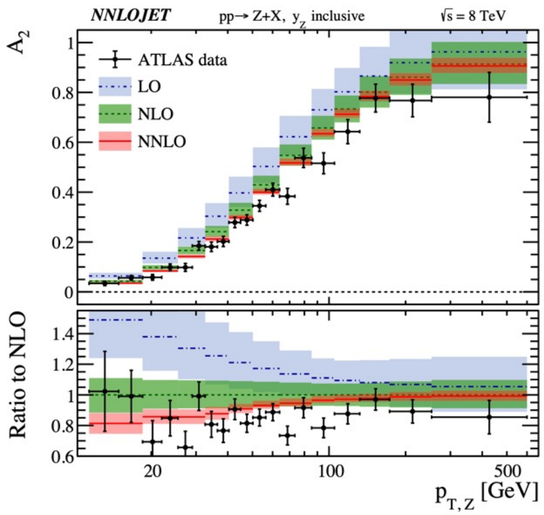
- Leading uncertainty in hard scatter amplitude calculate ('Matrix Element') is the **incompleteness of the perturbative expansion of the calculation**
  - Calculation is truncated in expansion loops or legs at some point and therefore incomplete.
  - Shape of missing part is – since it is presently uncalculable – unknown.
- Magnitude of effect of missing part of calculation can be *approximately* estimated through variation of 'scale parameters'
  - Factorization and renormalization scales ($\mu_{F,}\mu_R$) are unphysical parameters in the calculation, but the dependence of incomplete calculations on their value gives and indication of how far off the calculation is from the 'full answer'
  - Agreed evaluation procedure (empirical): consider for each separate 0.5x, 1x and 2x nominal (& product also in this range) → **7 ($\mu_{F,}\mu_R$) configurations**
  - **Envelope** spanned by 7 variants of calculation is **uncertainty prescription**
  - **No** assumptions on **correlation structure** inside phase-space should be made

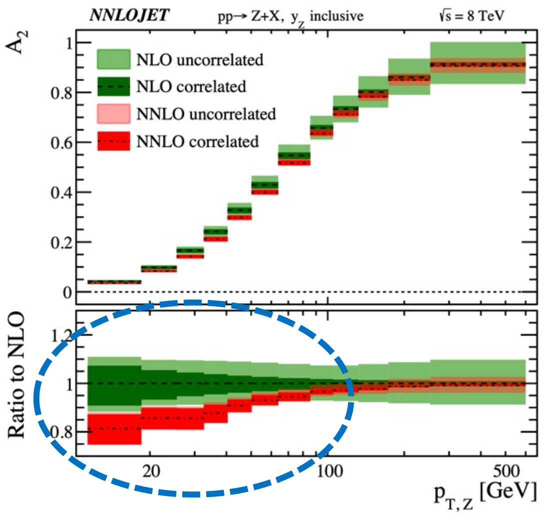




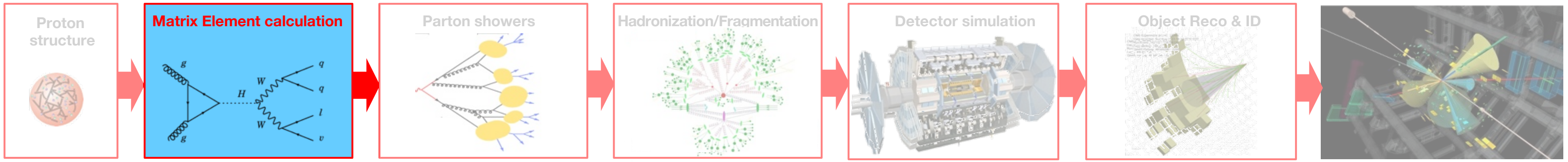
Hard Scatter – Missing Higher Orders

- Leading uncertainty in hard scatter amplitude calculate ('Matrix Element') is the **incompleteness of the perturbative expansion of the calculation**





