Mining the LHC data for excesses

Angelo Monteux

1707.05783 w/ P. Asadi, M. Buckley, A. DiFranzo, D. Shih

Rutgers, UC Irvine

CIPANP 2018

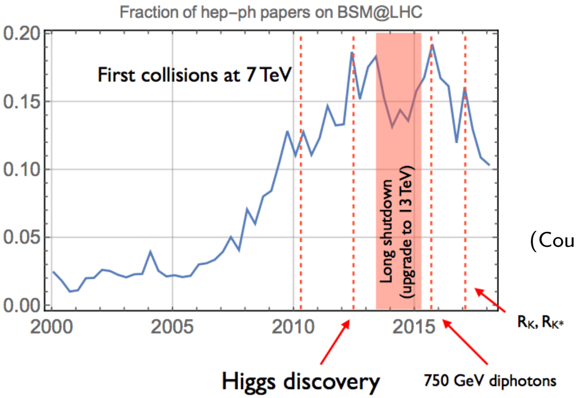
Palm Springs, 05/30/2018

Introduction

So far, 36 fb^{-1} of data released from 2015-2016 runs at 13 TeV (Moriond+summer 2017). More coming (2019?)

Clearly, we have not discovered the vanilla squark/gluino.

It seems that interest in the LHC has already declined in the pheno community.*



(Courtesy of D. Shih)

Is this justified?

For the most part, experimental collaborations only test their data against a (relatively small) set of simplified models.

Can we interpret/discover a complicated *unexpected* signal at a hadron collider?

Most discoveries start with a 2-3σ excess...
If we are to discover something before the HL-LHC, bumps should start appearing now.

It would be a shame if we did not make use of the LHC full potential by not looking in the right place.

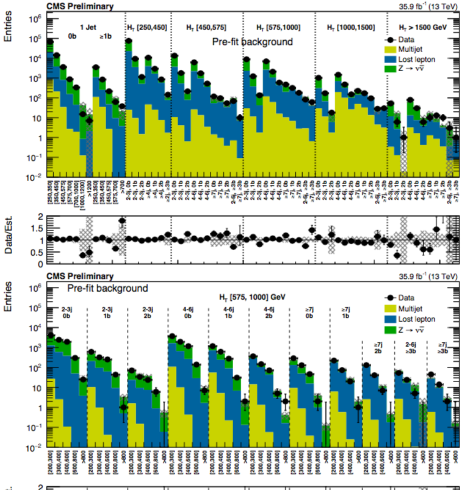
Avoiding bias

We should make sure to cast a net as wide as possible. ATLAS signal regions are optimized for a set of simplified models, while in CMS they form a (hyper)-grid by tiling the whole range of each kinematic variable.

First, take a detailed look at the data.
E.g. CMS-PAS-SUS-16-036 (jets + M_{T2})

213 signal regions!

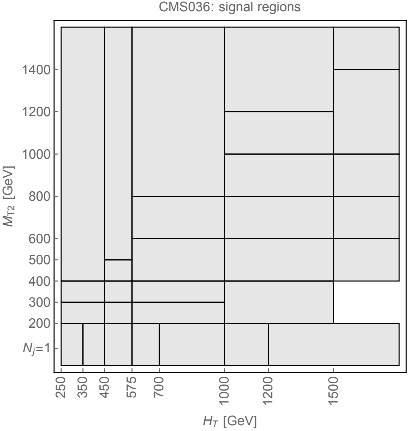
How to recognize an excess?



