Particle Hunters' Guide: how to discover or exclude a BSM model at the LHC

Márton Bartók[1,2], Péter Major[2], Gabriella Pásztor[2]

[1]Wigner RCP, Budapest

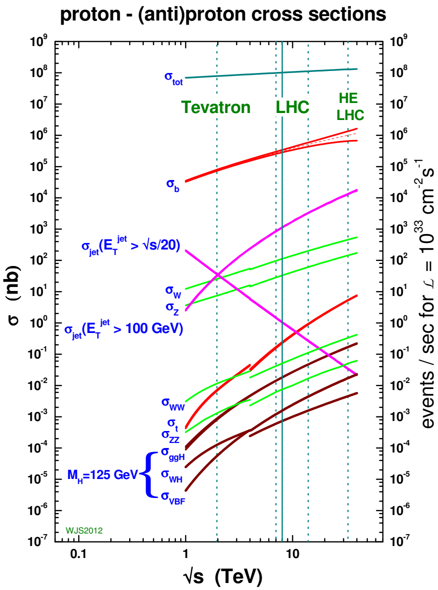
[2]MTA-ELTE Lendület Particle and Nuclear Physics Group,
Eötvös University, Budapest

ELFT Winter School 2021

Outline

Disclaimer: only SUSY analysis showed here, but generally true for any BSM search

Motivation



Physics processes at LHC

- SM physics well understood
- New physics processes (assumed to be) rare, typically buried under large backgrounds
- SUSY gluino pair production @13TeV $\approx 10^{-6} nb$ ($m_{\tilde{g}} = 2$ TeV)

Choosing the right BSM model

Ideal theory according to an experimentalist…

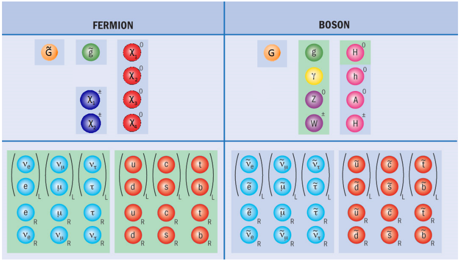
- Distinctive final state
- Easy/possible to simulate (i.e. MC generator exists)
- Only few model parameters
- Varying parameters does not (drastically) change final state
- Provable/falsifiable with available data (eg. ≈ 300 fb^{-1} at LHC, or 3000 fb^{-1} at HL-LHC)
- (Hopefully realized in nature…)

In practice…

- Choose a viable final state based on the prediction of a (few) model(s)
- Make the analysis as general as possible (quasi model-independent)
- Find models with similar final states and interpret the results in them

Quick introduction to SUSY

- Symmetry between fermions and bosons
- SUSY particles not seen at low energy → supersymmetry broken
- Breaking of the symmetry
  - Supergravity
  - Gauge Mediated Supersymmetry Breaking

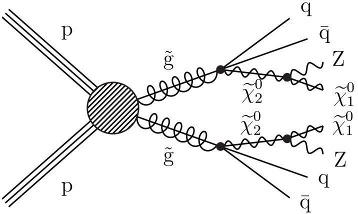


- In general: ≈ 100 extra free parameters
- Constrained Minimal Supersymmetric Standard Model: 5 parameters
- Phenomenological MSSM: 19 parameters
- etc. . .

So many versions, variations. Need to simplify for experiment!

Examples of simplified SUSY models

- Simplified models: bridge between theory and experiment
- Assume a low number of new particles and interactions (others e.g. assumed to have high mass)
- Few physics parameters
  - Particle masses
  - Production cross-sections
  - Branching fractions (BRs)
- Cross-section x BR limits apply to general models with same (similar) final state topology

