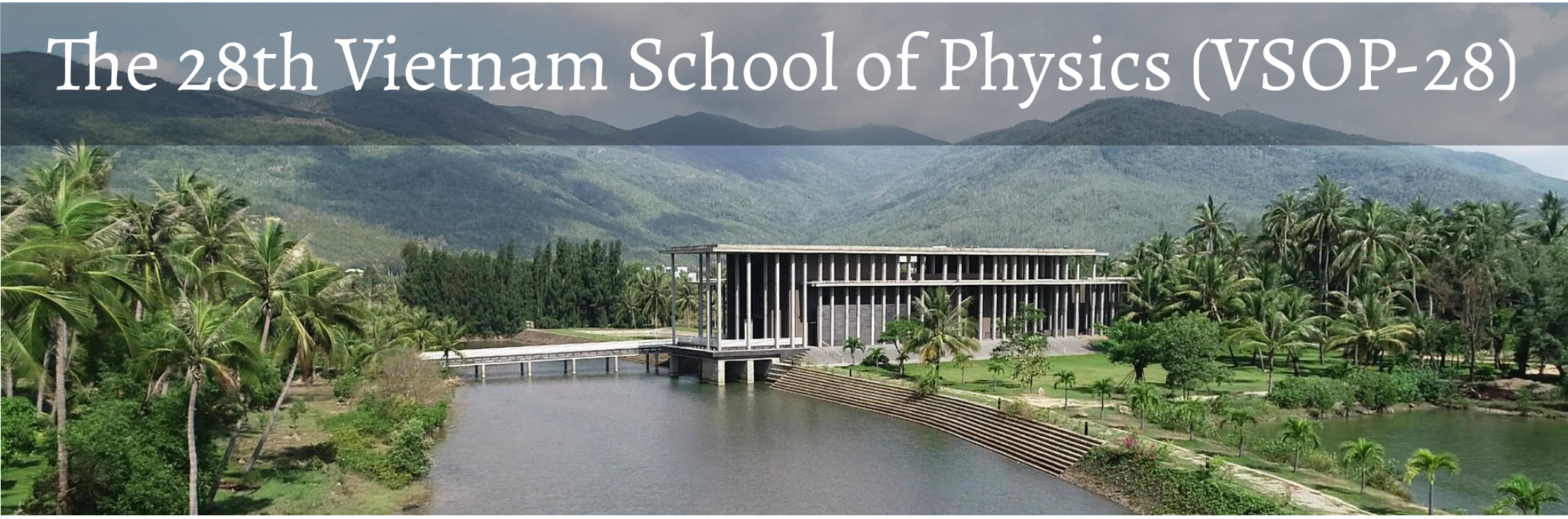


Experimental methods for physics at the LHC

Sourabh Dube

IISER PUNE

July 24, 2022 to
August 5, 2022

Lecture 5: Backgrounds

Background estimation

Typical analysis flow

1. Start with the process we want to study. This is signal.

   Identify its main features, final states.

2. Decide the event selection.

3. Other SM processes passing this selection are background.

4. Estimate the background. Gain confidence in estimate, assess uncertainties on estimate.

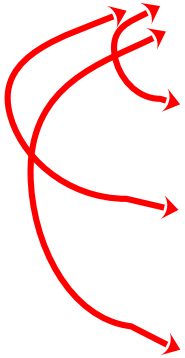
5. “Open the box” – check data.

6. Interpret the findings…

Background estimation

Typical analysis flow

1. Start with the process we want to study. This is signal.
   Identify its main features, final states.
2. Decide the event selection.
3. Other SM processes passing this selection are background.
4. Estimate the background. Gain confidence in estimate, assess uncertainties on estimate.
5. Assess the expected sensitivity…
6. “Open the box” – check data.
7. Interpret the findings…



Background estimation

is crucial.

Typical analysis flow

1. Start with the process we want to study. This is signal.

   Identify its main features, final states.

2. Decide the event selection.

3. Other SM processes passing this selection are background.

4. Estimate the background. Gain confidence in estimate, assess uncertainties on estimate.

5. Assess the expected sensitivity…

6. “Open the box” – check data.

7. Interpret the findings…

Using simulation

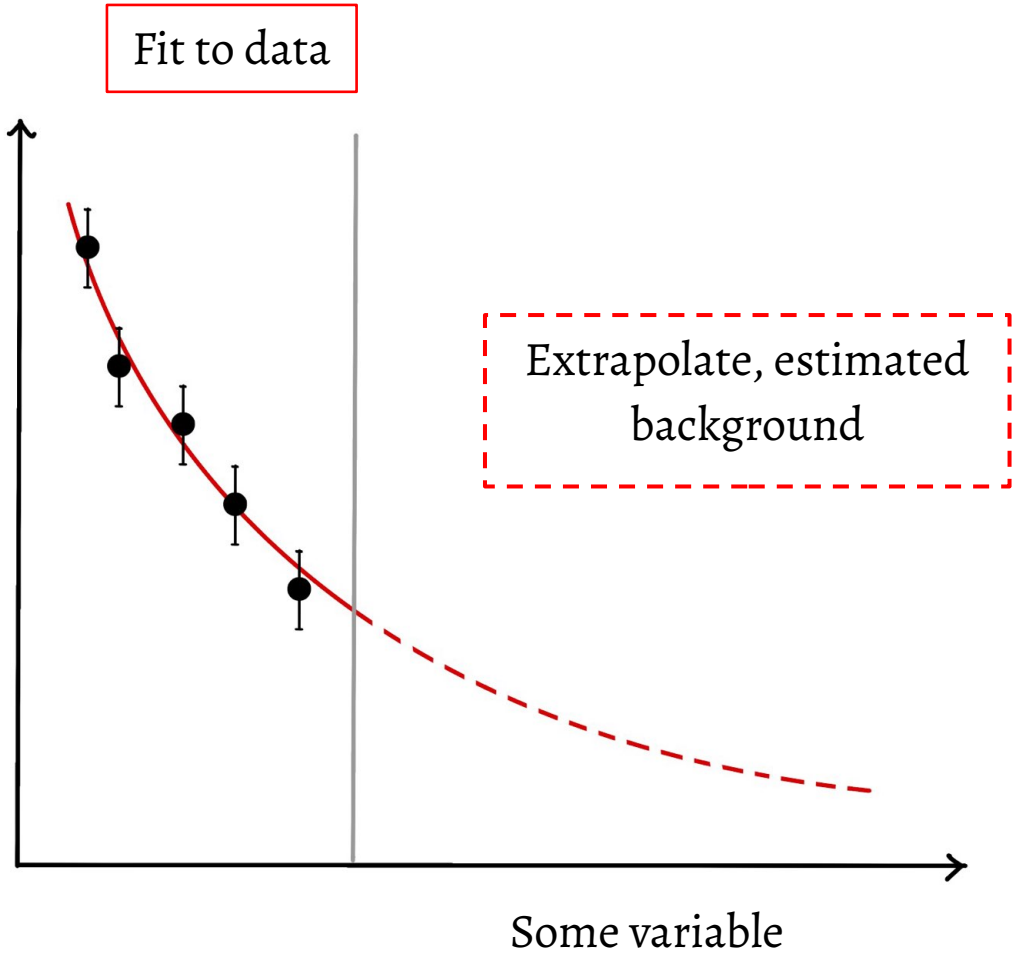
$$N = L\ \sigma\ B\ A\ \varepsilon$$

Suppose we want to estimate the yield from a particular process.

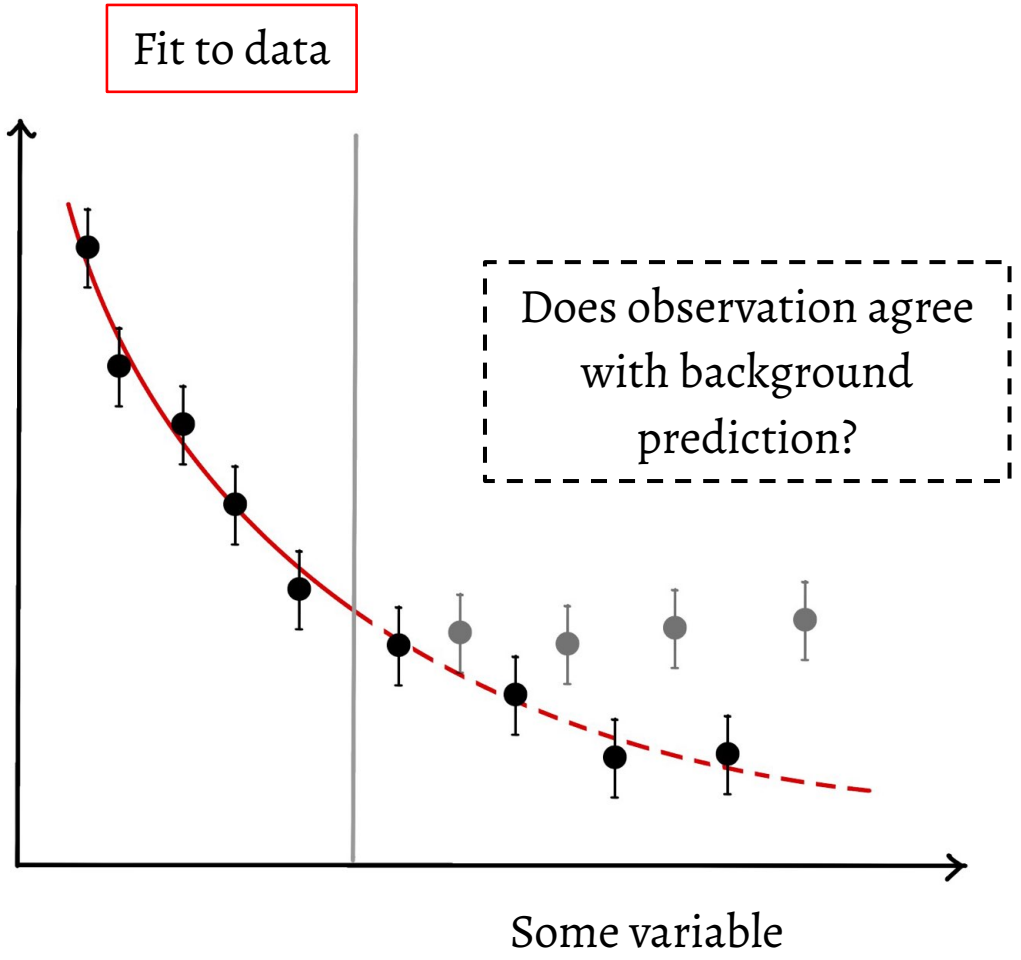
We generate a simulation sample. This allows us to determine the $A \cdot \varepsilon$, and thus we can estimate the yield.

In addition, we can also estimate the shape of distributions
(p_T, MET, H_T, $\Delta\phi$)

Using data

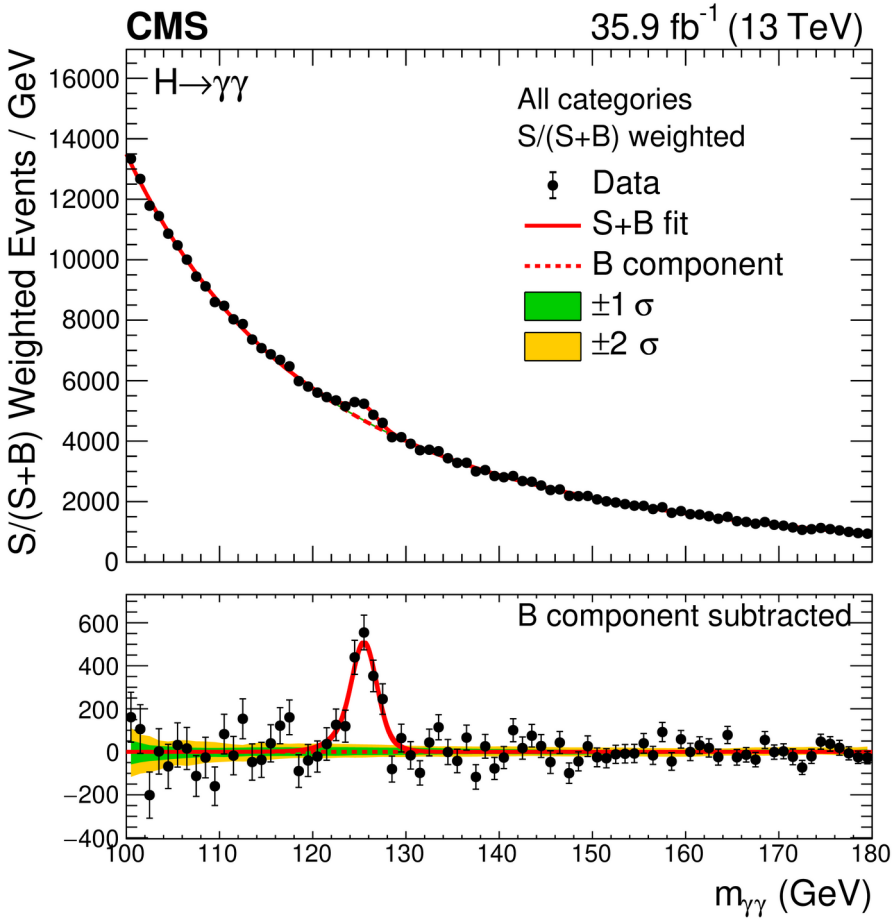

Fit to data
Extrapolate, estimated background
Some variable

Using data

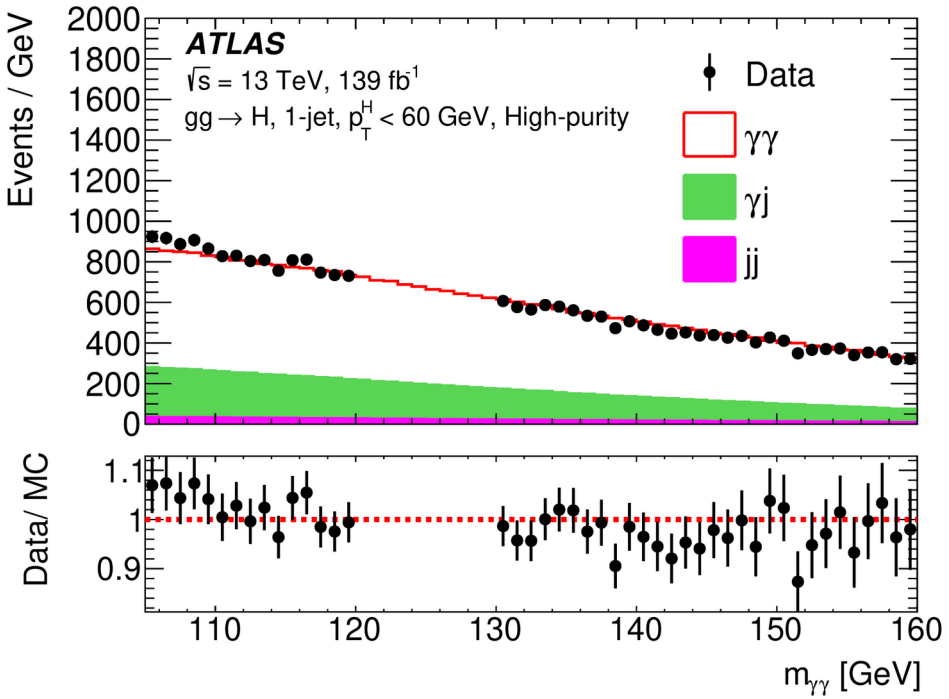

Fit to data
Does observation agree with background prediction?
Some variable

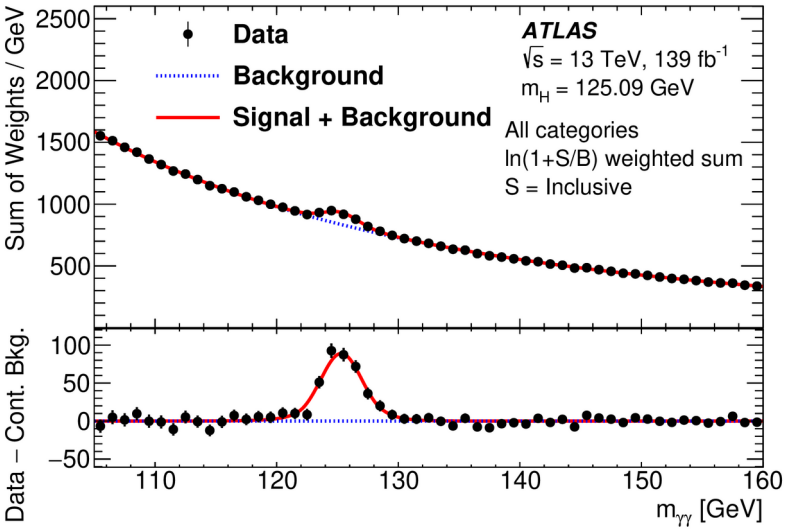
Higgs to diphoton

arXiv:2207.00348 [hep-ex]