Example:
Evolution of ME prediction with calculation order



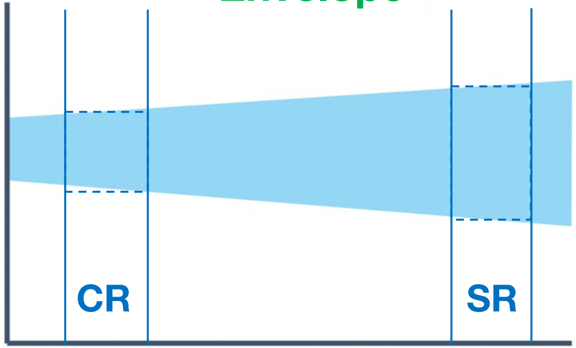
Hard Scatter – Missing Higher Orders

- Leading uncertainty in hard scatter amplitude calculate ('Matrix Element') is the **incompleteness of the perturbative expansion of the calculation**

Correlation model scale uncertainties

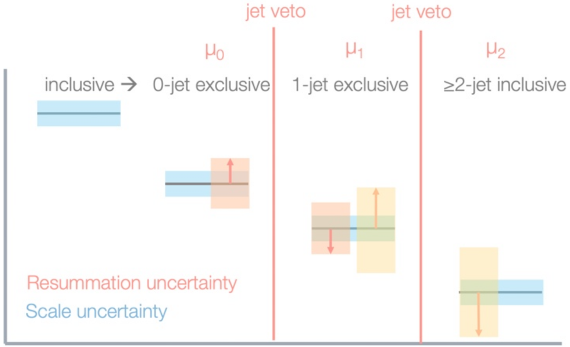
1 NP (constraining?), 2 NP (√2 red on int. unc?)

