BSM needs on MC matters: feedback from CMS

Valentina Dutta

(University of California, Santa Barbara)

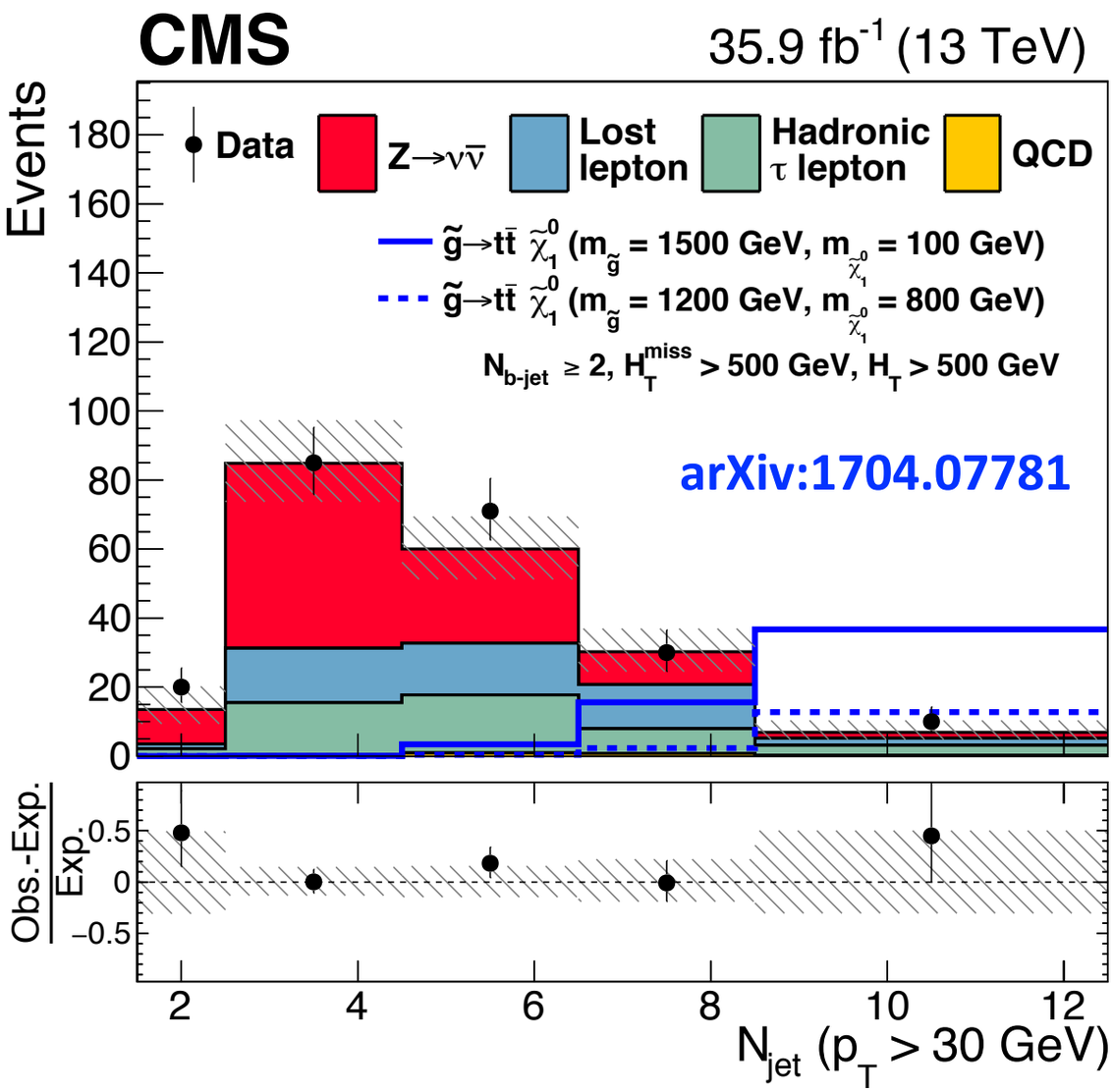
ATLAS-CMS Monte Carlo Generators Workshop