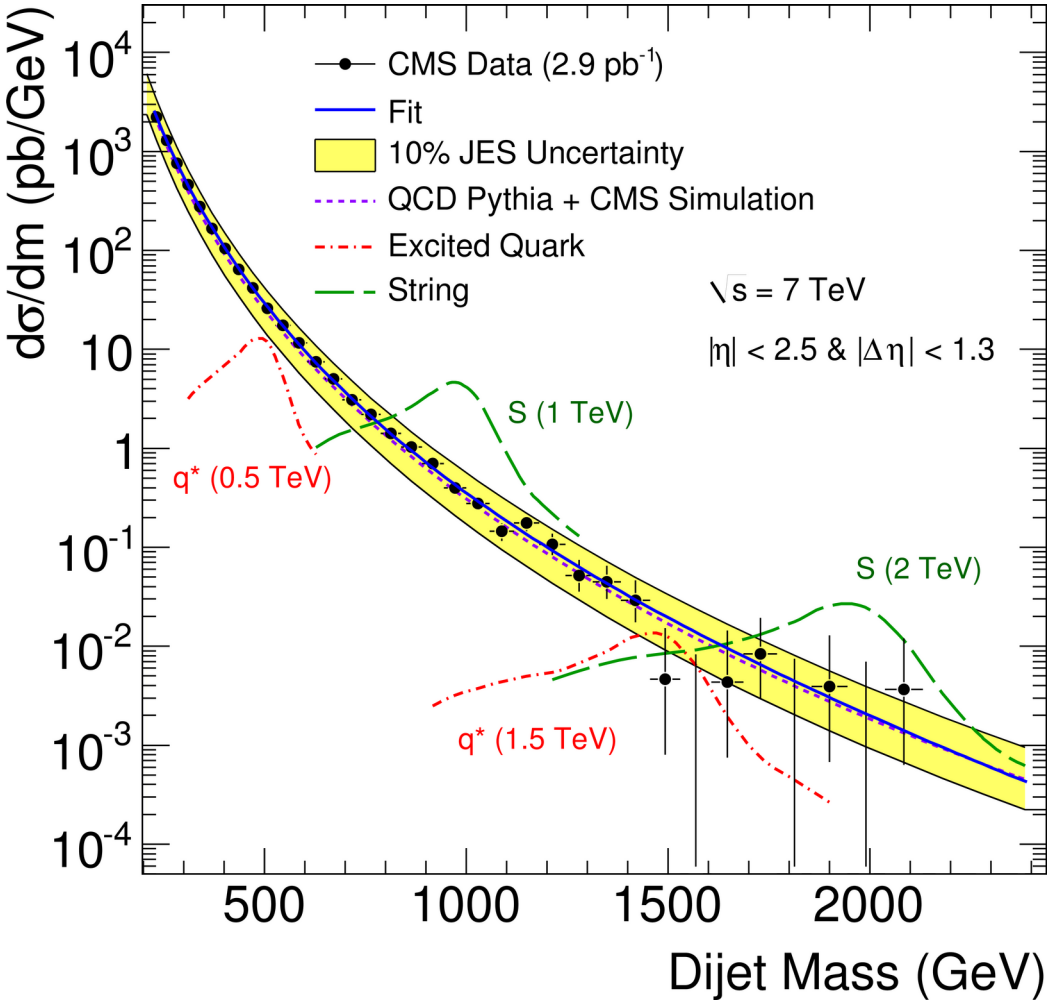


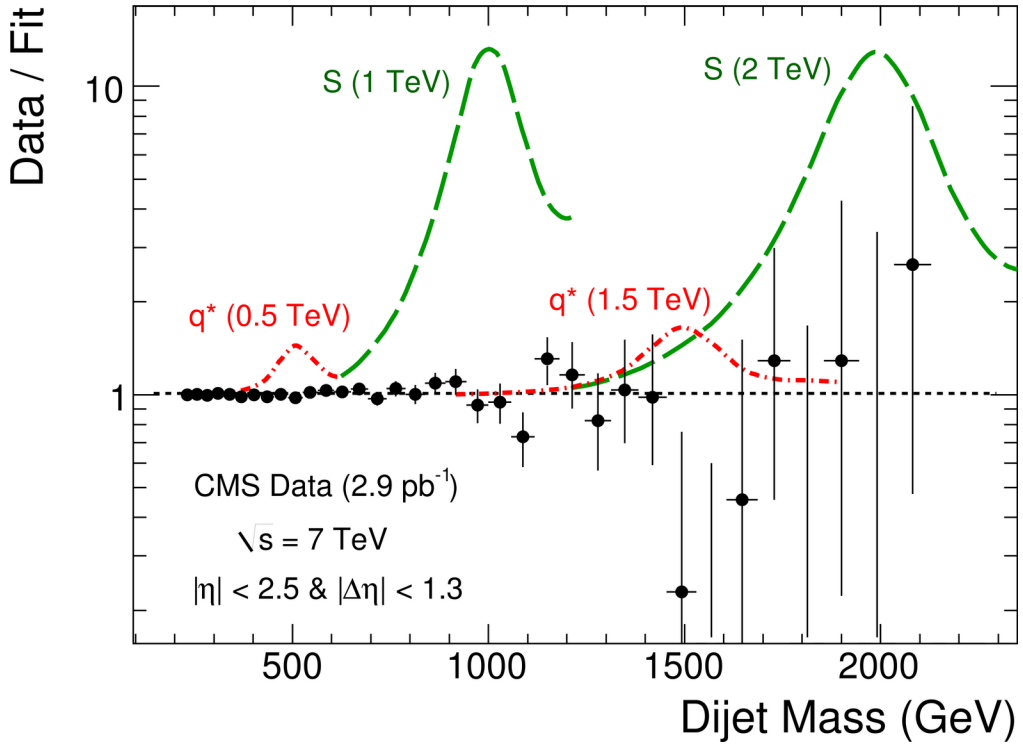
Phys. Lett. B 805 (2020) 135425





Dijet search

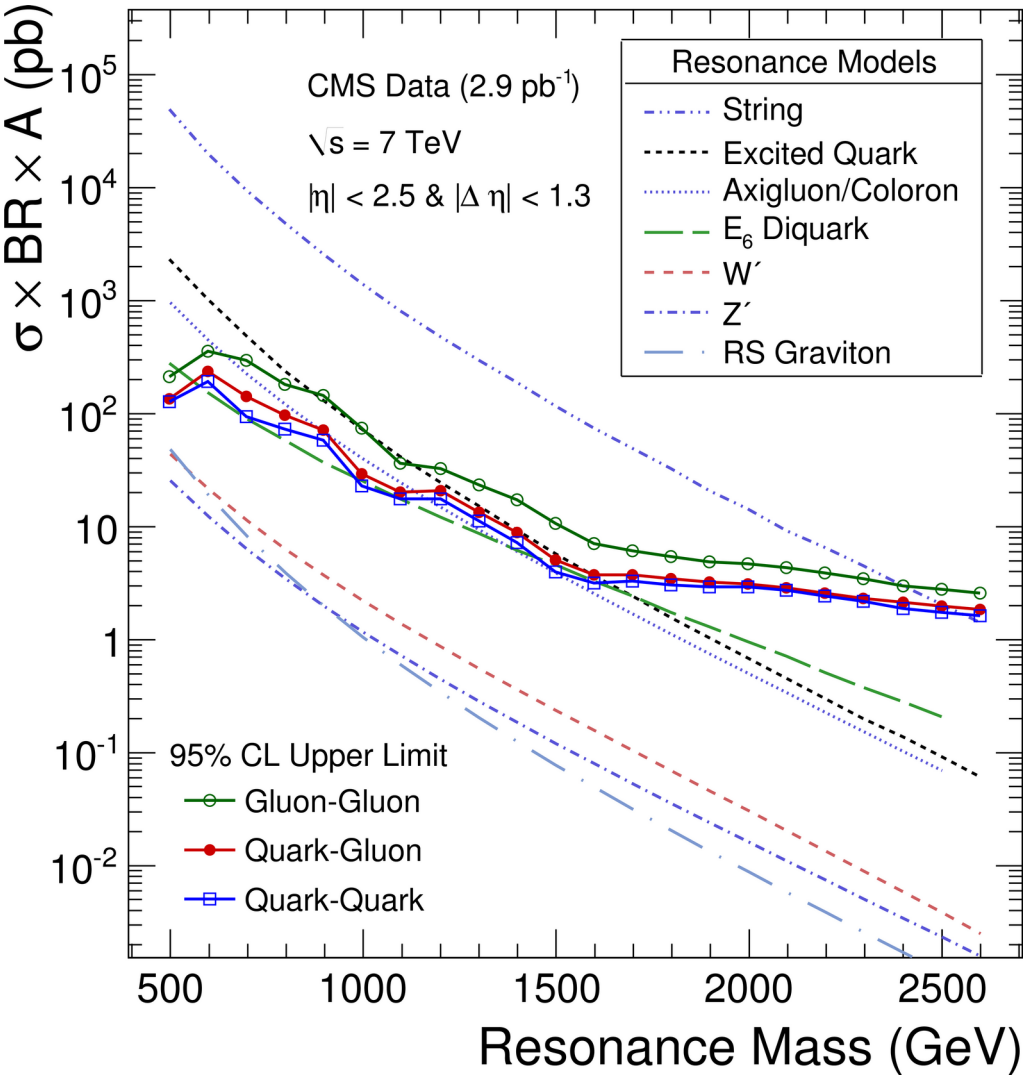


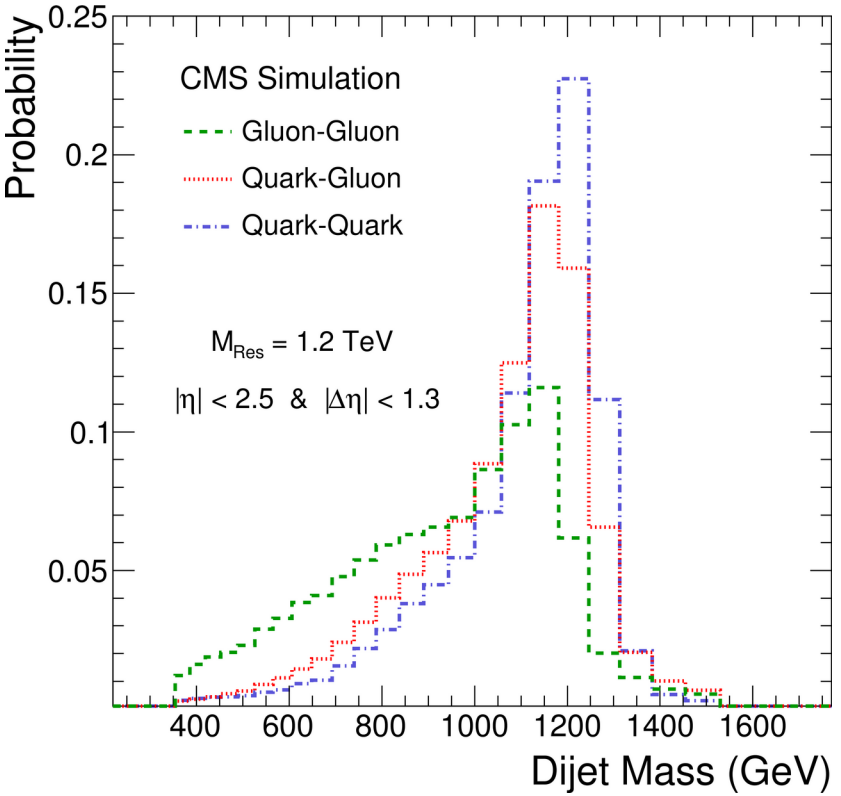


$$\frac{\mathrm{d}\sigma}{\mathrm{d}m} = \frac{P_0(1-m/\sqrt{s})^{P_1}}{(m/\sqrt{s})^{P_2+P_3\ln(m/\sqrt{s})}},$$

Phys. Rev. Lett. 105 (2010) 211801

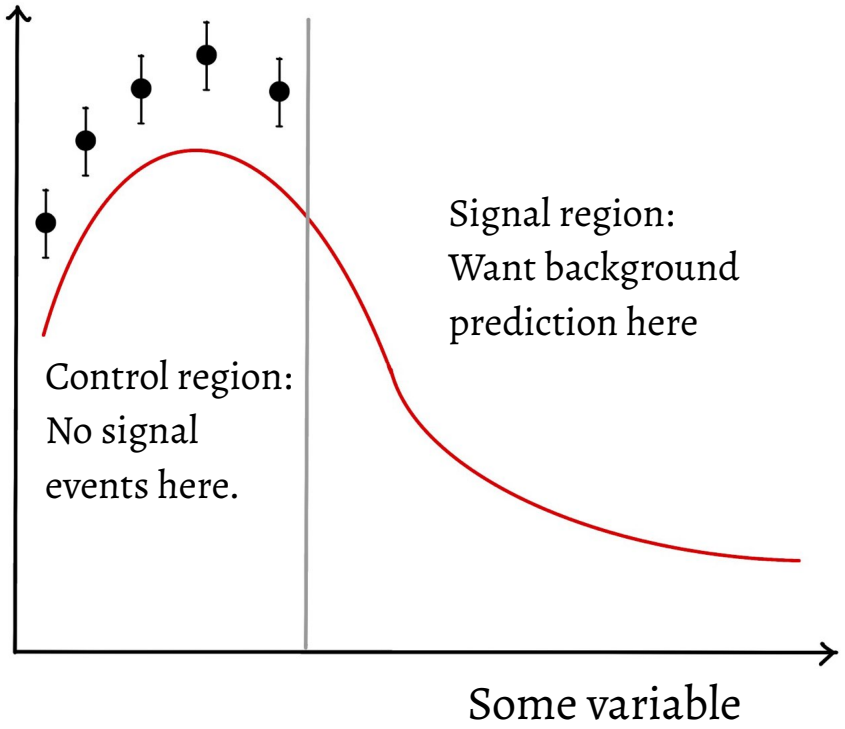
Dijet search





Phys. Rev. Lett. 105 (2010) 211801

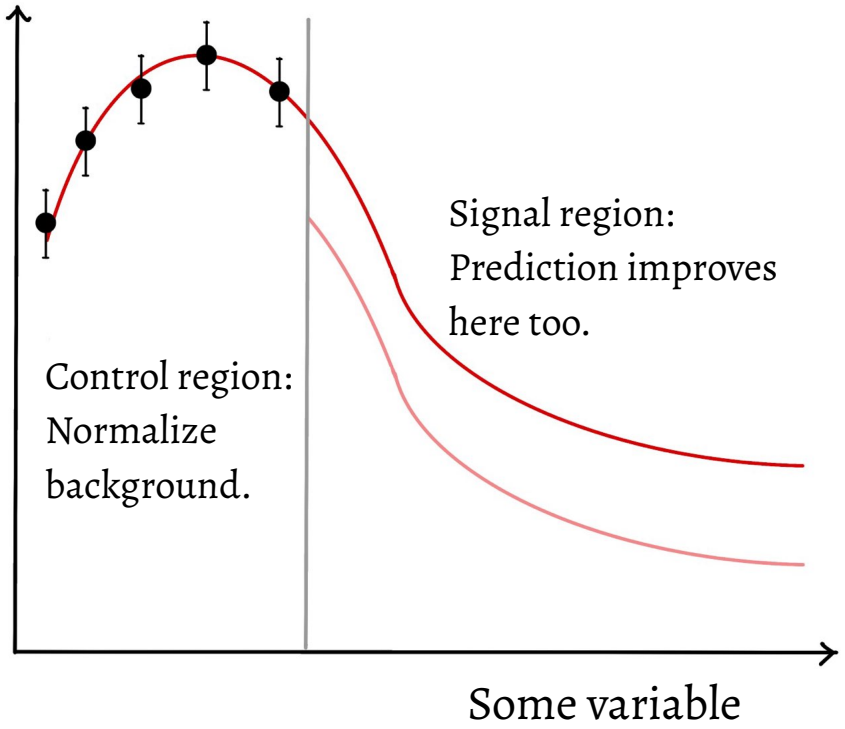
Background *normalized in CR*



Signal Region:
A set of selections we make to enhance signal over background.

Control Region:
A set of selections that signal is unlikely to pass (thus this selection is dominated by background)

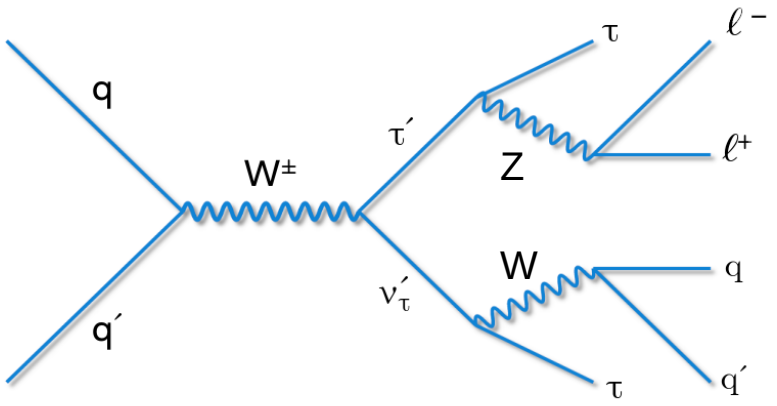
Background *normalized in CR*



Signal Region:
A set of selections we make to enhance signal over background.

