Higgs and exotics searches using jet substructure techniques

Petar Maksimovic, Johns Hopkins

PITT-PACC 2021

What we need to search for BSM

- Data
- Tools (taggers, new variables) to suppress background and isolate the signal
  - Most ML taggers still trained on MC...
- Background estimate (minimize uncertainty)
  - If dominated by ttbar, W+jets – get away with MC…
  - QCD: tricky and messy (after lots of work…)
- Signal efficiency (minimize uncertainty)
  - For top, W/Z, Higgs tagging – use standard candles
  - For exotic signatures – ???

Recent CMS results (to whet your appetite)

- Many new results, selected three to illustrate the issues:
  - $b^* \rightarrow tW$ (B2G-19-003)
  - W’ $\rightarrow$ bT’ or tB’ $\rightarrow$ tbH (B2G-20-002)
  - Triboson in lepton+jets (B2G-20-001)
- Not covered here:
  - $X \rightarrow WH$ in lepton+jets (B2G-19-002)
  - $X \rightarrow WZ(\rightarrow \nu\nu)$ in jet+MET (B2G-20-008)
  - T’$\rightarrow tZ(\rightarrow \nu\nu)$ in jet+MET (B2G-19-004)
  - $b^* \rightarrow tW$ in lepton+jets (B2G-20-010)
  - $X \rightarrow HY \rightarrow \tau\tau bb$ (no substructure) (HIG-20-014)

Recent CMS results (to whet your appetite)

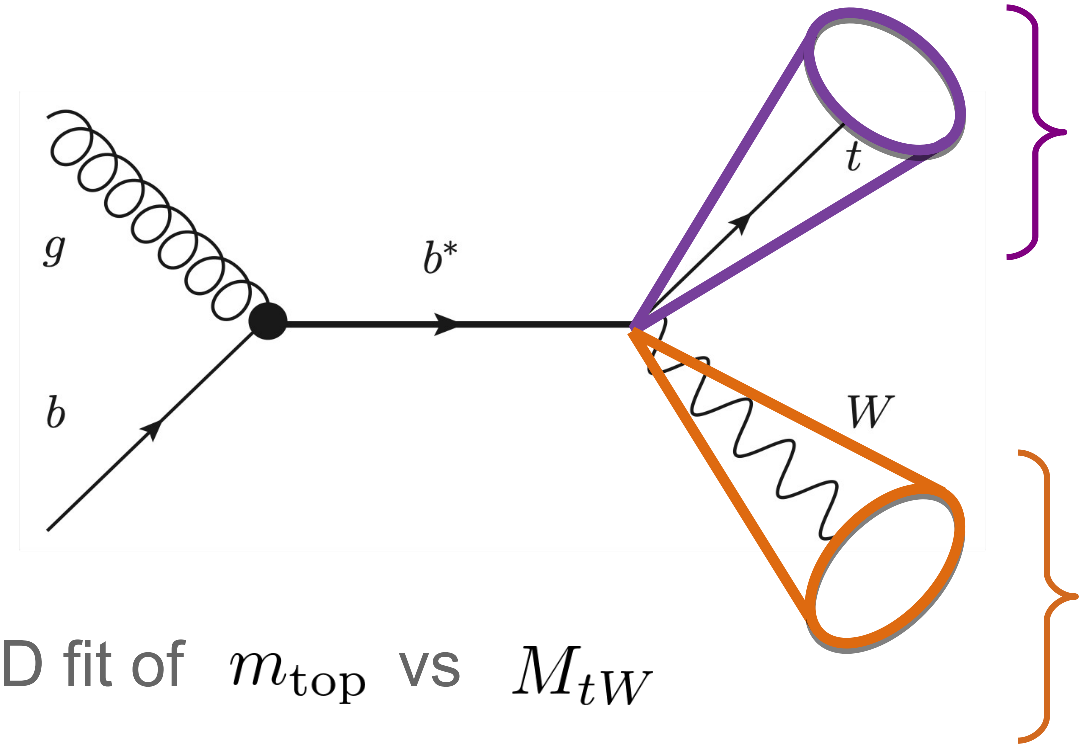
- Many new results, selected three to illustrate the issues:
  - b* → tW (B2G-19-003)
  - W' → bT' or tB' → tbH (B2G-20-002)
  - Triboson in lepton+jets (B2G-20-001)
- Not covered here:
  - X→WH in lepton+jets
  - X→WZ(→νν) in jet+MET (
  - T'→tZ(→νν) in jet+MET (B2G-
  - b*→ tW in lepton+jets (B2G-20-010)
  - X→HY→ττbb (no substructure) (HIG-20-014)

I was asked to cover CMS results,
But the issues I want to raise
are the same for ATLAS.

Exhibit A: b*→ tW

B2G-19-003

- Two large-R jets (mostly back-to-back)



top jet:
τ_3/τ_2
& b-tag one subjet

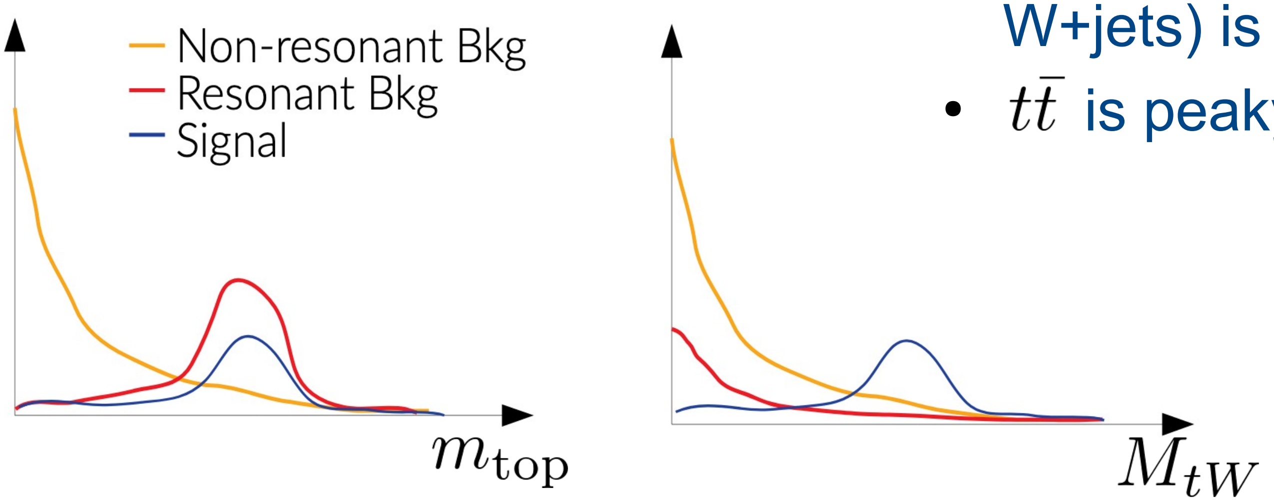
W jet:
τ_2/τ_1
& in W mass window

- 2D fit of m_{top} vs M_{tW}
- Main backgrounds:
  - QCD – from data
  - $t\bar{t}$ – from template-morphed MC

Exhibit A: b*→ tW

B2G-19-003

- 2D fit of m_{top} vs. M_{tW}



- Non-res bkg (QCD, W+jets) is *smooth*
- $t\bar{t}$ is peaky

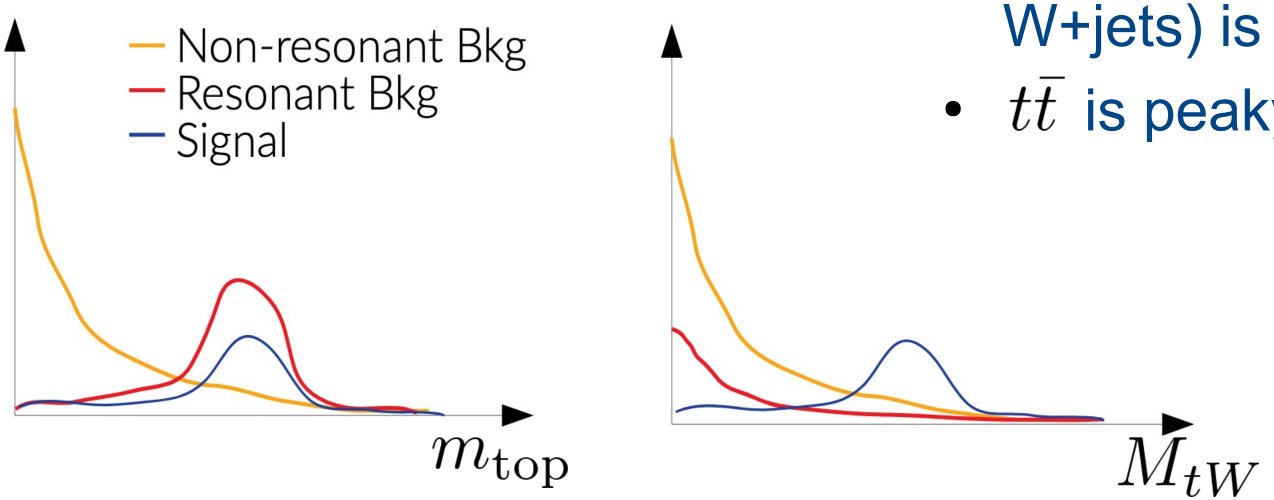
$$N_p^{\text{nonres}}[i] = N_f^{\text{nonres}}[i] \times R_{p/f}^{MC}[i] \times R_{\text{ratio}}(m_{\text{top}}, M_{tW})$$

Non-res bkg modeled from data events that fail top tagging

Exhibit A: b*→ tW

B2G-19-003

- 2D fit of m_{top} vs. M_{tW}



- Non-res bkg (QCD, W+jets) is *smooth*
- $t\bar{t}$ is peaky

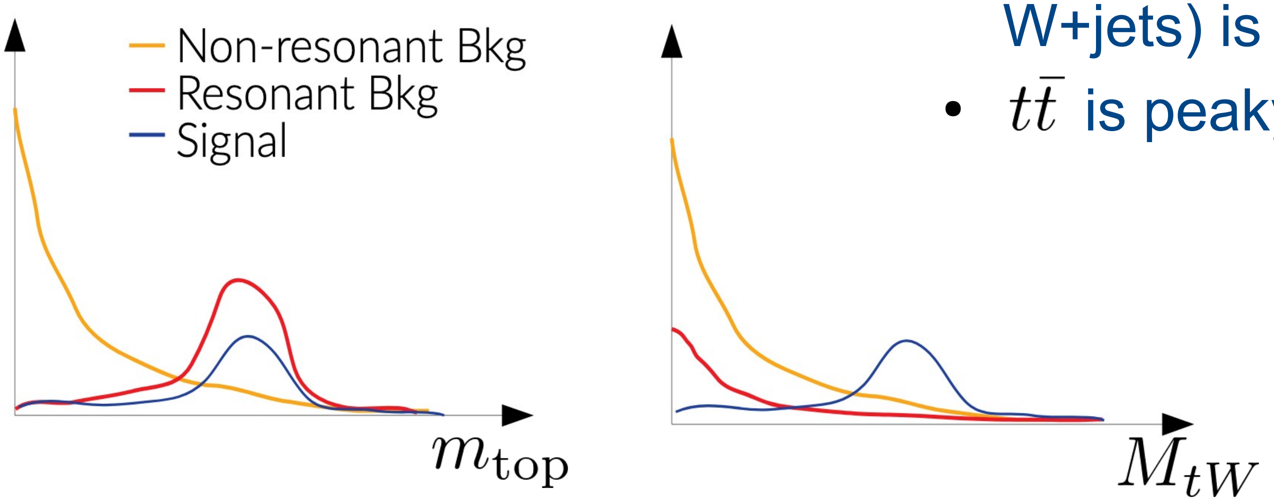
$$N_p^{\text{nonres}}[i] = N_f^{\text{nonres}}[i] \times R_{p/f}^{MC}[i] \times R_{\text{ratio}}(m_{\text{top}}, M_{tW})$$

To model corners of 2D plane, start from pass-fail ratio from MC

Exhibit A: b*→ tW

B2G-19-003

- 2D fit of m_{top} vs. M_{tW}



- Non-res bkg (QCD, W+jets) is *smooth*
- $t\bar{t}$ is peaky

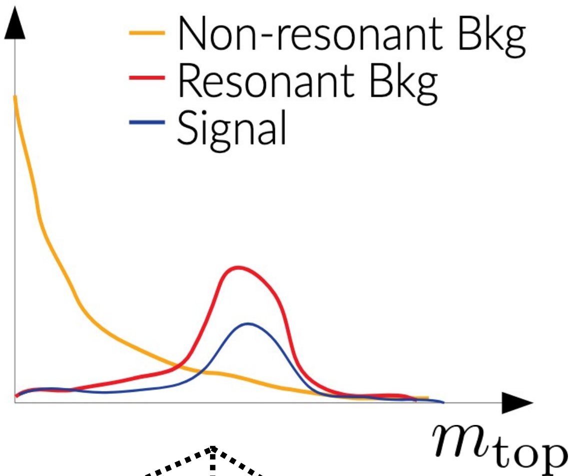
$$N_p^{\text{nonres}}[i] = N_f^{\text{nonres}}[i] \times R_{p/f}^{MC}[i] \times R_{\text{ratio}}(m_{\text{top}}, M_{tW})$$