Example: CMS-PAS-SUS-16-036 (arXiv:1705.04650 - jets+M_{T2})

213 signal regions binned in four variables:

$N_j \geq 1, \quad N_b \geq 0, \quad M_{T2} \geq 200$ GeV, $\quad H_T \geq 250$ GeV



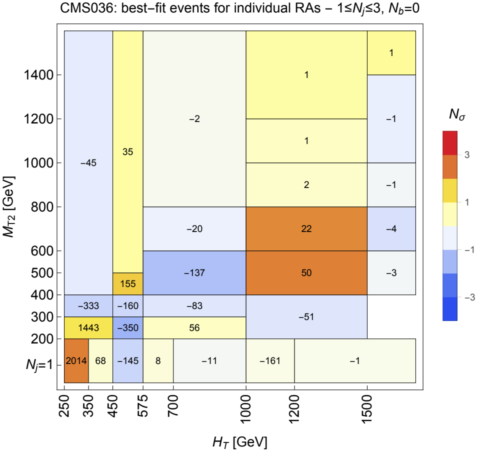
2D projection

each rectangle is one SR

Example: CMS-PAS-SUS-16-036 (arXiv:1705.04650 - jets+M_{T2})

213 signal regions binned in four variables:

$N_j \geq 1, \quad N_b \geq 0, \quad M_{T2} \geq 200$ GeV, $\quad H_T \geq 250$ GeV

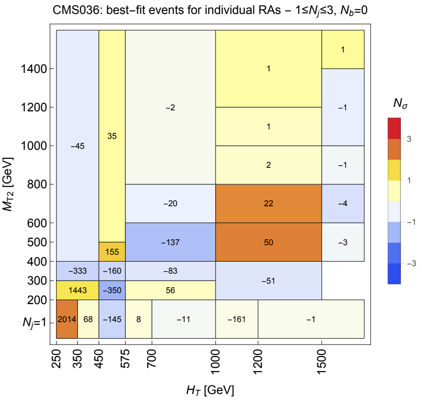


actual data

$N_j = 1,\, 2-3,\, N_b = 0$

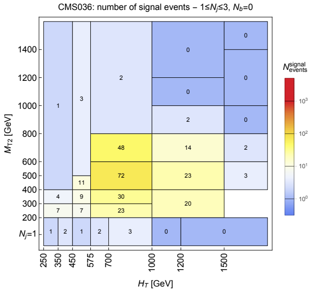
Color represents observed *single-bin* deviation from SM.
More details later on.

Example: CMS-PAS-SUS-16-036 (arXiv:1705.04650 - jets+M_{T2})

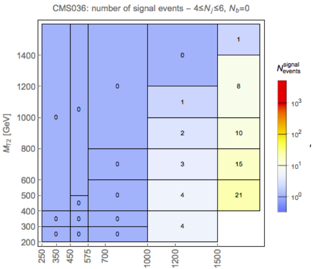


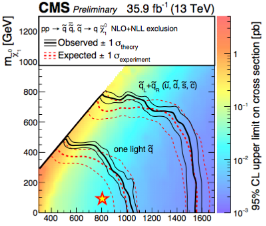
What does a signal look like?

$pp \to \tilde q \tilde q, \tilde q \to q\tilde\chi_1^0$: (800 GeV)



$pp \to \tilde g \tilde g, \tilde g \to qq\tilde\chi_1^0$: (1600 GeV)

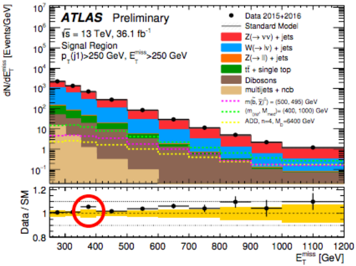


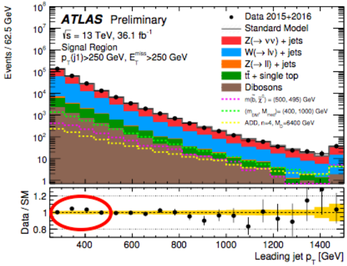


any model fitting the bumps in data? *Maybe. . .*

Are we blinding ourselves?

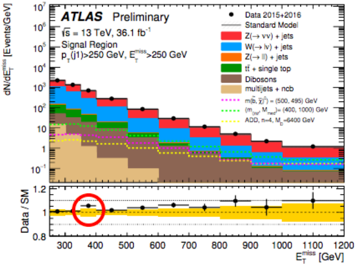
Example: ATLAS monojet

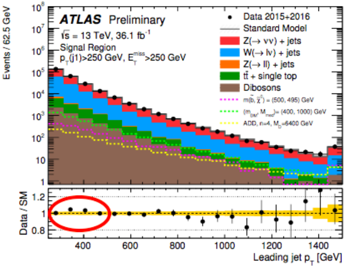




Are we blinding ourselves?