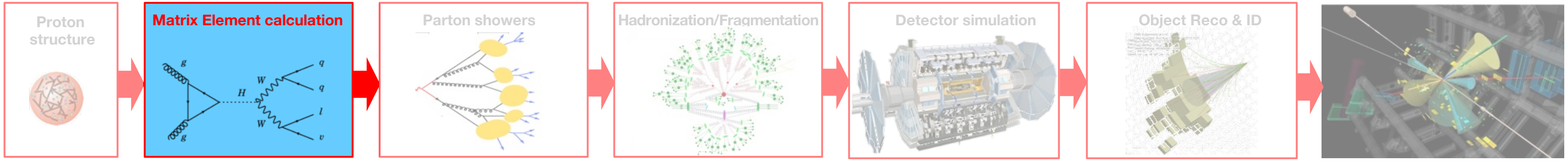
Envelope



Beware of special modeling situations

e.g. Stewart-Tackman prescription across jet-counting boundaries

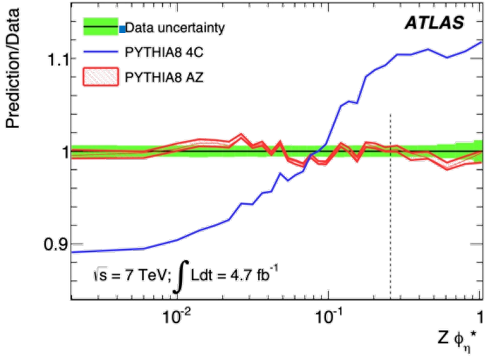




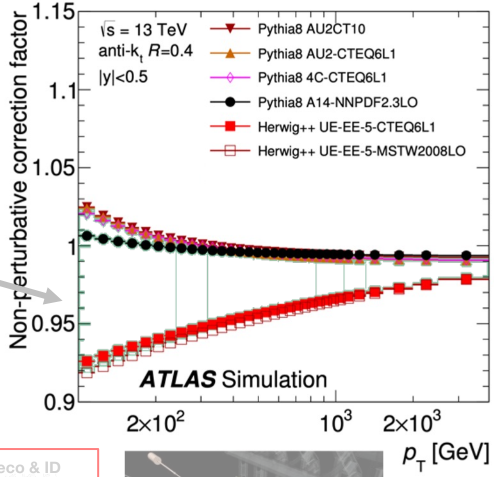
Showering Monte Carlo

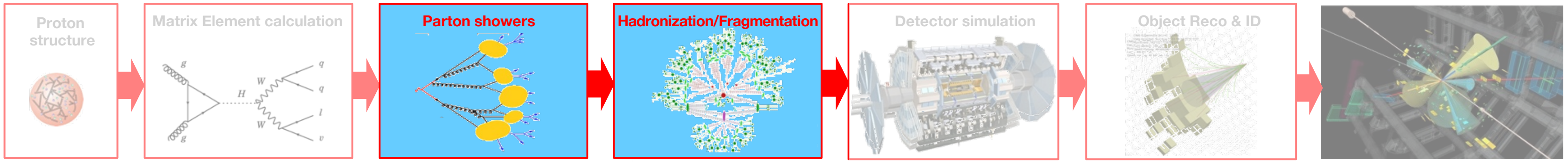
- Parton showers and Hadronization/Fragmentation typically integrated into a single package
- **Multiple equivalent implementations** available (Herwig, Pythia, Sherpa…)
  - Non-perturbative physics process is (semi-)empirically modeled, and extensively tuned to available data
  - **No (complete) set of internal systematic uncertainty prescriptions available for packages**
  - Prediction results can strongly disagree between packages (and sometimes even within version numbers of the same package)

Example 1
Pythia retuning or early LHC data



Example 2
Differences in Herwig/Pythia predictions for jet-gap fractions



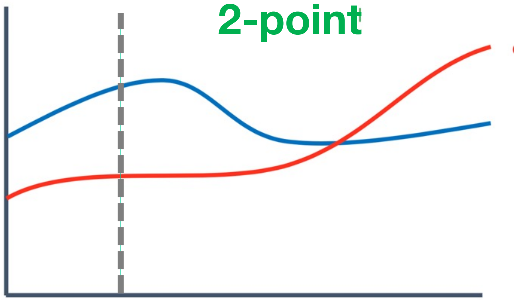


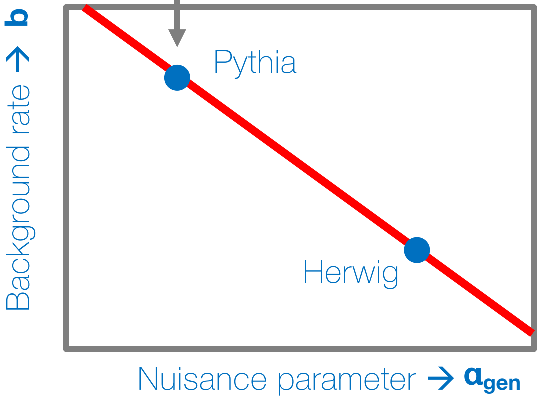
Showering Monte Carlo

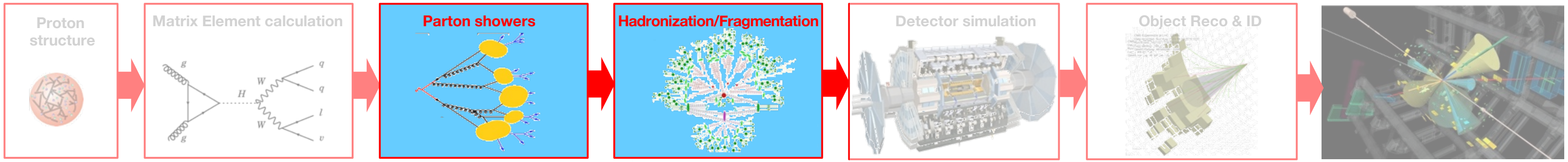
- Given that dominant effect is difference between packages, usually a **'2 point systematic'**
- Parametric implementation in likelihood models have *additional pitfalls*.
- For *scalar predictions* (counting experiments),

$$L(N_{SR}) = \text{Poisson}(N_{SR}|s+\mathbf{b}(\boldsymbol{\alpha})) \cdot \mathbf{Gauss}(\mathbf{0}|\boldsymbol{\alpha},\mathbf{1})$$

- The **response function** is trivial.
- The **subsidiary measurement** is not necessarily
  - *Common choice is a Gaussian centered on one prediction*, with alternative generator at 1 sigma away (symmetrized)
  - Probabilistic interpretation assigned to generators are usually assumptions

