Experimental HEP Analyses

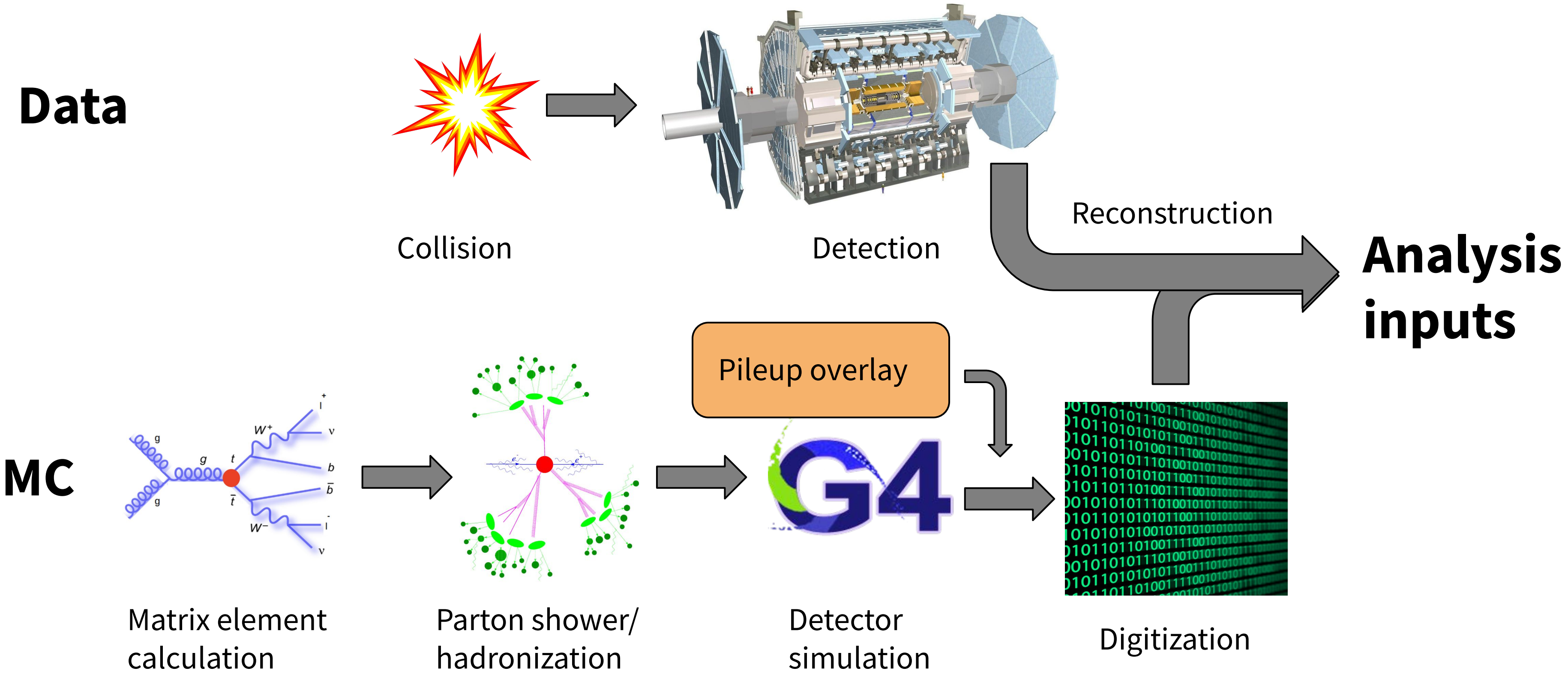
Overview of analyses

- Signal and background simulation
- Object definition/selection
- Event preselection
- Selection optimization
- Background estimation and validation
- Systematic uncertainty evaluation
- Statistical analysis
    - Fitting, bump hunting or setting limits
- Interpreting results

Monte Carlo

- Most analyses involve comparing collision data to simulated data
  - Monte Carlo (MC) method used
- Significant tuning is applied to MC sample parameters to match data
- MC generation is done in discrete steps:
  - Event generation - exact calculations of interactions and decays given initial conditions
  - Parton showering/hadronization - parton fragmentation and formation of hadronic showers
  - Detector simulation - parameterized or stepwise simulation of particles interacting with detector material and depositing energy
  - Pileup overlay - superimpose pileup events on single collision simulation
  - Digitization - conversion of deposited energy to digital signals
- Many third party tools used for first 2 steps (Pythia, Herwig, MadGraph, etc.)
- Analysis teams design and test commands to simulate signal processes
  - Request sent to central production to ensure correct settings for full sample
- Common Standard Model processes managed centrally