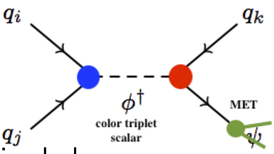
Example: ATLAS monojet





Is there a feature? The simplified models tested by ATLAS and CMS (DM mediator + jets) have smooth distributions. $\rightarrow$ a new model:

$$\mathcal{L} \supseteq \lambda\, \phi q_i^c q_j^c + g \phi^* q_i^c \psi + m_\psi \psi \psi' + g' \psi' N \tilde{N}$$



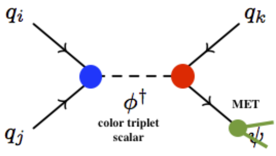
Possible justification:

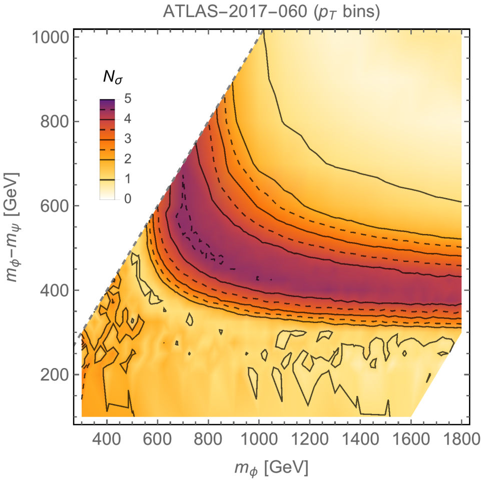
UV completion of "hylogenesis" model of asymmetric dark matter & baryogenesis.

[Davoudiasl,Morrissey,Sigurdson,Tulin, 1008.2399,1106.4320]

Mono-ϕ model

$$\mathcal{L} \supseteq \lambda\,\phi q_i^c q_j^c + g\phi^* q_i^c \psi + m_\psi \psi\psi' + g'\psi' N\tilde{N}$$



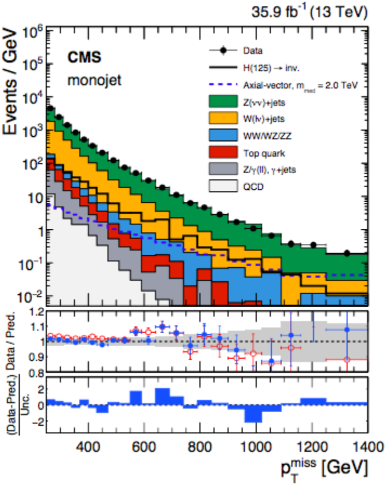


Local significance for this model goes up to 4.5σ!

Compatible with 3σ excess in orthogonal CMS SUSY search

Is this New Physics, or a red herring?

First, CMS monojet seems to see nothing:



The background fits are different between ATLAS and CMS:

- ATLAS takes the shape from theory and fits overall normalization to CRs.
- CMS fits floating bin-by-bin normalization to CRs.

Both have correlated nuisance parameters, and use NNLO predictions from theory.

[1705.04664]

Systematic errors dominate!

Previous 3.2fb^{-1}ATLAS monojet (bin-by-bin fit, but LO bg) seems to see no excess.

Waiting for data points from ATLAS to check compatibility. . .

Rectangular aggregations

We did not run into this excess by chance.

Instead, we developed a simple, model-independent, data-driven method to find significant excesses in the LHC data.