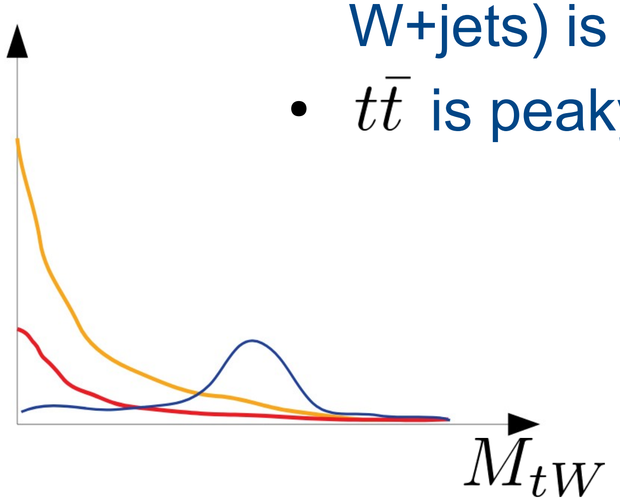
To correct pass-fail ratio from MC, multiply by a smooth 2D function floating in the fit to data

Exhibit A: b*→ tW

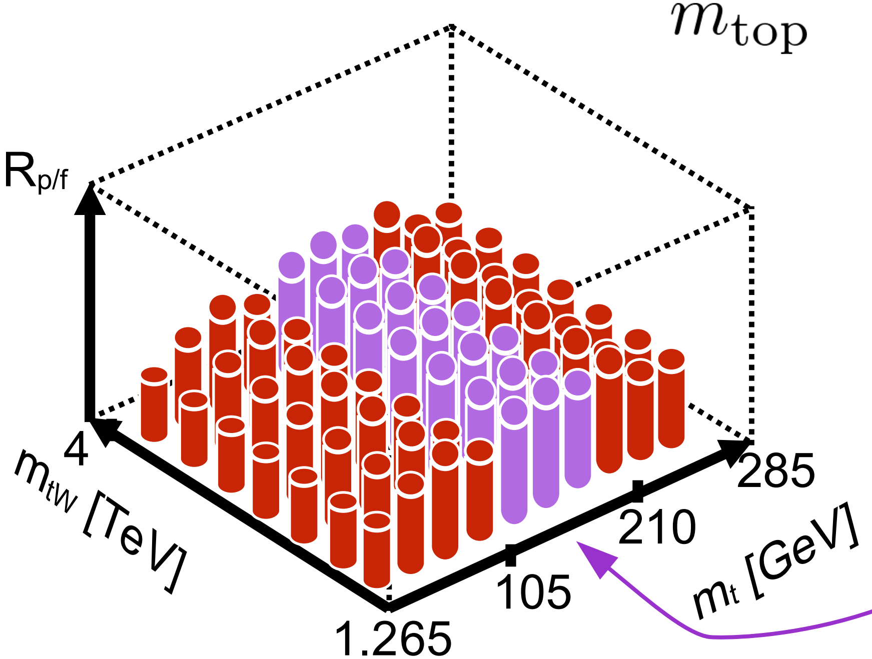
B2G-19-003

- 2D fit of m_{top} vs. M_{tW}





- Non-res bkg (QCD, W+jets) is ***smooth***
- $t\bar{t}$ is peaky



The basic idea is to *interpolate* through m_{top} signal region

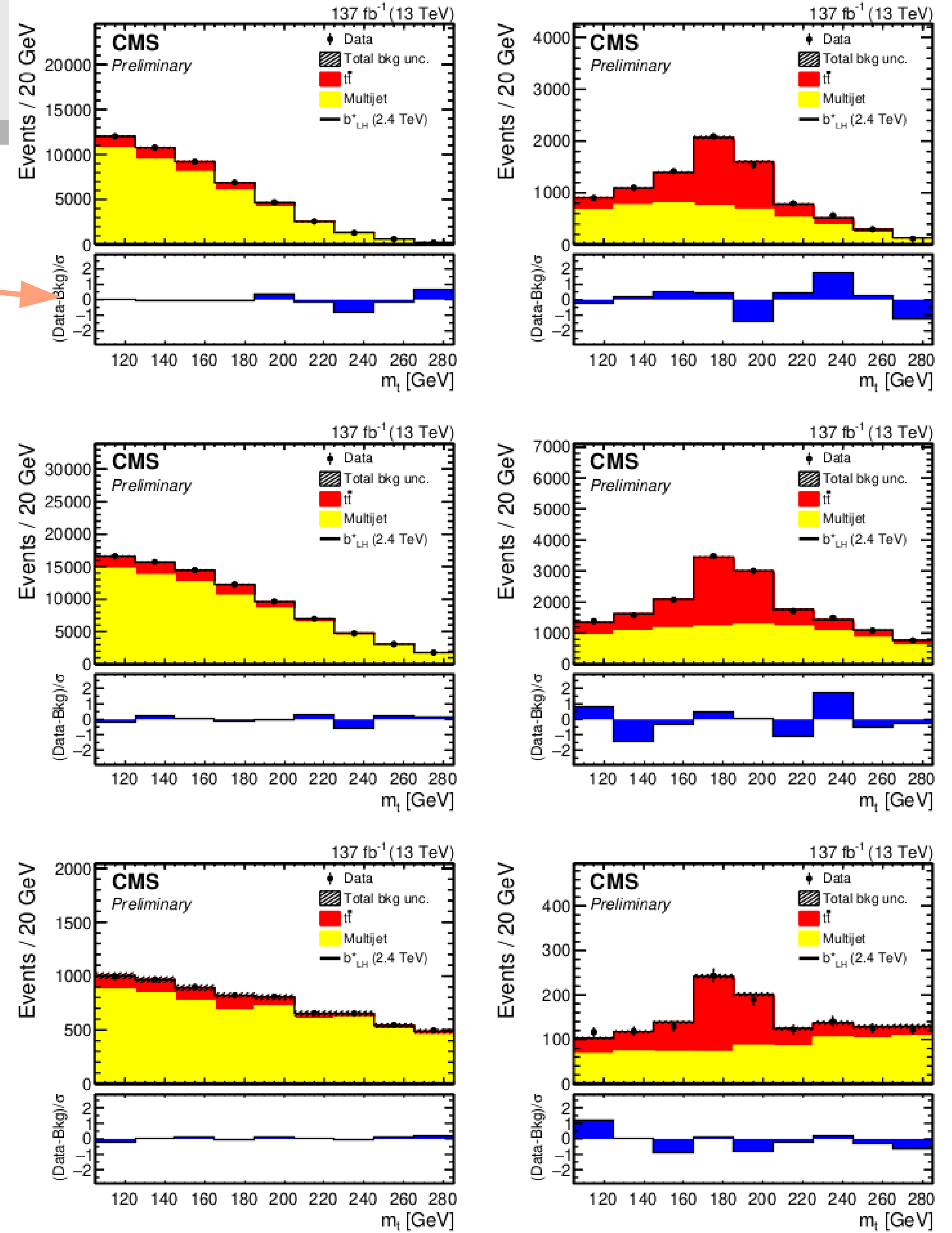
B2G-19-003

b*→ tW

- 2D fit of m_{top} vs. M_{tW}

m_{top} projection in three slices of M_{tW}

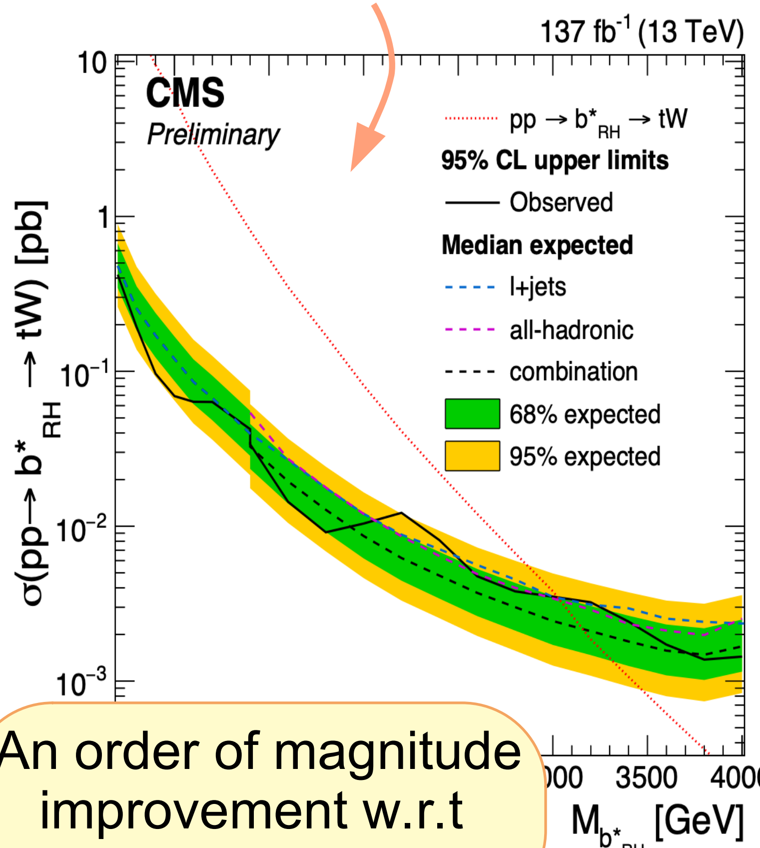
Left: fail.
Right: pass.



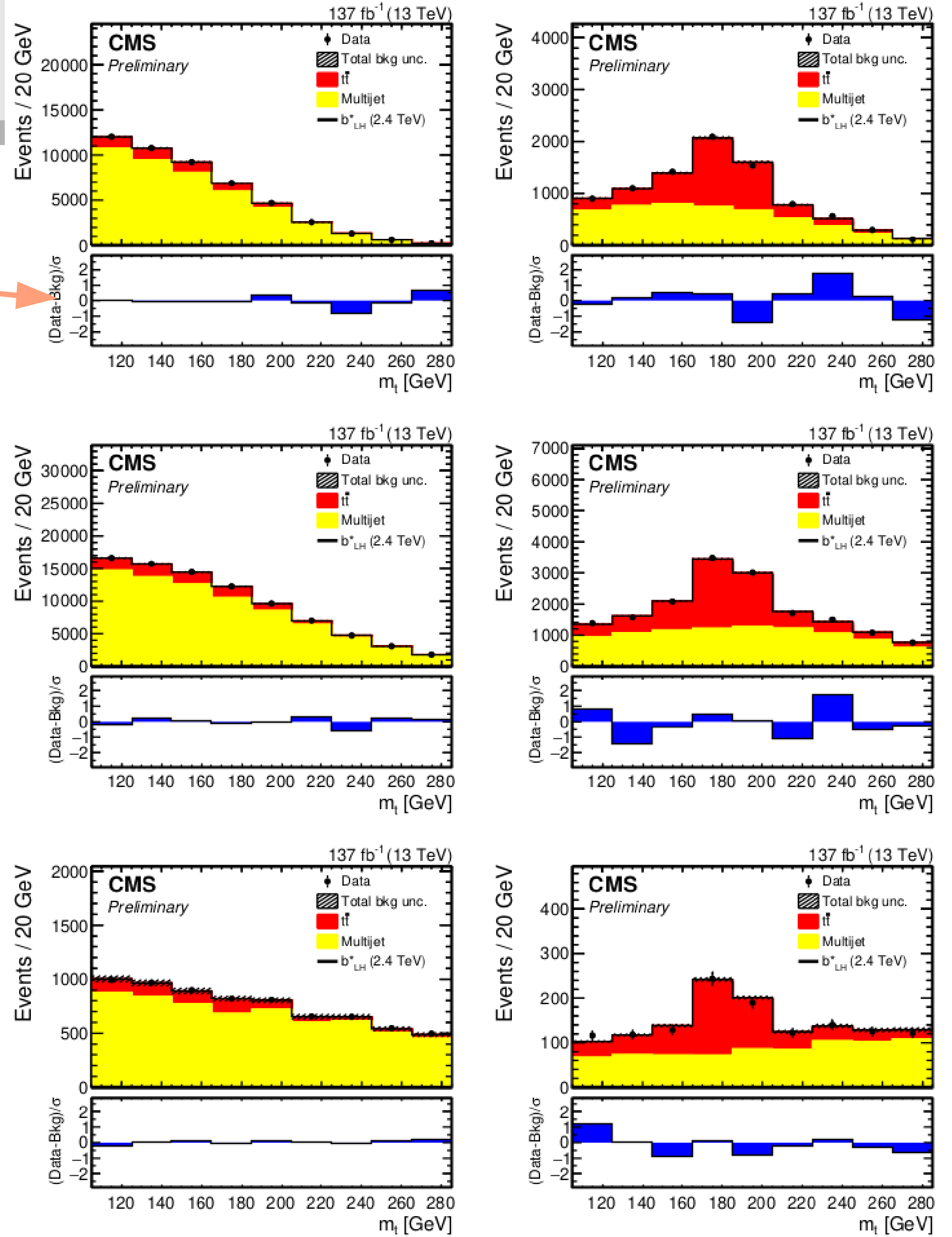
B2G-19-003

b*→ tW

- 2D fit of m_{top} vs. M_{tW}
- Limit (right-handed), comb. with 1lepton



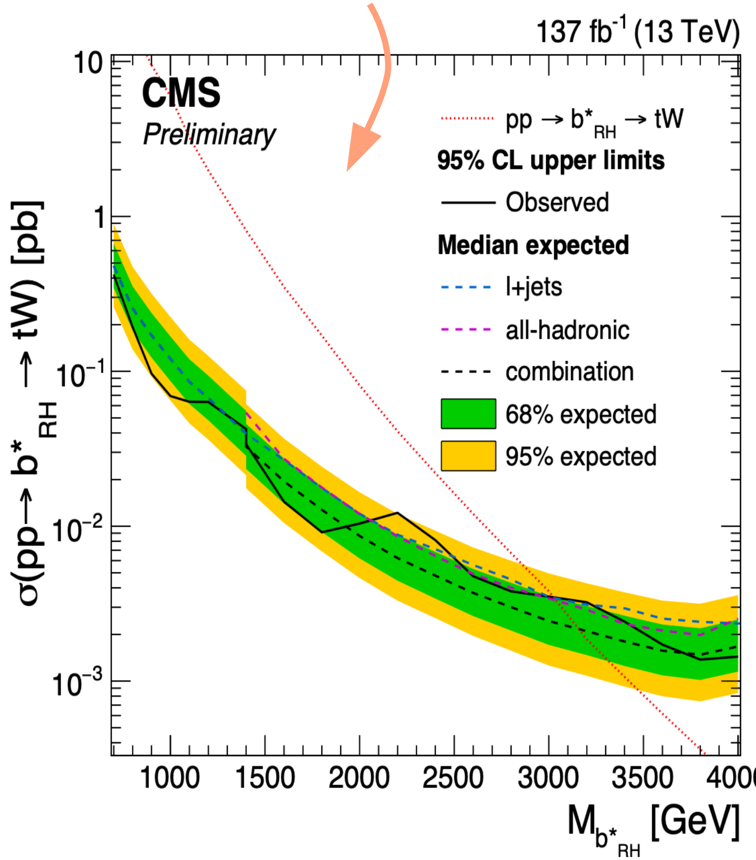
An order of magnitude improvement w.r.t previous searches