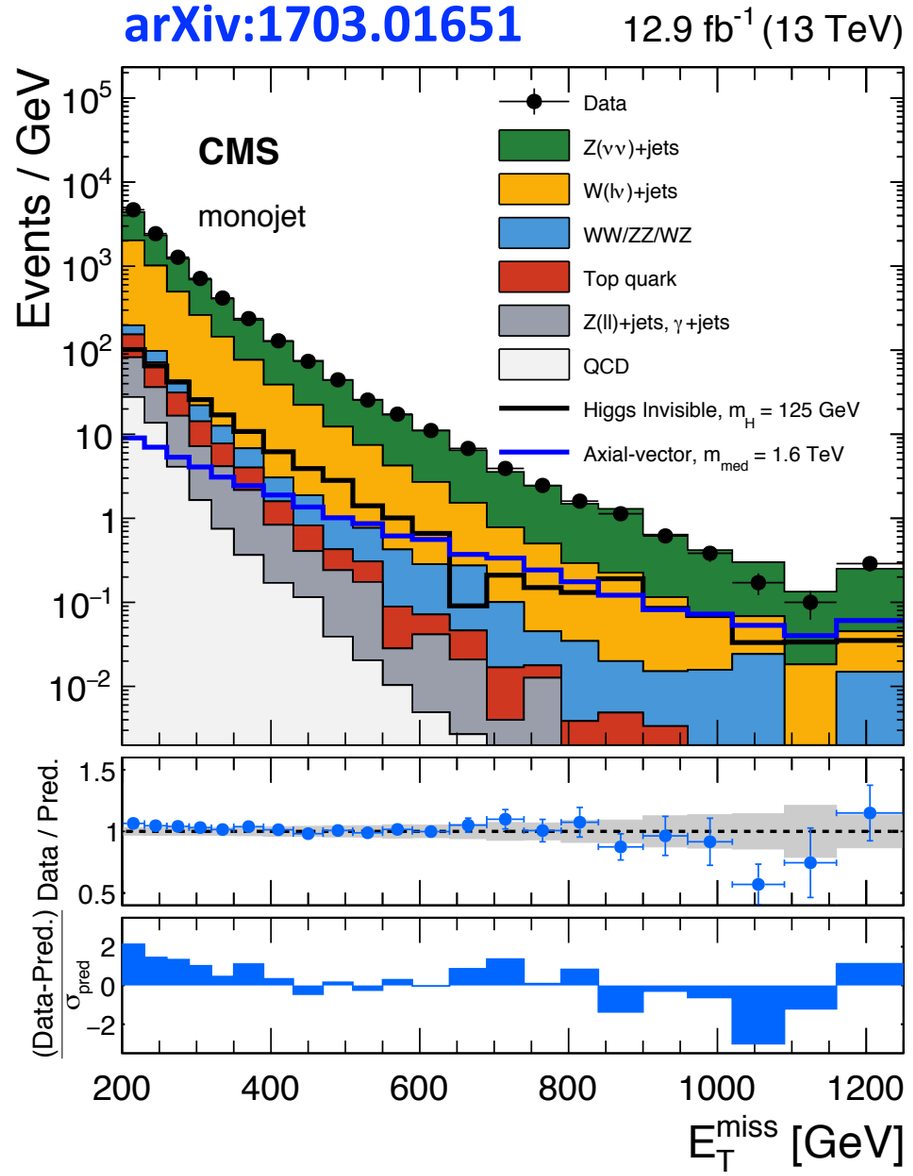
May 5, 2017

Backgrounds for BSM searches

Plenty of discussion of most common SM background processes in previous sessions

For many BSM searches, relevant areas of phase space are in tails of distributions like N(jets), V pT … challenging to model





Control regions and extrapolations

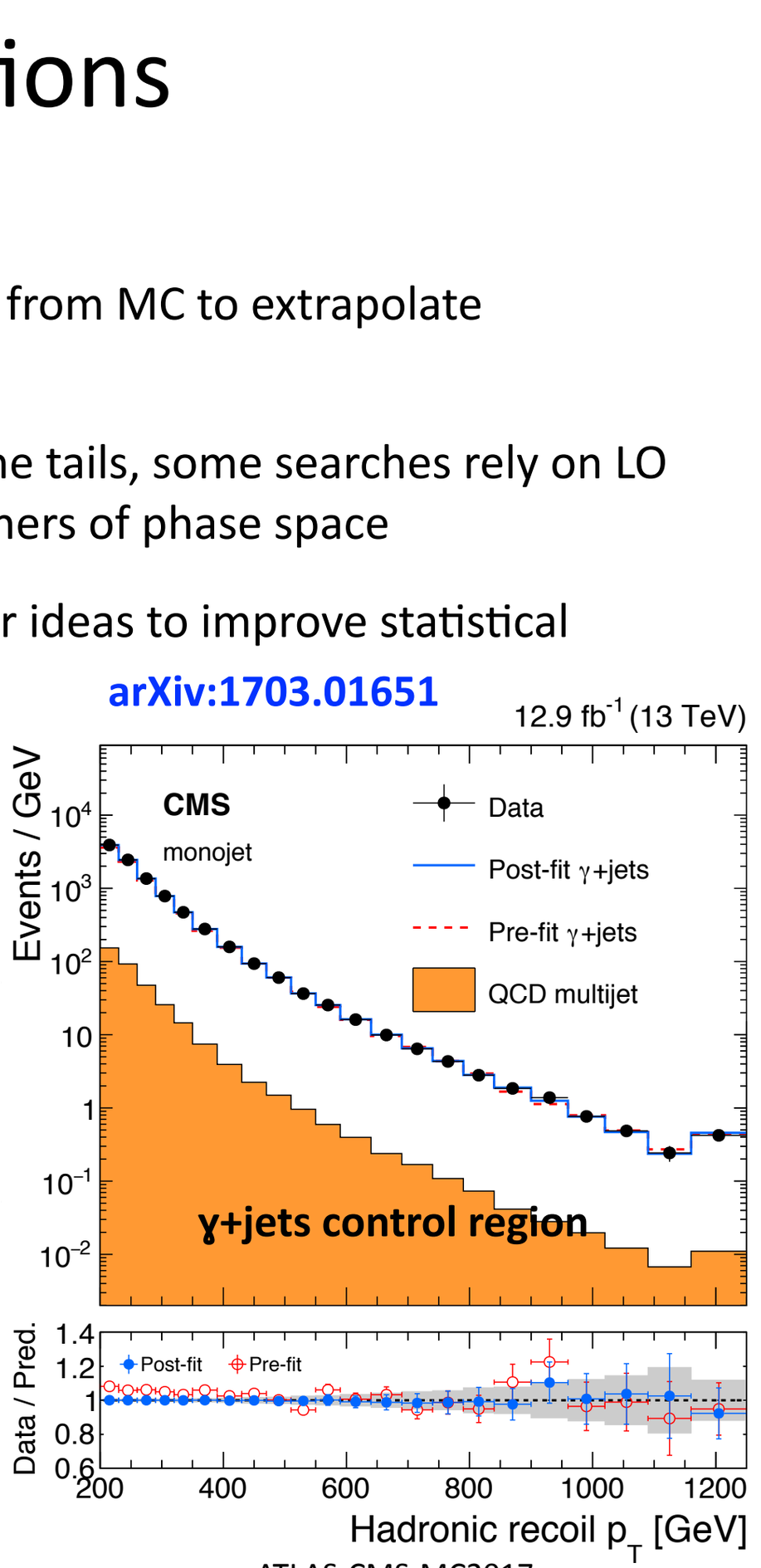
Searches often rely on data control regions and transfer factors from MC to extrapolate predictions into search regions for major backgrounds

Since NLO samples can be quite limited in statistical power in the tails, some searches rely on LO samples and correction factors from data to describe these corners of phase space

- Might benefit from clever binning of NLO samples or other ideas to improve statistical power