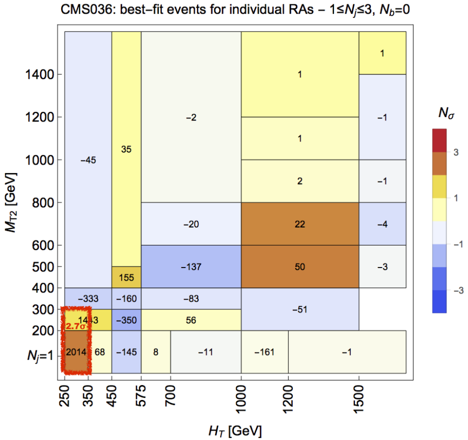
[Asadi,Buckley,Di Franzo,AM,Shih, 1707.05783]

Simple idea:*A true signal will usually populate mutiple "neighboring" signal regions, while background fluctuations are more often confined to individual bins.*

The only model-independent analysis is a single-bin analysis. Without having to assume a signal distribution over multiple bins, we can *aggregate* together nearby bins in a *rectangle* R.

Compute the likelihood of observing a deviation as large as observed in the data, assuming New Physics *only* contributes to that rectangle. Repeat for all aggregations. . .

Example: rectangular aggregations in CMS036

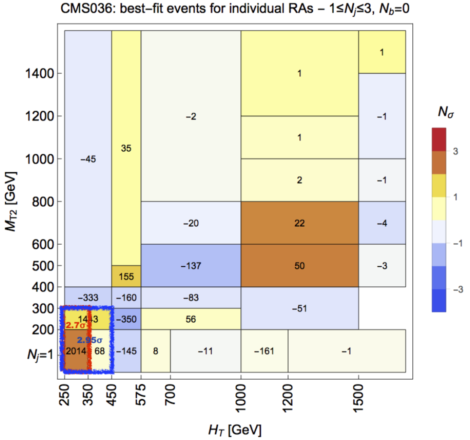


$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ

This example in 2D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Example: rectangular aggregations in CMS036

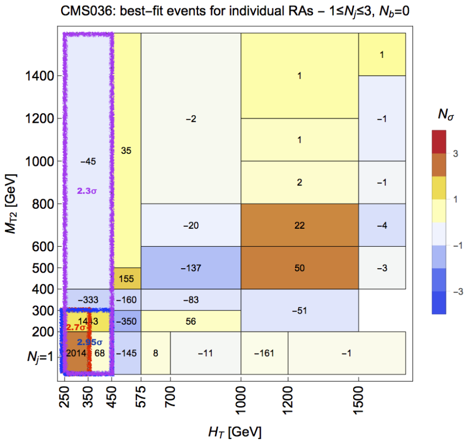


$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ
145 140	4 750	2.95σ

This example in 2D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Example: rectangular aggregations in CMS036

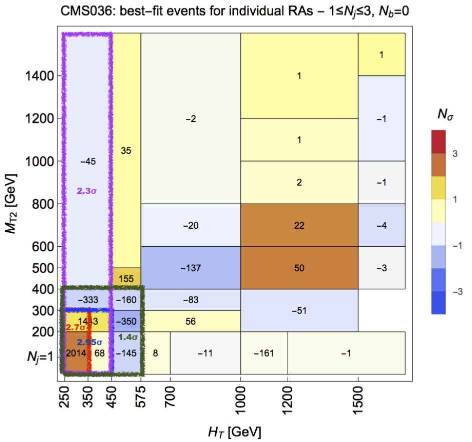


$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ
145 140	4 750	2.95σ
157 440	4 600	2.3σ

This example in 2D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Example: rectangular aggregations in CMS036

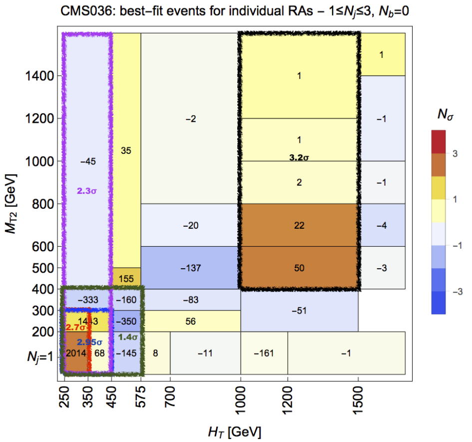


$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ
145 140	4 750	2.95σ
157 440	4 600	2.3σ
172 040	3 200	1.4σ