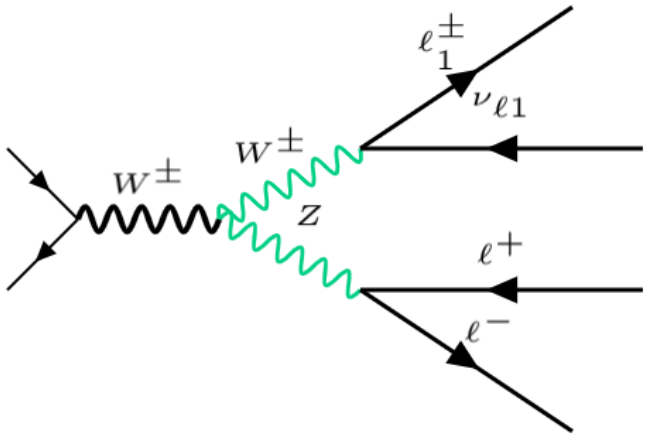
Control Region:
A set of selections that signal is unlikely to pass (thus this selection is dominated by background)

Vector-like tau search



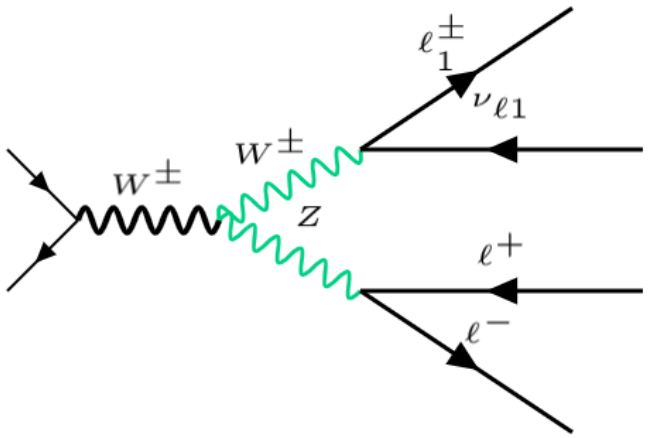
This is signal. (Hypothetical τ', ν')

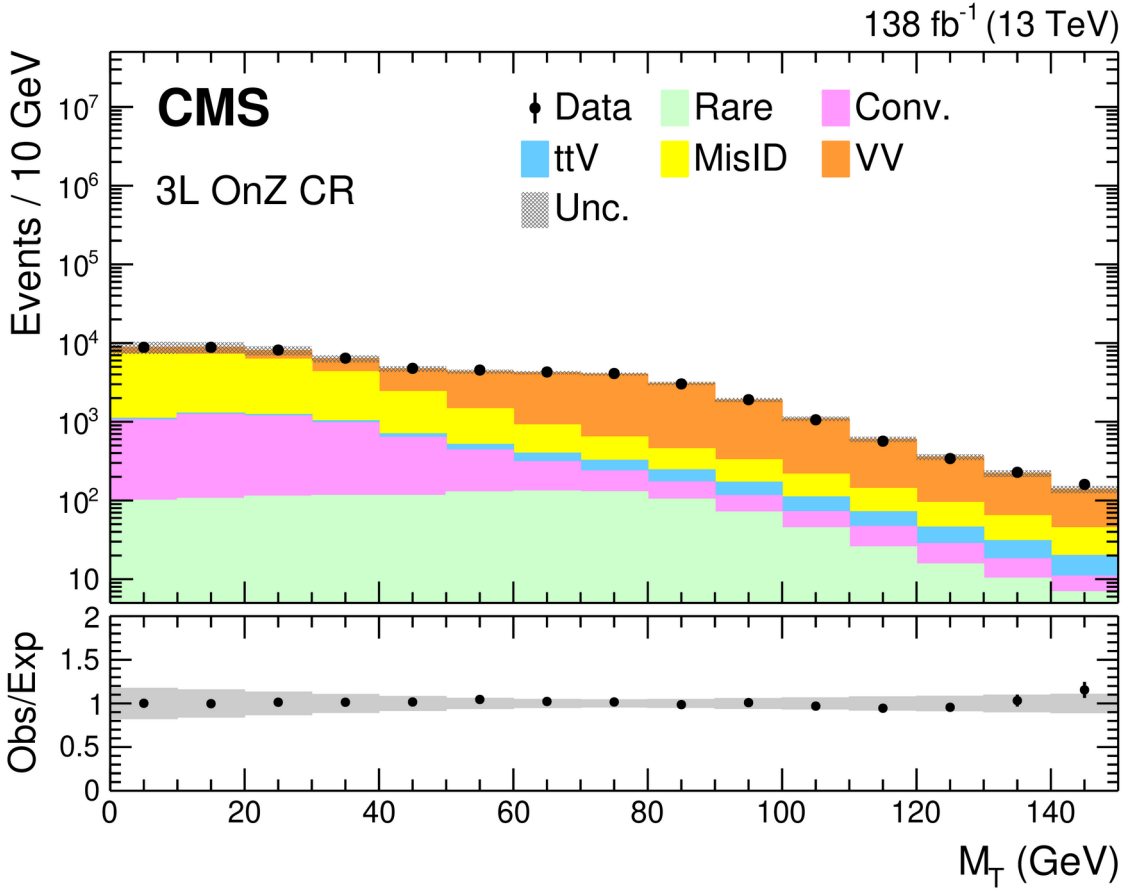
One possible selection is three leptons (e or μ)
(perhaps some p_T^{miss} from tau decay)



Thus this process WZ is background
(It gives three leptons and some p_T^{miss})

Vector-like tau search





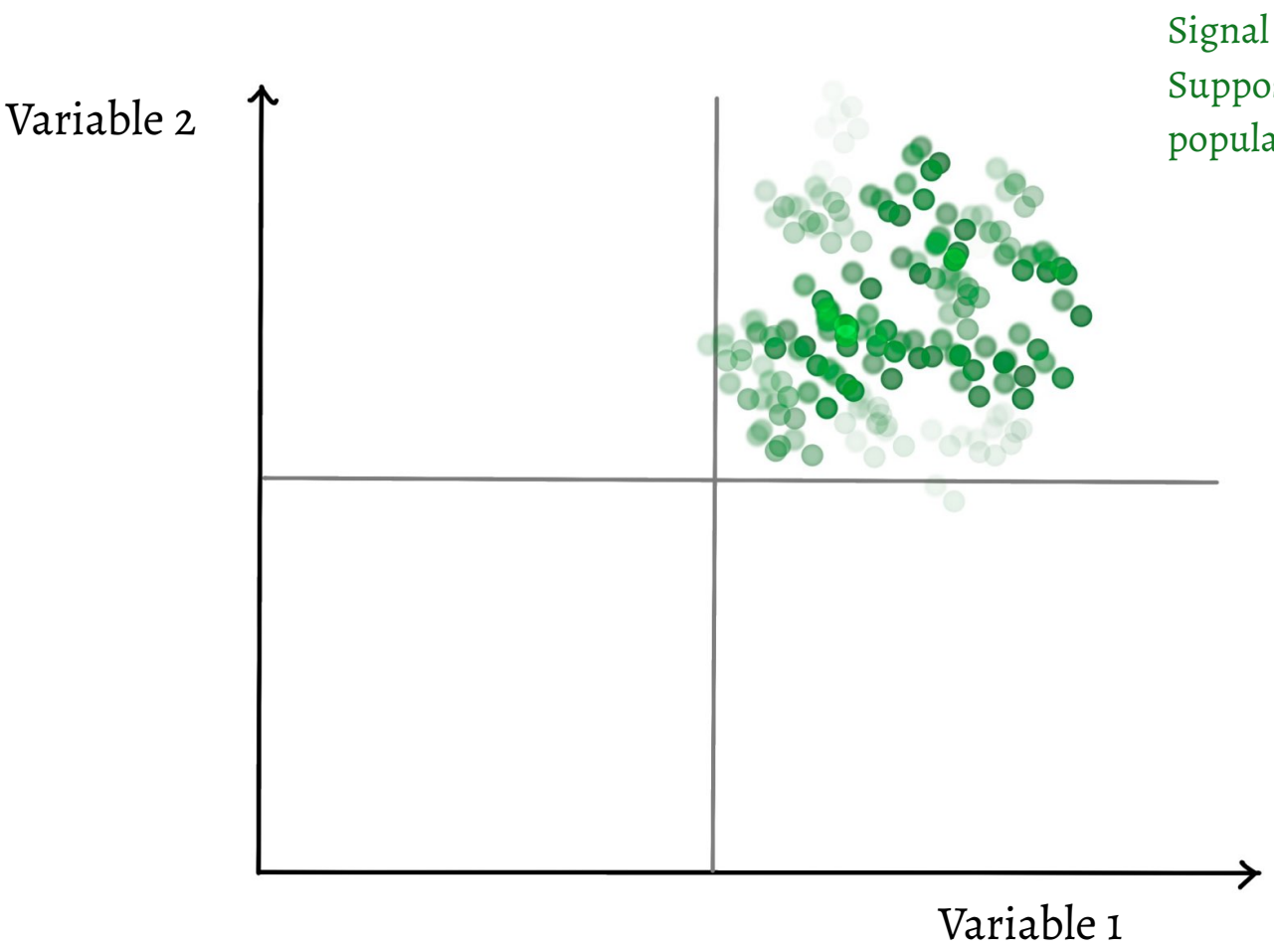
WZ Control region:

- Three leptons
- one pair consistent with 76 < $M_{\ell\ell}$ [GeV]< 106 – Z-tag
- 50 < M_T [GeV] < 150 for other lepton – W-tag

The normalization factor measured in this CR is then used everywhere to multiply the WZ prediction.

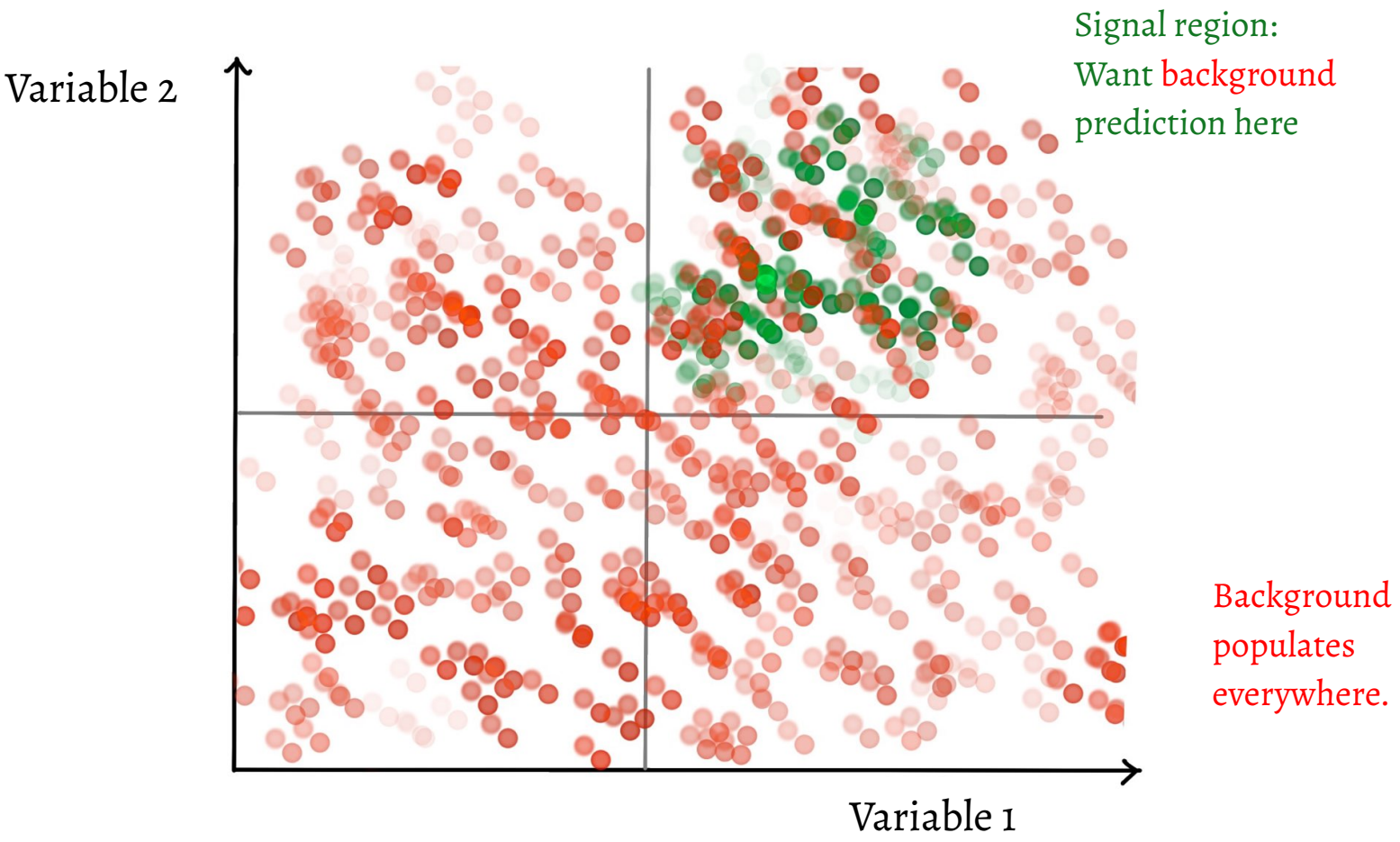
Phys. Rev. D 105 (2022) 112007

ABCD method

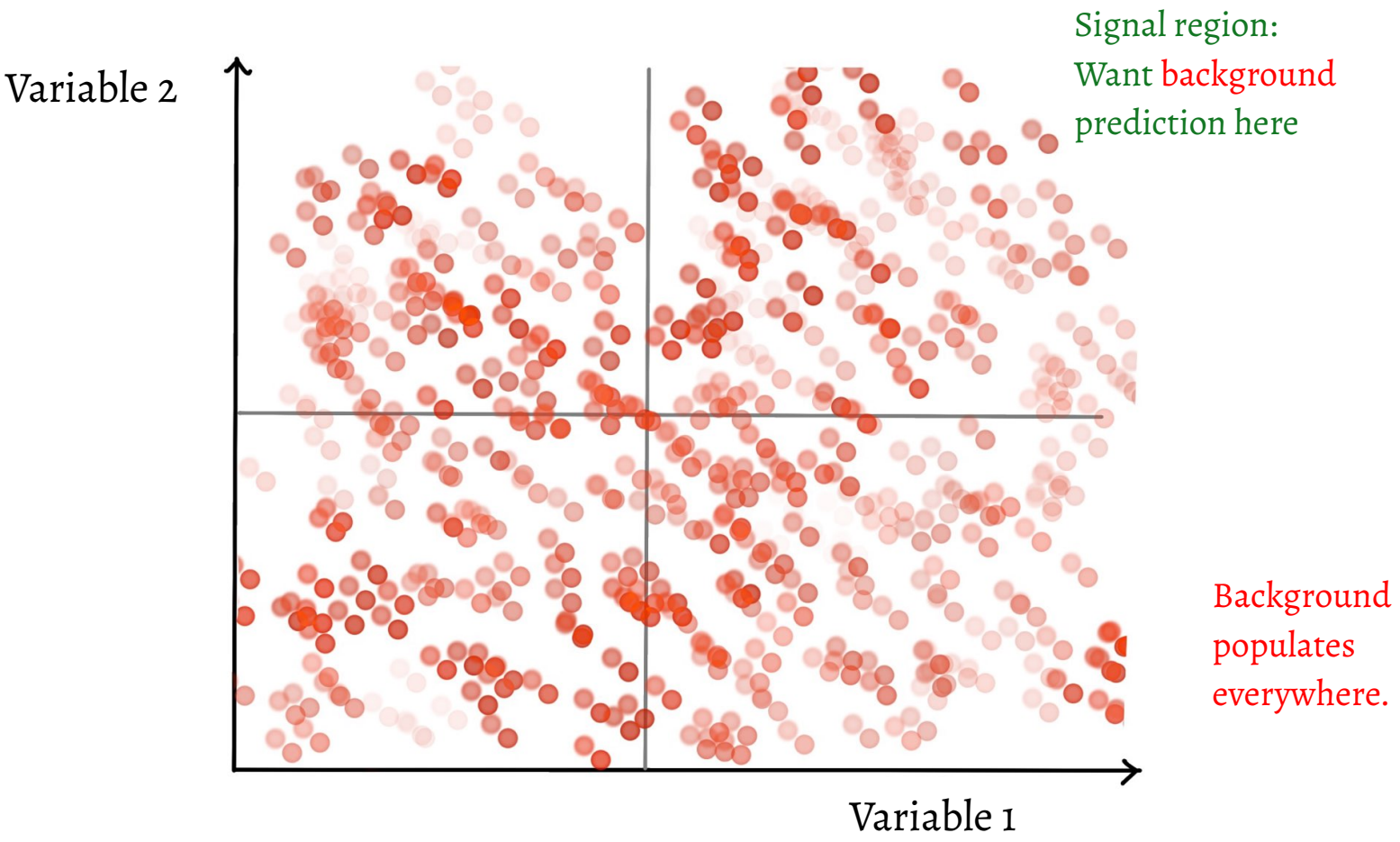


Signal region:
Suppose signal
populates here.