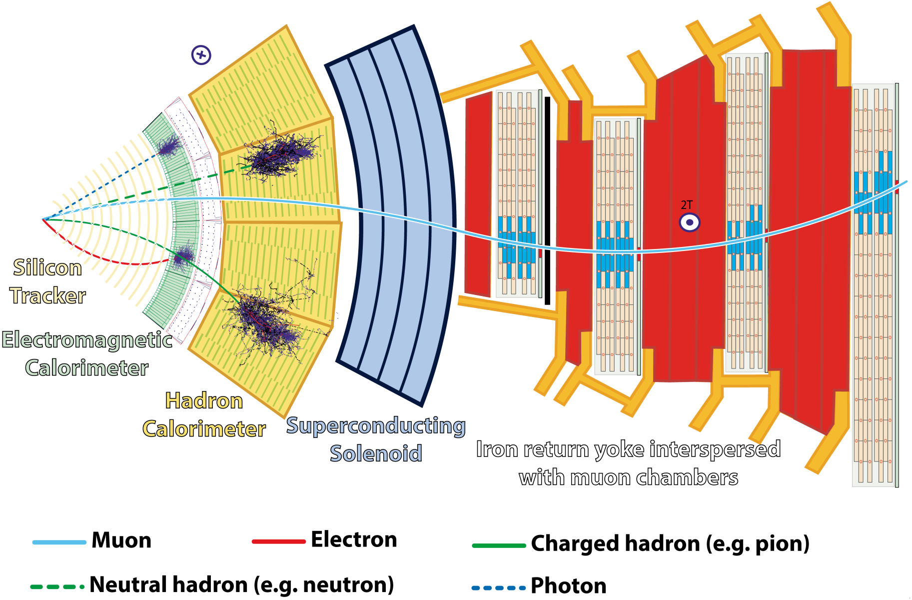


$\tilde{g} \xrightarrow{100\%} \tilde{\chi}_2^0 + q + \overline{q}$, $\tilde{\chi}_2^0 \xrightarrow{100\%} \tilde{\chi}_1^0 + Z$

$m_{\tilde{\chi}_1^0} = 1$ GeV, $m_{\tilde{\chi}_2^0} = m_{\tilde{g}} - 50$ GeV

→**1 free parameter**: $m_{\tilde{g}}$

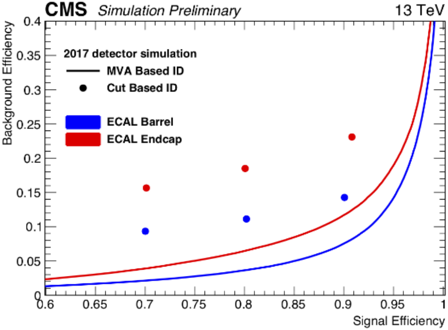
CMS detector



Object definitions

Imperfect reconstruction

- Need to define each object (leptons, photons, jets. . .)
- Higher purity definitions → lower statistics
- Usually 3 standard working points, with increasing purity (loose, medium, tight) and decreasing efficiency
- Different reconstructed objects can overlap → decide object priority order and remove overlap

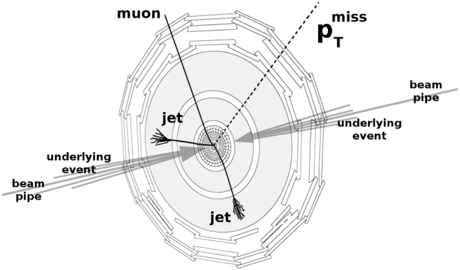


Photon recontruction ROC curve

Missing transverse momentum $\vec{p}_T^{miss}$

Definition

- Negative vector sum of transverse momenta of all reconstructed objects → projected to transverse plane
- $\vec{p}_T^{miss} = -\sum_i \vec{p}_T^i$ (i=all objects)
- magnitude: $p_T^{miss} = |\vec{p}_T^{miss}|$



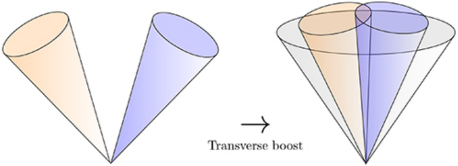
Important quantity for BSM searches
In SM: (should) only come from neutrinos
In reality: it also comes from mis-measured jets, etc.

Note: missing E_T (MET or E_T^{miss}) is a misnomer, but sometimes it's still used

Boosted object tagging

Different jet algorithms

- Standard jet (AK4): anti-k_T algorithm with $R = 0.4$
- Fat jet (AK8): anti-k_T algorithm with $R = 0.8$ →used to reconstruct boosted objects



Example: $Z \rightarrow q\bar{q}$ tagging

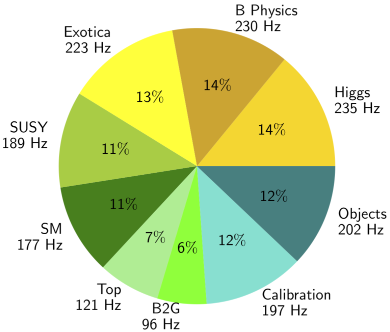
- AK8 jet, $p_T^{AK8} > 200$ GeV
- $70 < m_{jet} < 100$ GeV, consistent with m_Z

CMS Trigger system

LHC collisions every 25ns → 40 million bunch crossing per second
Impossible to fully process or store

- Trigger = real-time event selection
- Event not triggered → lost forever
- Shrink event rate to ≈kHz range
- Select only "interesting" events
- Design triggers before data taking

CMS *Preliminary* (13 TeV, 2018, 2.0×10^{34} $\mathrm{cm^{-2}s^{-1}}$)