Examples of CR —> SR translation

Final state	Background	Control sample	Features of extrapolation
0(1)-lepton + jets + MET	W+jets/ttbar with a "lost" lepton	1(2)-lepton	Correct lepton ID inefficiencies in MC to data, generally rely on MC to describe lepton acceptance
0-lepton + jets + MET	Z(νν)+jets	Z(ll)+jets or γ+jets	Hadronic recoil (MET excluding leptons or photon) in CR proxy for MET in SR, differences between Z and γ



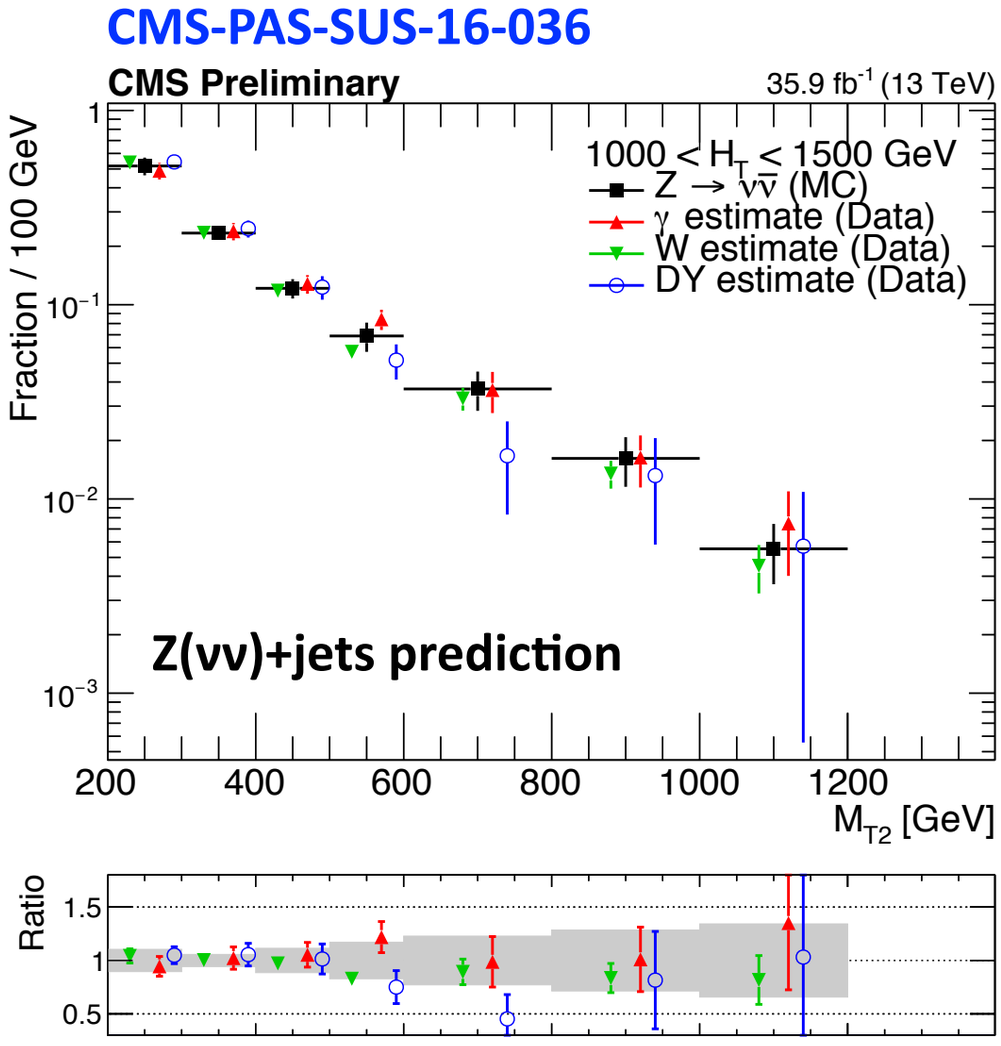
Extrapolation issues

Choice of control regions represents trade-off between statistical and systematic uncertainty

Example: Z(ll)+jets vs ɣ+jets to predict Z(vv)+jets

Larger extrapolations —> higher-order effects become important, fewer cancellations between control and search regions

With many-bin searches, not feasible to sub-divide CRs in as many dimensions as SRs —> rely on MC for some extrapolations, propagate experimental (e.g. jet energy scale/resolution, MET resolution, b-tagging description) and theoretical (factorization/ renormalization scales, PDF) uncertainties