Data flow



Data

Collision

Detection

Reconstruction

Analysis inputs

MC

Matrix element calculation

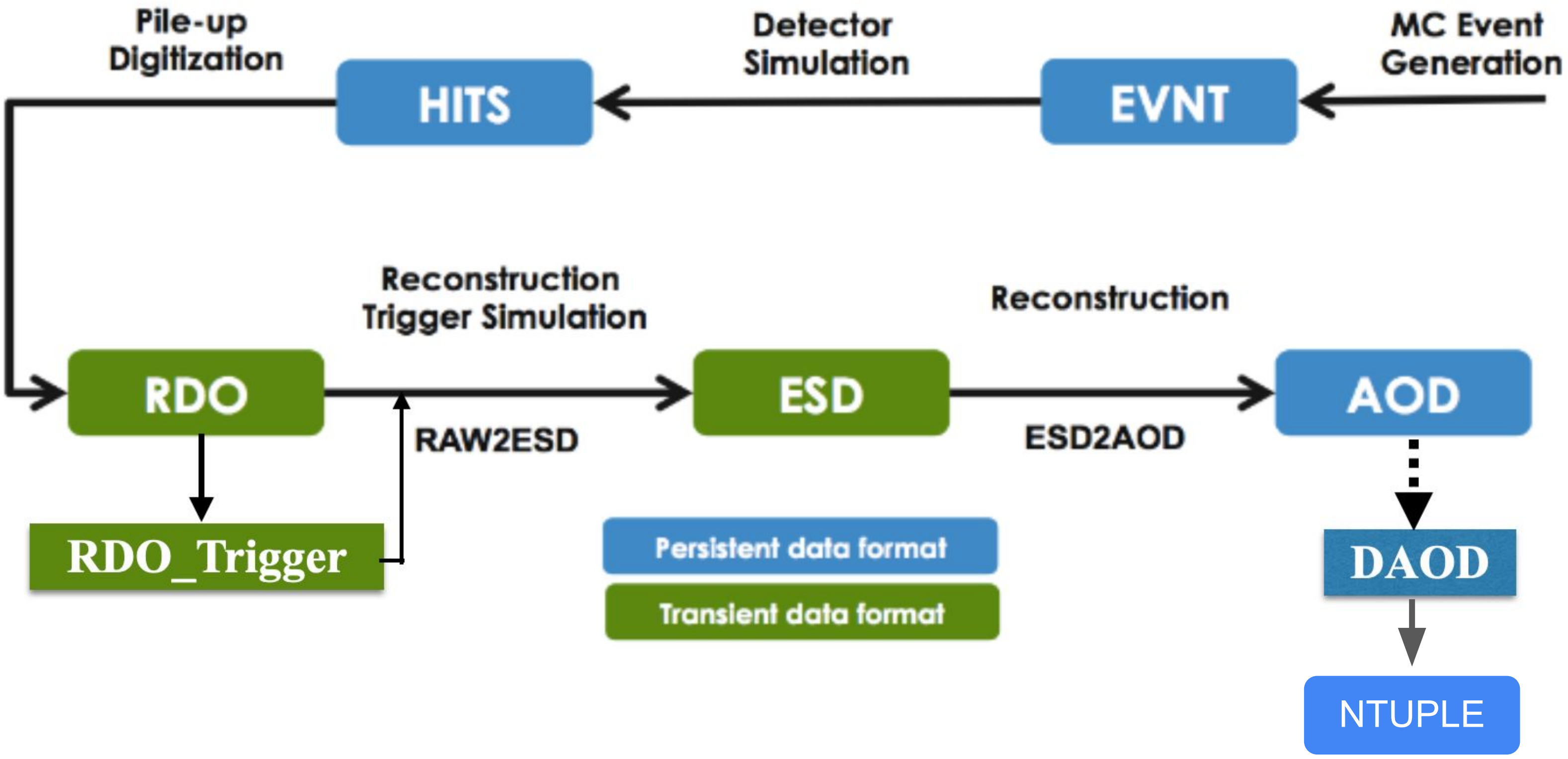
Parton shower/ hadronization

Pileup overlay

Detector simulation

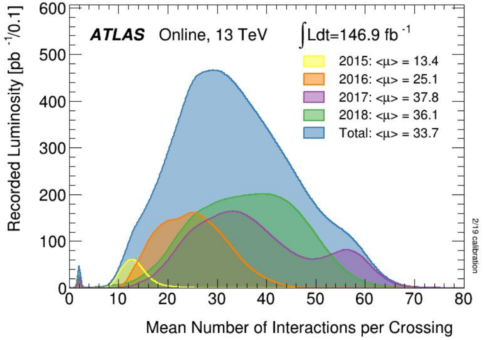
Digitization

Data formats



Applying corrections

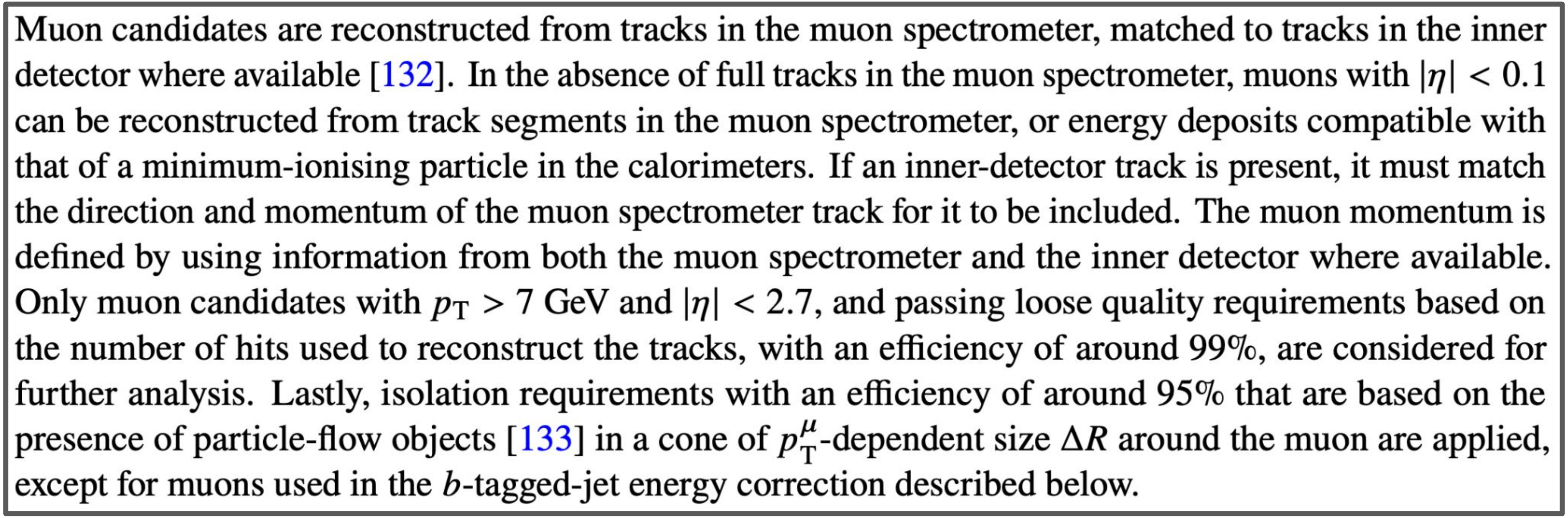
- It is critical that MC describes data well
  - Disagreements between MC and data can result in false observations
- Modeling is corrected/validated in phase space where no signal is expected
- Pileup reweighting:
  - Pileup condition profile of collisions is assumed for MC production
    - Allows MC to be produced before collision data is collected
  - MC events are reweighted to match the data pileup distribution
- Scale factors:
  - Detector response is not perfectly modeled
  - Object calibrations and characteristics can differ from data to MC
  - Per-object scale factors applied to reweight MC to match data response



Object selection

- Analysis is performed using physics objects
  - Electrons, photons, jets, etc.
- Necessary to define what constitutes an object for the analysis
  - Important to harmonize definitions between analyses that will be combined
- CP groups provide recommendations that need to be followed
- Minimum p_T, η range, and quality criteria for validity of recommendations
  - Limited by detector design and techniques to derive calibrations/uncertainties
- Stricter criteria can be used for analysis reasons such as trigger thresholds
- Working points (WPs) are provided for object identification
  - Loose, medium, tight, etc.
  - Tighter WPs reject more background as well as signal
- Choose overlap removal priority based on analysis signature

Example: muons



Muon candidates are reconstructed from tracks in the muon spectrometer, matched to tracks in the inner detector where available [132]. In the absence of full tracks in the muon spectrometer, muons with $|\eta| < 0.1$ can be reconstructed from track segments in the muon spectrometer, or energy deposits compatible with that of a minimum-ionising particle in the calorimeters. If an inner-detector track is present, it must match the direction and momentum of the muon spectrometer track for it to be included. The muon momentum is defined by using information from both the muon spectrometer and the inner detector where available. Only muon candidates with $p_{\mathrm{T}} > 7$ GeV and $|\eta| < 2.7$, and passing loose quality requirements based on the number of hits used to reconstruct the tracks, with an efficiency of around 99%, are considered for further analysis. Lastly, isolation requirements with an efficiency of around 95% that are based on the presence of particle-flow objects [133] in a cone of p_{T}^{μ}-dependent size ΔR around the muon are applied, except for muons used in the b-tagged-jet energy correction described below.