Trigger rate allocated to each physics group

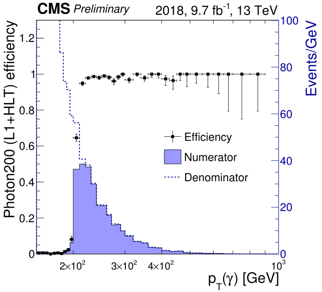
Choosing the trigger for an analysis

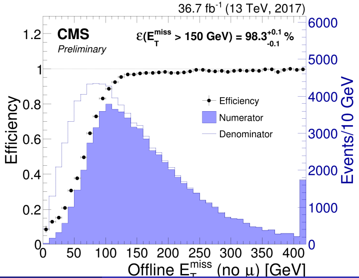
Important decision

- Choose "loosest" unprescaled trigger (eg. lowest pT threshold)
- Possible to use OR of triggers
- Trigger object reconstruction is somewhat different from "offline" object

Trigger efficiency measurement in data

- Orthogonal trigger
- Tag-and-probe method



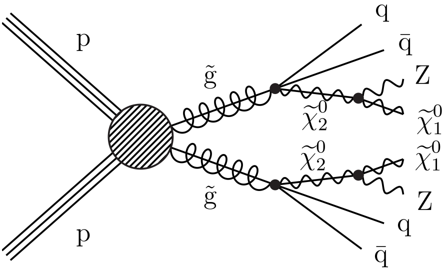


SUSY analysis example CMS-SUS-19-013

Toy model for a BSM search in this talk

Signal model

- Consider 1 simplified signal model
- Only $m_{\widetilde{g}}$ is free parameter
  Fixed: $m_{\widetilde{\chi}_1^0} = 1$ GeV,
  $m_{\widetilde{\chi}_2^0} = m_{\widetilde{g}} - 50$ GeV



Final state

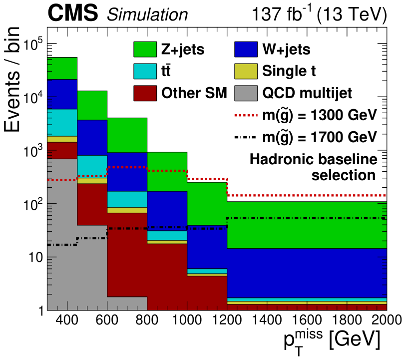
- $\Delta m(\widetilde{g}, \widetilde{\chi}_2^0)$ is small → only soft jets from $\widetilde{g} \rightarrow q\bar{q}\widetilde{\chi}_2^0$ decay
- $\Delta m(\widetilde{\chi}_1^0, \widetilde{\chi}_2^0)$ is large → highly **boosted Z bosons**
- High $\boldsymbol{p_T^{miss}}$ from $\widetilde{\chi}_1^0$

Analysis strategy: identify highly boosted $Z \rightarrow q\bar{q}$ in high p_T^{miss} region

Study simulation

Generate signal MC with full detector response

- Look at many variables and compare their distributions to SM MC
- Define dominant backgrounds
- Find important discriminating variables to suppress these backgrounds



Main backgrounds in this case

- $Z/W/t\bar{t}$ + jets: p_T^{miss} from ν and unreconstructed lepton
- QCD: p_T^{miss} from mismeasurement of jets

Choice of trigger: p_T^{miss} ($> 120 - 140$ GeV)
(single photon trigger used for Validation Region)

Analysis selections

Variable definitions

- $H_T = \sum_{jets} |\vec{p}_T|$
- $\vec{H}_T^{miss} = -\sum_{jets} \vec{p}_T$

- $\Delta\phi(obj_1, obj_2)$ – azimuthal angle between two objects
- transverse mass: $m_T(\vec{p}_T^{miss}$,isolated track$)=$ $= 2p_T^{miss}p_T^{track}[1 - \cos\Delta\phi(\vec{p}_T^{miss}, \vec{p}_T^{track})]$ $m_T \approx m_X$, when X→invisible + track

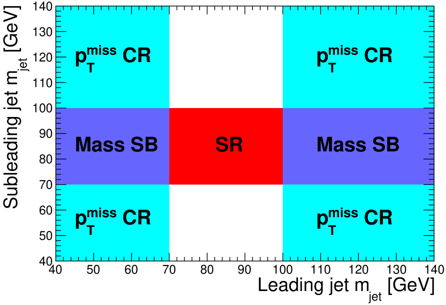
Baseline cuts

- $N_{jet} \geq 2$, $H_T > 400$ GeV
- $p_T^{miss} > 300$ GeV → fully on trigger plateau
- $\Delta\phi(jet, \vec{H}_T^{miss}) > 0.5(0.3)$ leading (subleading) → suppress QCD
- Lepton & photon veto
- $m_T > 100$ GeV (p_T^{miss}, any isolated tracks) → suppress $W \to l\nu$

More complex selection variables and methods are also used in BSM searches (e.g. usage of machine learning)

Analysis regions

- **Signal Region** (SR): most of the signal expected here – **blinded** until analysis approval
- **Control Region**(s) (CR): background rich, used for background estimation
- **Validation Region**(s) (VR): orthogonal region used to validate background estimations



x-y: m_{jet} of the 2 Z boson candidates

In this case:

- SR: **2 Z candidates**, with $70 < m_{jet} < 100$ GeV
  -Subdivided into 6 bins according to p_T^{miss}
- Mass SB CR: leading Z candidate mass in side band
- p_T^{miss} CR: both Z candidates' mass in side band
- VR (not shown here): require 1 lepton or photon (instead of veto)

Background estimation methods

Fully data-driven – the good

- Prediction from combination of different control regions
- Independent of the quality of physics model and detector simulations
- Can be limited by statistics
- Not possible in most cases without some MC input → `goto: the ugly`

Simulation – the bad

- Take into account all imperfections of MC
    - Data/MC corrections → extra systematic uncertainties
- Often MC is not reliable on the edges of the "phase space"
- Simulation is always there
    - Some backgrounds are very hard (or impossible) to estimate from data
    - If the MC is trustworthy, it's easier to use
- Statistics can be increased if needed but computing time is limited

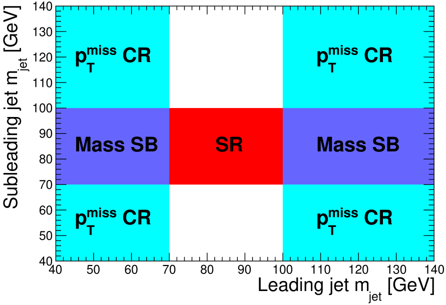
Data & simulation – the ugly

- Probably the most frequent method
- Less affected by the drawbacks of simulations

Background estimation strategy in example search

Estimate all background with a **fully data-driven method**

- Mass SB CR: Fit m_{jet} distribution and interpolate $\rightarrow$ $\mathcal{B}_{norm}$ = total number of background events in SR
- p_T^{miss} CR: Look at p_T^{miss} distribution shape $\rightarrow$ normalise integral to match $\mathcal{B}_{norm}$
- Use this normalised distribution as background prediction



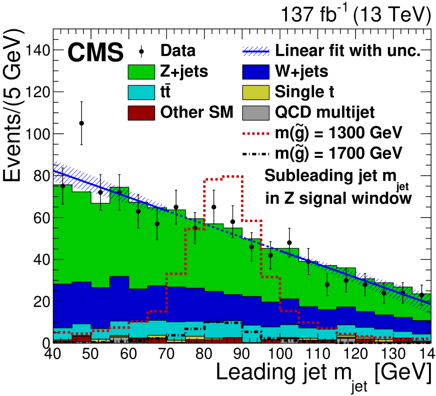
x-y: m_{jet} of the 2 Z boson candidates

Assumptions

- good fit of m_{jet} distribution
- m_{jet} and p_T^{miss} uncorrelated
  i.e.: p_T^{miss} shape looks the same in CR and SR

Data-driven: side-band fit

Note: data in Z window is not used for fitting



Mass SideBand CR

- Background smoothly falling under m_{jet}
- Fit with linear function (difference of higher order fits used for systematic uncertainty)
- Interpolation of fit $\mathcal{B}_{norm} = 325 \pm 15$

Shape of p_T^{miss} (6 bins)

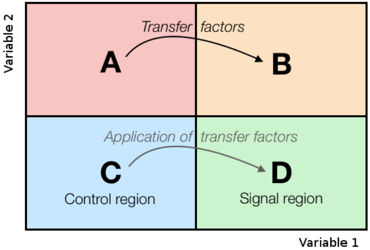
- Normalization factor for p_T^{miss} CR: $\mathcal{T} = \frac{\mathcal{B}_{norm}}{\sum_i N_i^{CR}} = 0.198 \pm 0.009$
- Bkg est. in each p_T^{miss} bin: $\mathcal{B}_i = \mathcal{T} N_i^{CR}$

Other background estimation methods

Few commonly used methods

Data-driven: ABCD method

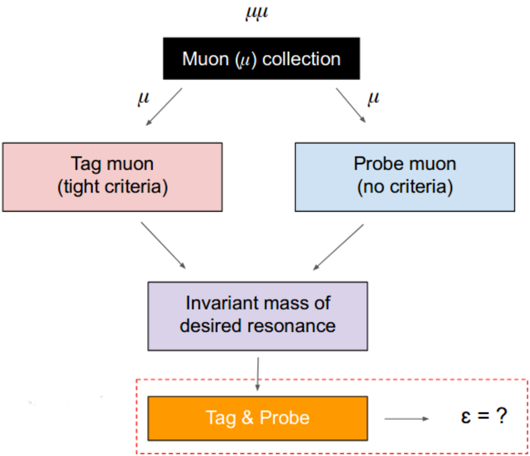
- Two uncorrelated variables
- Signal = "D" region
- ABC regions rich in background
- $N(A)/N(B) = N(C)/N(D) \rightarrow N(D) = \frac{N(B)N(C)}{N(A)}$