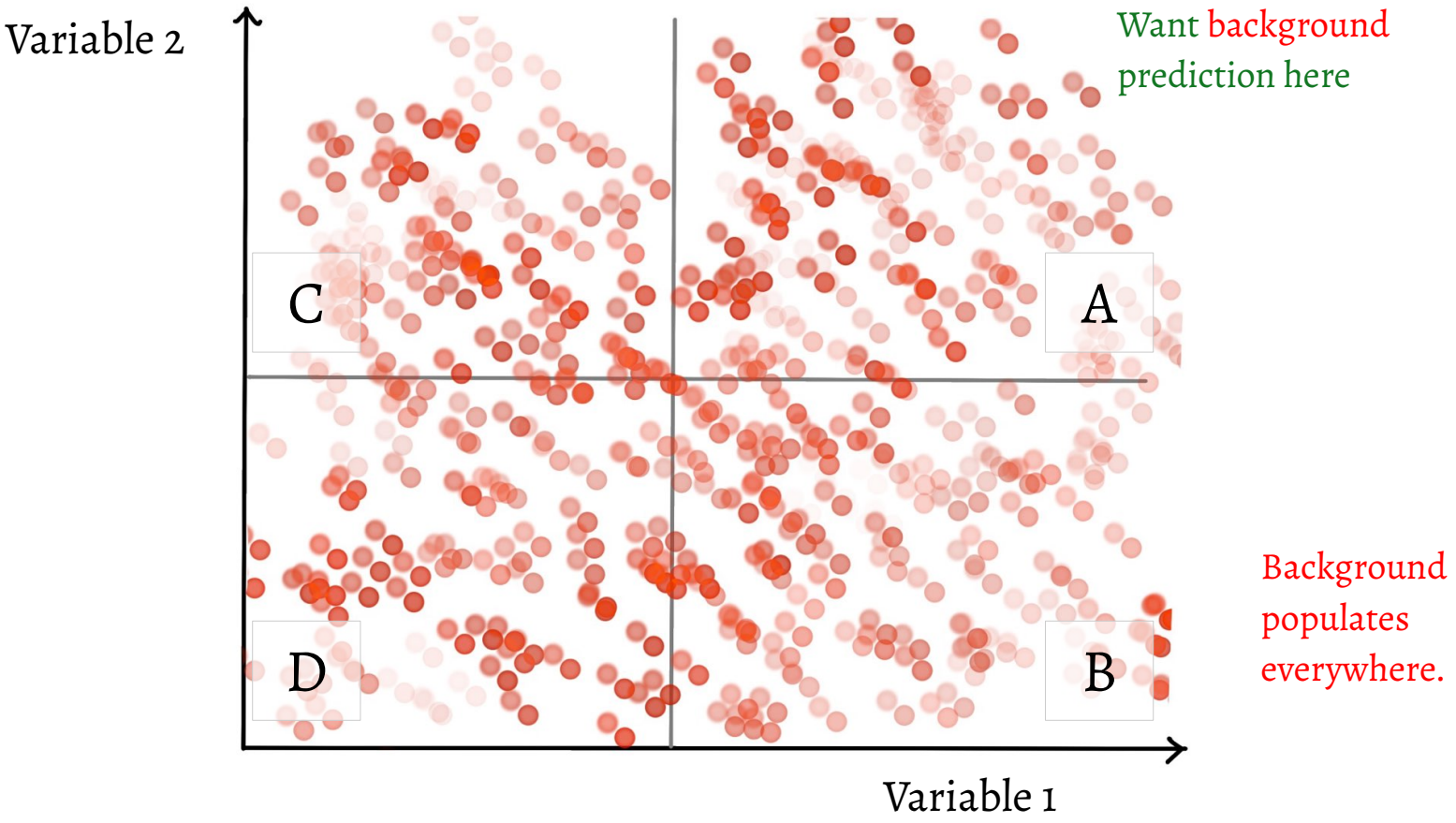
ABCD method


Variable 2
Variable 1
Signal region:
Want background
prediction here
Background
populates
everywhere.

ABCD method


Variable 2
Variable 1
Signal region:
Want background
prediction here
Background
populates
everywhere.

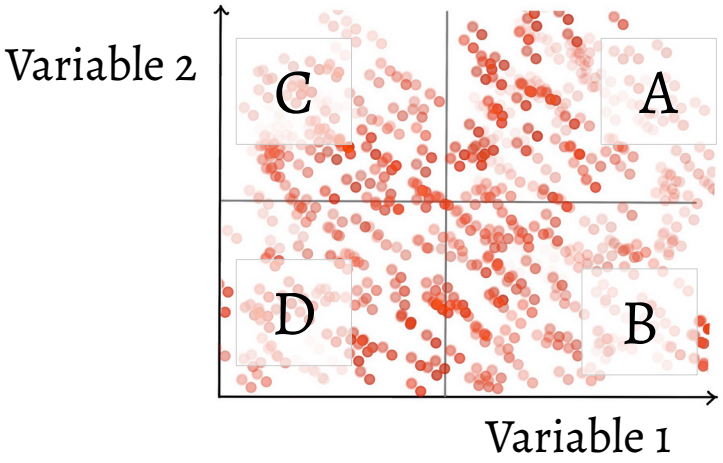
ABCD method



Signal region:
Want background prediction here

Background populates everywhere.

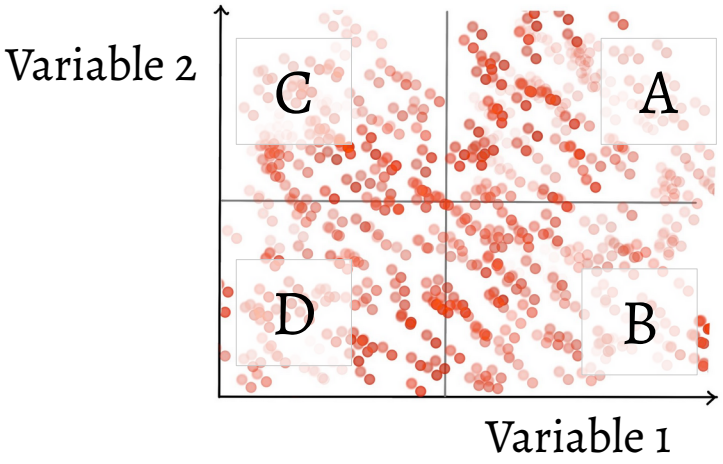
ABCD method



$$\frac{A}{B} = \frac{C}{D}$$

Thus $A = B \cdot C/D$

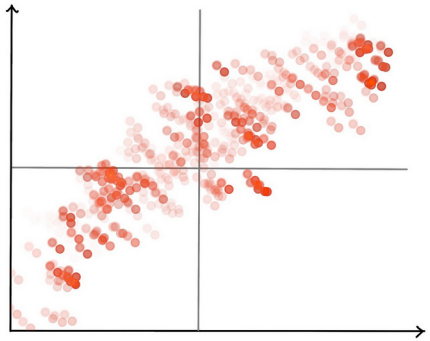
ABCD method



$$\frac{A}{B} = \frac{C}{D}$$

Thus $A = B \cdot C/D$

Of course, this relies on the background evenly populating the plane.
Will not work in cases like this.



Fix: find different variables!

Long-lived Multi-charged particles

Search for particles with electric charge > 1

This search for charge $2 \leq z \leq 7$, and 500 < mass < 2000 GeV

Experimental signature: MCPs are highly ionizing and create large dE/dx

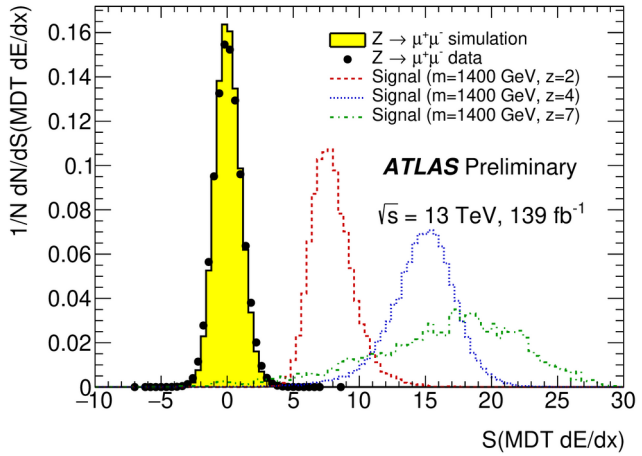
Thus muon-like tracks, but with high dE/dx values in subdetector systems

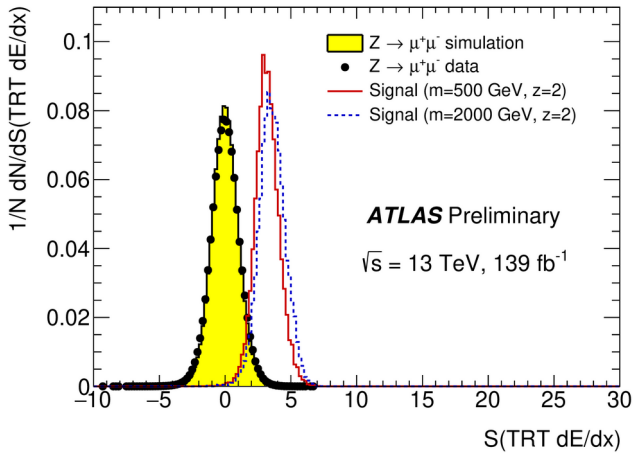
(Muons typically lose 3 GeV in ATLAS calorimeter, MCPs lose z^2 times that)

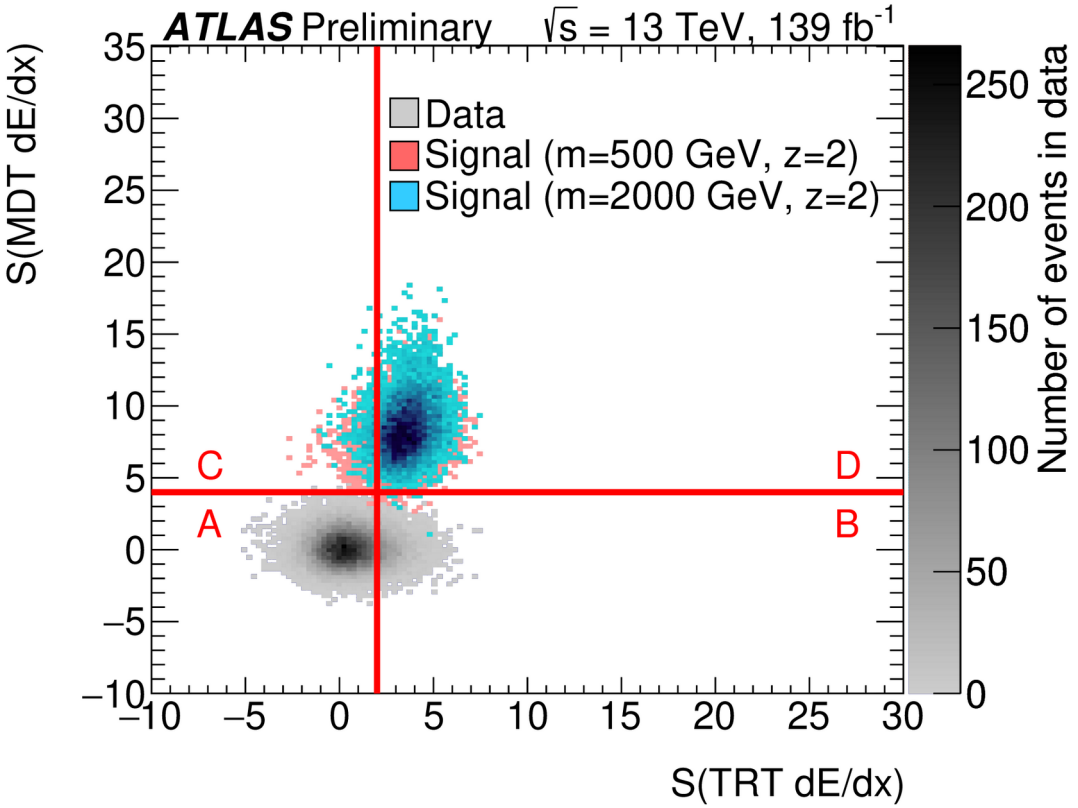
ATLAS-CONF-2022-034

Long-lived Multi-charged particles

$$S(dE/dx) = \frac{dE/dx - \langle dE/dx \rangle_{\mu}}{\sigma(dE/dx)_{\mu}}$$



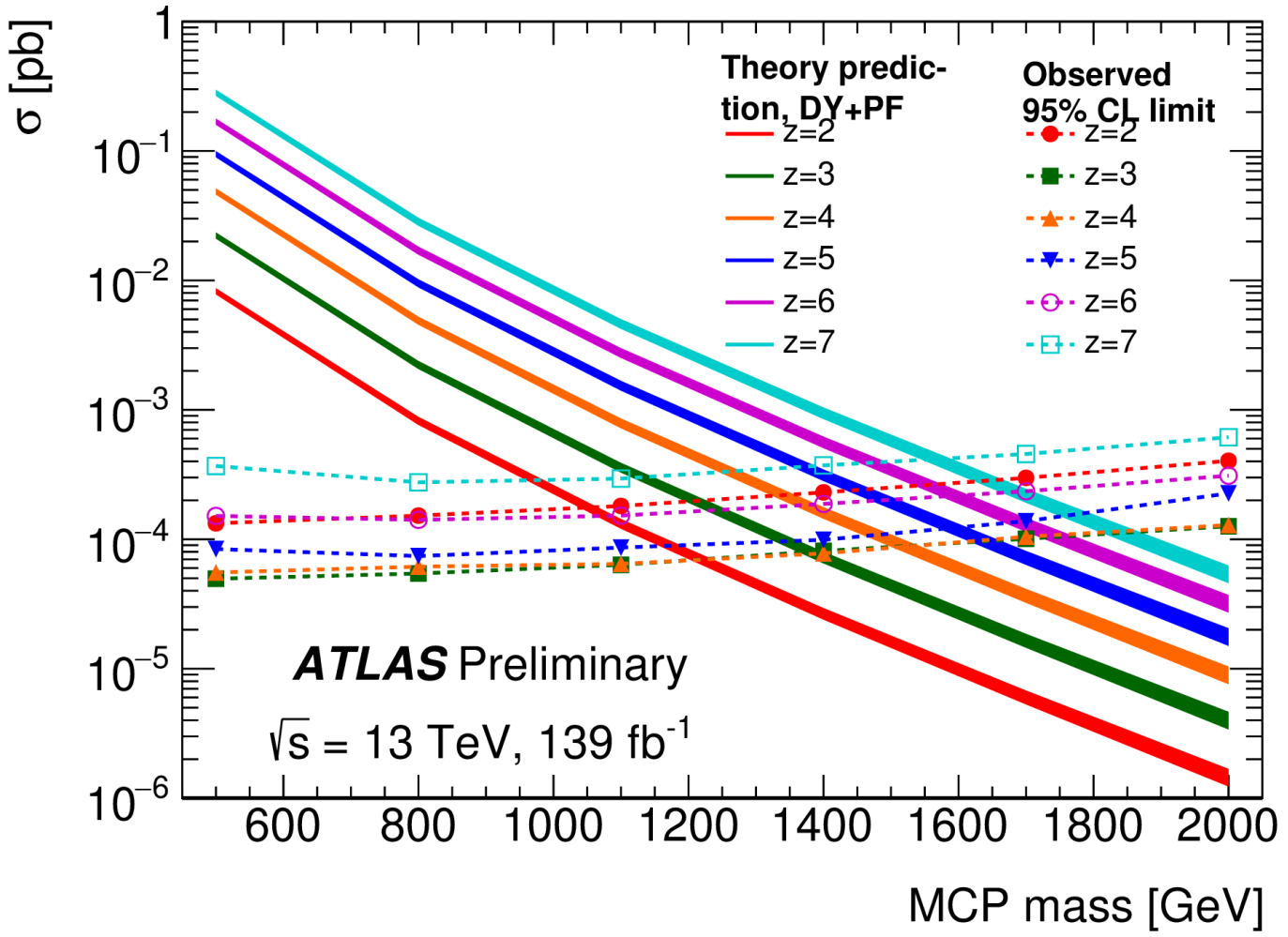




$$N_{\text{data}}^{\text{D expected}} = \frac{N_{\text{data}}^{\text{B observed}} \times N_{\text{data}}^{\text{C observed}}}{N_{\text{data}}^{\text{A observed}}}$$

ATLAS-CONF-2022-034

Long-lived Multi-charged particles



ATLAS-CONF-2022-034

Background estimation

Is as diverse as the analyses are.

Will hear lots of descriptions, data-driven, irreducible, etc.
But go through the description, it will be one of these

- Simulation based
- Simulation corrected using data (i.e. normalized in a CR)
- Data-driven (fit, extrapolation, 2D-extrapolation)

Fakes

All lepton analyses will have this background (and photon analyses too).
It is alternatively called “fakes”, “MisID”, “Non-prompt” etc.