Event preselection

- Preselection refers to basic selection criteria that define your signal signature
- Begins with one or more triggers
  - Select triggers with lowest thresholds that are sensitive to signal
  - Multiple triggers can target different regions of phase space (1 high-p_T μ vs 2 med-p_T μ)
- Select basic set of objects that define analysis signature
  - Example: 2 opposite sign leptons, ≥ 3 jets, MET > 100 GeV, and 0 b-tagged jets
- Basic kinematic selections to remove significant backgrounds
  - Example: $|m_{\ell\ell} - 91.2\ \mathrm{GeV}| > 10$ GeV to remove most $Z \to \ell\ell$ events
- Criteria are often placed on objects ordered by p_T (leading means highest p_T)
- Multiple channels can be defined, such as electron and muon channels
  - Generally useful because background composition can differ

Regions

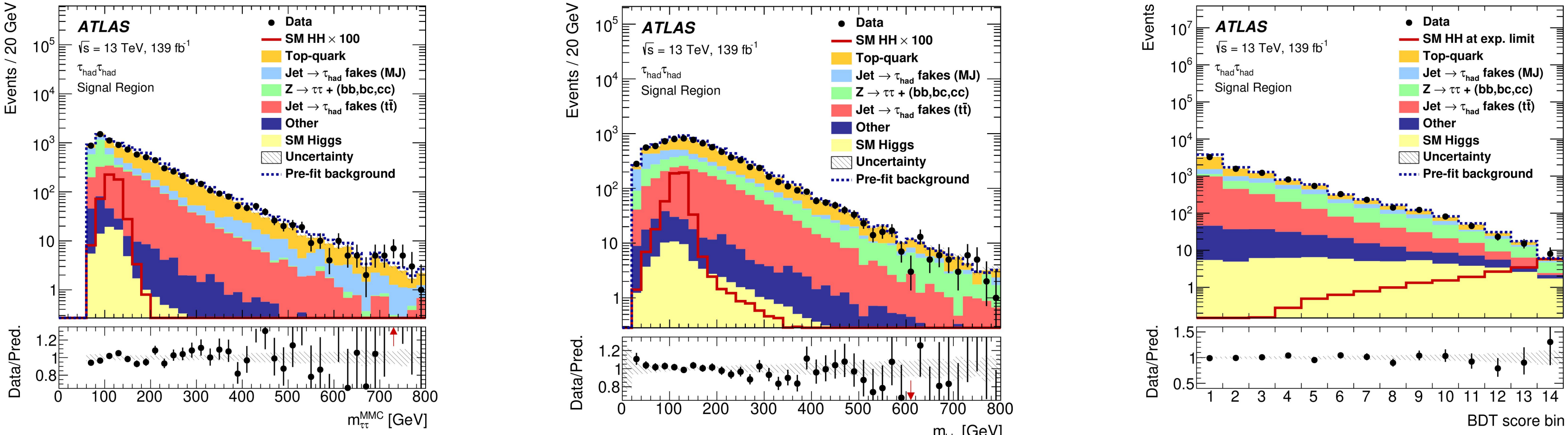
- Multiple regions of phase space are used in an analysis
  - A region is defined with a set of selection criteria (cuts) on various object/event quantities
  - Regions should usually be defined to be mutually exclusive (orthogonal)
- Signal region (SR):
  - Region which contains the majority of the expected signal
  - Where the final fit or statistical analysis is performed
- Control region (CR):
  - Region that is enriched in some background or depleted of signal
  - Useful for constraining/correcting/deriving background estimate
- Validation region (VR):
  - Region that is reasonably close to the signal region
  - Used to validate methods involving CRs and to derive associated uncertainties

Analysis optimization

- Preselection is usually insufficient to maximize sensitivity to signal
- Signal significance can be used as a proxy for actual sensitivity. One version:

$$z_0 = S/\sqrt{B}$$

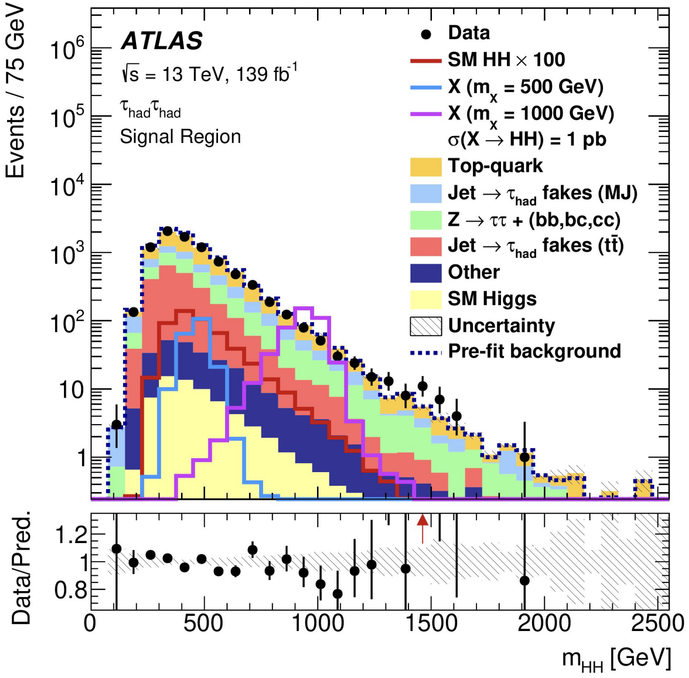
  - S is number of expected signal events, B is number of expected background events
- Can be done using individual variables or multivariate method outputs
  - Machine learning techniques are very useful for distinguishing signal from background