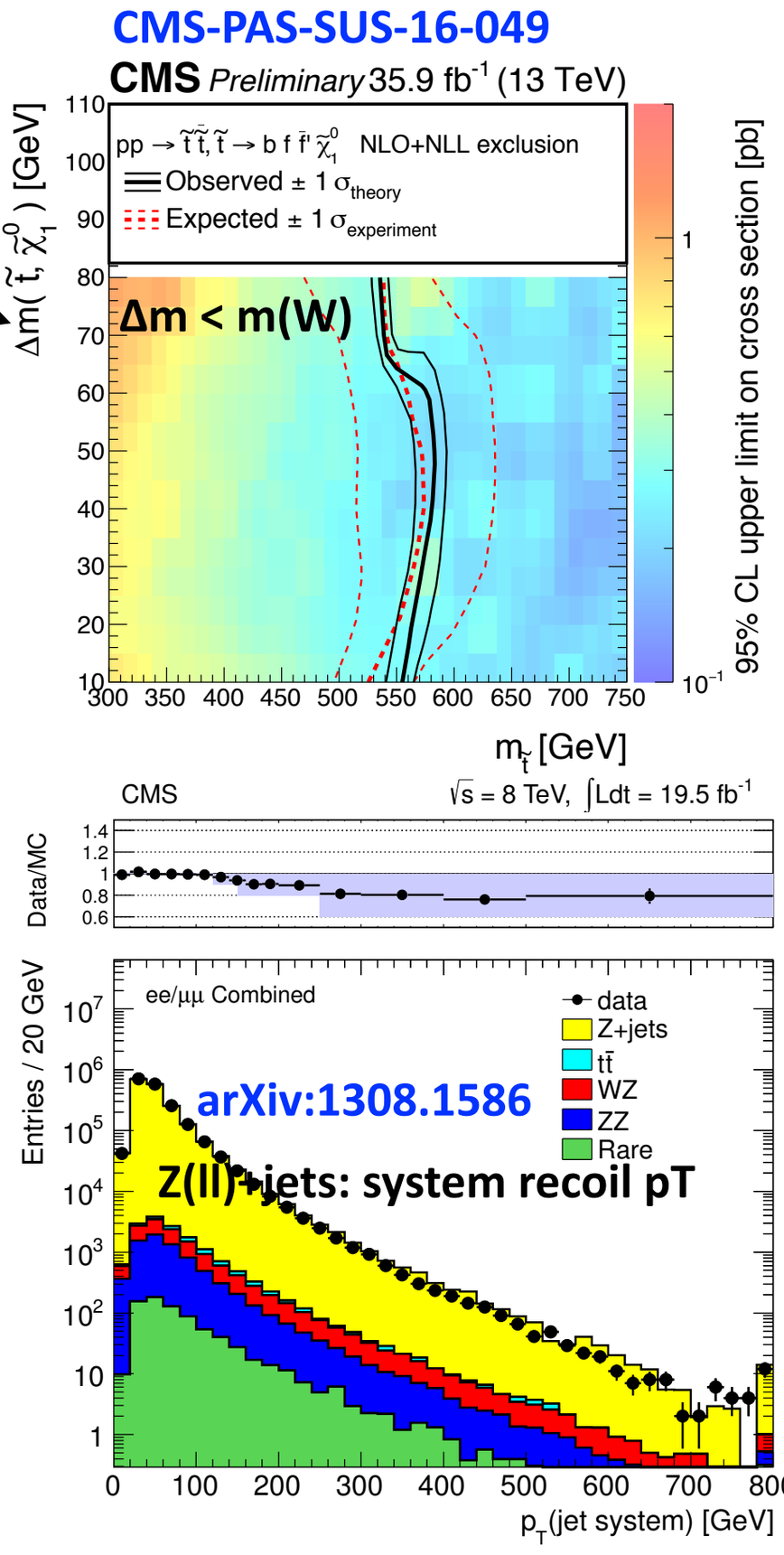
Initial-state radiation

Take advantage of system recoil (i.e. against ISR) to access otherwise difficult signal models, e.g. compressed spectra

Reweight either the ISR jet multiplicity or ISR pT based on studies of ttbar/Z+jets, propagate corresponding uncertainties

ISR reweighting for signal in SUSY searches

- same MC (Madgraph5) used for signal samples as for background samples used to derive weights
- possible limitations: calibrated at top mass and extrapolated to very heavy objects (e.g. gluinos at 2TeV)



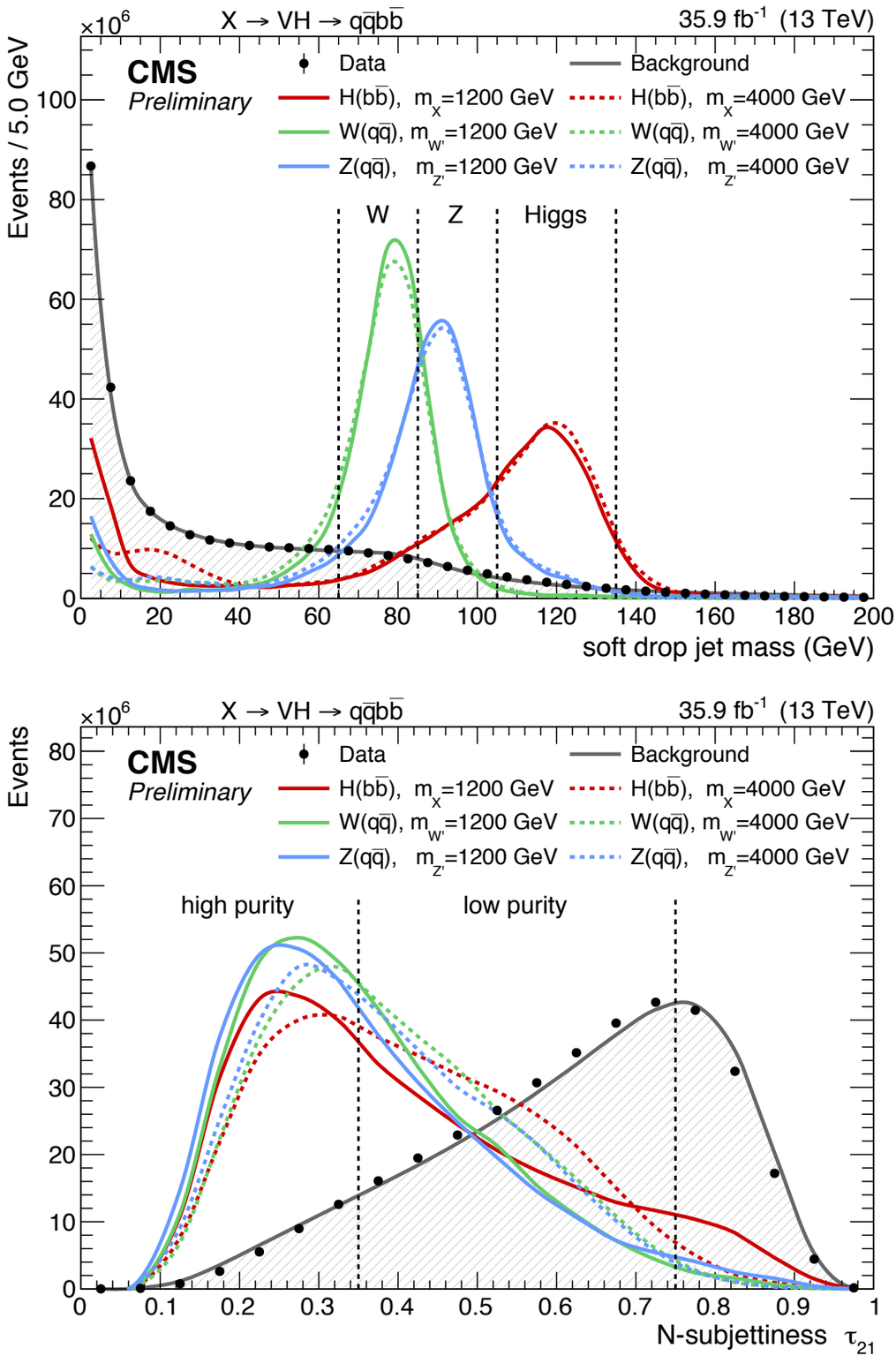
Boosted object tagging

Use of large-radius jets and substructure to identify boosted tops, vector bosons, or Higgs increasingly common in BSM searches