B2G-19-003

b*→ tW

- 2D fit of m_{top} vs. M_{tW}
- Limit (right-handed), comb. with 1lepton

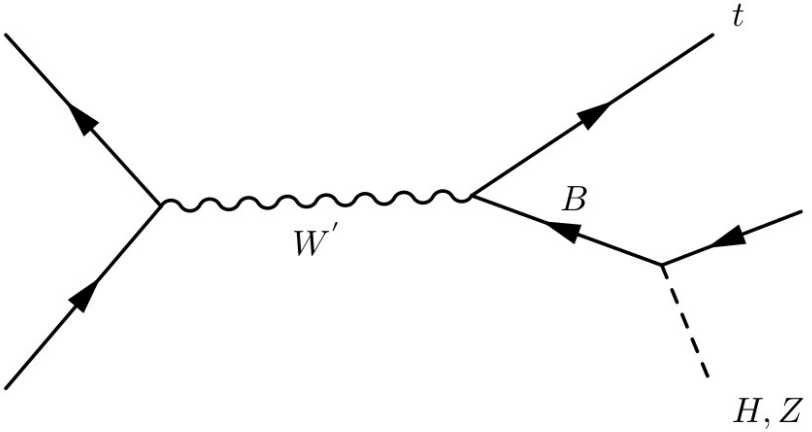


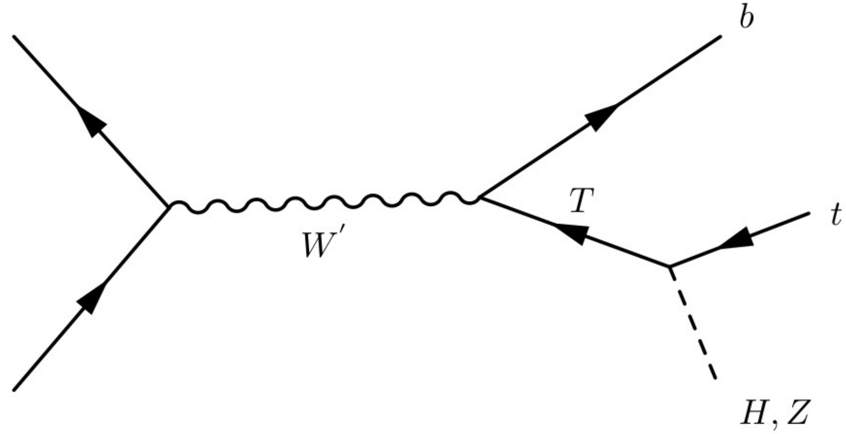
Lessons learned:

- QCD MC is used
- A better MC would be great!
- Substructure techniques used are simple
- But it would be easy to upgrade to cutting-edge ML algorithms
- …that preserve shape!
- (= pass-fail ratio is flat)
- (called “mass decorrelated”)

Exhibit B: W' → VLQ + t/b → tbH

B2G-20-002





- Same signature! Assume 50% BR to each.
- For relevant param. space, results in a "Mercedes-sign" topology
- Using a new CNN-based "ImageTop" tagger
  - Mass-decorrelated (by training)

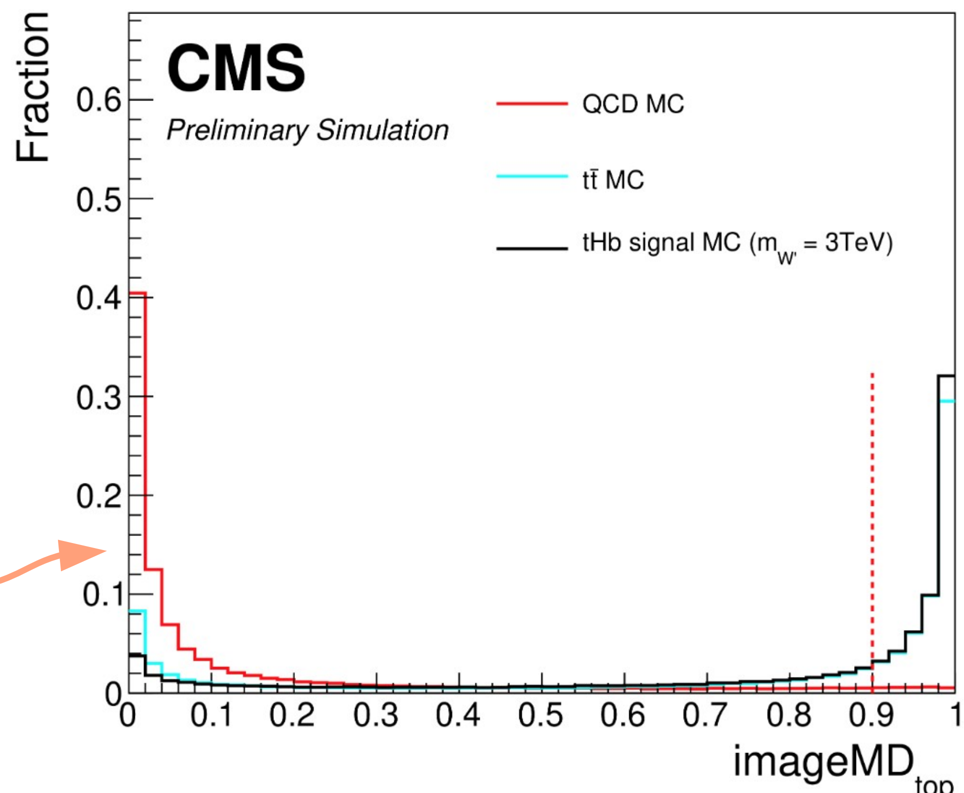
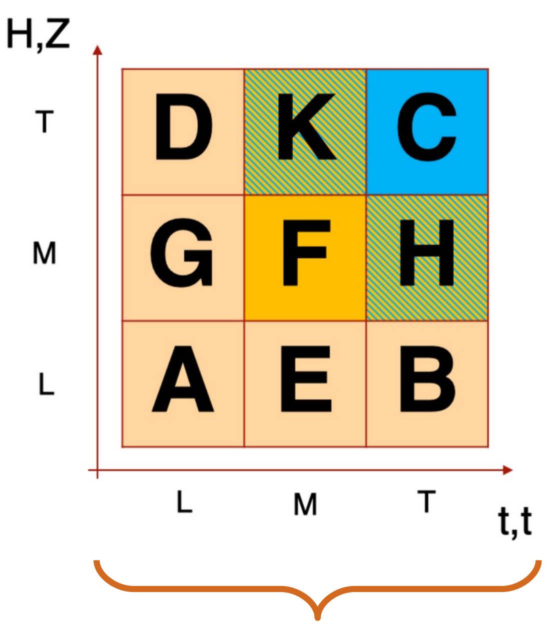


Exhibit B: W' → VLQ + t/b → tbH

B2G-20-002

- Measure per-jet "mistag rate" = B/A (vs p_{T})
- Apply to jets in D to get C



("Loose", "Medium", "Tight" = working points of the top tagger)

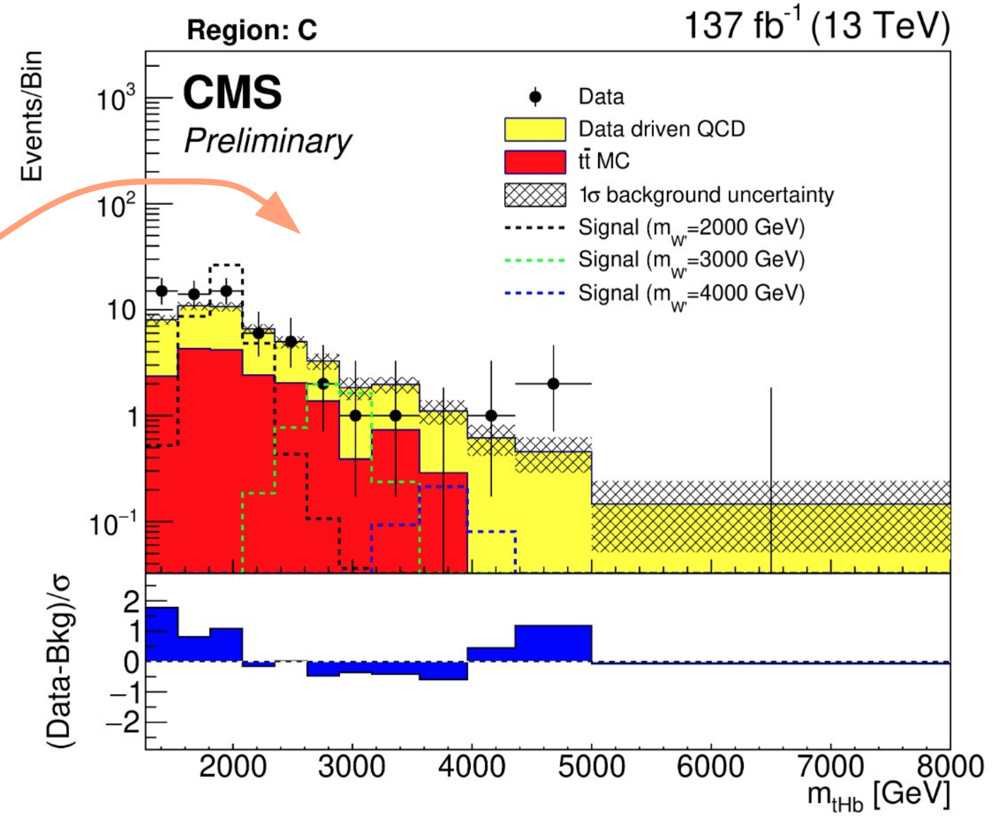
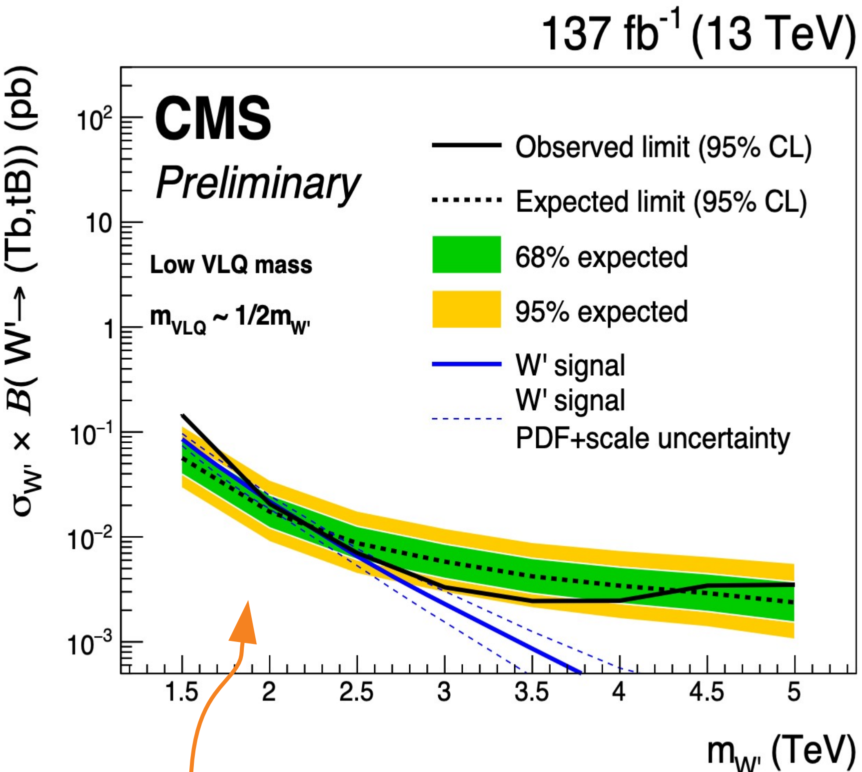