Typically leptons from three sources

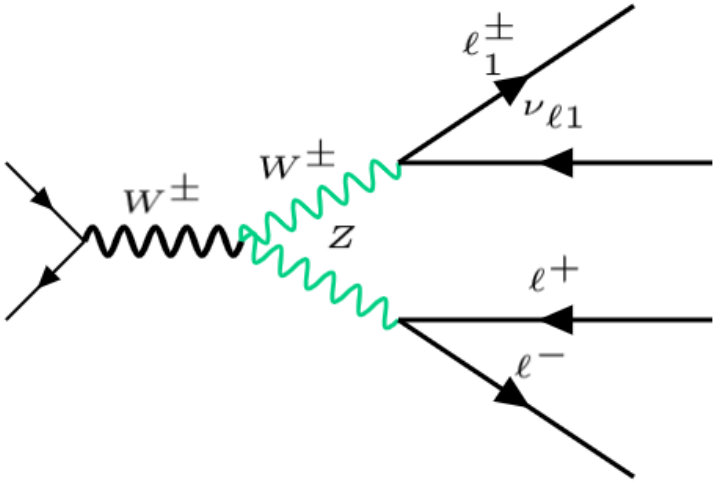
“Prompt and isolated” – directly from collision, or from W,Z,H

From hadron decay – these occur inside jets (since hadrons usually come in jets)

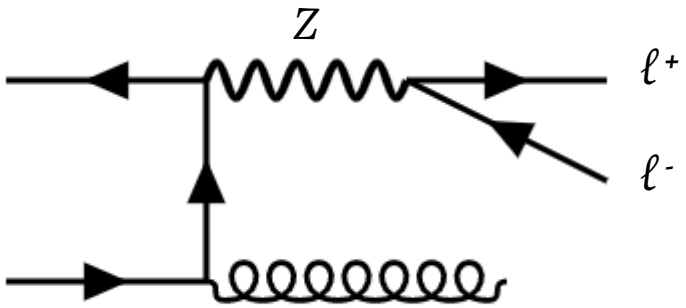
Pure fakes – a detector signature that looks like a lepton, but isn’t. This actually happens mostly when a jet mimics the signature of a prompt lepton

} Fakes

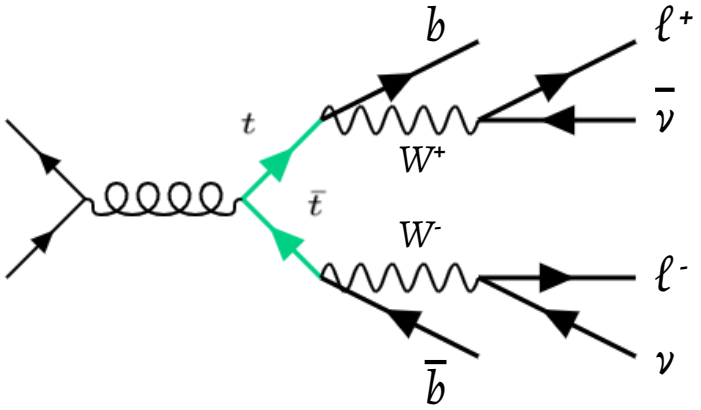
Fakes



WZ producing 3 leptons
σ ~ 5 pb



Z + jets σ ~ 900 pb
Z gives 2 leptons + 1 fake



$t\bar{t}$ gives 2 leptons + 1 fake, σ ~ 90 pb

How often a jet gives rise to a fake lepton depends on specific selections and lepton flavors

Typical fake estimation

Use a control region to measure f = rate of obj → lepton
Here obj can be track or jet etc.

To estimate $3L_{(1\ fake)}$ events,
Select 2L+obj events, and multiply by f

$N_{2L+obj} \times f$ = predicted $N(3L_{1\ fake})$

To estimate $3L_{(2\ fake)}$ events,
Select 1L+2obj events, and multiply by f^2

$N_{1L+2obj} \times f^2$ = predicted $N(3L_{2\ fake})$

Typical fake estimation

Use a control region to measure f = rate of obj → lepton
Here obj can be track or jet etc.

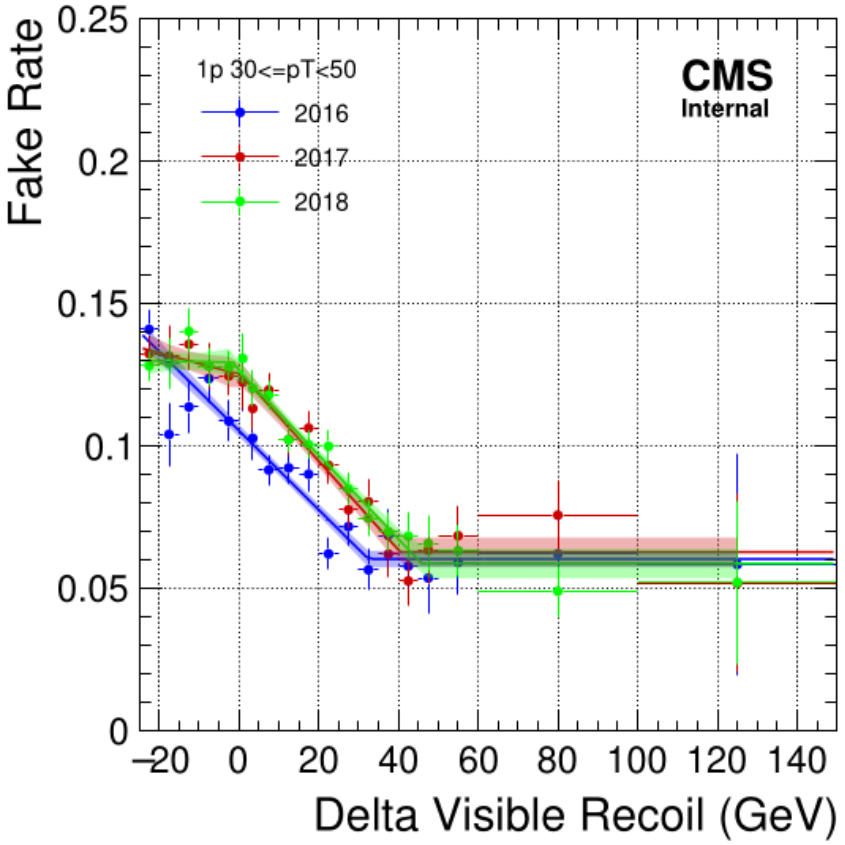
To estimate $3L_{(1\,fake)}$ events,
Select 2L+obj events, and multiply by f
$N_{2L+obj} \times f$ = predicted $N(3L_{1\,fake})$

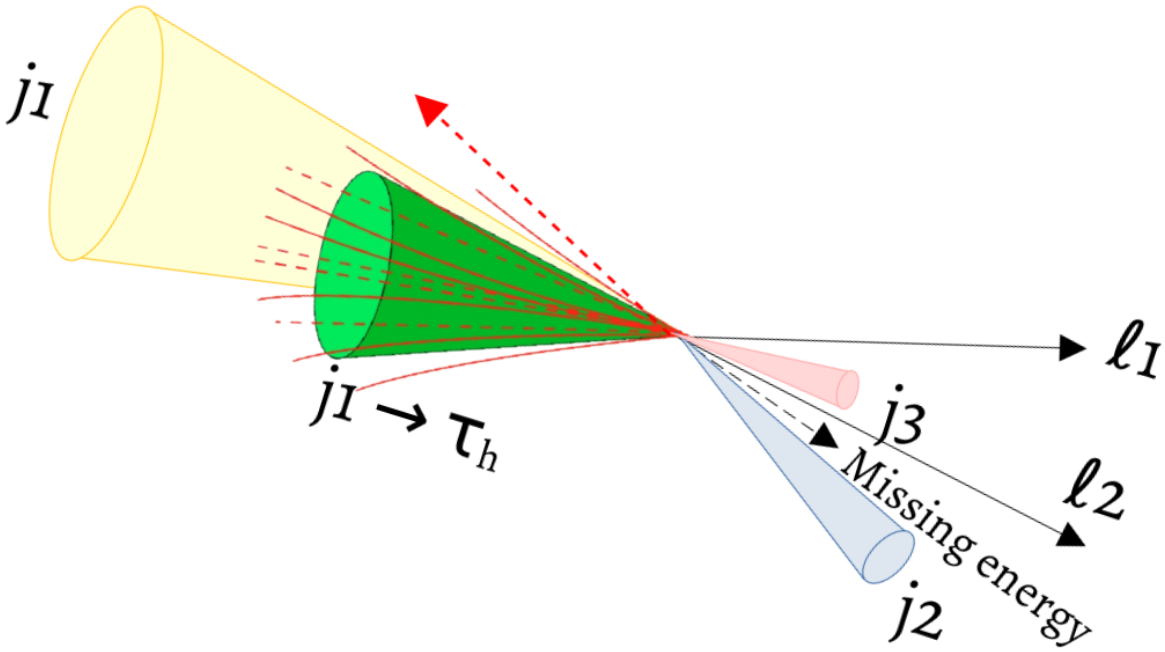
To estimate $3L_{(2\,fake)}$ events,
Select 1L+2obj events, and multiply by f^2
$N_{1L+2obj} \times f^2$ = predicted $N(3L_{2\,fake})$

Can generalize to predict $2L_{1fake}$, $4L_{2fake}$ etc.

Typically f will depend on lepton flavor, p_T, η as well as the obj chosen.

Inclusive nonresonant multileptons



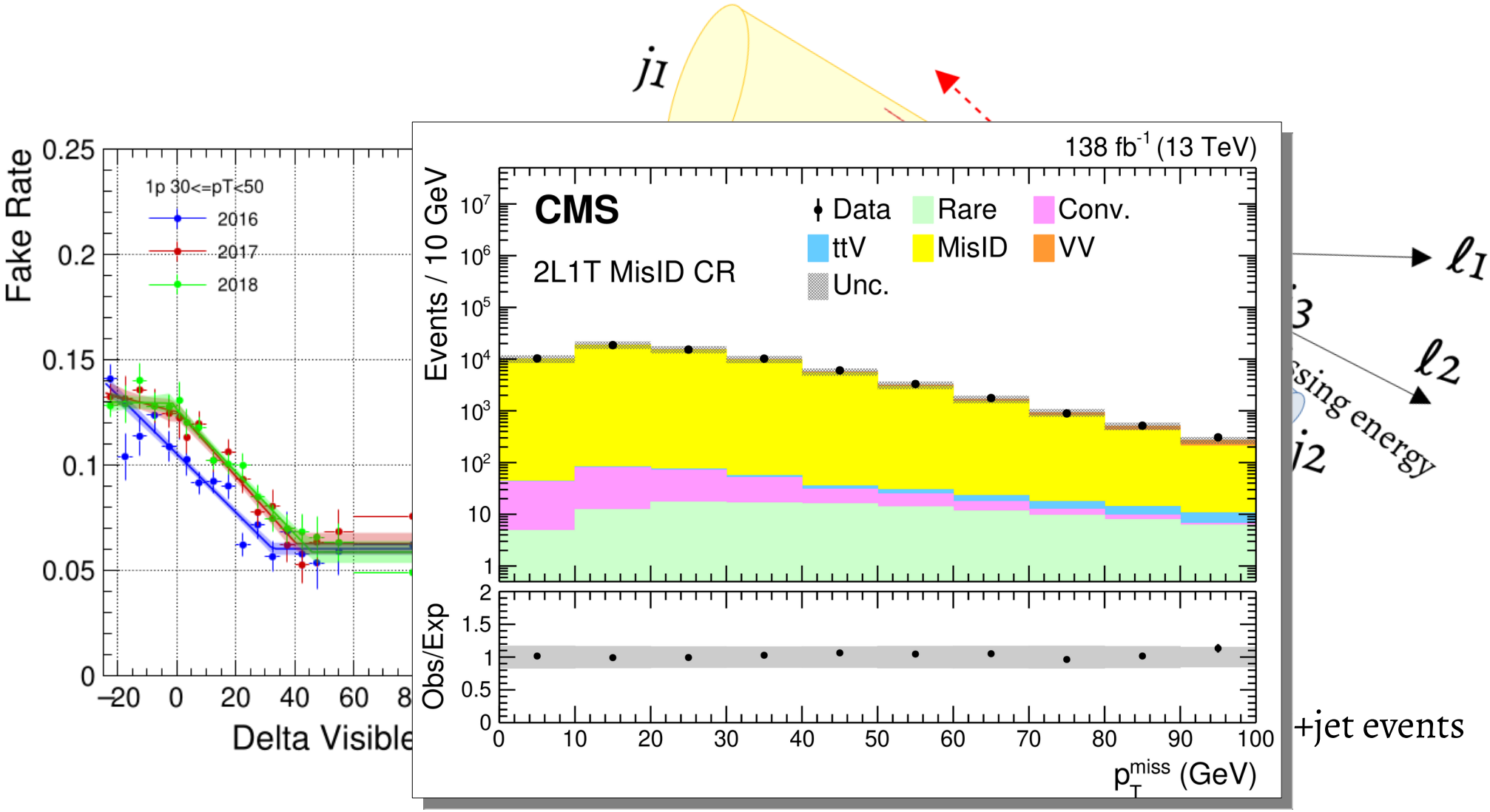


Measuring fake rate of jet → τ in Z+jet events

Rastogi, A 2022, PhD thesis

Phys. Rev. D 105 (2022) 112007

Inclusive nonresonant multileptons



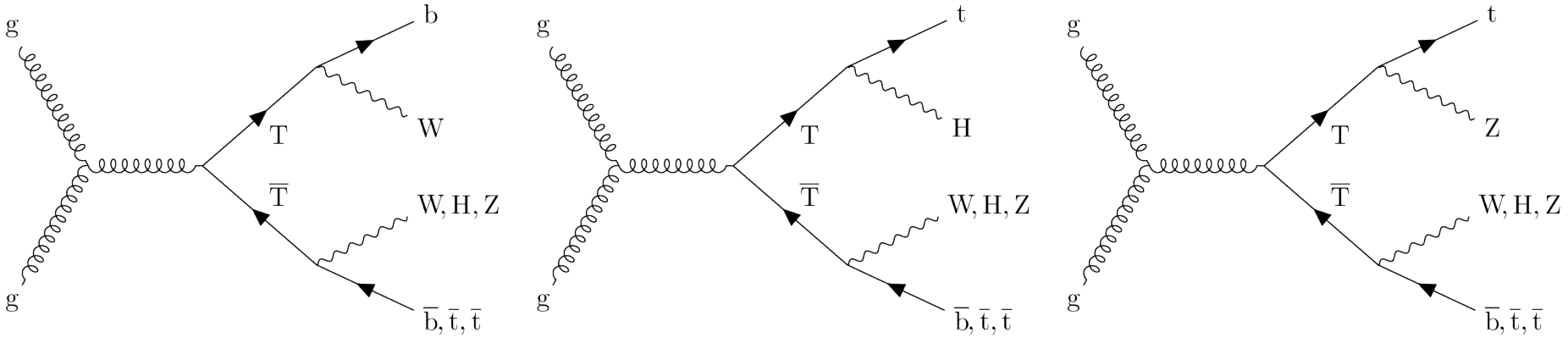
Rastogi, A 2022, PhD thesis

Phys. Rev. D 105 (2022) 112007

Jets and substructure

Search for vector-like T and B

JHEP 11 (2017) 085



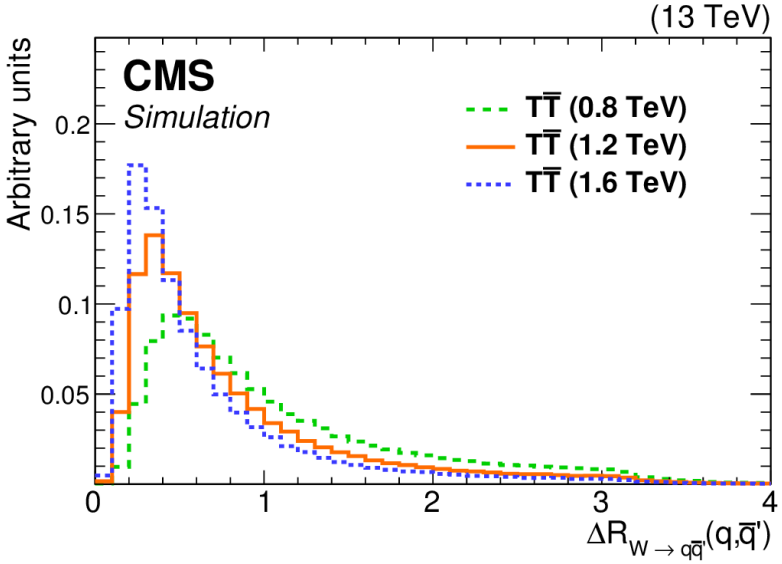
T additional vector-like quark with charge ⅔ e

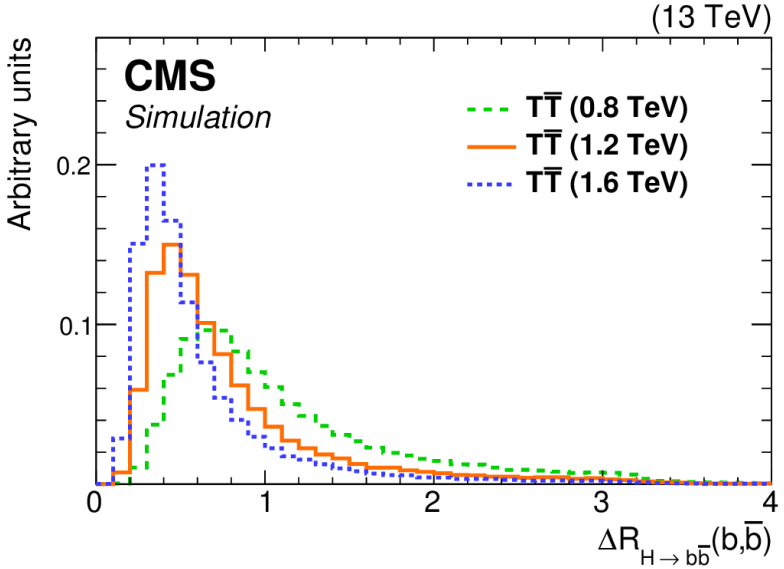
B additional vector-like quark with charge -⅓ e

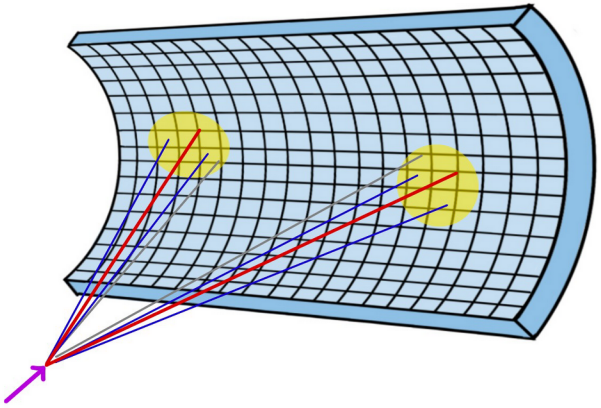
Masses more than 700 GeV considered here.