In practice: often derive correction for correlation from simulation

More sophisticated versions exist using simultaneous fits of signal and background

Data-driven: tag-and-probe method



$\mu\mu$

Muon (μ) collection

μ

μ

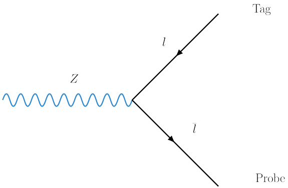
Tag muon
(tight criteria)

Probe muon
(no criteria)

Invariant mass of
desired resonance

Tag & Probe

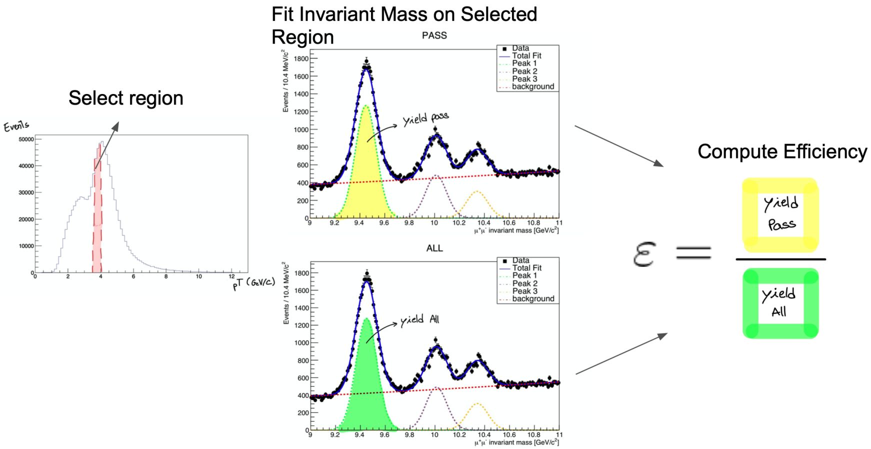
ε = ?



- Based on the decay of resonances to particle pairs (e.g. $J/\Psi, \Upsilon, Z$)
- Tag: well reconstructed triggered object
- Probe: loose selection, pass/fail the criteria for efficiency measurement
- Invariant mass of tag+probe consistent with resonance: $m_{TP} \approx m_X$

Data-driven: tag-and-probe method

- Measures the detection efficiency
- Fit and subtract side-bands then fit peak(s)



Select region

Fit Invariant Mass on Selected Region

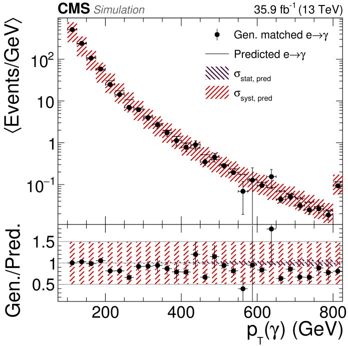
Compute Efficiency

$$\varepsilon = \frac{\text{Yield Pass}}{\text{Yield All}}$$

Using tag-and-probe for background estimation

Electron faking a photon ($e \rightarrow \gamma$)

- Use $Z \rightarrow e^+e^-$
- Tag: tight electron identification with trigger matching (e)
- Probe:
  - a) photon identification (γ)
  - b) fake photon (electron like photon) (f)
- Fake rate: $f_{(e\rightarrow\gamma)} = \frac{N(Z\rightarrow e\gamma)}{N(Z\rightarrow ef)}$
- Apply fake rate to fake photon CR
  - $N(SR) = N(CR) \cdot f_{(e\rightarrow\gamma)}$



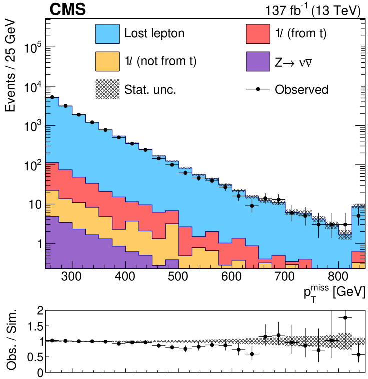
Semi data-driven: transfer factor from MC

Control Region and transfer factor

- Define a CR by inverting a cut
- Calculate transfer factor in MC $TF^{MC} = N^{MC}(SR)/N^{MC}(CR)$
- Apply transfer factor in data $N^{Est.}(SR) = N^{Data}(CR) \cdot TF^{MC}$

Example: lost lepton (not reconstructed)

- SR: lepton veto, CR: require lepton(s)
- In MC: require a truth lepton (both SR & CR)
- TF^{MC} : probability of not reconstructing a lepton
- Apply transfer factor in data