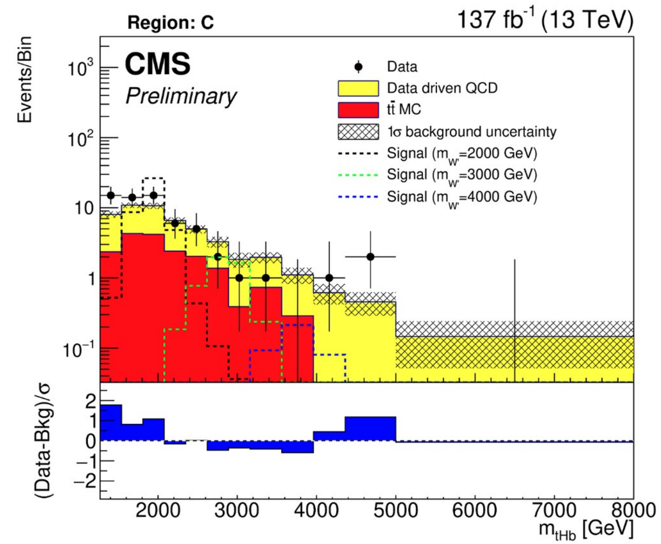
Exhibit B: W’ → VLQ + t/b → tbH

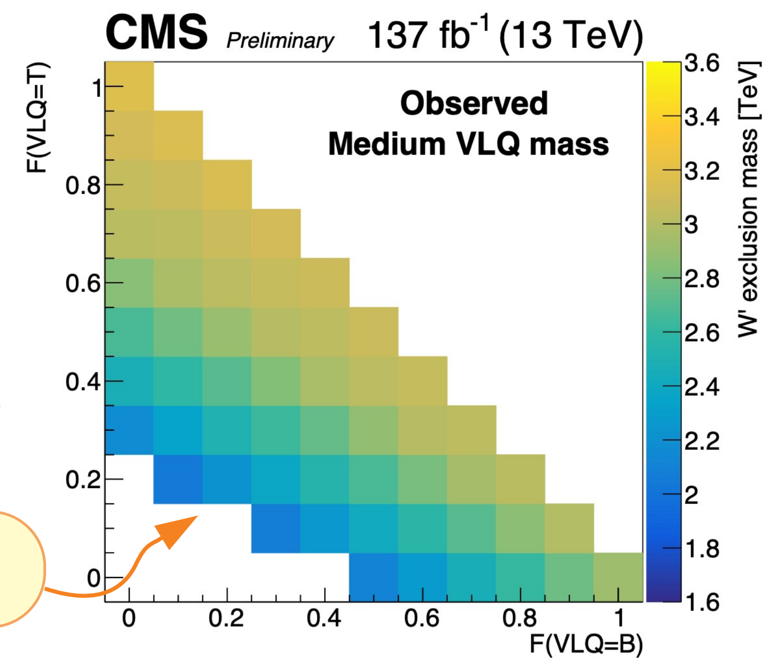
B2G-20-002

- Limits depend on $m_{W'}, m_{\mathrm{VLQ}}$



Limit vs $m_{W'}$ for $m_{\mathrm{VLQ}} \sim m_{W'}/2$



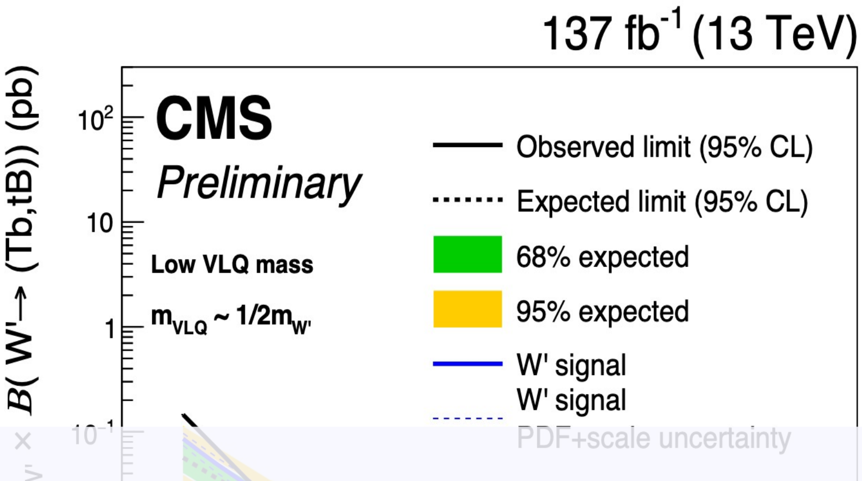


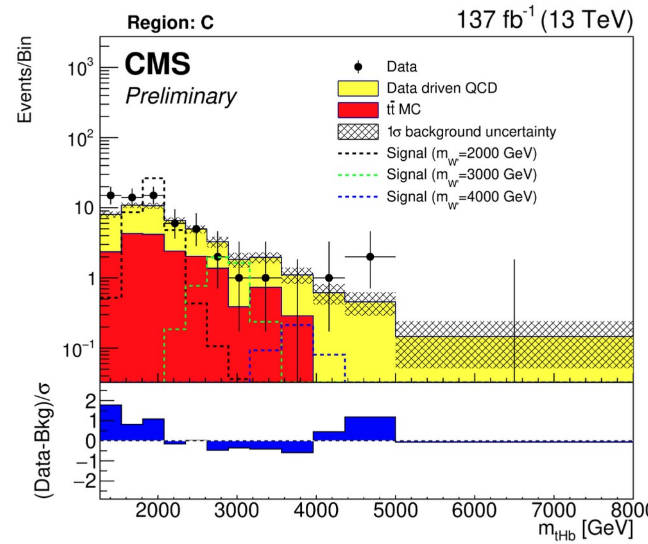
Changing fraction of tB’ and bT’

Exhibit B: W' → VLQ + t/b → tbH

B2G-20-002

- Limits depend on $m_{W'}, m_{\mathrm{VLQ}}$





Lessons learned:

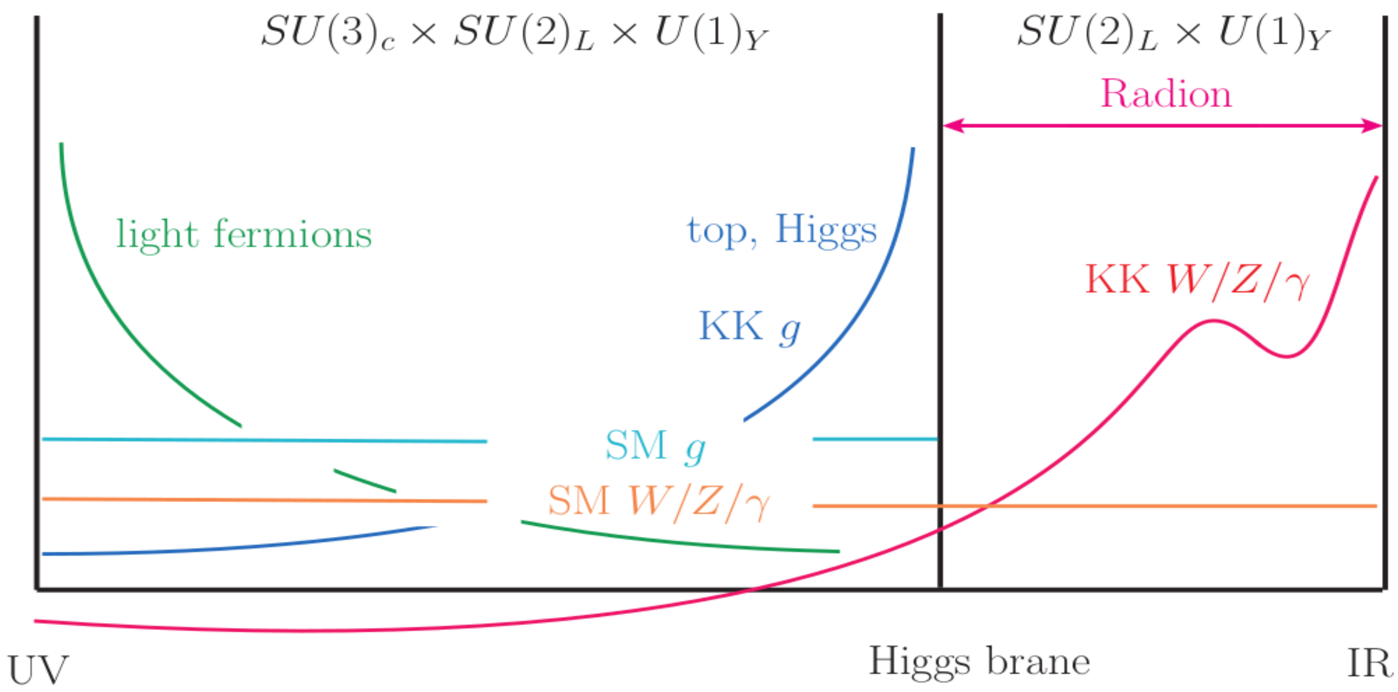
- ImageTop tagger: powerful, mass-decorrelated
- Calibrated on semileptonic $t\bar{t}$ events (standard candle)

- An off-the-beaten-path signature. Need more!

Exhibit C: Triboson resonances

B2G-20-001

- Warped extra dimensions, with only SM EW gauge fields in the bulk. [arXiv:1711.09920,1809.07334]

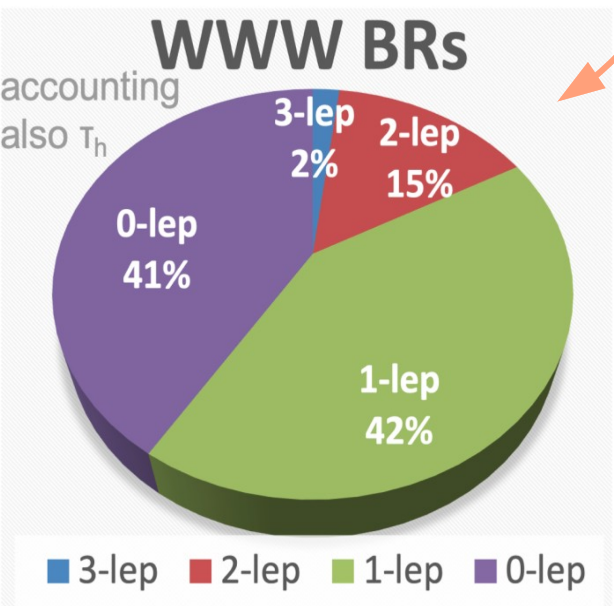


- KKg is too heavy, but KK of $W/Z/\gamma$ are open
- Cascade decay: $KKV \to V\phi \to V(VV)$

Exhibit C: Triboson resonances

B2G-20-001

- $KKV \to V\phi \to V(VV)$
  - Focus on $V = W,Z$
  - One $W \to \nu\ell$



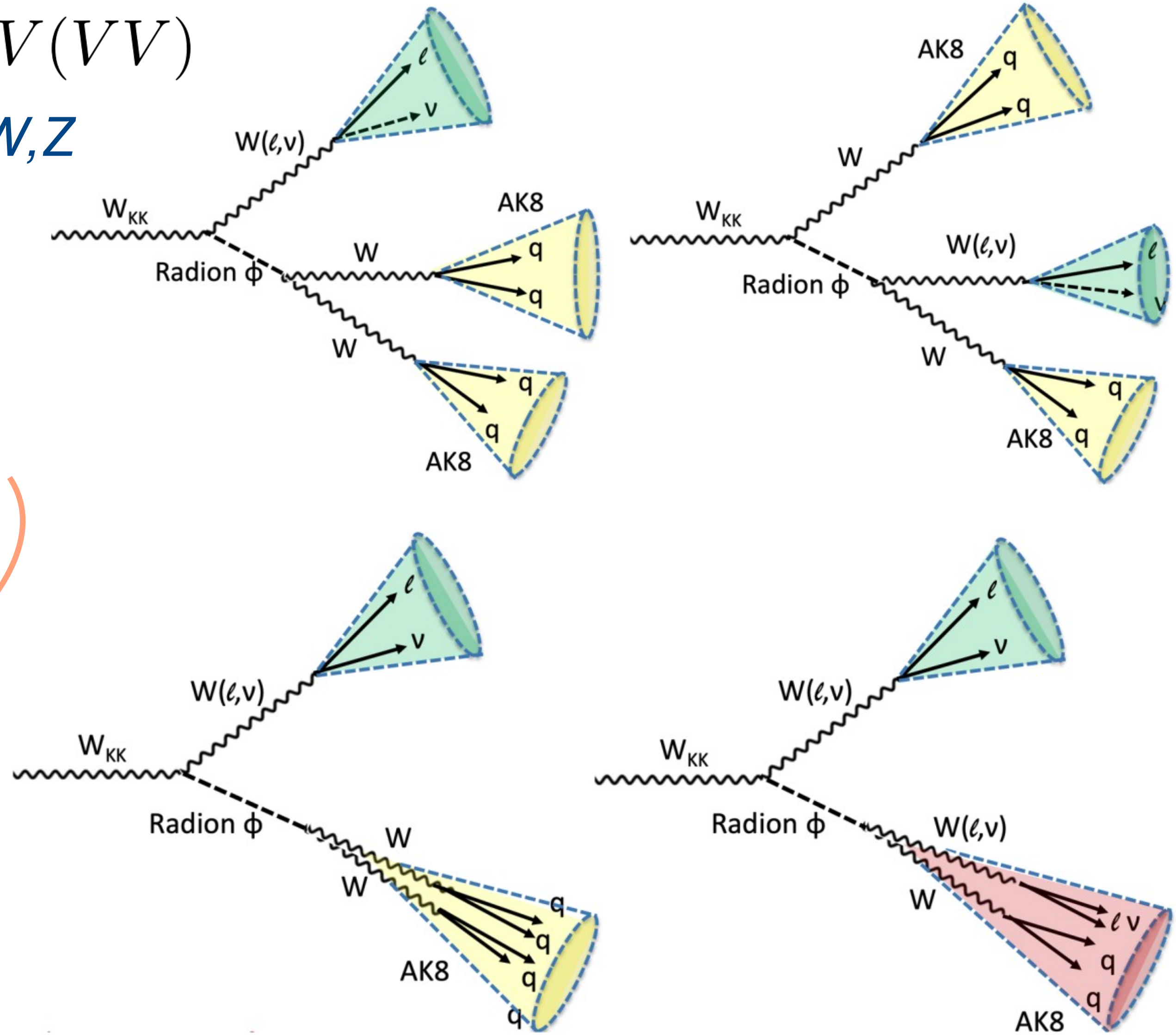
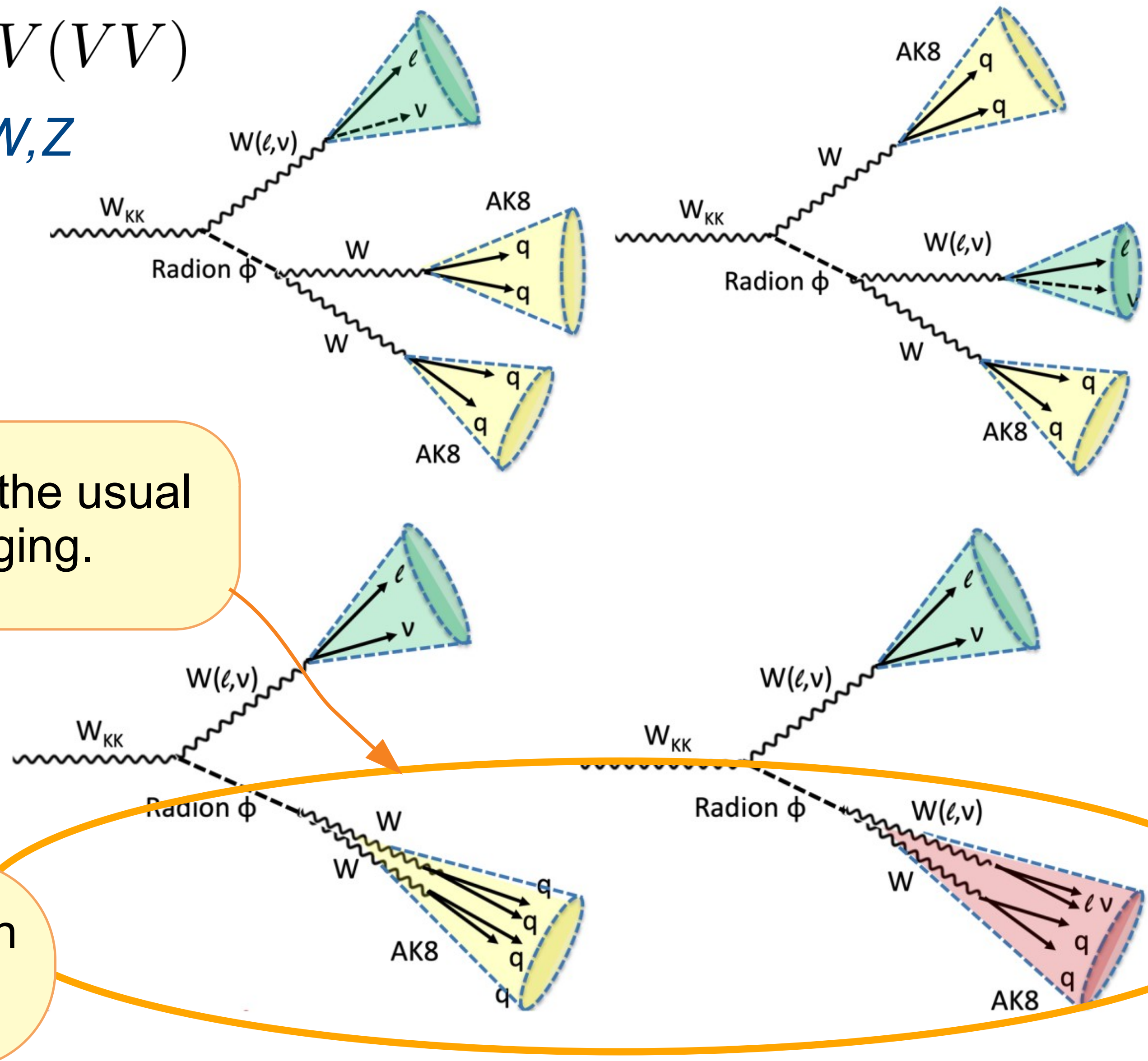