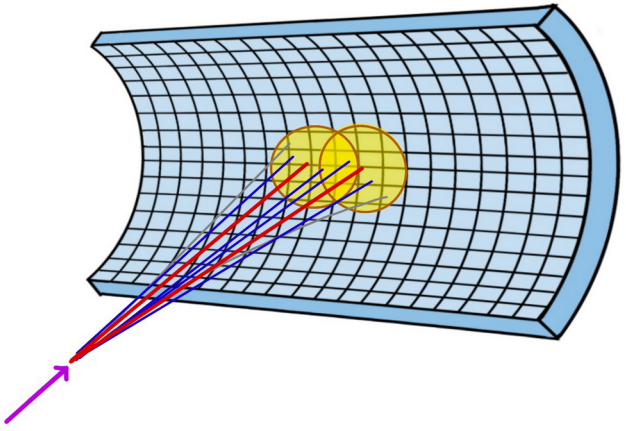
Search for vector-like T and B

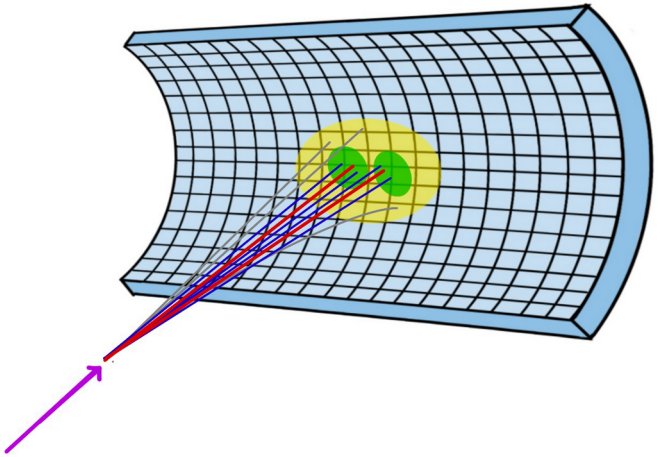
JHEP 11 (2017) 085





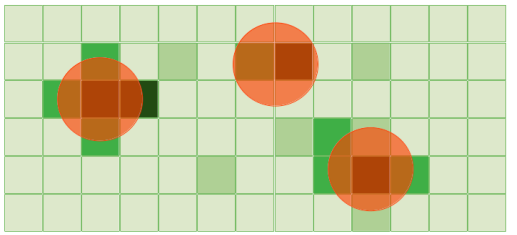






Recall jets

$$d_{ij} = min(k_{ti}^{2p}, k_{tj}^{2p})\frac{\Delta_{ij}^2}{R^2}$$

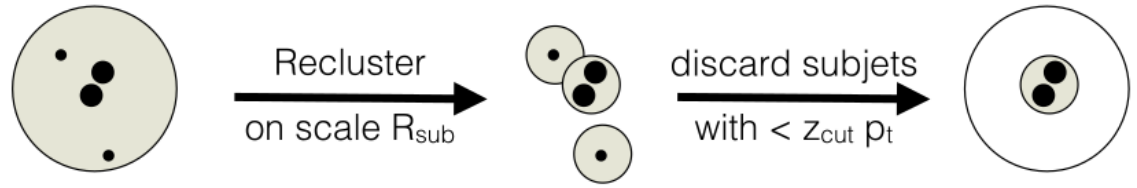


After we have a jet, and its 'constituents' we can study several variables and experimentally measure them

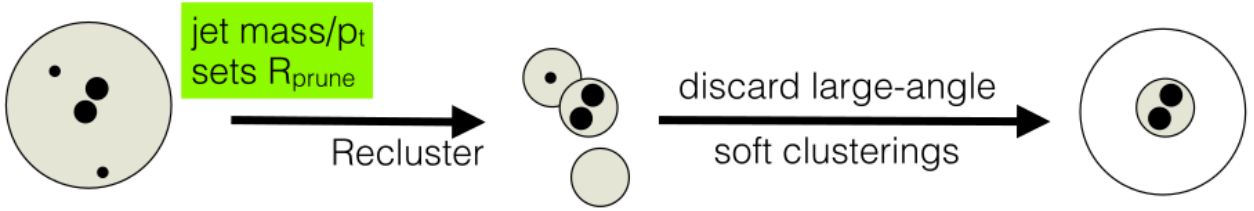
0. Grooming a jet
1. Particle multiplicity, taking into account pT fraction of each particle, and angular separation of each particle from jet axis.
2. The overall shape of the jet (circular → elliptical) based on energy distribution
3. Decluster the jet into subjets –
   Study these subjets: multiplicity, momentum fraction, angular behavior
   N-subjettiness and ratios
4. Energy correlation functions

Phys. Rev. D 98 (2018) 092014

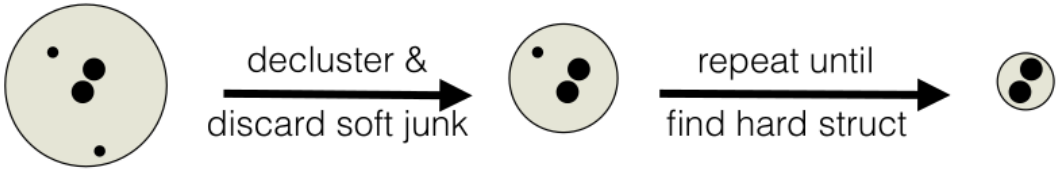
Grooming



Trimming

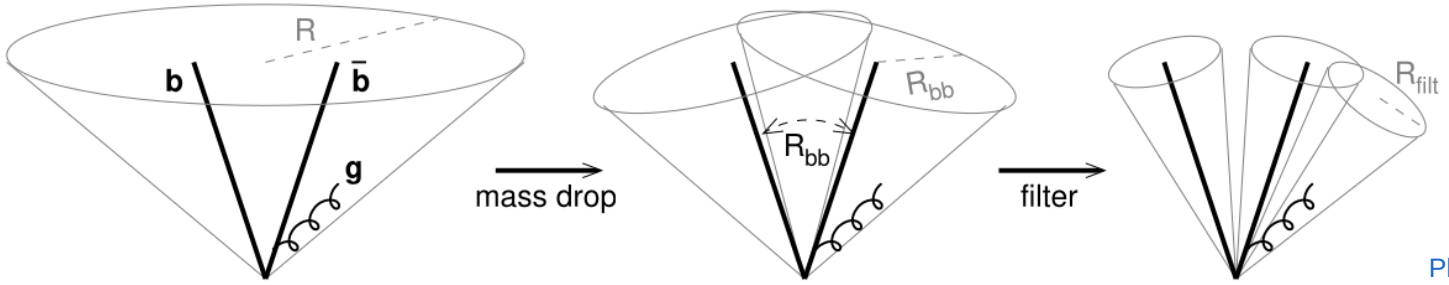


Pruning



Mass drop

Gavin Salam, @IHEP, 2014



Phys.Rev.Lett.100:242001,2008

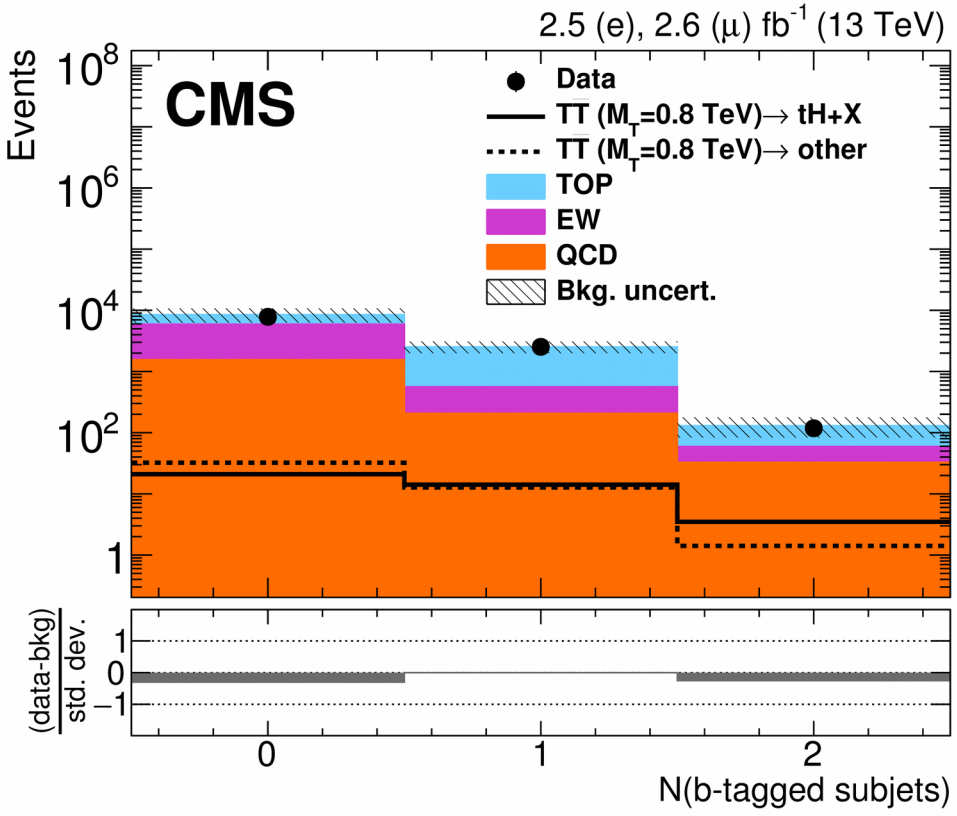
Search for vector-like T and B

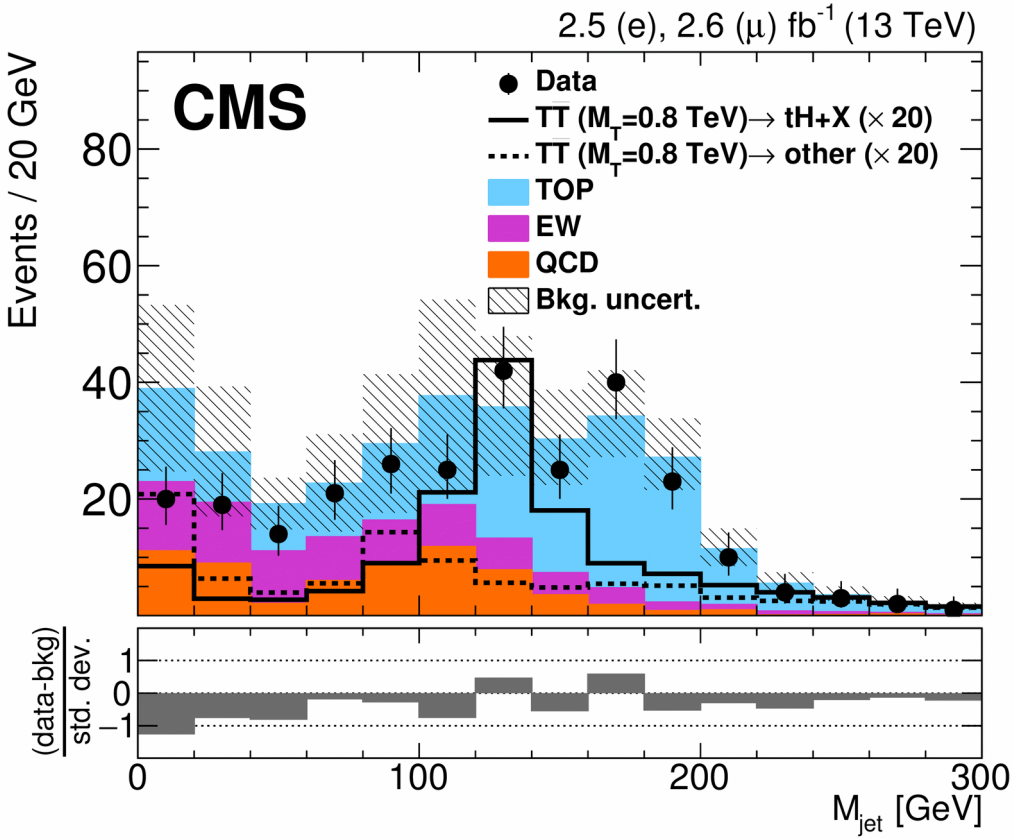
JHEP 11 (2017) 085

Consider AK8 jets – groom them using pruning and soft drop.

Tag the subjets as b-tags.

Tag jet as H-jet if p_T>300 GeV, 60<jet_{mass}<160 GeV, and one 1 b-tagged subjet





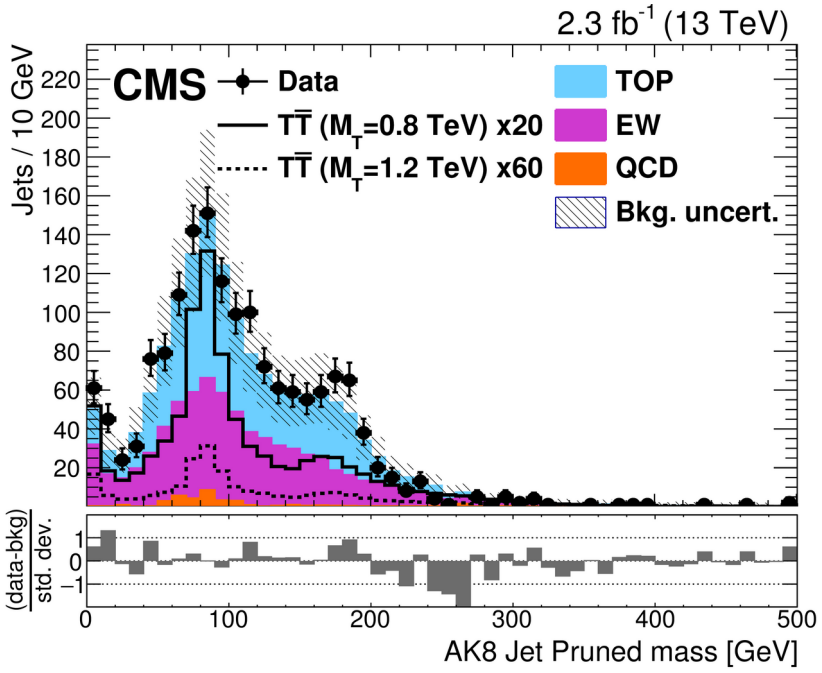
Search for vector-like T and B

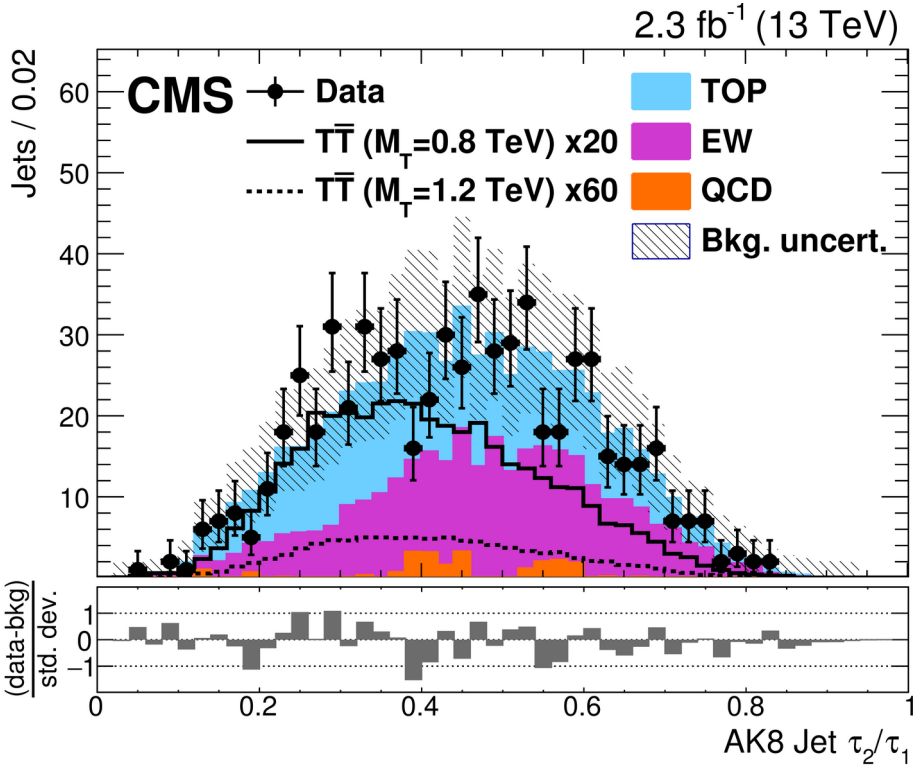
JHEP 11 (2017) 085

Consider AK8 jets – groom them using pruning and soft drop.