This example in 2D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Example: rectangular aggregations in CMS036

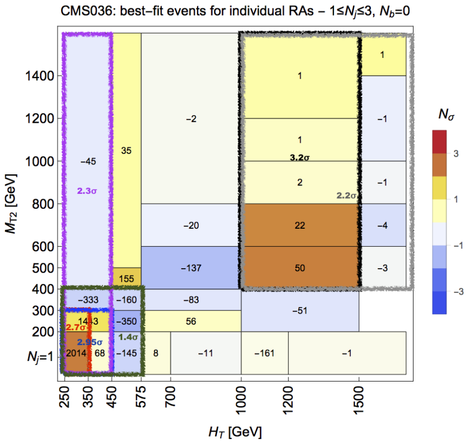


$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ
145 140	4 750	2.95σ
157 440	4 600	2.3σ
172 040	3 200	1.4σ
259	80	3.2σ

This example in 2D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Example: rectangular aggregations in CMS036

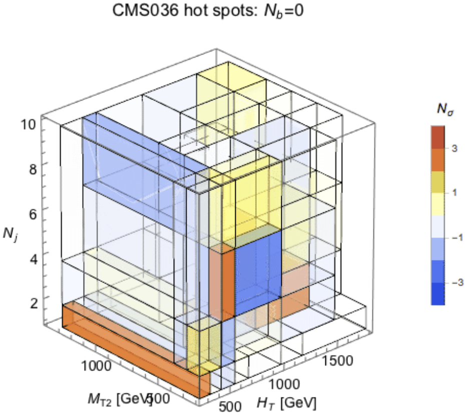


$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ
145 140	4 750	2.95σ
157 440	4 600	2.3σ
172 040	3 200	1.4σ
259	80	3.2σ
311	62	2.2σ

This example in 2D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Example: rectangular aggregations in CMS036

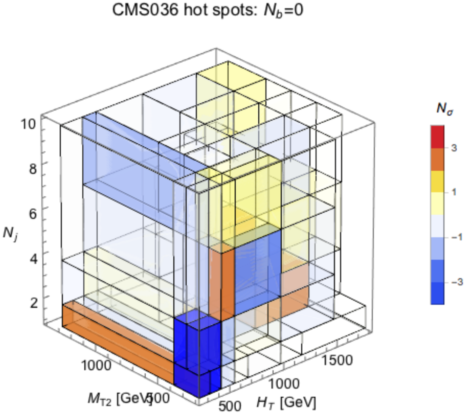


$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ
145 140	4 750	2.95σ
157 440	4 600	2.3σ
172 040	3 200	1.4σ
259	80	3.2σ
311	62	2.2σ

This example in 2D/3D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Example: rectangular aggregations in CMS036

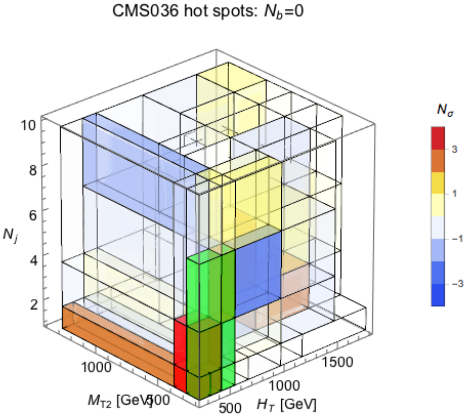


$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ
145 140	4 750	2.95σ
157 440	4 600	2.3σ
172 040	3 200	1.4σ
259	80	3.2σ
311	62	2.2σ

This example in 2D/3D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Example: rectangular aggregations in CMS036



$$b_{RA}^{postfit} = b_{RA}^{prefit} + \hat{\theta}_{RA}$$

n_{RA}	$n_{RA} - b_{RA}^{post}$	N_σ
130 840	4 000	2.74σ
145 140	4 750	2.95σ
157 440	4 600	2.3σ
172 040	3 200	1.4σ
259	80	3.2σ
311	62	2.2σ
169 100	4 800	2.48σ

This example in 2D/3D. Can't plot in 4D/4+D, but algorithm can be applied automatically.