A number of different approaches:

- Cut-based using jet mass and substructure variables
- Multivariate discriminator
- Categorization using shapes

CMS-PAS-B2G-17-002

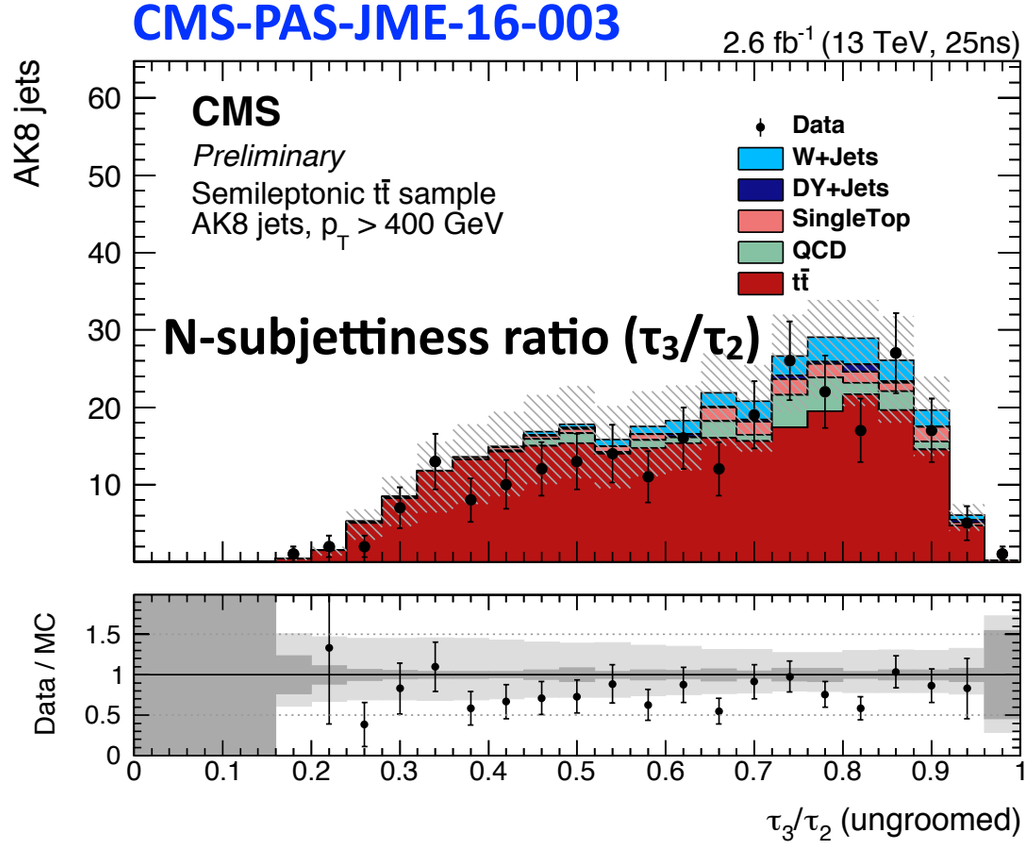


Boosted object validation

Performance validated in data and correction factors applied to MC

Validation can be limited at high pT and in some phase spaces

Various sources of systematic uncertainty considered, e.g. choice of generator (Madgraph vs Powheg), parton shower description (Pythia8 vs Herwig++), pT dependence, purity of control sample