Tag jet as W-jet if p_T>200 GeV, 65<jet_{mass}<105 GeV, and $\tau 2/\tau 1 < 0.6$

N-subjettiness $\tau_N = \frac{1}{R_0 \sum_k p_{T,k}} \sum_k p_{T,k} \min(\Delta R_{1,k}, \Delta R_{2,k}, \ldots, \Delta R_{N,k})$





Lets talk Uncertainties

Statistical

Statistical uncertainties arise because of stochastic fluctuations in a measurement given that it is based on a finite set of observations.

Make some measurement with 10 data points – get an answer.
Repeat with some other 10 data points – won't get identical answer.
(Presumably measurements with 100 points is better than one with 10 points!)

In our case, there are two obvious places where this happens
1. Data: given that we observe some events, there is an inherent statistical uncertainty in that… for example data in a control region (as well data in signal region!)
2. Predictions from simulations: $N = L\ \sigma\ B\ A\ \varepsilon$
Here $A \cdot \varepsilon$ are based on simulation.
Given the finite size of the simulation sample, there is a statistical uncertainty.

Systematic

Systematic uncertainties arise because of the specific nature of procedures (experiments, algorithms) and their limitations, assumptions made, or from inadequacy of the precise underlying theoretical model used.

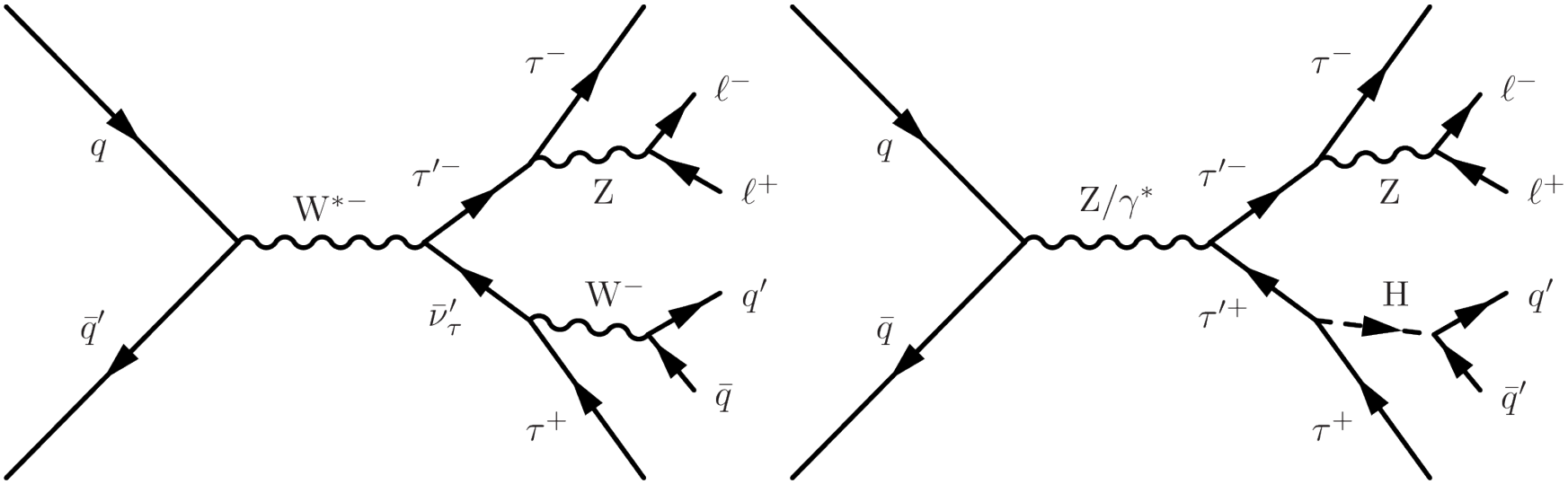
Typical sources:
Estimated luminosity of the experiment
Efficiency measurement (trigger, object reconstruction, identification)
Jet energy scale and resolution
Background estimation uncertainties (normalization, fake rates, different methods)
And others…..
In the theory:
Uncertainties in cross section (and BR) calculations
Parton distribution functions (can affect cross section, but also)
Amount of radiation, renormalization/factorization scales
And others……

Example

Uncertainty source	Magnitude	Type	Processes	Variation	Correlation
Statistical	1–100%	per event	All MC samples	1–100%	No
Integrated luminosity	1.2–2.5%	per event	Conv./Rare/Signal	1.2–2.5%	Yes
Electron/Muon reco., ID, and iso. efficiency	1–5%	per lepton	All MC samples	2–5%	No
τ_h reco., ID, and iso. efficiency	5–15%	per lepton	All MC samples	5–25%	No
Lepton displacement efficiency	1–2%	per lepton	All MC samples	3–5%	No
Trigger efficiency	1–4%	per lepton	All MC samples	<3%	No
b tagging efficiency	1–10%	per jet	All MC samples	2–5%	No
Pileup	5%	per event	All MC samples	<3%	Yes
PDF, fact./renorm. scale	<20%	per event	All MC samples	<10%	Yes
Jet energy scale	1–10%	per jet	All MC samples	<5%	No
Unclustered energy scale	1–25%	per event	All MC samples	<2%	No
Electron energy scale and resolution	<2%	per lepton	All MC samples	<5%	Yes
Muon energy scale and resolution	2%	per lepton	All MC samples	<5%	No
τ_h energy scale	<10%	per lepton	All MC samples	<5%	No
Electron charge misidentification	30%	per lepton	All MC samples	<25%	No
WZ normalization	3–5%	per event	WZ	3–5%	No
ZZ normalization	4–5%	per event	ZZ	4–5%	No
$t\bar{t}Z$ normalization	15–25%	per event	$t\bar{t}Z$	15–25%	No
Conversion normalization	10–50%	per event	$Z\gamma$/Conv.	10–50%	No
Rare normalization	50%	per event	Rare	50%	No
Prompt and misidentification rates	20–60%	per lepton	MisID	20–50%	No
DY-$t\bar{t}$ process dependence	5–25%	per lepton	MisID	5–25%	Yes
Diboson jet multiplicity modeling	<30%	per event	WZ/ZZ	5–30%	No
Diboson p_T modeling	<30%	per event	WZ/ZZ	5–15%	No

Phys. Rev. D 105 (2022) 112007

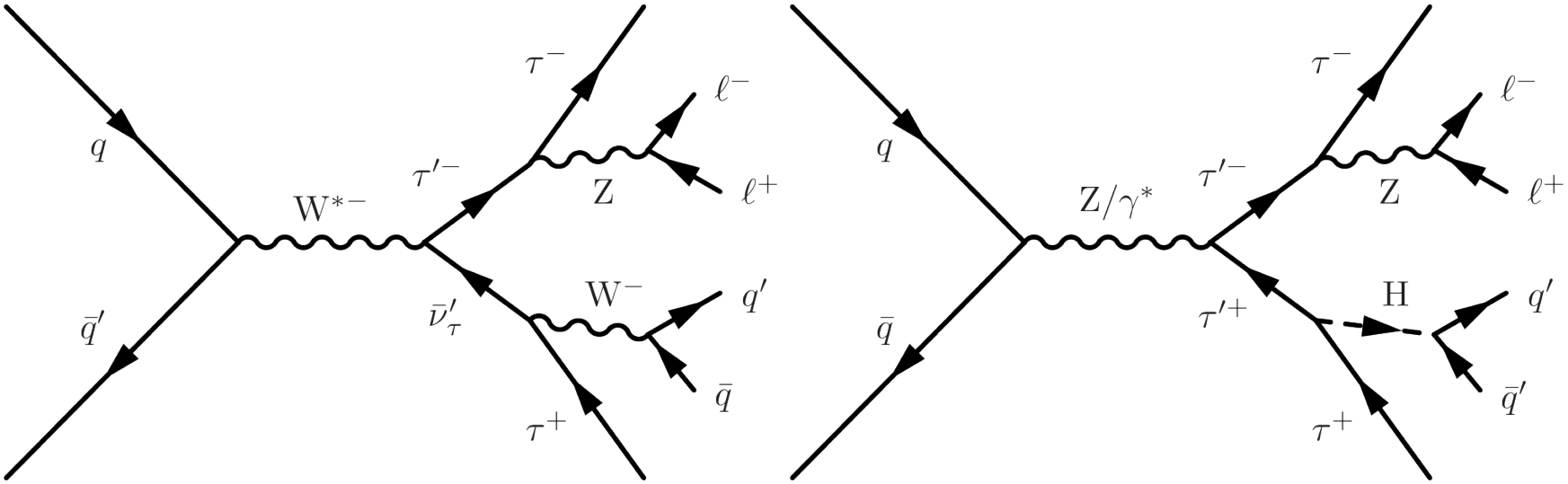
An example analysis: VLL



- Hypothesis: two new particles τ' and ν' and corresponding antiparticles (all have same mass)
- They carry same quantum numbers as τ and ν
- They are pair produced $\tau'\bar{\tau}'$ and $\nu'\bar{\nu}'$ or associated $\tau'\nu'$
- The mass of τ' and ν' can be anything > 150 GeV
- Possible decays: $\tau' \to Z\tau$ and $\tau' \to H\tau$, $\nu' \to W\tau$

What possible final states?

An example analysis: VLL



- Hypothesis: two new particles τ' and ν' and corresponding antiparticles (all have same mass)
- They carry same quantum numbers as τ and ν
- They are pair produced $\tau'\bar{\tau}'$ and $\nu'\bar{\nu}'$ or associated $\tau'\nu'$
- The mass of τ' and ν' can be anything > 150 GeV
- Possible decays: $\tau' \to Z\tau$ and $\tau' \to H\tau$, $\nu' \to W\tau$

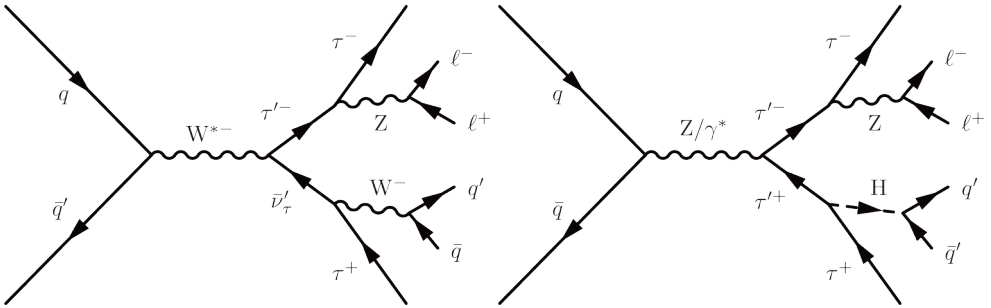
What possible final states?

ZτZτ , ZτHτ , HτHτ

WτWτ

ZτWτ, HτWτ

An example analysis: VLL

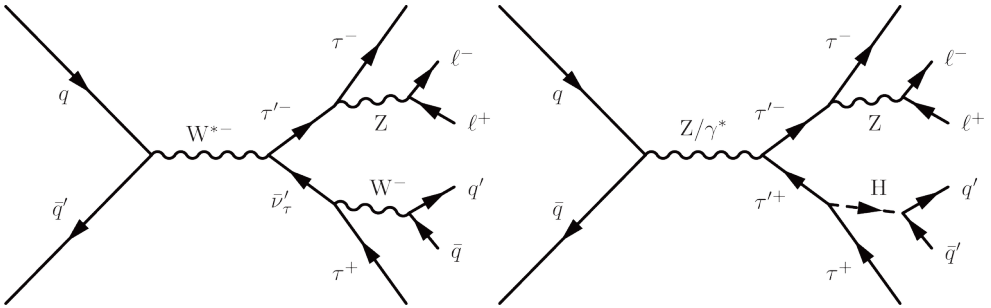


$N_{leptons}$
$\geq 4e/\mu$
$3e/\mu$
$2e/\mu$ OS (or SS) + $\geq 1\tau_h$

So what will be the backgrounds?

Phys. Rev. D 100, 052003 (2019)

An example analysis: VLL



$N_{leptons}$
$\geq 4e/\mu$
$3e/\mu$
$2e/\mu$ OS (or SS) + $\geq 1\tau_h$

So what will be the backgrounds?

Processes that give 4 or more leptons

ZZ, ttZ, H→ZZ

Processes that give 3 or more leptons

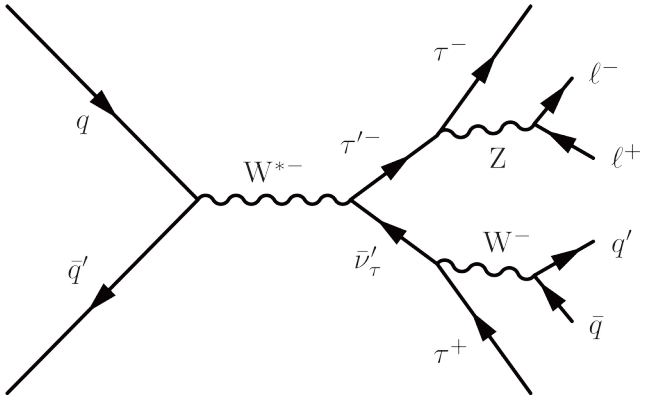
WZ, ttW

Processes that give 2 or more leptons (+fake)

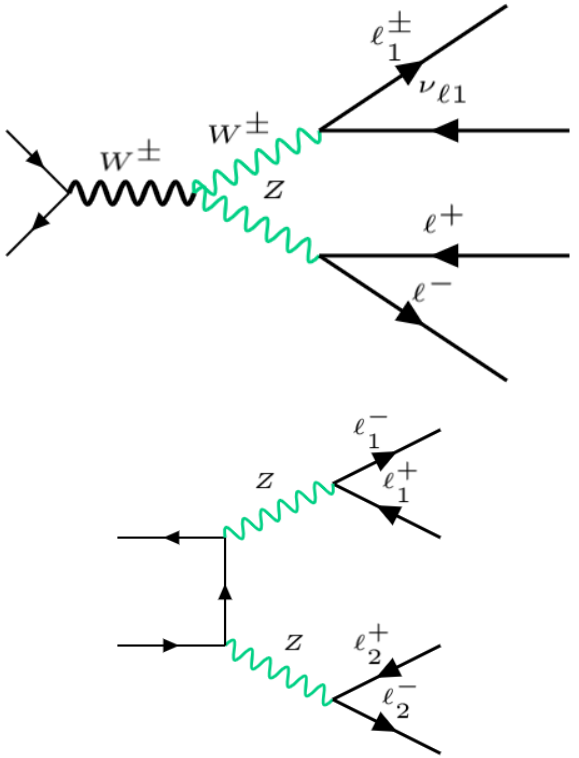
Z+jets, tt+jets, WW+jets

Phys. Rev. D 100, 052003 (2019)

An example analysis: VLL



Suggested selections?



$N_{leptons}$
$\geq$4e / μ
3e / μ
2e / μ OS (or SS) + $\geq 1\tau_h$

What possible final states?

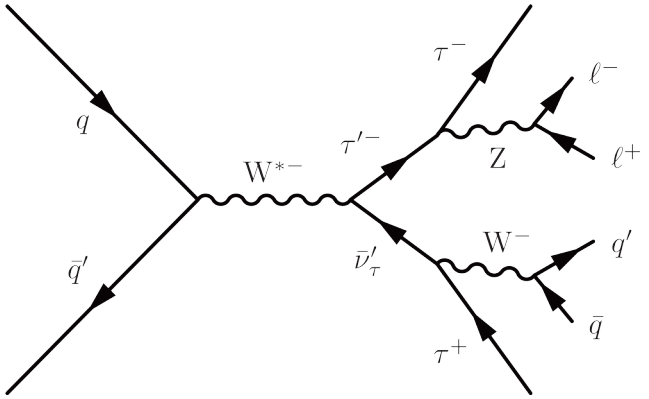
ZτZτ , ZτHτ , HτHτ

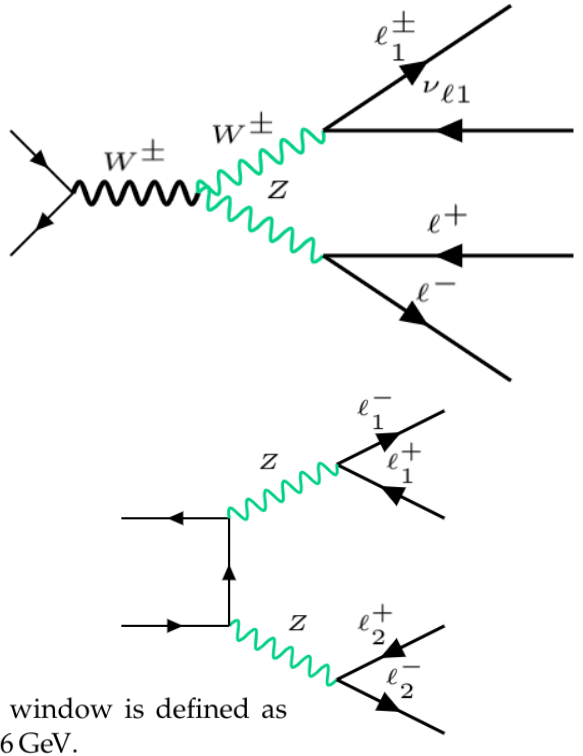
WτWτ

ZτWτ, HτWτ

Phys. Rev. D 100, 052003 (2019)

An example analysis: VLL





$N_{leptons}$
$\geq$4e / μ
3e / μ
2e / μ OS (or SS) + $\geq 1\tau_h$

Suggested selections?