Background estimates

- Any event selection will retain background events
- It is crucial that the background be well-modeled
    - Incorrect modeling can lead to false differences w.r.t. collision data
- Begin by creating a list of all SM processes that can give same signature
    - Account for possibility of particles being incorrectly reconstructed/identified
- Significant effort is spent thoroughly verifying background modeling
- Typically a combination of MC simulation and data-driven methods are used

Example: HH→bbττ



Top quark processes

Shape from MC and normalization from fit

Z→ττ + heavy flavor

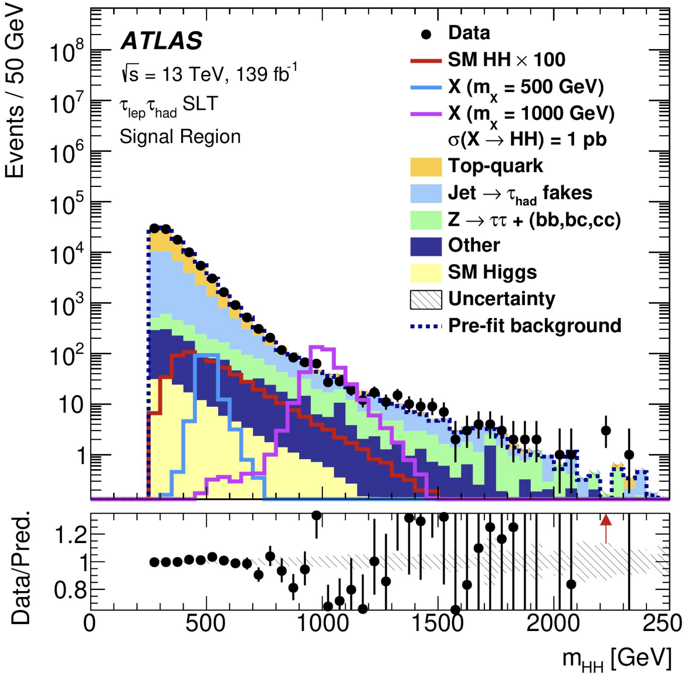
Shape from MC and normalization from Z→ee/μμ+HF control region

Single Higgs and others

Estimate from MC

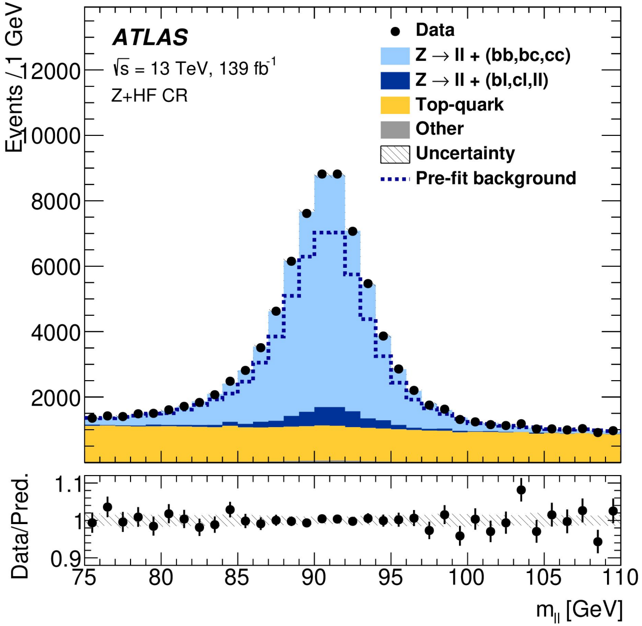
Fake τ backgrounds

Data-driven estimate



Data-driven MC corrections

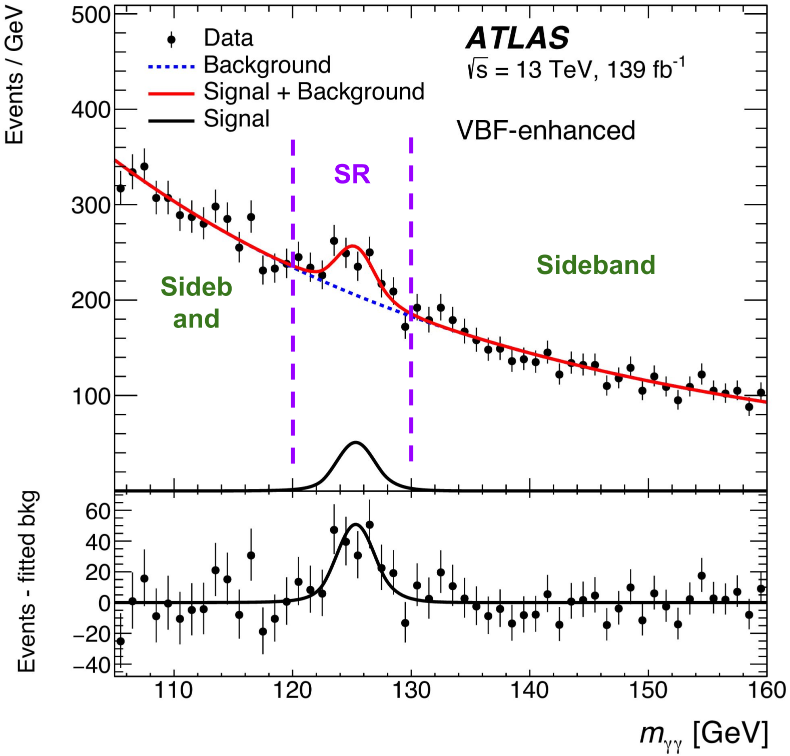
- MC can sometimes fail to describe some aspect of a background process
  - Normalization in analysis phase space or parton flavor composition
- Data is used to fix relative or absolute normalization of MC sample
- Example: Z(→μμ) + b-jets can be difficult to model
  - Total σxBR of Z→μμ is known, but the subset including a b-jet is poorly constrained
  - Use CR that is dominated by background process of interest
  - Float normalization of background(s) to fit to data