Exhibit C: Triboson resonances

B2G-20-001

- $KKV \to V\phi \to V(VV)$
  - Focus on $V = W,Z$



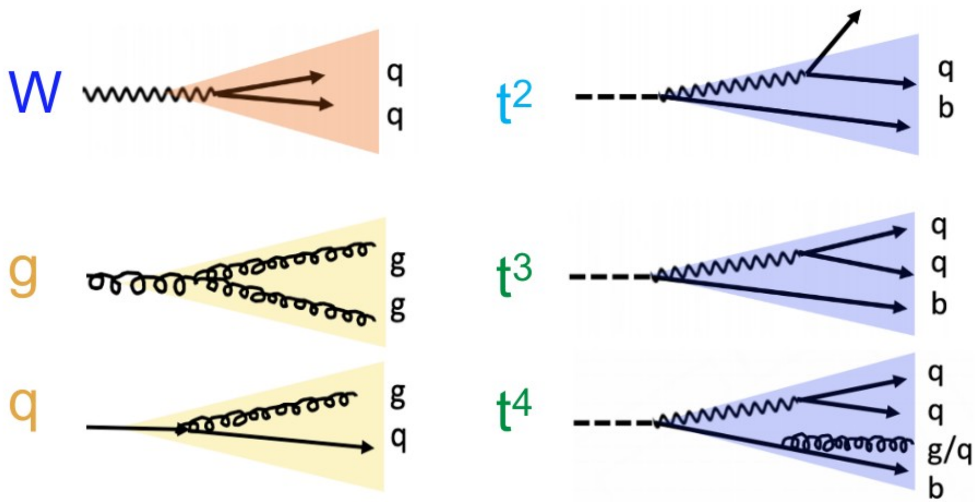
"Radion jets"! Beyond the usual top/W/Z/Higgs tagging.

Need special calibration – no standard candle

Exhibit C: Triboson resonances

B2G-20-001

- Calibrate deep tagger discriminant shape using SM proxies:



Used as a proxy for 4q Radion jet (with ***large*** systematics)

deep−WH @ Mj > 120

PTj: 200-500

PTj >500

- Correct MC shapes, bin by bin

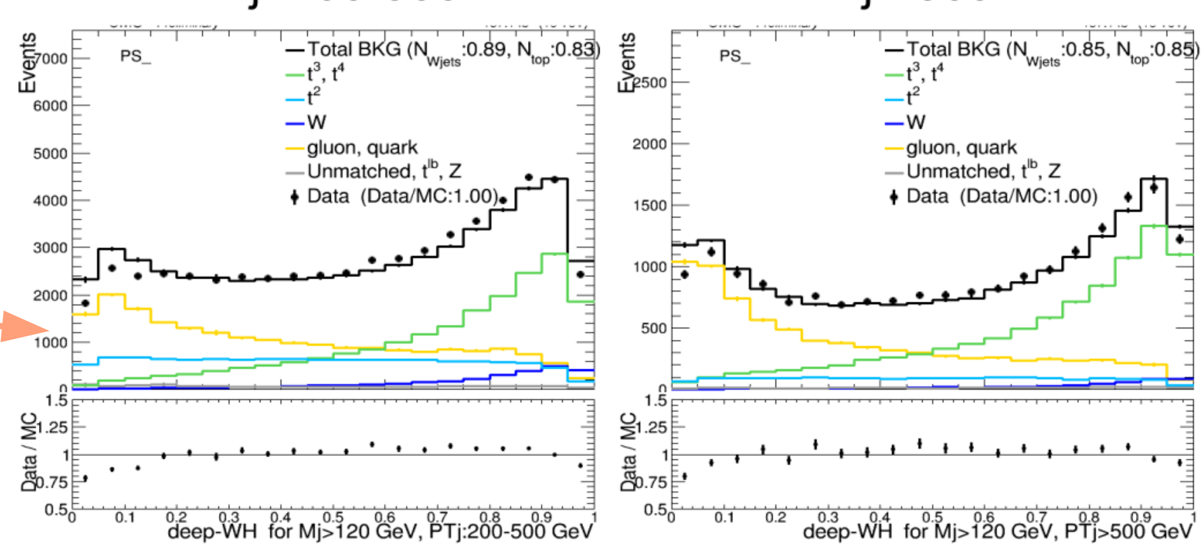


Exhibit C: Triboson resonances

B2G-20-001

- The proof is in the pudding: the method works!

All SFs derived for all 4 bins (2 Mj , 2 pTj bins) and for all types of jets W, t^2, $t^{3,4}$, g/q

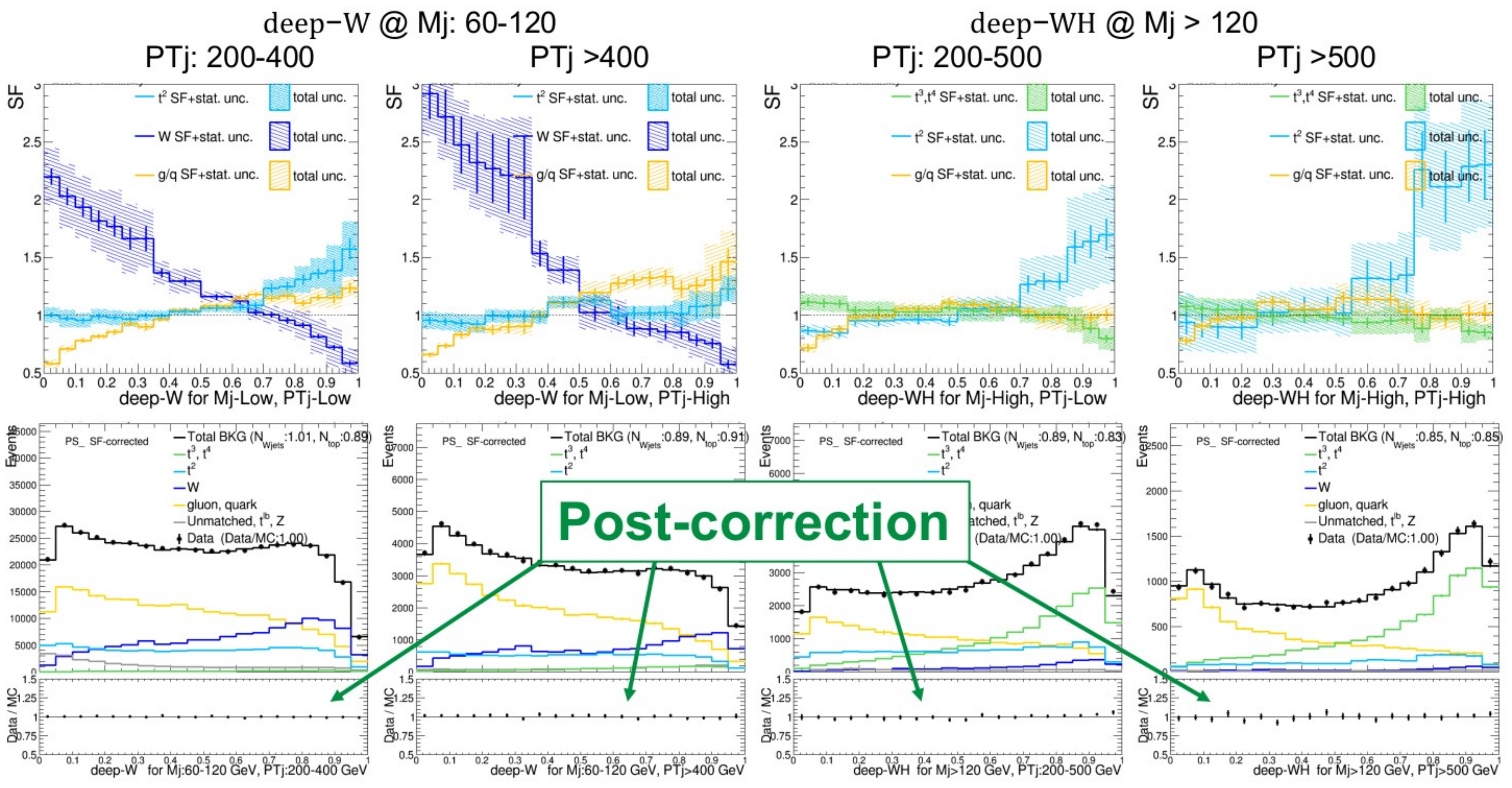
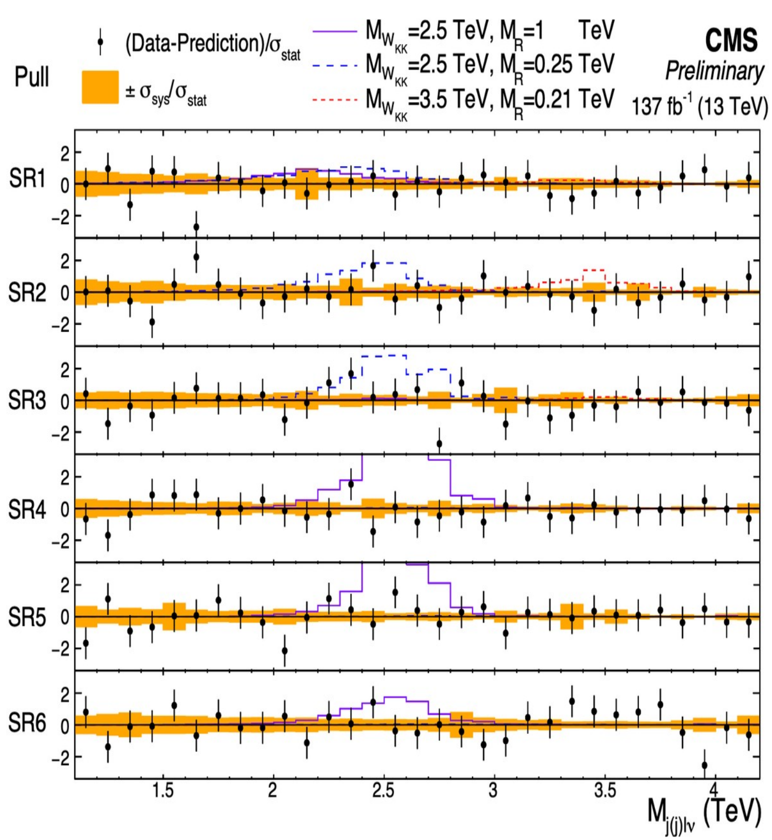
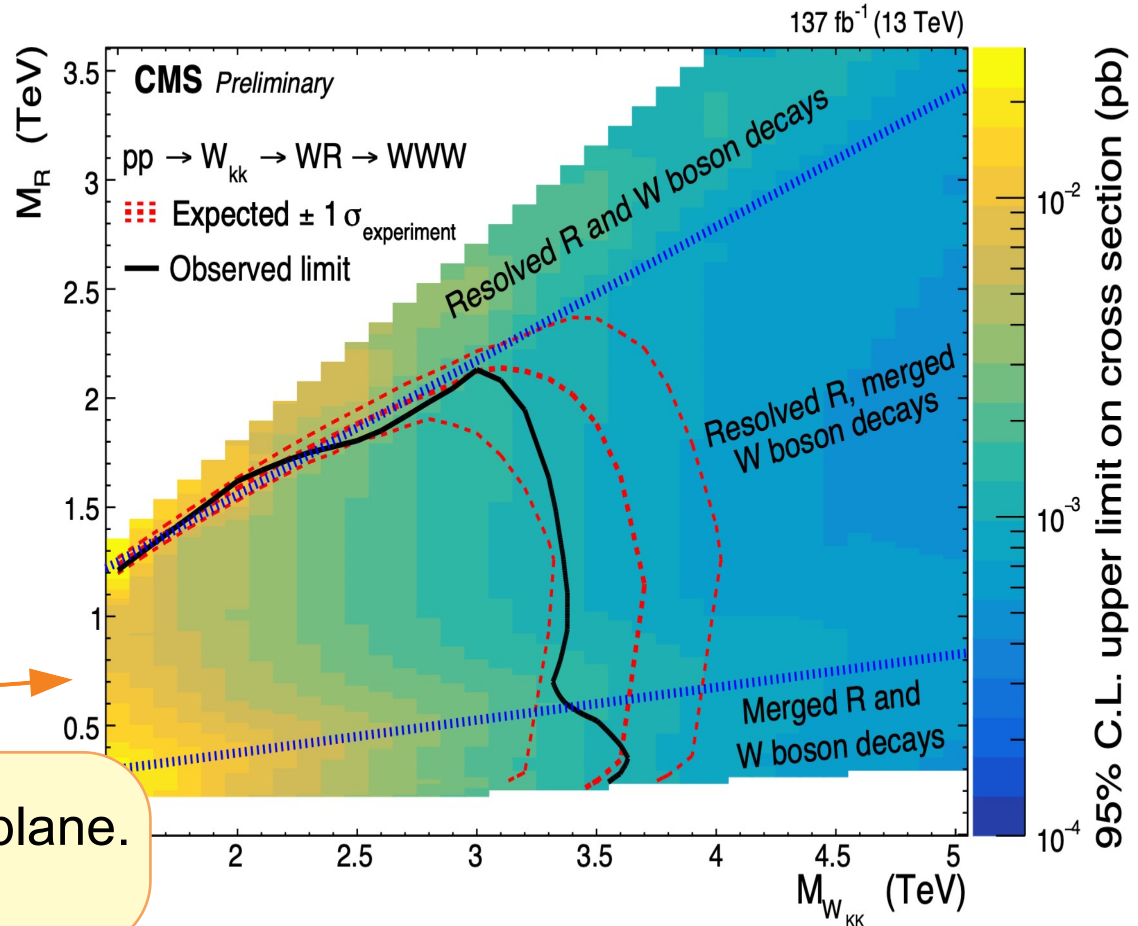