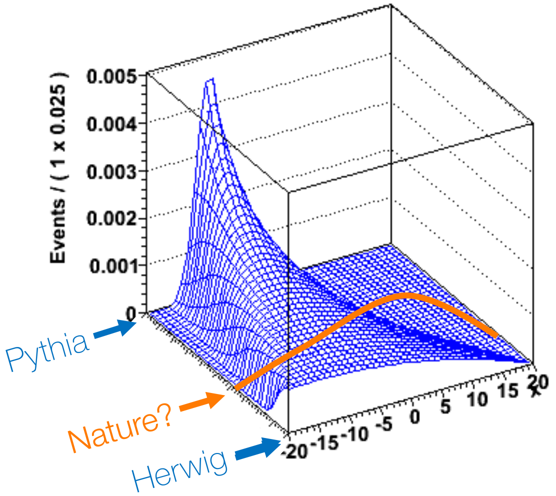


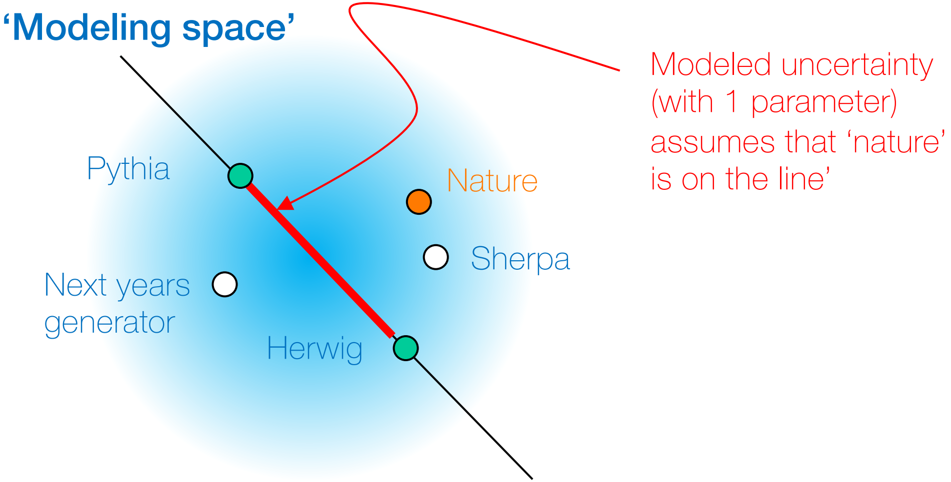


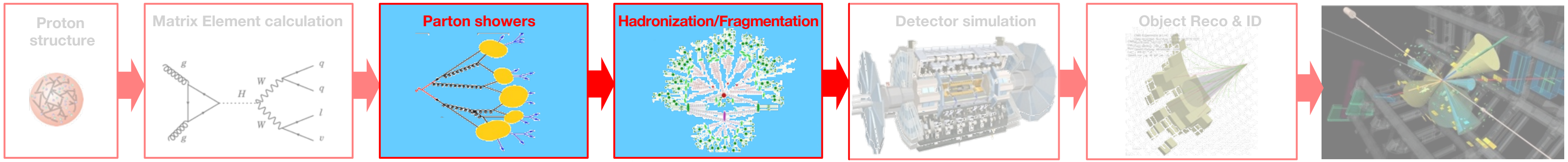


Showering Monte Carlo

- Modeling of 2-point systematics for **differential predictions** (shapes) fraught with many more issues
- The **response function** is *also not trivial*. Brute-force 1-parameter shape interpolation common choice, but *no guarantee that has the flexibility to cover Nature or alternative predictions*