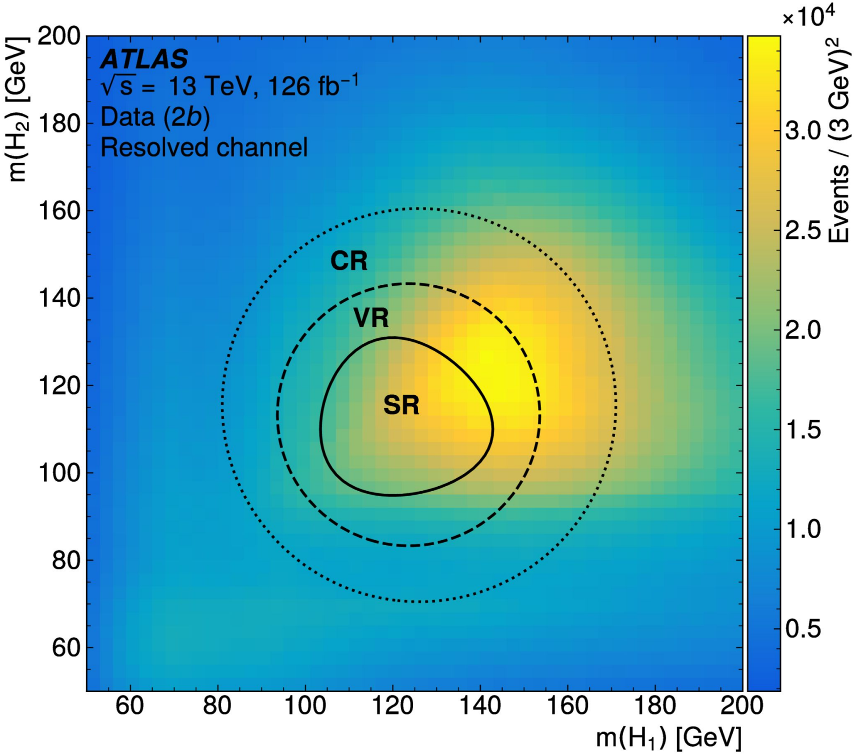
Data-driven background estimates

- It is often possible/necessary to estimate background directly from data
  - Avoids issues of detector response modeling and background composition in MC
- Sideband fits:
  - If background shape is expected to be smooth in SR, it can be derived from nearby data
  - Define sideband control regions with very little expected signal
  - Fit smooth function to data in sidebands and interpolate to signal region
  - Additional validation regions can be used to validate extrapolation and derive uncertainties
- Fake factors and fake rates:
  - MC generally models detector response to incoming particles well
  - It poorly models response for incorrectly identified (fake) objects
  - Composition of backgrounds with fake objects is very difficult to get right
  - Use data with objects that fail identification/quality criteria to estimate fakes background
  - Techniques involve multiple control and validation regions