Exhibit C: Triboson resonances

B2G-20-001

Combined fit of six signal regions.
(Shown: pulls in all bins.)





Limits in 2D KKW vs R mass plane.
The first of their kind!

Exhibit C: Triboson resonances

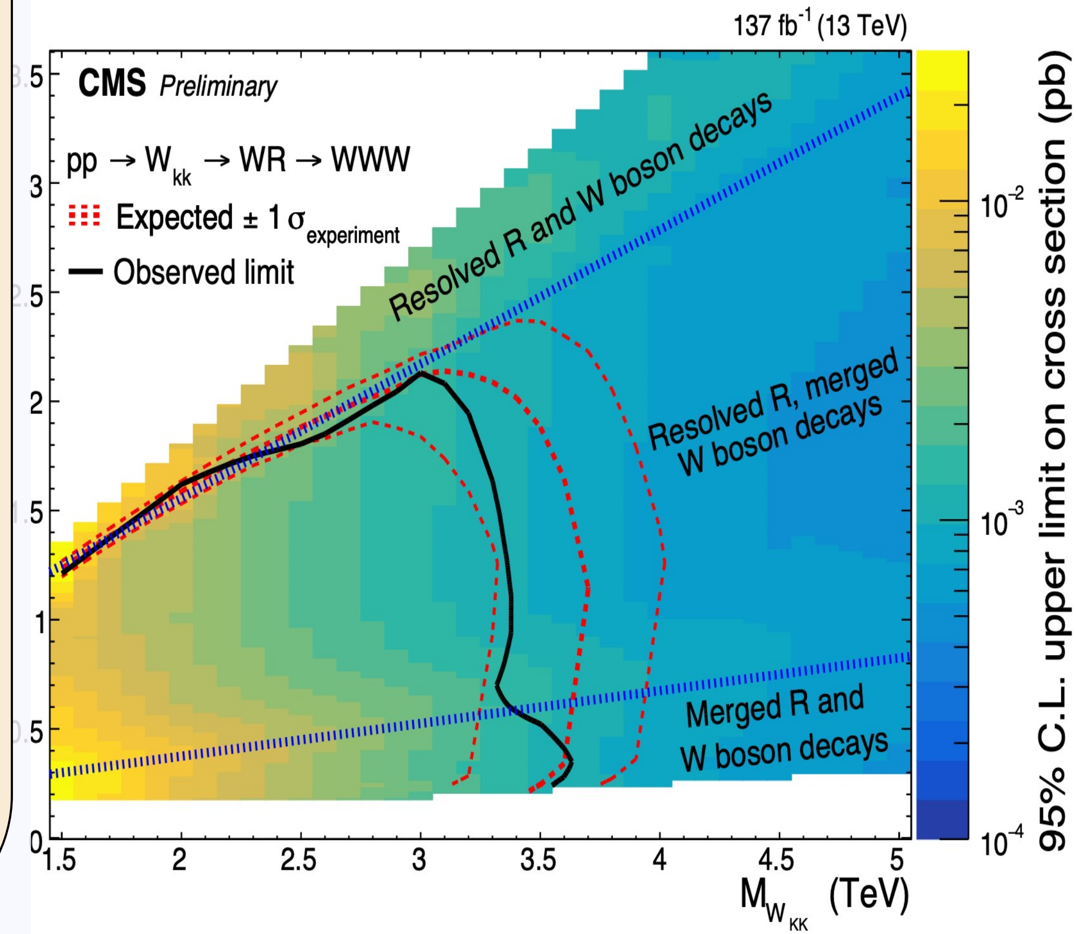
B2G-20-001

Lessons learned:

- A novel signature!
- Doing quite well in “boosted Radion” regime.

- Extensive use of ML taggers.
- Whole shape is calibrated

- However, systematics for R→ 4q signal are very large, due to lack of good standard candles.



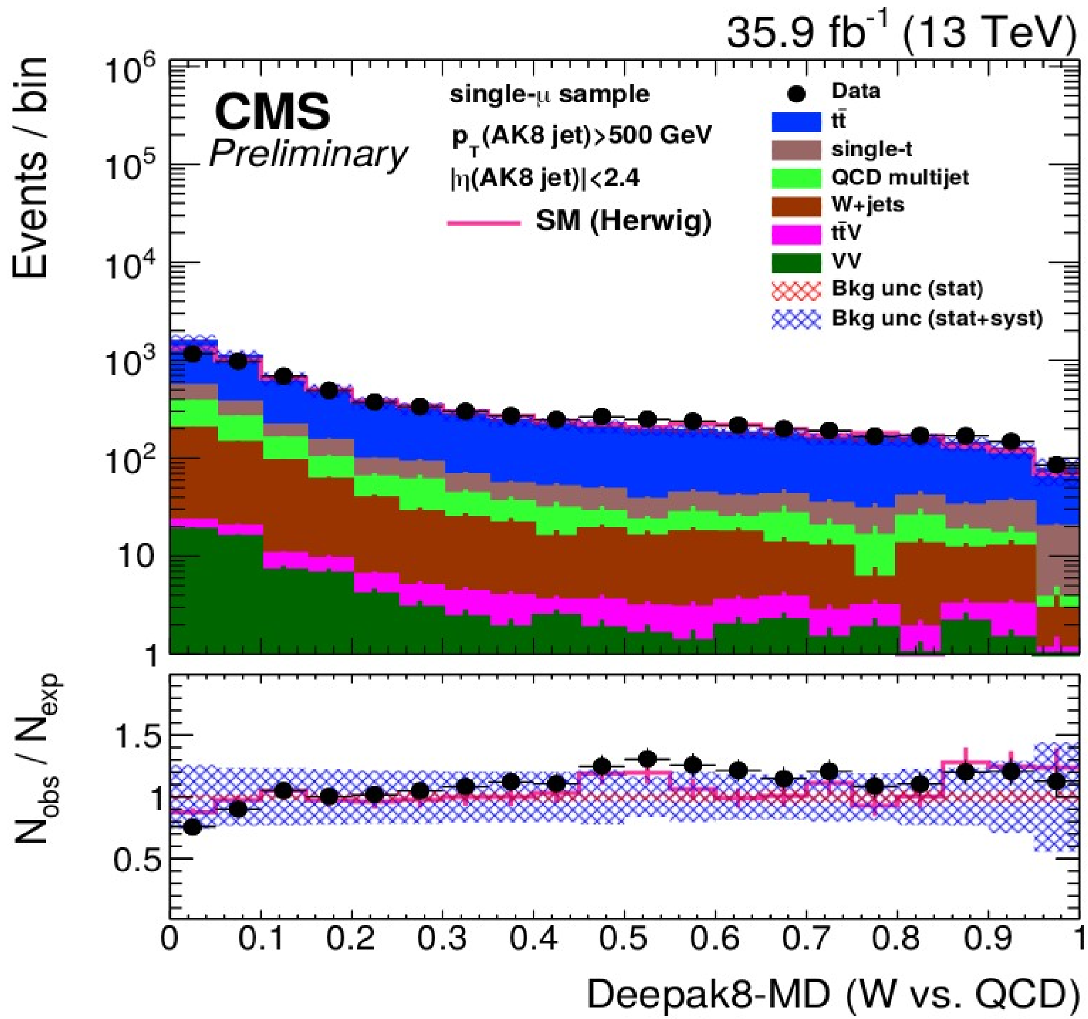
Summary so far: there's still work to do!

- Data ✓
- Tools (taggers, new variables) to suppress background and isolate the signal ✓
  - Most ML taggers still trained on MC...
- Background estimate (minimize uncertainty)
  - If dominated by ttbar, W+jets – get away with MC... ✓
  - QCD: tricky and messy (after lots of work… ✓)
- Signal efficiency (minimize uncertainty)
  - For top, W/Z, Higgs tagging – use standard candles ✓
  - For exotic signatures – ??? ✗

The topics of the rest of the talk.

Imperfect MC is used to train ML

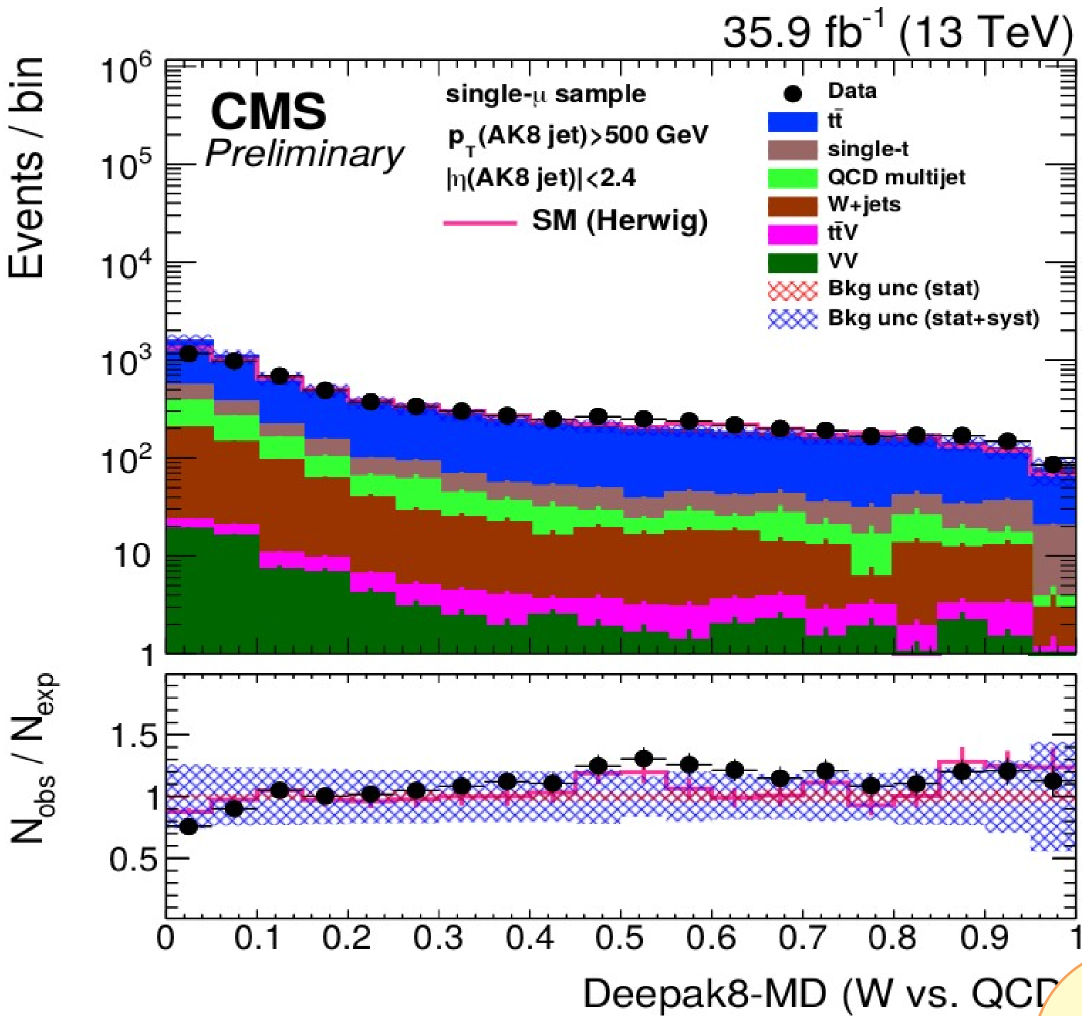
JME-18-002



- Powerful taggers, but…
- Nominally "within errors" from data
- Need to be careful:
  - Scale factors must be measured...
  - And they may be different from 1…
  - … with large error bars