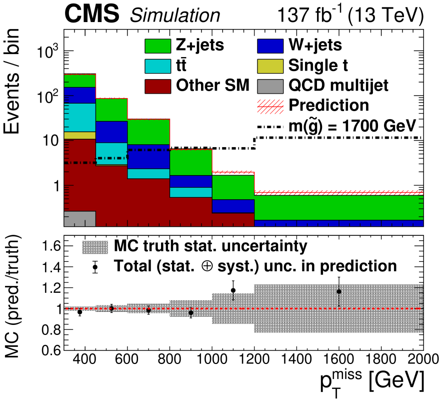
Validation of background estimation

Going back to the boosted Z search

Closure test – check bkg estimation with MC

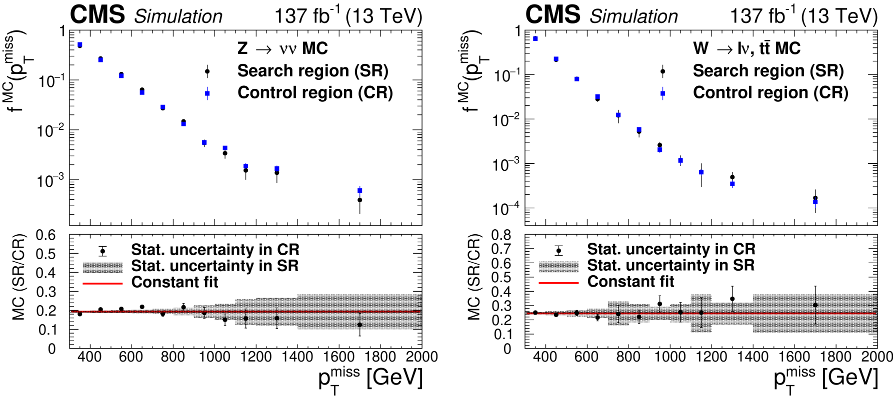
Analysis still blinded: can't look at signal region!

- Redo background estimation with MC
- Compare predicted events to observed events
- Prediction in agreement with background yield



Relative difference taken as a systematic uncertainty on the shape $(1 - 20\%)$

Check $p_T^{miss} - m_{jet}$ correlation in MC

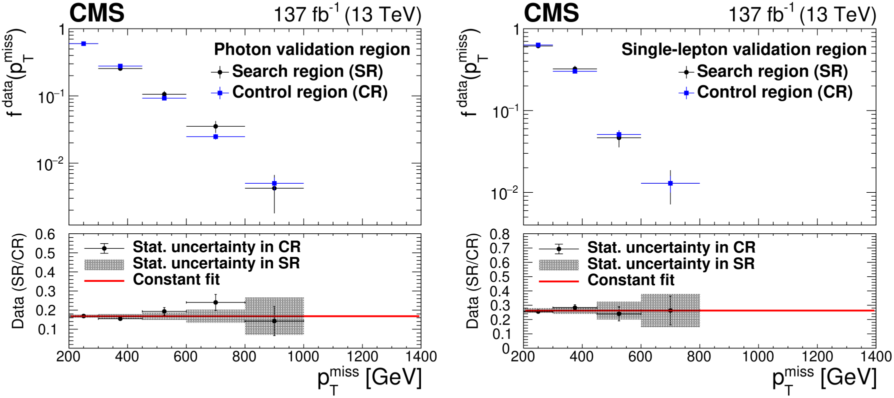


2 main Bkgs: $Z \rightarrow \nu\nu$ and $W \rightarrow l\nu$ (including $t\bar{t}$)

- p_T^{miss} distributions, normalised to 1
- SR and p_T^{miss} CR is consistent

Check $p_T^{miss} - m_{jet}$ correlation in data

p_T^{γ} treated as p_T^{miss} $(Z \to \nu\nu)$; "SR" means here: photon+SR or lepton+SR



- SR/CR is consistent
- Fit ratio with constant and linear function
- Difference of fits → systematic uncertainty on shape

Corrections, data/MC Scale Factors

For MC (affects only signal in this analysis):

- Different efficiency or resolution in Data/MC → SF for almost every reconstructed object
- Corrections for event generator (e.g. initial state radiation modeling)

For data:

- Few object corrections (e.g. jet energy correction)
- Detector or data taking issues (something happens every year)

Example: CMS Hadron calorimeter sector failure in 2018

- Power interruptions by false fire alarm
- 2 sectors (40 degree section) could no longer be operated
- Affects 65% of data taken that year

Systematic uncertainties

In general

- Redo analysis with corrections modified by $\pm 1\sigma$
- Check relative difference wrt nominal event yields

Examples

- Trigger, reconstruction and identification efficiencies (and their Data/MC scale factors)
- Energy and momentum scales (eg. muon p_T, jet E_T, ...)
- Luminosity determination
- Theory (e.g. cross sections)
- etc.