Results

We apply this technique to two “big” CMS SUSY searches, and the same excess seems to appear:

- CMS-PAS-SUS-16-036 (CMS036): jets+M_{T2}

ROI		N_j	N_b	H_T (GeV)	M_{T2} (GeV)	N_σ
2	b	$1-3$	0	$250-450$	$200-300$	2.95
	d	$1-3$	0	$250-350$	$200-300$	2.74

- CMS-PAS-SUS-16-033 (CMS033): jets+$\not{E}_T$

ROI		N_j	N_b	H_T (GeV)	$\not{E}_T$ (GeV)	N_σ
	a	$2-6$	0	$300-500$	$300-500$	2.96
2	c	$2-4$	0	$300-500$	$300-500$	2.64
	d	$3-6$	0	$300-500$	$300-500$	2.57

NB: same dataset, we cannot combine the significances.

Here we found the monojet excess in the CMS data!!

NB: background fit different than CMS monojet! LO Monte-Carlo, data-driven, bigger uncertainties. *nothing there?* ☹

- ATLAS060 (monojet):

ROI	N_j	p_T (GeV)	N_σ
	1	$300-450$	4.66

Results

We apply this technique to two "big" CMS SUSY searches, and the same excess seems to appear:

- CMS-PAS-SUS-16-036 (CMS036): jets+M_{T2}

ROI		N_j	N_b	H_T (GeV)	M_{T2} (GeV)	N_σ
2	b	$1-3$	0	$250-450$	$200-300$	2.95
	d	$1-3$	0	$250-350$	$200-300$	2.74

- CMS-PAS-SUS-16-033 (CMS033): jets+$\not{E}_T$

ROI		N_j	N_b	H_T (GeV)	$\not{E}_T$ (GeV)	N_σ
	a	$2-6$	0	$300-500$	$300-500$	2.96
2	c	$2-4$	0	$300-500$	$300-500$	2.64
	d	$3-6$	0	$300-500$	$300-500$	2.57

NB: same dataset, we cannot combine the significances.

Here we found the monojet excess in the CMS data!!

NB: background fit different than CMS monojet! LO Monte-Carlo, data-driven, bigger uncertainties. *nothing there?* ☹

- ATLAS060 (monojet):

ROI	N_j	p_T (GeV)	N_σ
	1	$300-450$	4.66

Conclusions

We have introduced a new technique to sift through the CMS datasets in search for deviations from the SM background.

The aggregation strategy itself is simple and yet powerful. Takes 5 minutes to write a script, seconds to hours to run with Minuit to find significance (minimize $\Delta \ln \mathcal{L}$).
It only needs event counts (and error/covariance matrix). No Madgraph/Pythia/Delphes needed until model-building step.

Python/Jupyter notebook code on GitHub, anyone can play with it!
https://github.com/ilmonteux/LHC_rectangular_aggregation/

We have shown that there are interesting *previously unidentified* excesses. 3σ fluctuations come and go all the time, but the only way to know is to keep looking (another $\sim$100fb^{-1} recorded this year!).

There is value in keeping an eye on these hot spots, mostly to avoid raising thresholds and blind ourselves in the future.

Conclusions

We have introduced a new technique to sift through the CMS datasets in search for deviations from the SM background.

The aggregation strategy itself is simple and yet powerful. Takes 5 minutes to write a script, seconds to hours to run with Minuit to find significance (minimize $\Delta \ln \mathcal{L}$).
It only needs event counts (and error/covariance matrix). No Madgraph/Pythia/Delphes needed until model-building step.

Python/Jupyter notebook code on GitHub, anyone can play with it!

`https://github.com/ilmonteux/LHC_rectangular_aggregation/`

We have shown that there are interesting *previously unidentified* excesses. 3σ fluctuations come and go all the time, but the only way to know is to keep looking (another $\sim$100fb^{-1} recorded this year!).