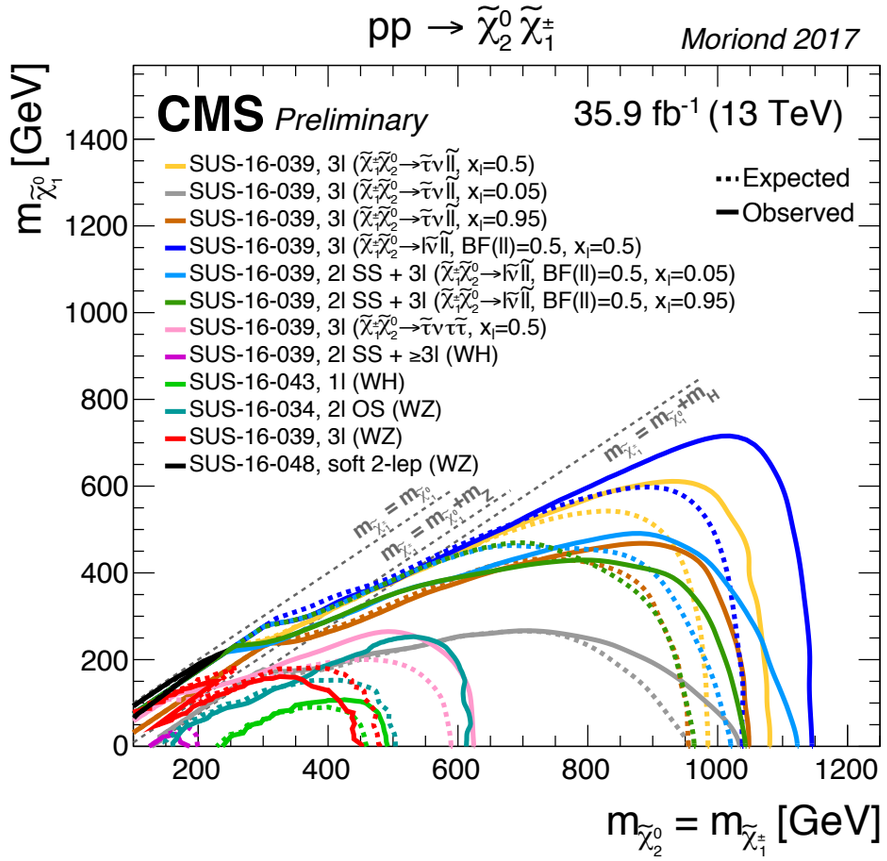


Signal samples

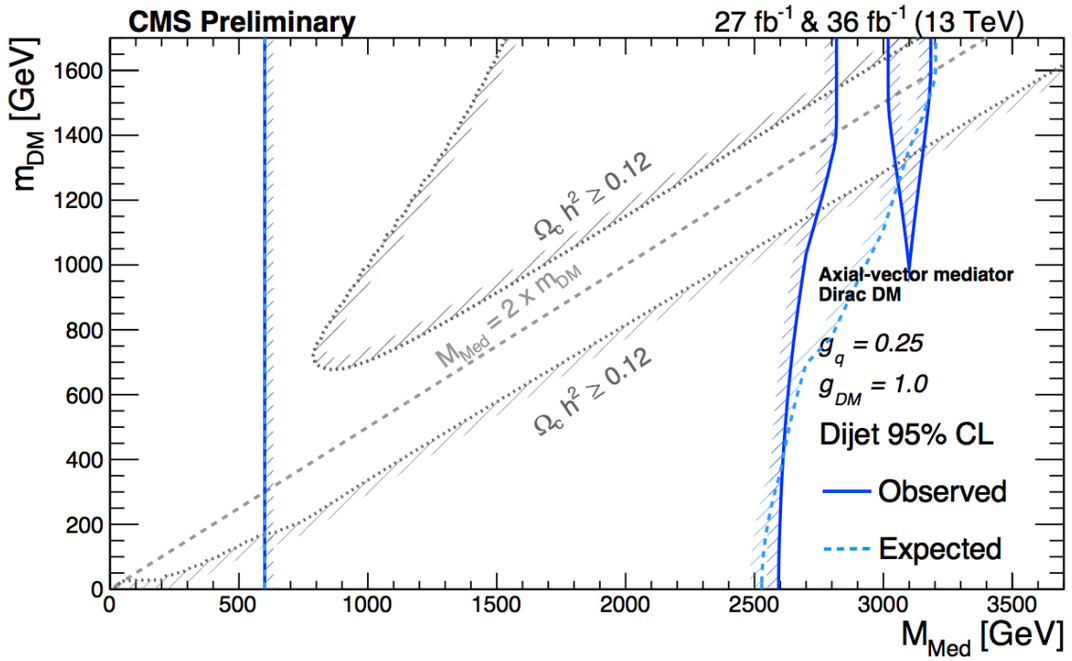
Variety of generators used for generating different types of signal models: Madgraph5 is a common choice, other options include Pythia8, Powheg, CompHEP, JHUGen

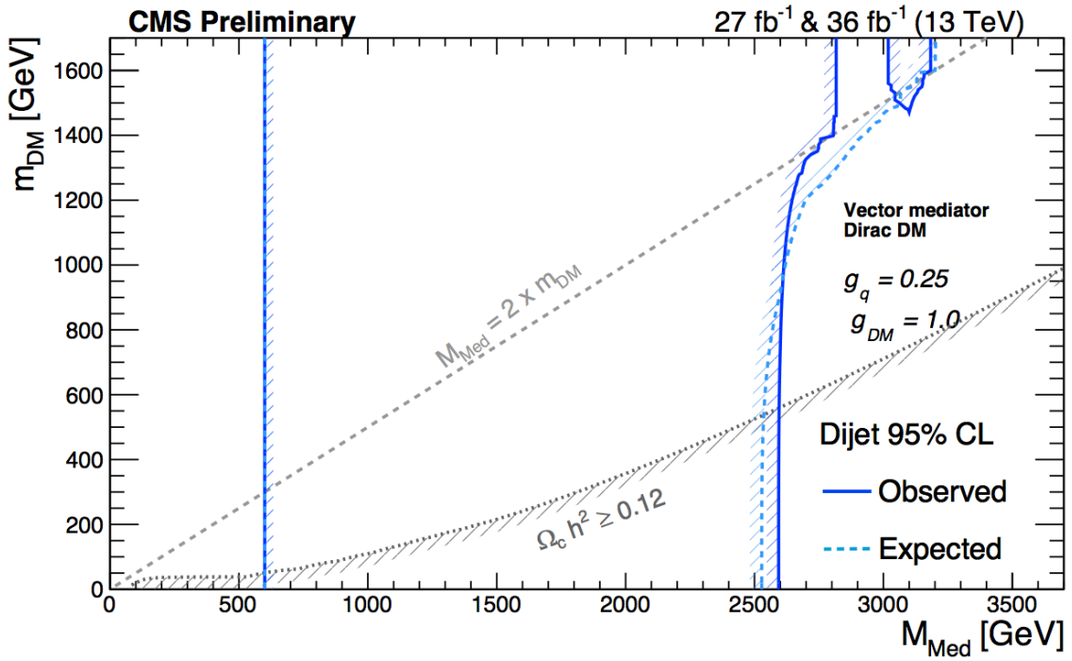
Usually generated at LO and scaled to NLO cross sections … no NLO generation e.g. for SUSY processes

Extensive use of simplified models for SUSY, dark matter searches



CMS-PAS-EXO-16-056





Signal sample generation challenges

Depending on fineness of signal parameter scan, can require producing a large number of gridpacks

- Single-gridpack parameter scan if possible would enormously simplify things
- Exploring matrix element re-weighting feature in Madgraph to scan physics parameters, e.g. couplings, instead of generating multiple samples, could significantly reduce overhead. One possibility under discussion: generating resonance samples with large width, re-weighting to smaller widths keeping fixed resonance mass

Due to computational challenges, “fast” simulation (FastSim) is used extensively in SUSY searches to scan large parameter space for signal

Additional validation for FastSim with respect to full GEANT4-based detector simulation