Imperfect MC is used to train ML

JME-18-002

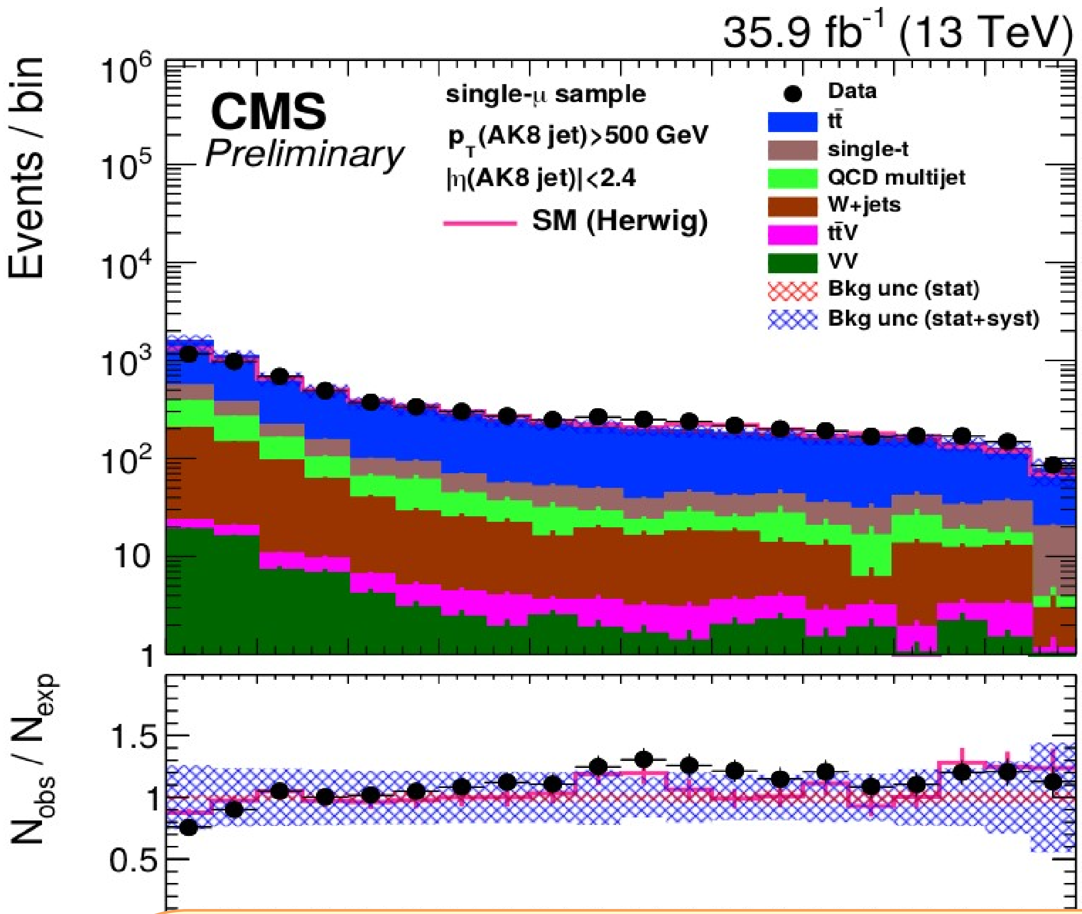


- Powerful taggers, but…
- Nominally "within errors" from data
- Need to be careful:
  - Scale factors must be measured...
  - And they may be different from 1…
  - … with large error bars

Can erase some of the gains from an improved tagger!

Imperfect MC is used to train ML

JME-18-002



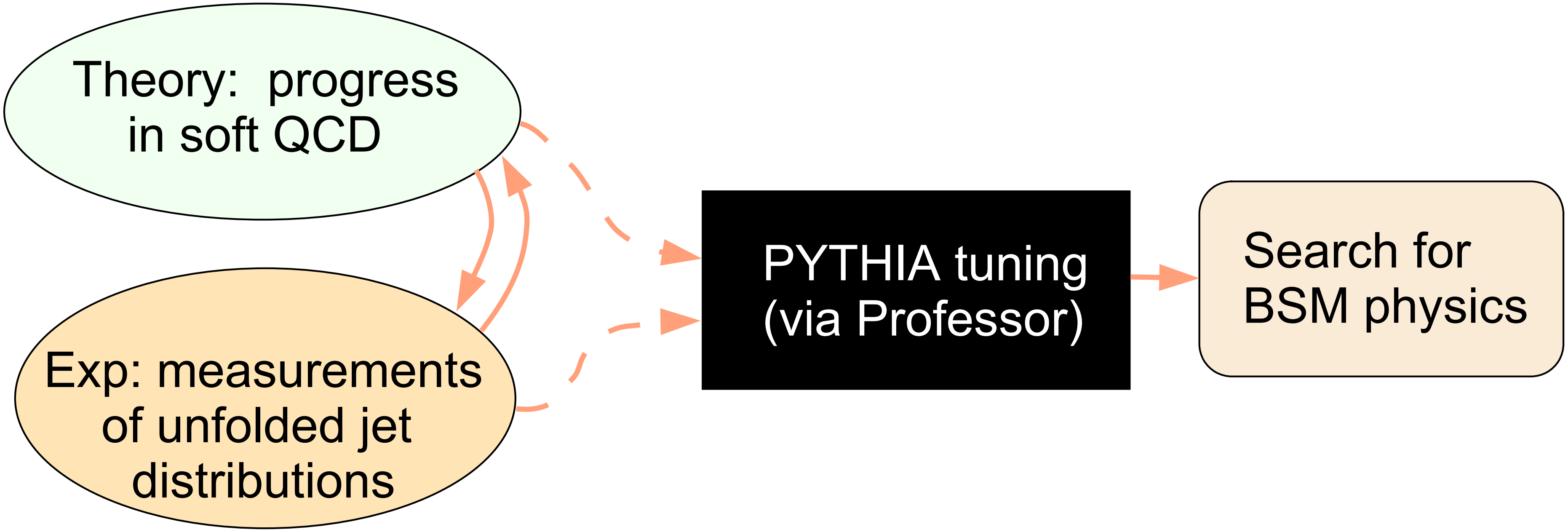
- Powerful taggers, but…
- Nominally "within errors" from data
- Need to be careful:
  - Scale factors must be measured...
  - And they may be different from 1…
  - … with large error bars

Hard to tell whether DNN is focusing on features poorly modeled in top/W/Z/H MC...

QCD modeling for the future

- With a better QCD modeling, we could:
  - Train ML algorithms
    - → better data/MC agreement
    - → minimize signal efficiency systematics
  - Decorrelate taggers
    - → well-behaved background shapes → better bkg estimates
    - → if there's a BSM excess, it would be "easier" to see
  - Estimate efficiencies of tagging N-prong jets
    (e.g. H→ WW*→ 4q, or BSM)
- In general, experimentalist's life would become a lot easier

Why can't I have that?!

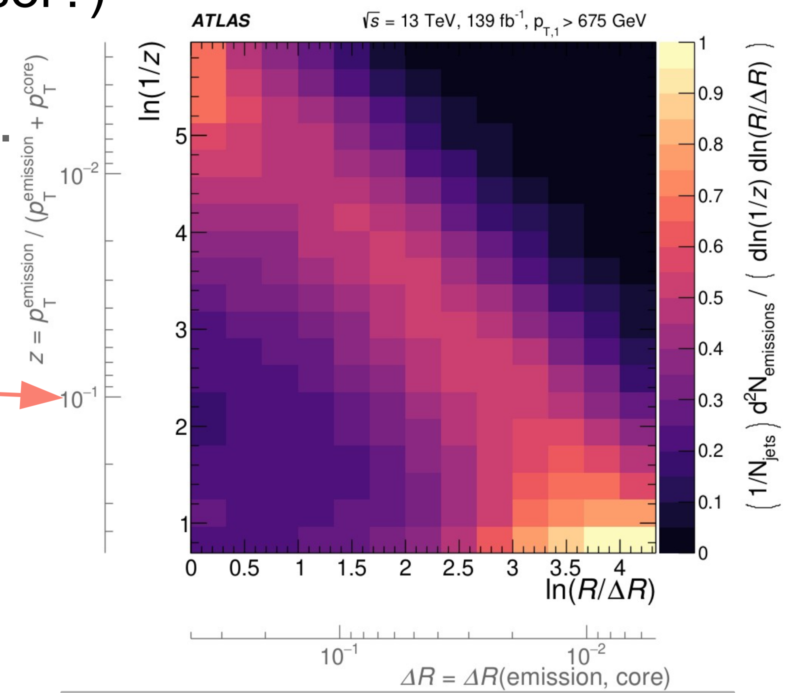


- Somehow, theoretical and experimental progress in soft QCD does not seem to propagate to PYTHIA we use.
  - Not enough measurements fed into "Professor"?
  - Can't tune both UE and substructure???
  - PYTHIA is insufficient for shower/hadronization?

What if PYTHIA doesn't have enough knobs?

- Make a better PYTHIA, or another shower program. Then tune to data. (Repurpose the Professor?)
- Or, correct simulation *a posteriori*.
  - Reweight using the Lund Plane?
  - Use ML: JUNIPR, DCTR?

    PRD 101, 091901(R) (2020)
  - …. or something else?



ATLAS: PRL 124, 222002 (2020)

  - Maybe the best:
    measure → tune PYTHIA → reweight residual differences.
  - Experimentally, **the key question**: what are the *uncertainties* on the result of this procedure?

Do we need QCD MC at all?

- For multijet background estimates, we don't need MC
  - Have been data-driven anyway
  - Although there could be subtle correlations…
- Unsupervised learning from data…
  - Learns QCD: e.g., AutoEncoders
  - Learns QCD in the presence of other backgrounds: e.g., CWoLa
- Can we interpolate between two sidebands
  - e.g., CWoLa hunting
- Can we *extrapolate* from one CR to another???

Do we need QCD MC at all?

- For multijet background estimates, we don't need MC
  - Have been data-driven anyway
  - Although there could be subtle correlations…
- Unsupervised learning from data…
  - Learns QCD: e.g., AutoEncoders
  - Learns QCD in the presence of other backgrounds: e.g., CWoLa
- Can we interp
  - e.g., CWoLa
- Can we *extrapolate* from one CR to another???

Covered by talks by Ben Nachman and David Shih in the ML session TOMORROW

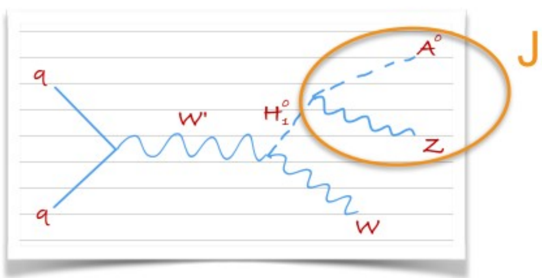
N-pronged jets: would be nice...

- The future of searches with substructure?