There is value in keeping an eye on these hot spots, mostly to avoid raising thresholds and blind ourselves in the future.

Conclusions

We have introduced a new technique to sift through the CMS datasets in search for deviations from the SM background.

The aggregation strategy itself is simple and yet powerful. Takes 5 minutes to write a script, seconds to hours to run with Minuit to find significance (minimize $\Delta \ln \mathcal{L}$).
It only needs event counts (and error/covariance matrix). No Madgraph/Pythia/Delphes needed until model-building step.

Python/Jupyter notebook code on GitHub, anyone can play with it!
`https://github.com/ilmonteux/LHC_rectangular_aggregation/`

We have shown that there are interesting *previously unidentified* excesses. 3σ fluctuations come and go all the time, but the only way to know is to keep looking (another $\sim$100fb^{-1} recorded this year!).

There is value in keeping an eye on these hot spots, mostly to avoid raising thresholds and blind ourselves in the future.

Thank you!

The important point is that the LHC dataset is not as bleak as many assume!

We are grateful to our ATLAS and CMS colleagues for making re-interpretating their results possible and (relatively) easy (e.g. correlation matrices).

And for being receptive to discuss with theorists finding excesses in their data.

Thanks for listening! Stay tuned for more data!

Thank you!

The important point is that the LHC dataset is not as bleak as many assume!

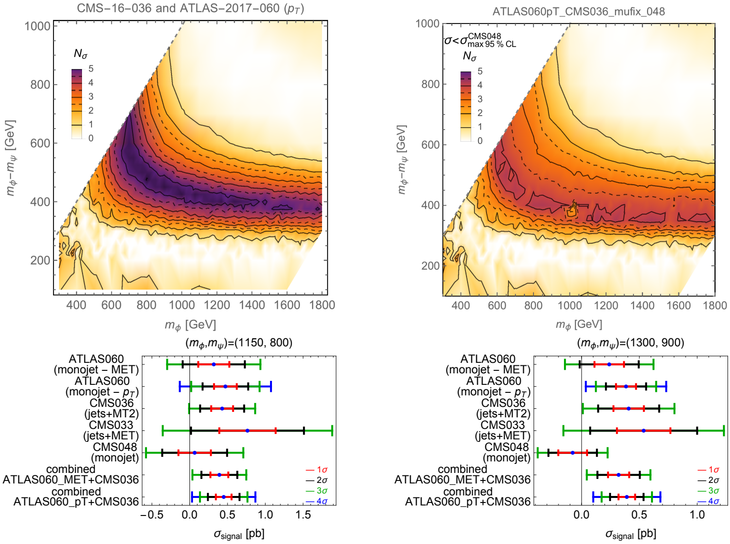
We are grateful to our ATLAS and CMS colleagues for making re-interpretating their results possible and (relatively) easy (e.g. correlation matrices).

And for being receptive to discuss with theorists finding excesses in their data.

Thanks for listening! Stay tuned for more data!

Extra slides

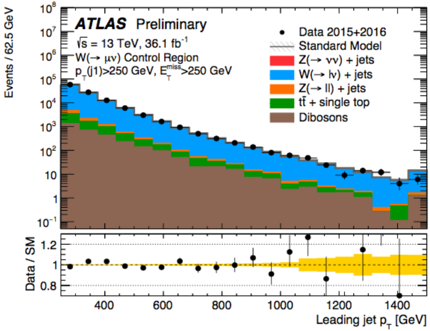
Cross-search compatibility

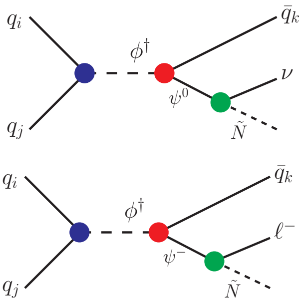


Tension with CMS048 not too bad in remaining regions (by construction). All other searches consistent in same range.

Control regions of ATLAS monojet

The extrapolation to signal regions depends on the CRs ($W \to \ell\nu$+jets, $Z \to \ell\ell$+jets). Is there a bump in the control regions?