- Correction factors applied for differences in reconstruction efficiencies, e.g. b-tagging, lepton efficiencies, jet energy scale
- Some substructure variables, e.g. n-subjettiness and subject b-tagging discriminants not very well described, accounted for in correction factors and uncertainties

Reinterpretation of results

Complicated for phenomenologists to re-interpret many of our results because of

- Binning in many (sometimes O(100)) exclusive search regions
- Can be difficult to reproduce object reconstruction, e.g. when using multivariate techniques
- Use of control regions in data to constrain backgrounds in search regions, correlations between different bins

In the past, we supplied the following types of information

- Observed event counts and predicted background + acceptance*efficiency for different signal models in each bin
- Efficiencies for reconstructing leptons, photons, b-jets vs pT (and η, where relevant)
- Cut-flow tables
- Instructions (e.g. standalone code) for calculating more complicated kinematic variables

Still not sufficient information to reasonably approximate more complicated searches

Simplified likelihood approach

Goal: provide sufficient information for re-interpreters to use a simplified likelihood approach to approximate results obtained by searches using full likelihood

Details in CMS-NOTE-2017-001

In addition to observed event counts and estimated background (+ uncertainty) in each bin, also provide background covariance matrix for full set of bins

Relies on a few assumptions:

- Constraints on background expectations (modeled by nuisance parameters reflecting uncertainties) are Gaussian
- Covariance matrix, i.e. just linear correlations between background estimates in different bins, sufficient to get a good enough approximation of full correlation model

Re-interpreters can inject their favorite signal model (+ uncertainties if desired) into likelihood

For analyses with many bins, also useful to define a set of “aggregated” search regions

- Formed by combining subsets of search regions
- Provide sensitivity to different ranges of signal models
- Can then be used with simplified likelihood approach, but with reduced complexity

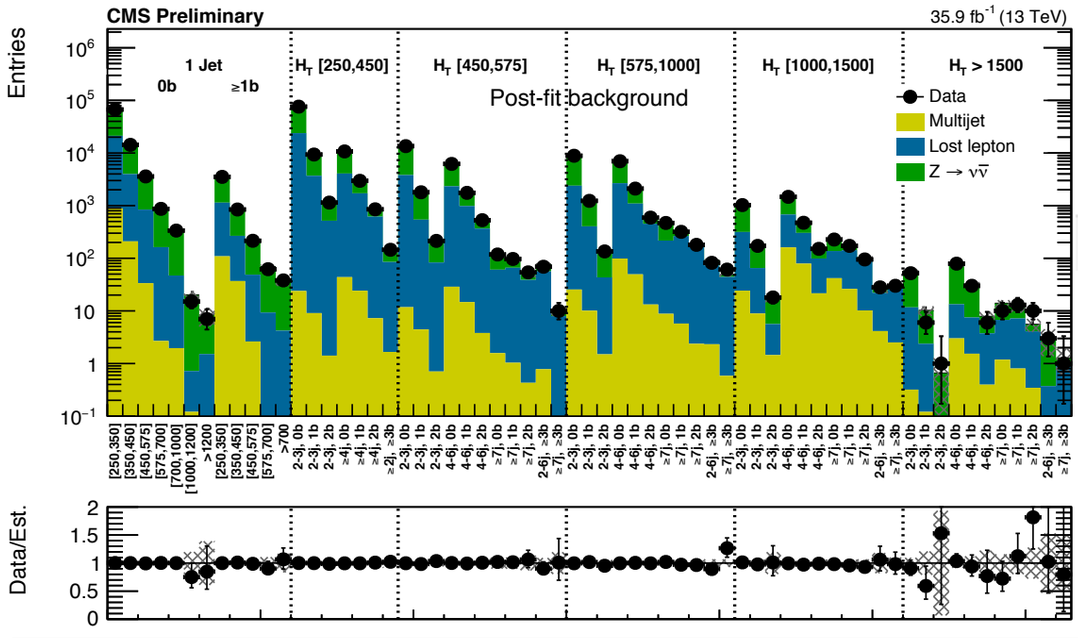
Example: SUSY search with MT2 variable

CMS-PAS-SUS-16-036

Search for new physics in the all-hadronic final state with the M_{T2} variable

CMS Collaboration

March 2017



From public web-page

Additional information on efficiencies needed for reinterpretation of these results are available here.

B-tagging Efficiency

B-tagging efficiency for the WPs of the discriminators used in SUS-16 analyses:

- the efficiency is given for both the CSVv2 and the DeepCSV b-tagging algorithms for each of the Loose, Medium and Tight working points;
- the efficiency is computed for generator-level b-jets in a simulated sample of ttbar events;
- the efficiency is corrected with the corresponding data/simulation scale factors extracted from 2016 data;

Efficiency plot: .pdf

Efficiency root file: .root

CADI	Analysis	Used WP
SUS-16-033	0L + jets with MHT	CSVv2 medium
SUS-16-036	0L + jets with MT2	CSVv2 medium