(Juan-Antonio Aguilar Saavedra, BOOST 2018)

Case II: Merged multibosons

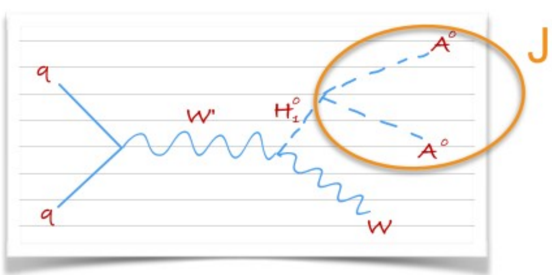
If intermediate particles are `light´, their decay products are merged



$$M_{W'} \gg M_{H_1^0} \gtrsim M_Z + M_{A^0}$$

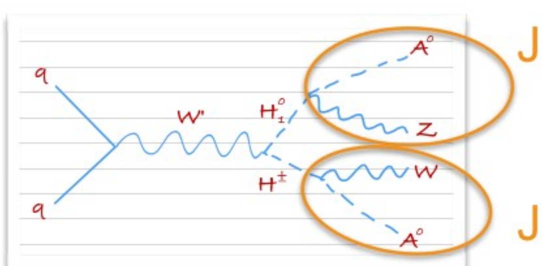
$Z \to qq$ / …

$A^0 \to bb$ → W + fat jet J



$$M_{W'} \gg M_{H_1^0} \gtrsim 2M_{A^0}$$

$A^0 \to bb$ → W + fat jet J



$$M_{W'} \gg M_{H_1^0}, M_{H^\pm} \gtrsim M_Z + M_{A^0}$$

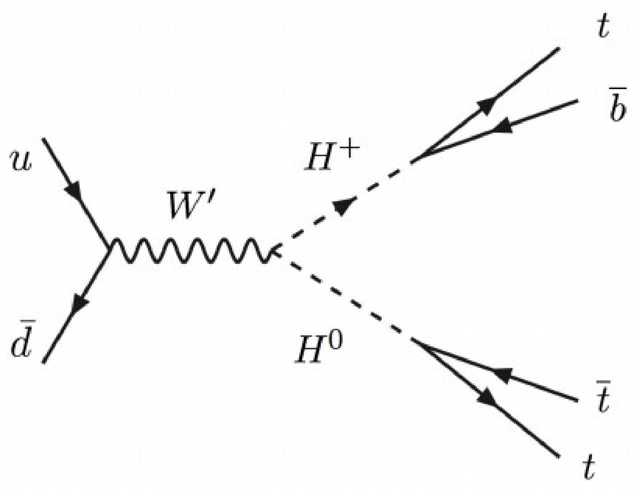
$W, Z \to qq$ / …

$A^0 \to bb$ → two fat jets J

More N-pronged jets

- More cool signatures with 4- and 6-pronged jets

Heavy Higgs bosons decay directly into a pair of the heaviest fermions:



For $M_{W'} \gg M_{H^+}$:

$(t\bar{b})$-tagged jet + $(t\bar{t})$-tagged jet

For $M_{W'} > M_{H^+} \gg m_t$:

three t-tagged jets + b

(Bogdan Dobrescu, BOOST 2017)

- Easy to do a cut-based analysis (let alone a DNN)
- But how to get the efficiency?

Conclusions

- Where we are, at the end of Run-2:
  - Boosted objects are useful and necessary
    - integral part of the LHC program
  - Lots of progress in understanding substructure and (sub)jet physics
  - Powerful new taggers and variables
  - top, W, Higgs taggers calibrated in data
- But, in some aspects, not enough progress:
  - Many of these improvements do not percolate to better/easier measurements
  - More complicated jets are under question until we figure out how to estimate signal efficiencies
- We still have (so much) work to do!

BACKUP MATERIAL

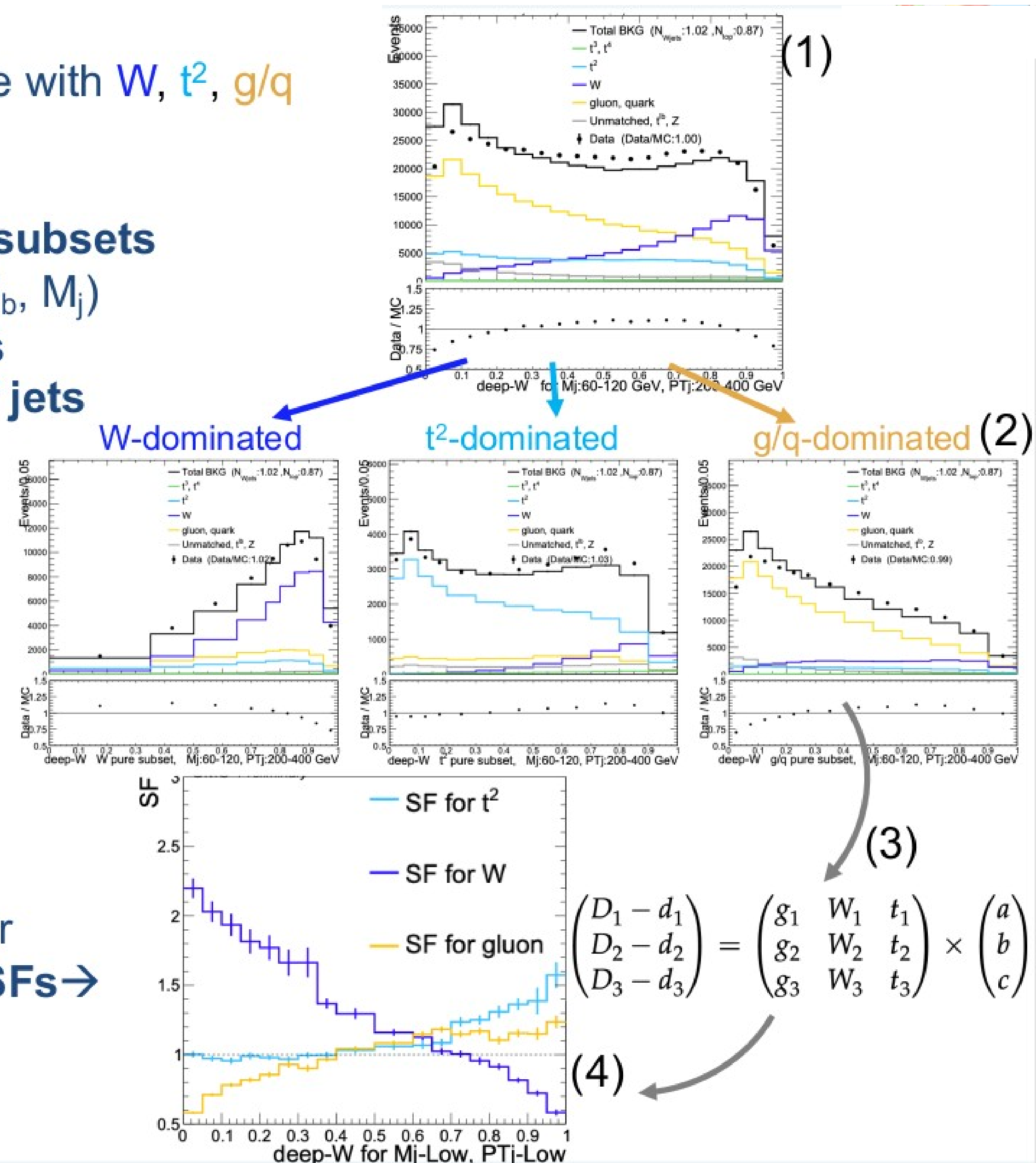
Exhibit C: Triboson resonances

1. Focus at low-Mj, low-PTj sample with W, t^2, g/q (first column of last slide)

2. **Split the samples into 3 pure subsets** (applying cuts on τ_{ij}, deep-x/y, N_b, M_j) **in a way where each subset is dominated by a single type of jets**
   → mismodeling revealed:

3. Demand:
   Data = scaled sum of yields
   $$D_1 = a g_1 + b W_1 + c t_1 + d_1$$
   Define system of 3 equations, 1 per each subset, and per tagger score bin

4. **Solve a 3x3 system** for SFs per each tagger score bin and **get SFs**→
   - **Known** yields: D, W, t, g/q, d
   - **Unknown** SFs: a, b, c

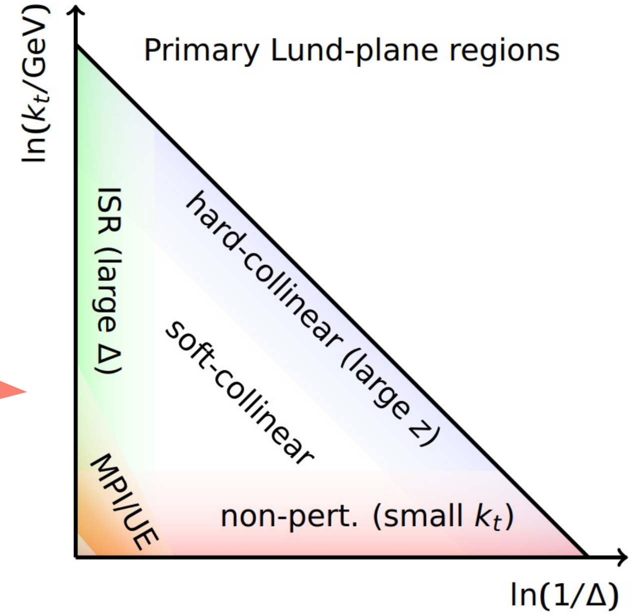


Primary Lund Jet Plane

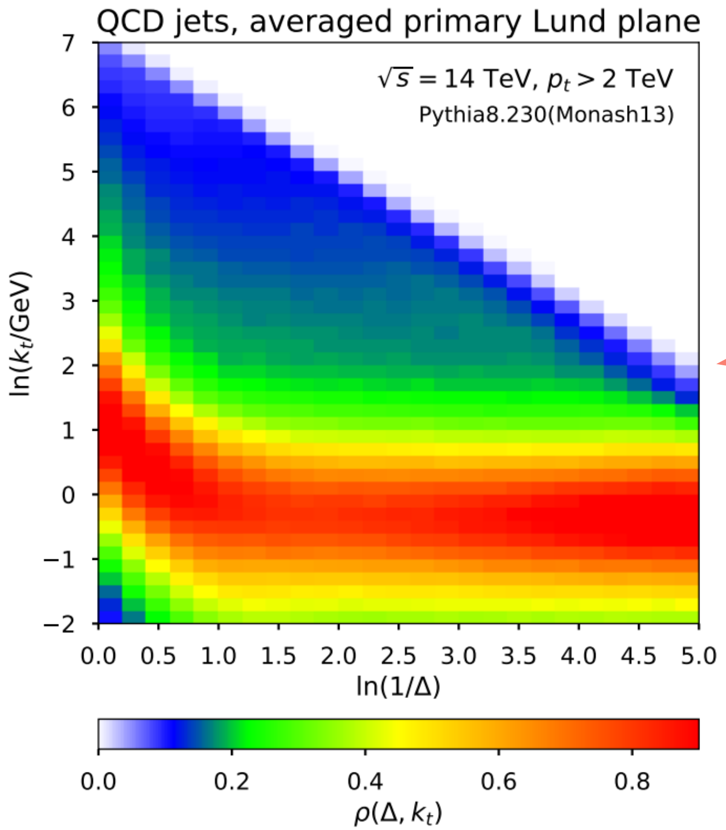
- “Jet is an unordered set”, I know… but…

- Access to low-level physics directly

- Intuitive and thus appealing



(Frederic Dreyer, BOOST’18)



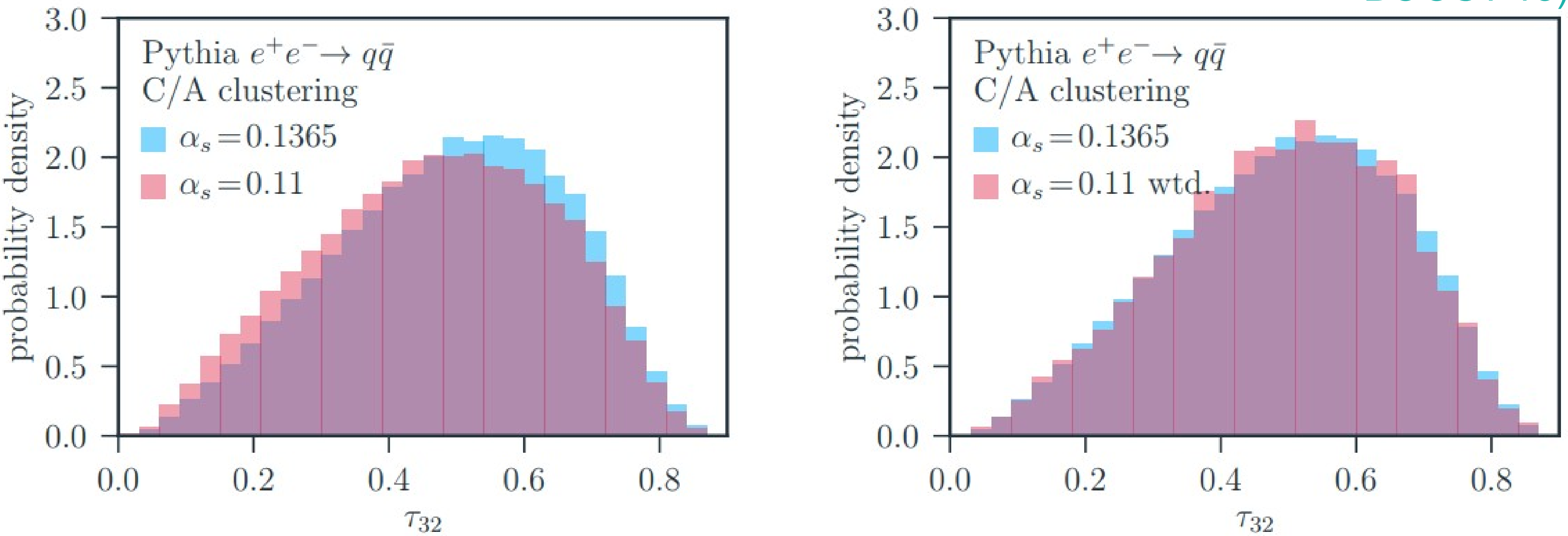
- Average Lund Plane (many jets)
  - one jet = set of point in this plane

- Can we reweight jets using the ratio of Lund Planes in data and MC?

JUNIPR

- Recursive NN, unsupervised learning on data
  - (A talk on more advance JUNIPR reweighting later this week!!!)
- Data/MC reweighting was one of its main goals!

(Cris Frye, BOOST'18)



- Works in MC: turns one PYTHIA into another.
- Will it work in data?

What to do about N-pronged jets?

- Give up, can't be done!
  - Can't measure efficiency in data
  - These analyses are always going to be out of reach...
- Report limit on $\sigma_X \cdot \mathcal{B}(X) \cdot \underline{\epsilon_X}$
  - Let the consumers of the paper worry about the signal efficiency
  - Would not affect the discovery, only limits
  - May actually spur progress in this area :-/
- Or try to make it work?
  - Learn how to reweight single quark jets from MC
  - Verify that the procedure works for W and top (2,3-prong)
  - Assign further systematics for 4,5,6-prong...