Systematic uncertainty on background estimation method

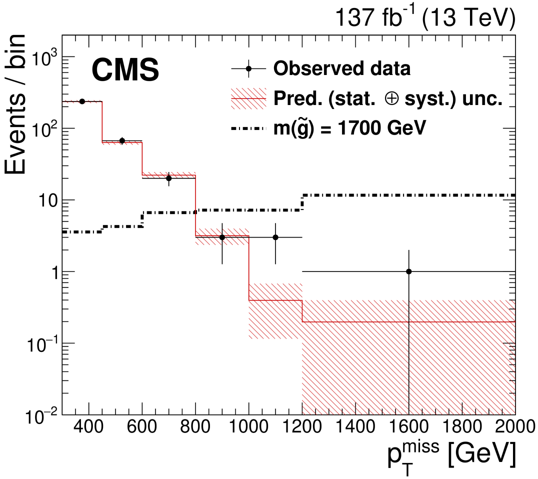
- Quantify how "robust" the estimation
- Important (and difficult) part of background estimation
- No clear rules how to calculate
- Examples on slides 19, 27, 29

Uncertainties in bkg prediction and signal yields

Source of uncertainty	Effect on yields (%)	norm. or shape
Uncertainties in the background predictions		
Fit, normalization	3.3	norm.
Fit, shape	3.4	norm.
m_{jet} CR statistics	3–100	shape
MC closure	2–13	shape
Data validation	2–30	shape
Uncertainties in the signal yields		
Integrated luminosity	2.3–2.5	norm.
Trigger efficiency	2.0	both
Isolated lepton and track vetoes	2.0	norm.
Jet quality requirements	1.0	norm.
ISR modeling	1–2	both
μ_{R} and μ_{F} scales	0.2–0.5	both
JEC	2–4	both
JER	5–6	both
MC statistics	1–2	both
m_{jet} resolution	1–3	norm.

Results in Signal Region

- Background prediction with stat. and syst. errors
- Unblinding: observed data points
- Data consistent with bkg prediction



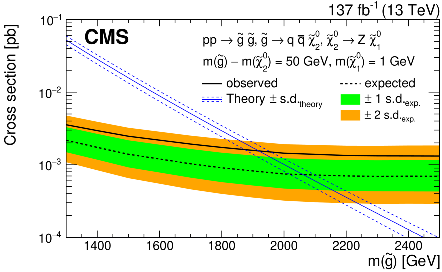
Interpretation of results – expected exclusion

In signal region

- Background expectation of N_{bkg}, Signal expectation (from MC) of S

If only background → how much can we constrain signal strength?

- Depends on uncertainties
$N_{bkg} = 100 \pm 1,\ S = 20 \pm 1$ vs.
$N_{bkg} = 100 \pm 10,\ S = 20 \pm 10$
- Statistical hypothesis testing is performed using CLs test statistics
- Family of signal models described by a continuous variable (e.g. $m_{\widetilde{g}}$) → expected upper limit on $pp \to \widetilde{g}\widetilde{g}$ cross section



σ_{theory} crosses expected curve → expected excl. limit on $m_{\widetilde{g}} \approx 2$ TeV

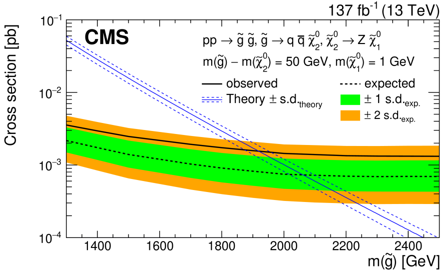
Interpretation of results – observed exclusion

After unblinding (look at observed data)

- Excess of data? Consistent with background? How significant?
- Statistical hypothesis testing done taking observed data into account

Exclusion at 95% significance level can "fail" due to

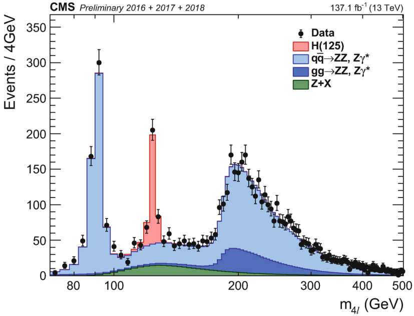
- Large excess in one or more bins
- Large uncertainties
- Too small signal

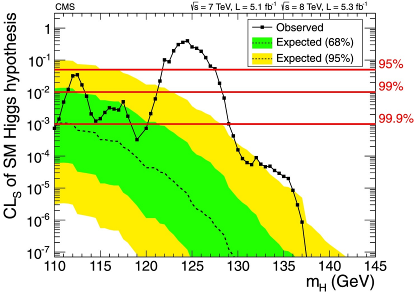


σ_{theory} crosses observed curve → observed excl. limit on $m_{\tilde{g}} \approx 1.9$ TeV

Result of the analysis: a discovery?

Unfortunately no BSM plots here. . .