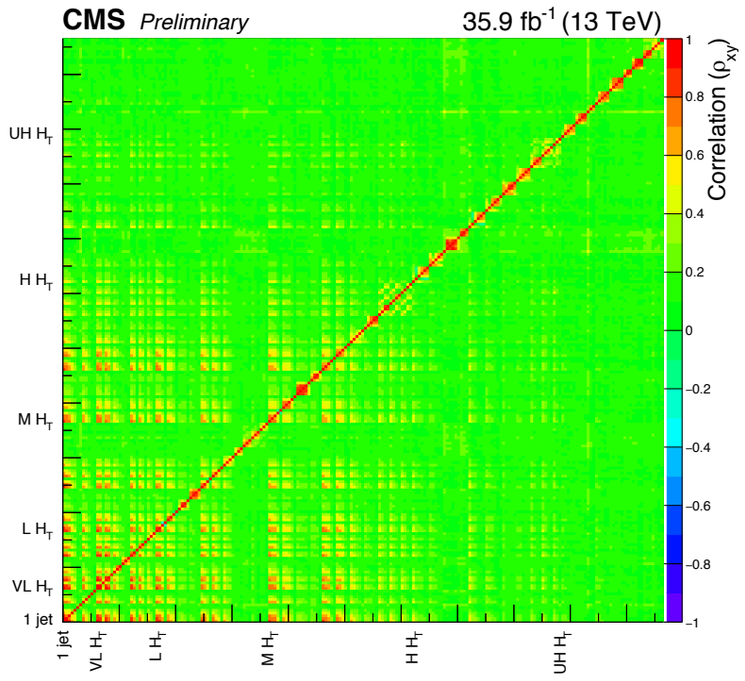
Example: SUSY search with MT2 variable

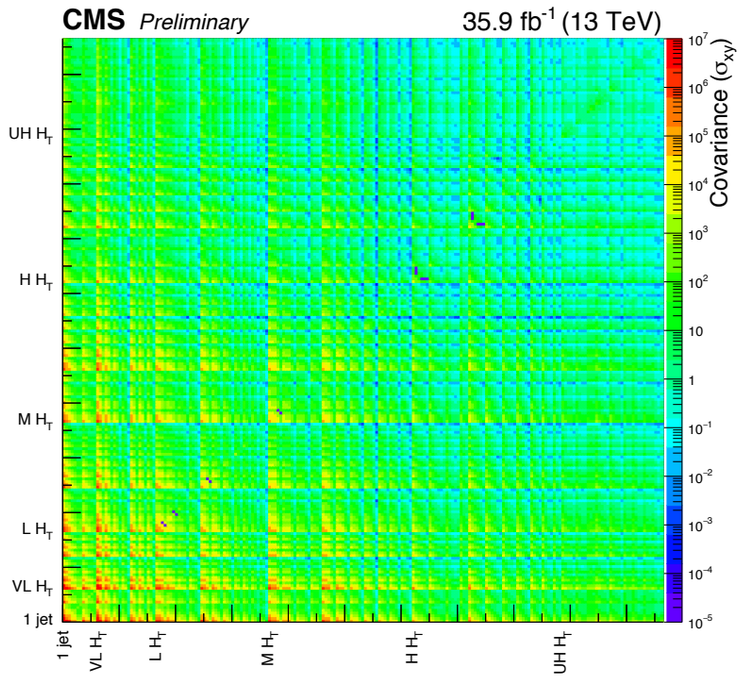
Region	N_j	N_b	H_T [GeV]	M_{T2} [GeV]	Prediction	Data	N_{95}^{obs}
2j loose	≥ 2	–	> 1000	> 1200	38.9 ± 11.2	42	26.6–27.8
2j tight	≥ 2	–	> 1500	> 1400	2.9 ± 1.3	4	6.5–6.7
4j loose	≥ 4	–	> 1000	> 1000	19.4 ± 5.8	21	15.8–16.4
4j tight	≥ 4	–	> 1500	> 1400	2.1 ± 0.9	2	4.4–4.6
7j loose	≥ 7	–	> 1000	> 600	$23.5^{+5.9}_{-5.6}$	27	18.0–18.7
7j tight	≥ 7	–	> 1500	> 800	$3.1^{+1.7}_{-1.4}$	5	7.6–7.9
2b loose	≥ 2	≥ 2	> 1000	> 600	$12.9^{+2.9}_{-2.6}$	16	12.5–13.0
2b tight	≥ 2	≥ 2	> 1500	> 600	$5.1^{+2.7}_{-2.1}$	4	5.8–6.0
3b loose	≥ 2	≥ 3	> 1000	> 400	8.4 ± 1.8	10	9.3–9.7
3b tight	≥ 2	≥ 3	> 1500	> 400	2.0 ± 0.6	4	6.6–6.9
7j3b loose	≥ 7	≥ 3	> 1000	> 400	5.1 ± 1.5	5	6.4–6.6
7j3b tight	≥ 7	≥ 3	> 1500	> 400	0.9 ± 0.5	1	3.6–3.7

Aggregate search regions

Table 2:
Definitions of super signal regions, along with predictions, observed data, and the observed 95% CL limit on the number of signal events contributing to each region (N_{95}^{obs}). No uncertainty on the signal acceptance is assumed in calculating these limits. A dash in the selections means that no cut is applied.

Background covariance matrix





Additional Figure 1:
Full correlation (a) and covariance (b) matrices.

Summary

BSM searches often target more extreme corners of phase space

- Statistical power of NLO samples can be limited in these regions

Extensive use of data control regions along with MC

Exploit features of signal topologies, e.g. ISR recoil or presence of boosted tops/bosons in search strategies

- Validated in data control regions and corrections applied to MC
- Can sometimes involve significant extrapolations from control regions

Signal MC generation can be challenging due to large parameter space under study

- Simplifications include use of simplified models, fast simulation
- Generation usually at LO

Re-interpretations of more complicated search results can be tricky

- Simplified likelihood approach shown to provide a good approximation to full likelihood model, necessary information already available for several SUSY and Exotica searches

Backup

Simplified likelihood approach

Full likelihood

$$\mathcal{L}(\mu, \boldsymbol{\theta}) = \mathcal{P}(\text{data}|\mu \cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta})) \cdot p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$$

$\mu = \sigma/\sigma_{\text{theory}}$ is the "signal-strength" modified

θ are the nuisance parameters which encode systematic uncertainties. There can be O(100) of these in a typical search

$s(\theta), b(\theta)$ are functions which yield the signal (assuming some input theory model) and SM backgrounds for a given value of the nuisance parameters.

$p(\tilde{\theta}|\theta)$ is the nuisance parameter pdf encoding (co)variances of nuisance parameters - "constraint term"

Simplified likelihood approach

Full likelihood

$$\mathcal{L}(\mu, \boldsymbol{\theta}) = \mathcal{P}(\text{data}|\mu \cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta})) \cdot p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$$

Simplified likelihood

$$\mathcal{L}_S(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu \cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu \cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{V}^{-1} \boldsymbol{\theta}\right)$$

Background uncertainty (*)

Nominal expected background in each bin (*)

Observed event counts in each bin (*)

Background covariance matrix (NEW!)

(*) Already provided in most cases