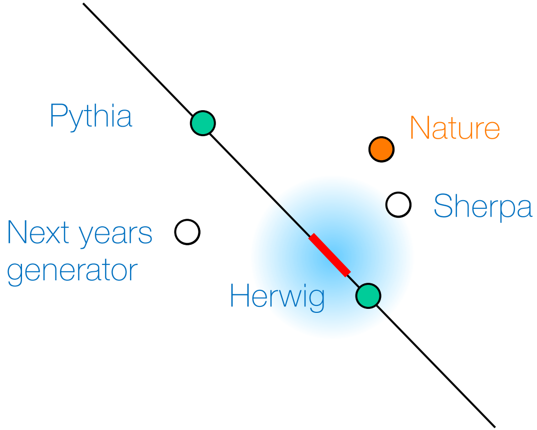


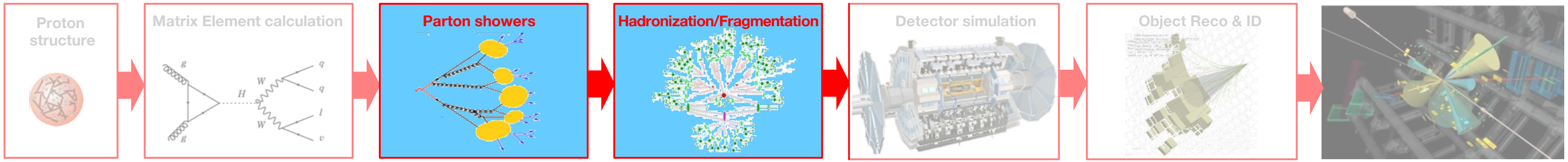


Showering Monte Carlo

- Modeling of 2-point systematics for **differential predictions** (shapes) fraught with many more issues
- **Constraining** of 2-point systematic nuisance parameters 'doubles down' on assumption that all modeling uncertainty can be captures with an empirical 1-parameter model. *Rarely justifiable*
- Yet constraining of MCgenerator systematics from the data often occurs in analysis → almost all SRs, CRs are sensitive to it
  - Introducing separate NPs for regions helps, but is not ideal
- Way forward is development of full prescription of modeling uncertainties for each generator
  - There is progress, but slow

'Modeling space'



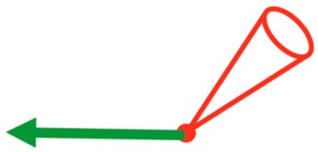


Experimental systematic uncertainties

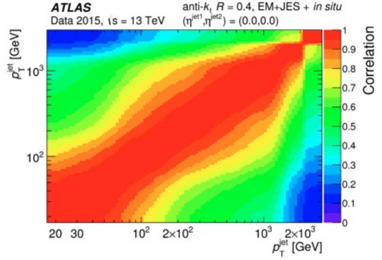
- Experimental systematic uncertainties relate **data/simulation differences**.
  - **Almost always based on measurement of (high-statistics) control samples**
  - Data/simulation differences removed (to 1st order) through correction functions
  - Measurement uncertainties propagated as experimental uncertainties

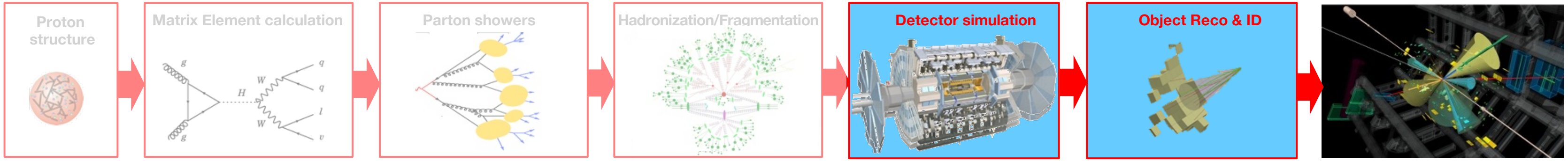
- Experimental systematics **mostly of the 'good' type**
  - Parametric structure largely motivated by physics/detector considerations
  - Uncertainties on parameters have clear probabilistic interpretation
  - **But beware some 'ugly' corners.**
    Difficult simulation uncertainties (b-quark fragmentation)
    may influence measurement of certain experimental uncertainties

Measurement on calibration data (e.g. jet-γ balance)



Calibration with parameterization and correlation structure motivated by underlying measurement

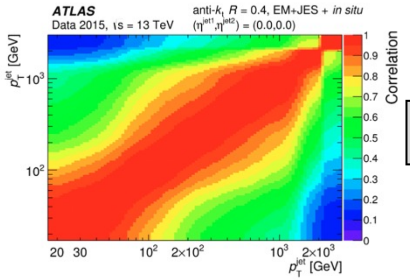




Experimental systematic uncertainties

- But **beware of (intentional) limitations to accuracy** of nuisance parameter model
  - Underlying model of calibration uncertainties often highly complex (>100 NP no exception)
  - But for many analyses high level of complexity not needed (e.g. a 1-bin counting experiments can use 1 NP)
  - (Multiple) simplified representations of uncertainty model are often provided