Sideband fits



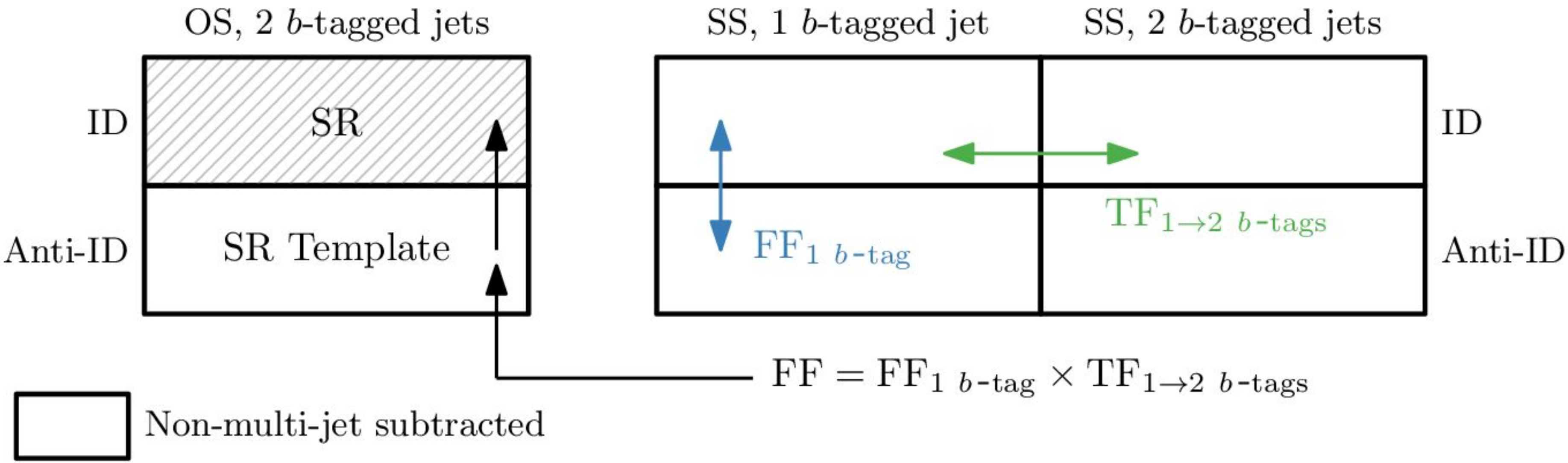
H→γγ



HH→4b

Fakes estimates

$\tau_{\mathrm{had}}\tau_{\mathrm{had}}$ channel

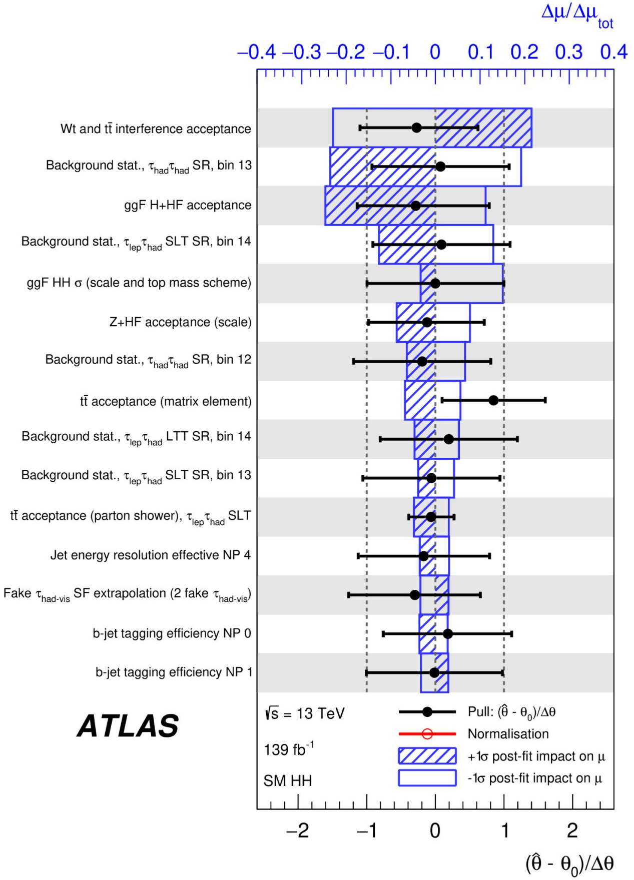


Evaluating uncertainties

- Careful evaluation of uncertainties is crucial for scientific rigor
- Statistical uncertainties
    - Fluctuations in data that would change with a repetition of the experiment (different dataset)
- Theoretical uncertainties
    - Uncertainty on any theoretical quantity such as cross-sections, PDF, hadronization scales, etc.
- Systematic uncertainties
    - Errors that will not change with a different dataset such as:
    - Differences in detector response between collision data and MC (e.g., calibrations)
        - Carefully derived by CP groups and provided as recommendations
    - Analysis techniques such as method to estimate fakes background
        - Often derived in CR as non-closure (difference between MC and data) in VR
    - Choice of MC generators
        - Comparison between two different generators - not actually based on data