What possible final states?

ZτZτ , ZτHτ , HτHτ

WτWτ

ZτWτ, HτWτ

Table 1: The signal regions defined in this analysis. The on-Z mass window is defined as $76 < m_{\ell\ell} < 106$ GeV, while the below-Z condition is defined as $m_{\ell\ell} < 76$ GeV.

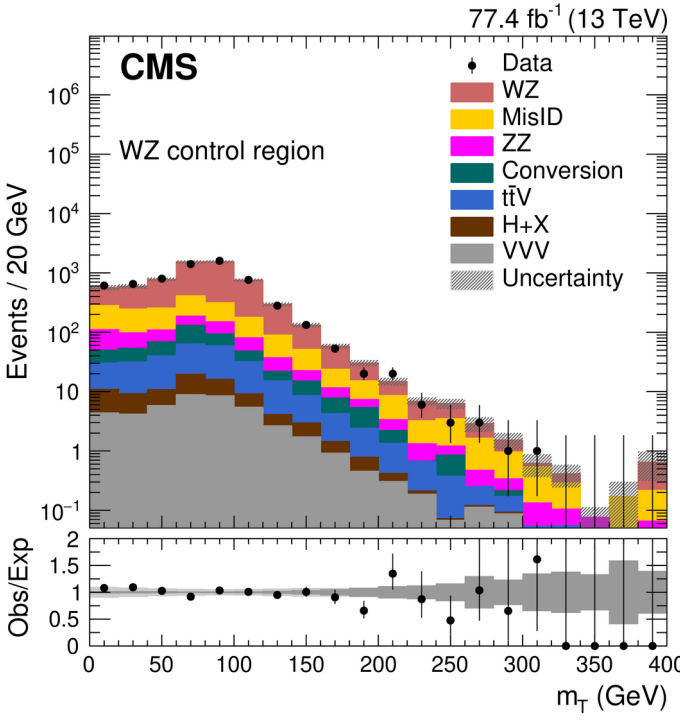
$N_{leptons}$	p_T^{miss} (GeV)	CR veto
$\geq$4e / μ	<50 >50	2 OSSF on-Z pairs and $p_T^{miss} < 50$ GeV
3e / μ	<150 >150	OSSF on-Z pair and $p_T^{miss} < 100$ GeV, or OSSF below-Z pair and $p_T^{miss} < 50$ GeV, or OSSF below-Z pair and on-Z $m_{3\ell}$
2e / μ OS (or SS) + $\geq 1\tau_h$	<150 >150	$p_T^{miss} < 50$ GeV

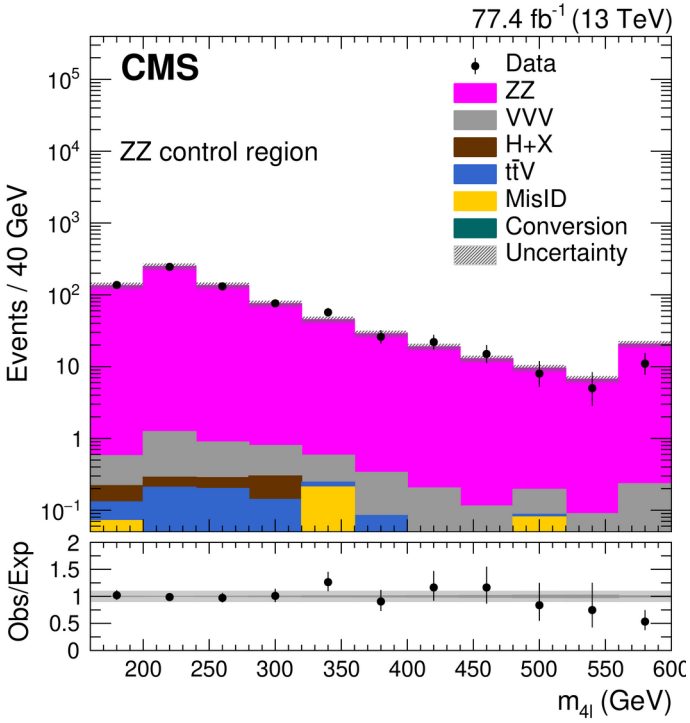
Phys. Rev. D 100, 052003 (2019)

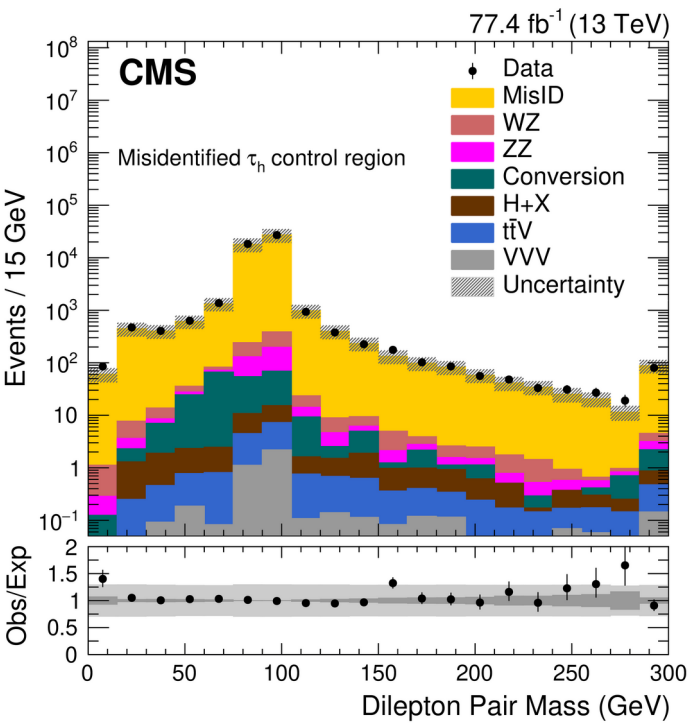
Main discrimination based on

$$L_T = \Sigma\, p_T^{\ell}$$

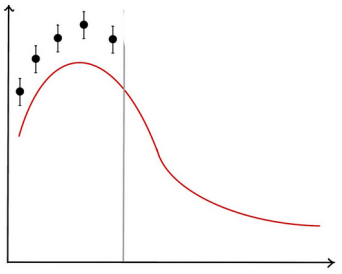
An example analysis: VLL





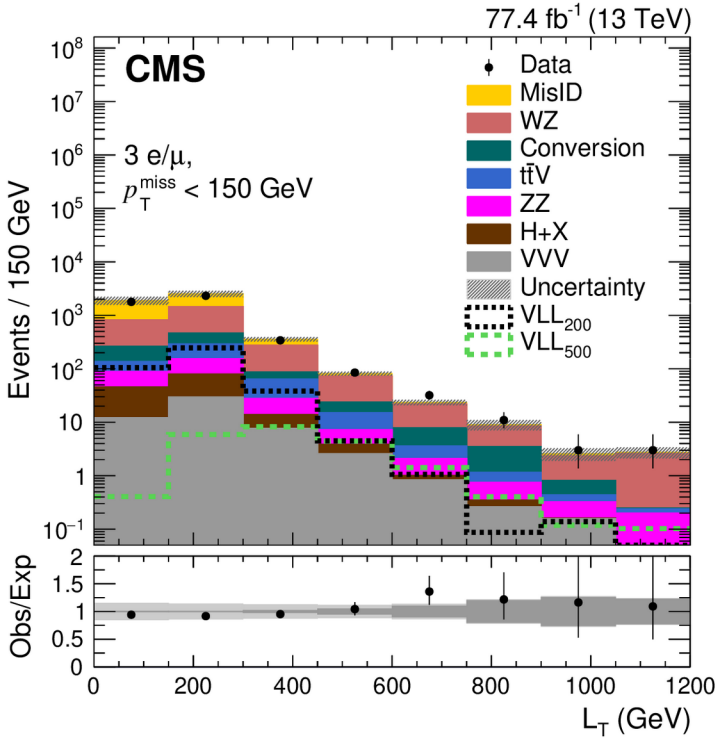


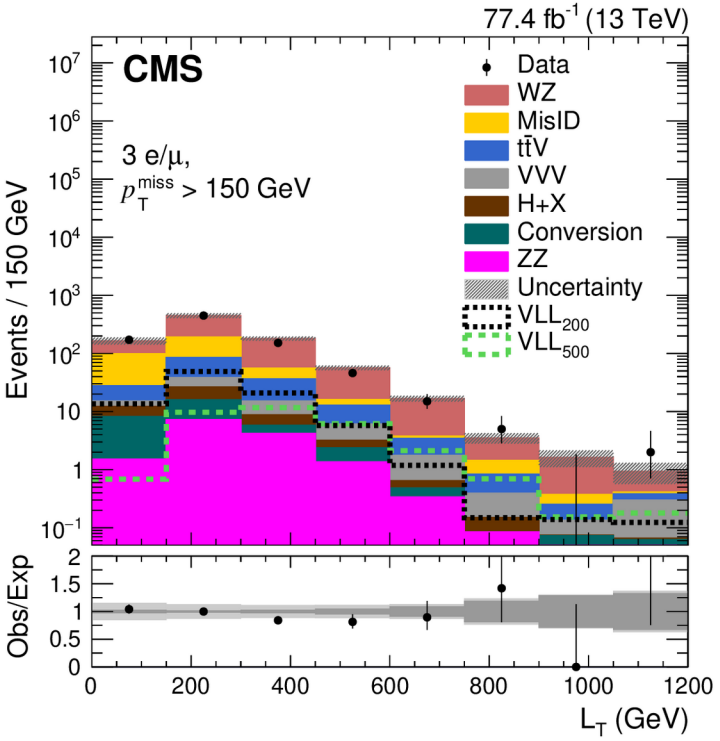
Simulation normalized to data in these control regions.

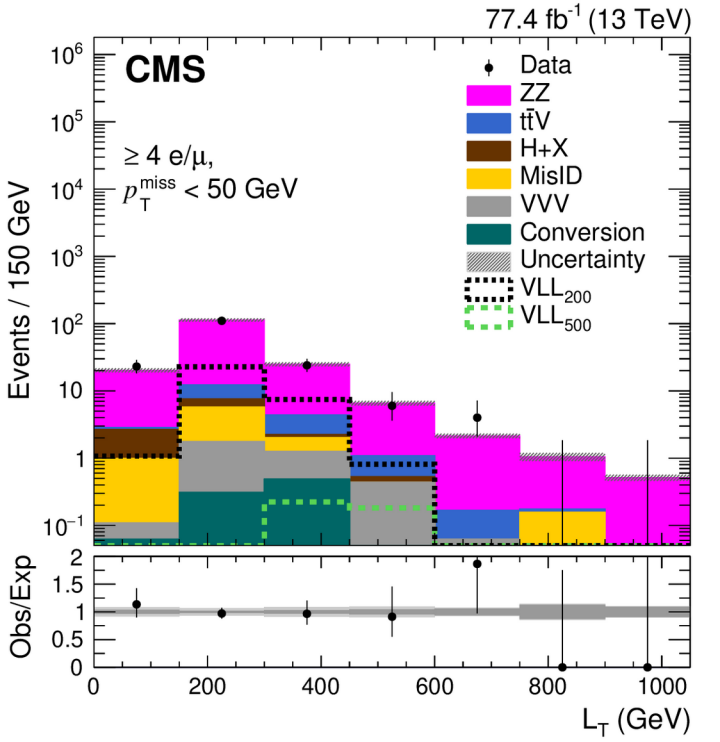


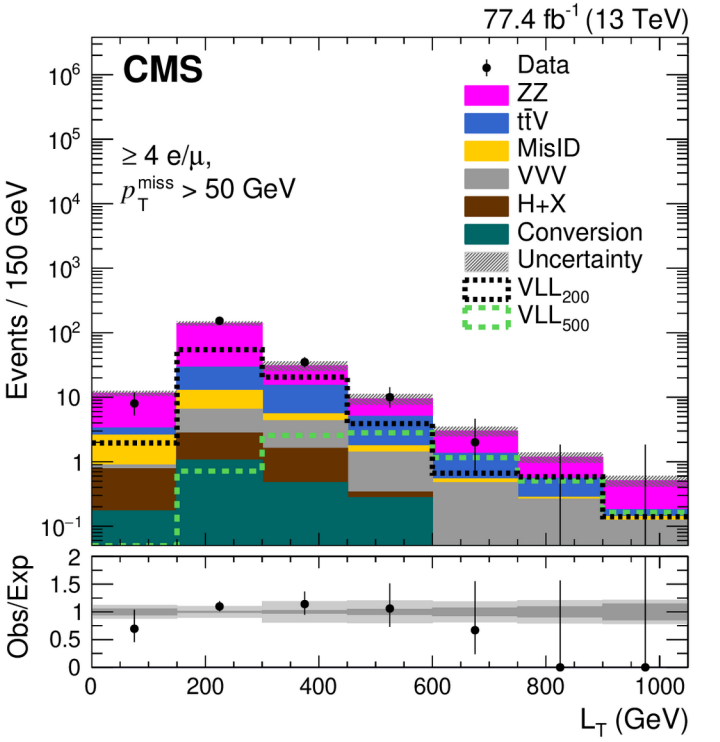
Matrix method: kind of like the fake rate method.

$N_{2L+obj} \times f$ = predicted $N(3L_{1\,fake})$









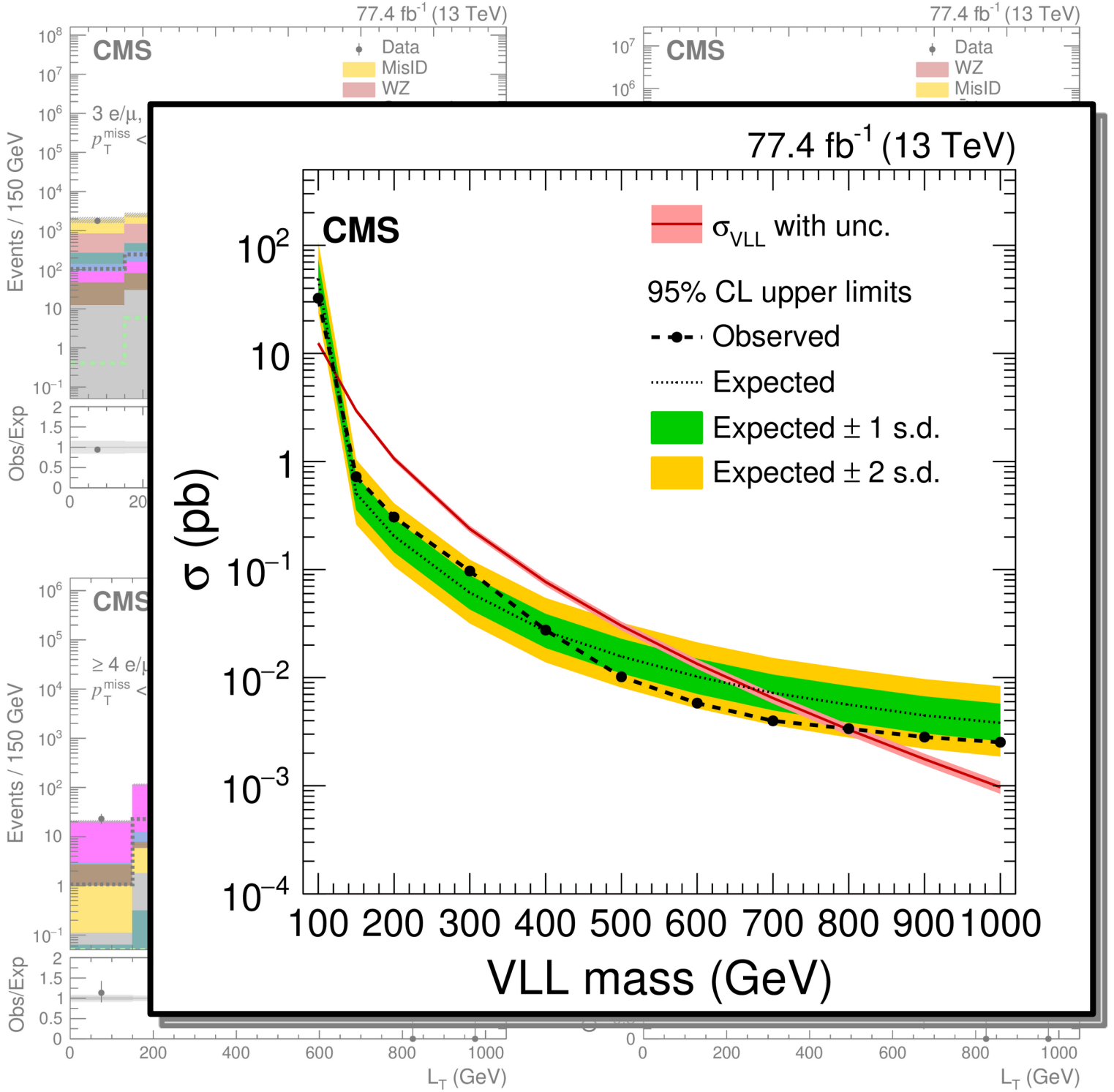
An example analysis: VLL

8x4 = 32 counting experiments here + 32 from the 2L1T channels

Total 64 signal regions

Phys. Rev. D 100, 052003 (2019)

An example analysis: VLL



Phys. Rev. D 100, 052003 (2019)

Thank you!

My group at IISER Pune

Email: sdube@iiserpune.ac.in

IISER PUNE