Original representation



Eigenvector transformed

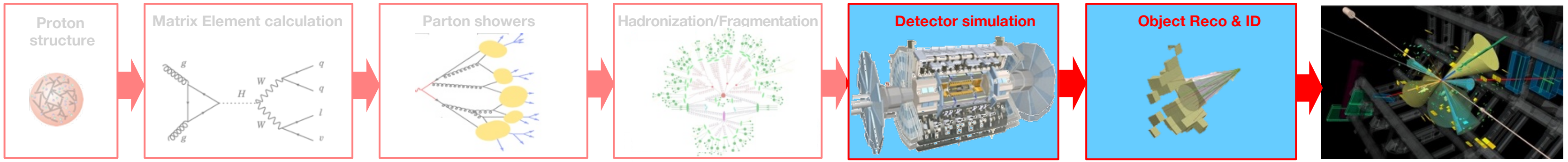
$$V' = \begin{pmatrix} V_{11} & & & \\ & V_{22} & & \\ & & V_{33} & \\ & & & V_{44} \end{pmatrix}$$

Simplified model (truncated EV basis)

$$V'' = \begin{pmatrix} V_{11} & \\ & V_{mm} \end{pmatrix}$$

Merger of weakest modes

- Beware that *not everything is quantified or measured*
  - For example “correlation of systematic uncertainties between 65% and 75% b-tagging operating points” may not be known.



Validation & Diagnostics

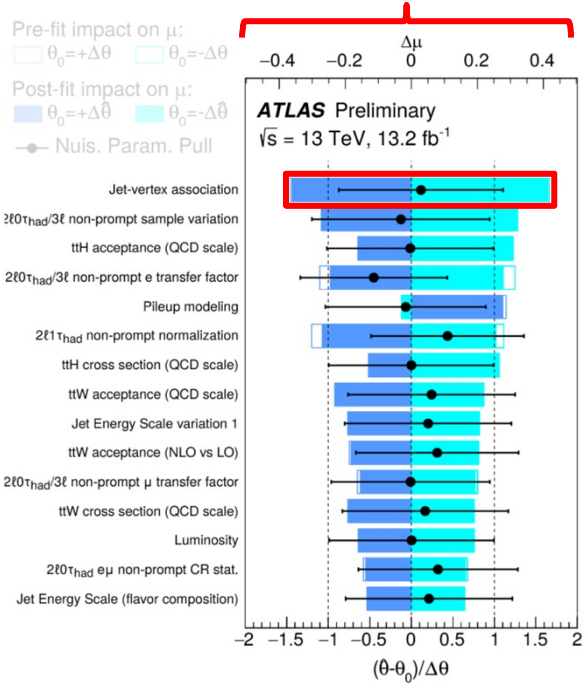
- **Extensive diagnostics of (often complex) fits to LHC data crucial for validation**

Nuisance parameter ranking plots

Scale for POI impact

Relative impact of ±1σ NP variation on POI

NPs ranked in order of impact

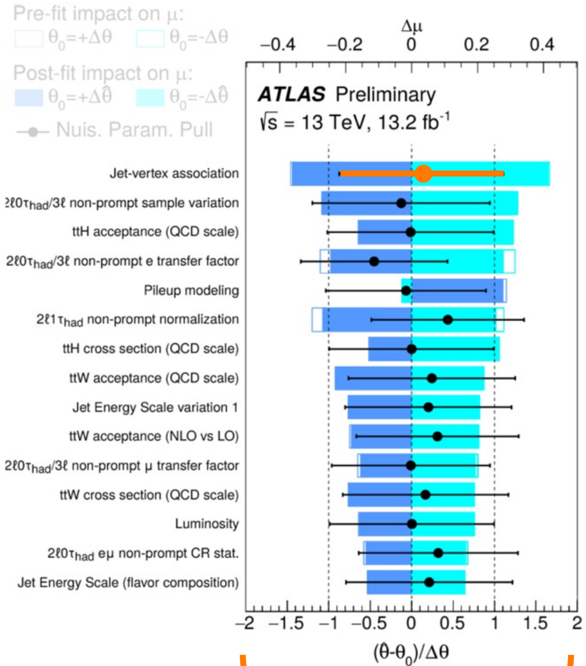


Validation & Diagnostics

- **Extensive diagnostics of (often complex) fits to LHC data crucial for validation**