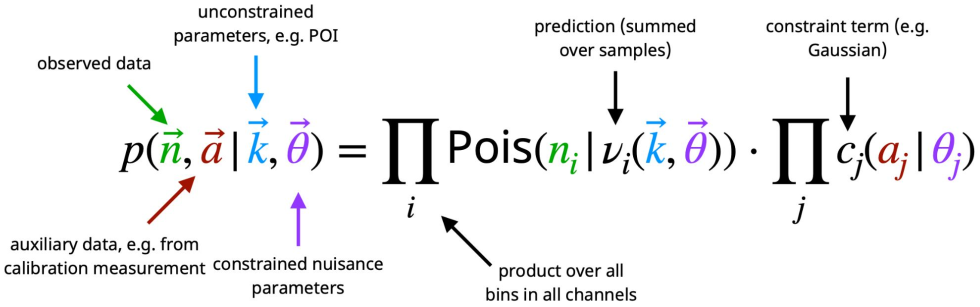
Using systematic uncertainties

- Usually up/down variations are used
    - Some can only be varied in one direction
- Correlated systematics should be varied up or down together, increasing impact
    - E.g., jet energy scale for multiple collections
- Uncorrelated systematics are independent
    - E.g., electron energy and b-tagging efficiency
- Each systematic differs in impact
- Effects below a threshold can be removed or combined into a single systematic
- Used as nuisance parameters (NP) in statistical analysis



Statistical analysis

- Performed in some binned distribution (invariant mass, number of events, etc.)
- Probability density function created in each bin using all contributions



$$p(\vec{n}, \vec{a} \mid \vec{k}, \vec{\theta}) = \prod_i \mathrm{Pois}(n_i \mid \nu_i(\vec{k}, \vec{\theta})) \cdot \prod_j c_j(a_j \mid \theta_j)$$

observed data

unconstrained parameters, e.g. POI

prediction (summed over samples)

constraint term (e.g. Gaussian)

auxiliary data, e.g. from calibration measurement

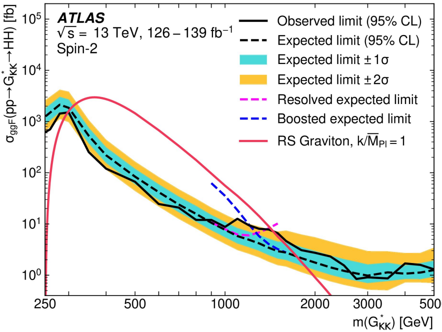
constrained nuisance parameters

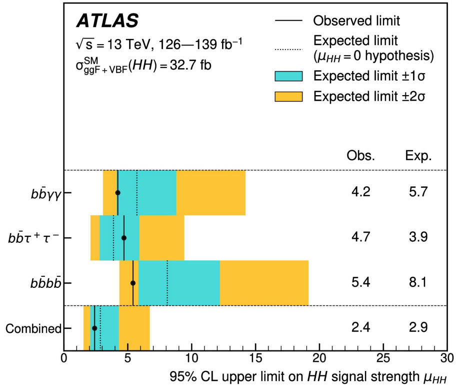
product over all bins in all channels

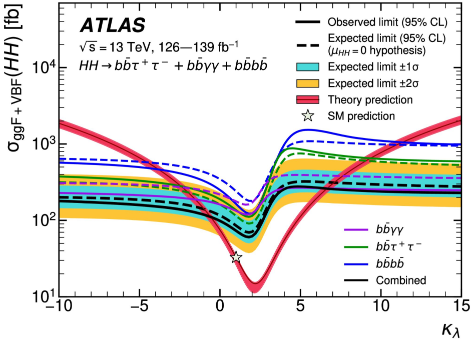
- Uses information from signal and control regions
- Signal and (some) background normalizations allowed to float
- Nuisance parameters represent a penalty for varying systematic by too much
- Vary normalizations and nuisance parameters to maximize likelihood function
- Discovery significance or limits can be calculated from likelihood

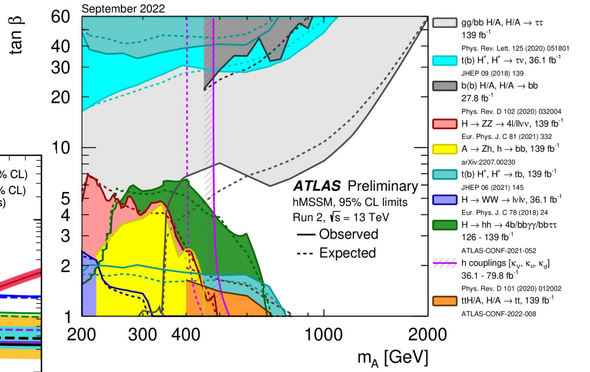
Interpreting results

- Results of statistical analysis need to be interpreted
- Almost all searches result in limits
- Cross-section limits on SM processes are often normalized by prediction
- Can set limits on parameters other than cross-section
- Exclusions of regions of theory parameter space









ATLAS approval procedures

- Analysis kickoff, formation of team and creation of Glance entry
  - Make analysis known to the collaboration
- Regular updates in subgroup meeting
  - Feedback from a larger audience
- Editorial Board (EB) request - need complete INT note
- Regular EB meetings to discuss all aspects of analysis
- Subgroup approval to unblind analysis
- Group approval of analysis and paper draft
- Circulation of paper draft to entire collaboration for feedback
- Paper approval meeting to get sign-off from Physics Coordination (PC)
- Language editor feedback
- Spokesperson sign-off and journal submission
- Feedback from journal reviewers and finally publication