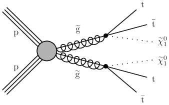




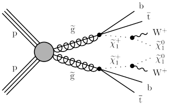
Last big discovery: the Higgs boson in 2012 (predicted in 1964)
Patience is part of the game. . .

A more complicated analysis CMS-SUS-19-008

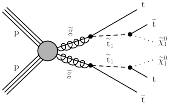
This search considers 9 simplified models. . .



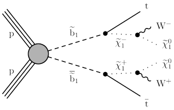
(a) T1tttt



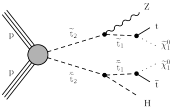
(b) T5ttbbWW



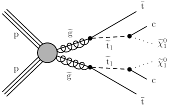
(c) T5tttt



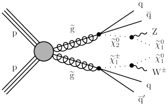
(a) T6ttWW



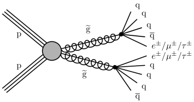
(b) T6ttHZ



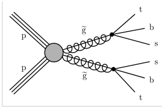
(d) T5ttcc



(e) T5qqqqWZ



(a) T1qqqqL

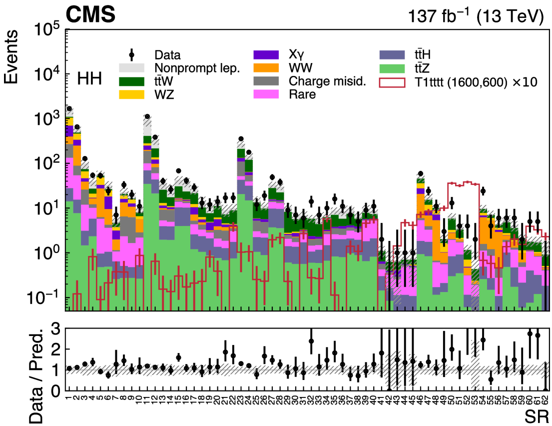


(b) T1tbs

5 $\tilde{g}$ pair productions, 2 $\tilde{q}$ pair productions

2 R-parity violating models

Results of CMS-SUS-19-008

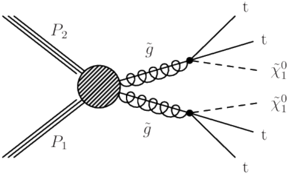


168 search regions

In outline: leptons ≥ 2, jets ≥ 2, b-jets and p_T^{miss}

No significant excess in any of the bins

Interpretation of results CMS-SUS-19-008



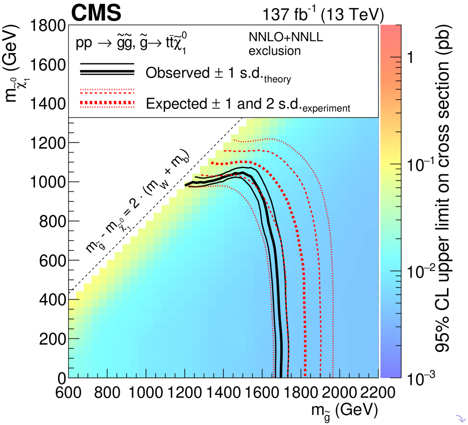
Both $m_{\tilde{g}}$ and $m_{\tilde{\chi}_1^0}$ are free parameters

- 2-dimensional exclusion curve
- colour scale: upper limit on SUSY cross section
- $m_{\tilde{g}}$ excluded up to 1.3 - 1.7 TeV

In other models:

$m_{\tilde{g}}$ excluded up to 2.1 TeV

$m_{\tilde{t}}$ and $m_{\tilde{b}}$ excluded up to 0.9 TeV



Summary

Example of a BSM search analysis shown

- Exclude/discover BSM theory is not easy
- Many complicated theory models exist
- Experimental aspects are challenging
- Complex Monte Carlo tools needed
- No sign of BSM in any of the searches

But there is hope!

- Analysises are getting more sophisticated
  - E.g. boosted boson tagging
- There are searches for every "corner of phase space"
- If new physics can be discovered at LHC → it will be discovered

Backup

Backup slides

Other tasks of experimentalists

Ensure high quality data-taking

- Efficient detectors (regular calibrations, monitoring, tests)
- Minimize downtime (not taking data while LHC is colliding)
- Data acquisition (DAQ)
- Trigger system
- Data quality monitoring
- Luminosity measurement
- etc. . .

Data reconstruction

- Tracking
- Particle flow
- Physics objects ($\mu, e/\gamma, \tau$, jets, b-tagging, p_T^{miss})
- MC
- etc. . .

Every author of CMS has to dedicate 1/3 of their work to these kind of "central tasks"

Bkg estimation: simulation with correction from data

- Control Region: derive data/MC scale factors (SF)
- Apply SF to MC in Signal Region

Example: fit MC to data

- Template fit 2 different MC SFs to best describe data in CR
- Use the SFs in SR to correct MC