→ **Nuisance parameter ranking plots**

NPs ranked in order of impact



Pull of NP in fit to observed data

Scale for NP pull

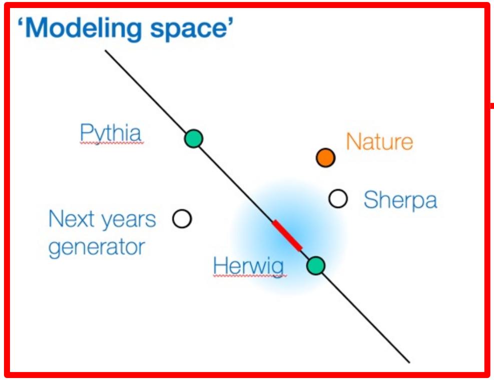
value 0 → no bias

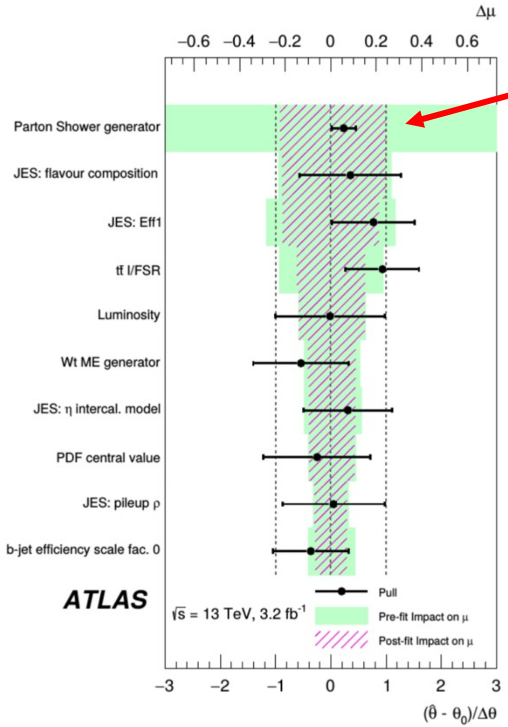
uncertainty ±1 → no reduction of uncertainty w.r.t input spec.

Validation & Diagnostics

- **Extensive diagnostics of (often complex) fits to LHC data crucial for validation**

Nuisance parameter ranking plots



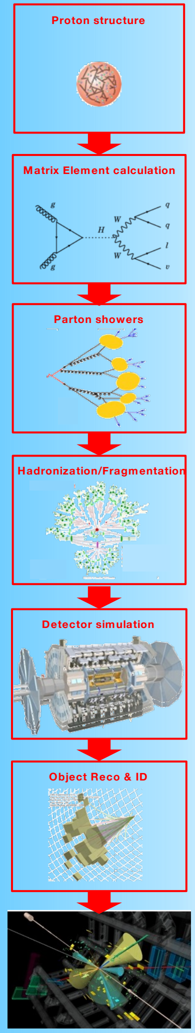


NP with 10x reduced uncertainty

Highlights point of attention:
difficult modeling uncertainty strongly reduced in fit to data
→ *investigaye*

Overview

1. Source of uncertainty in the particle physics simulation chain
2. Anatomy of a typical LHC analysis – minimizing depedence on uncertainties
3. Statistical treatment of uncertainties - Frequentist concepts
4. Modeling of simulation uncertainties in the likelihood – general approach
5. Common issues with modeling of specific uncertainties
6. Summary & conclusion



Summary & conclusions

- Simulation of LHC events incredibly powerful tool, driving analysis design & inference
  - Despite a decade of use, with advances in tools & methods, and very extensive validation efforts, still many corners of phase-space where modeling is quite imperfect
  - Some are known since years and simply hard to fix, but new ones are being discovered all the time as new analysis rely ever more on the details of simulated events. Use of ML/AI will accelerate this trend
- Extensive strategies exist to minimize dependence on simulation modeling uncertainties
  - Clever formulation of analysis goals ('fiducial regions'), clever use of theoretical predictions ('ratio corrections')
  - Extensive use of control regions to validate and correct for any mismodellings at the analysis level
  - Extensive use of object-level correction functions correct for data/simulation disagreements
- Detailed modeling of simulation systematics in inference stage indispensible
  - Fairly straightforward for 'good' type of systematics (based on measurement)
  - Thorny issues on definition and interpretation for 'ugly' type of systematics (mostly of a theoretical nature)
- Validation of results & statistical models indepensable for robust results
  - Exploitation of simplistic parametrizations of 'ugly' systematics can easily lead to spurious improvements of results
  - But careful design of analysis strategy can help to avoid 'getting stuck' being dominated by 'ugly' systematics