A simple extension of the mono-ϕ model could populate the W control regions... Harder, need to find other control regions or cut much harder on the "W-ness" of the background.

$$\mathcal{L} \supseteq \lambda\,\phi q_i^c q_j^c + g\phi^* Q_L \psi_L + m_\psi \psi\psi' + m_\phi^2 |\phi|^2 + g' \psi_L L \tilde{N}$$

Look-elsewhere effect

What is the probability of seeing such a greater or equal fluctuation, after having looked *everywhere*?

- ATLAS monojet: simply a 4.5σ bump in 26 bins. Highly significant even after look-elsewhere. The important thing is to control *systematics*.
- model-independent: with 213 SRs (33,000 RAs), how often will the SM fluctuate in such a way to give at least a 3.5σ excess in at least one aggregation?

  pseudo-experiments $\rightarrow 15\% \simeq 1.5\sigma$ global

  passing plausibility tests (?) $\sim 8\% \simeq 1.75\sigma$ global
- model-dependent: given this model, how often will the SM fluctuate in such a way to give at least a 3σ excess anywhere in the mass plane?

  pseudo-experiments $\rightarrow 5 - 3\% \simeq 1.95 - 2.2\sigma$ global
- what is the likelihood of a *compatible* fluctuations in the underlying data of two CMS searches, or between ATLAS and CMS? cannot answer at our level

Only further scrutiny/data can tell us if it was a fluctuation or not.

Statistics

We use the standard LHC profile likelihood approach:

$$\mathcal{L}(\mu,\theta) = \prod_i \frac{(\mu s_i + b_i + \theta_i)^{n_i} e^{-(\mu s_i + b_i + \theta_i)}}{n_i!} \exp\left(-\frac{1}{2}\theta^T V^{-1}\theta\right)$$

[Cowan, Cranmer, Gross, Vitells, 1007.1727]

- n_i is the number of observed events in each bin.
- s_i is the number of BSM signal events, for a reference xsec
- μ is a cross section multiplier.
- b_i is the expected background count in the bin (extrapolated from control regions). θ_i are *nuisance parameters* for the background b_i, and their variation is modulated by the covariance matrix V.

[CMS-NOTE-2017-001] great!

Maximizing the likelihood we get:

- local maximum for given μ: $\mathcal{L}(\mu, \hat{\theta}_\mu)$ - SM= $\mathcal{L}(0, \hat{\theta}_0)$
- global maximum: $\mathcal{L}(\hat{\mu}, \hat{\hat{\theta}})$

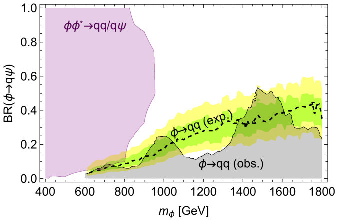
construct delta log-likelihood
$$q_0 \equiv \begin{cases} -2\ln\frac{\mathcal{L}(0,\hat{\theta}_0)}{\mathcal{L}(\hat{\mu},\hat{\hat{\theta}})} & \hat{\mu} \geqslant 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

χ^2-distributed with 1 dof in the large N limit, $N_\sigma = \sqrt{q_0}$.

Additional model signatures

A priori, initial parton flavors and the branching ratio into $q\psi$ are undefined:

- dijet resonance: $qq \to \phi \to qq$ (interestingly, a 2σ deviation in last CMS search near 1.2 TeV)
- pair production: $gg \to \phi\phi^*$: $2j + \not{E}_T$, $3j + \not{E}_T$, $(2j)(2j)$.



$$\sigma_\phi \times BR_{q\psi} \sim \frac{\lambda^2 g^2}{\lambda^2+g^2} \overset{\lambda \gg g}{\to} const.$$

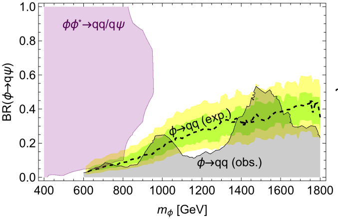
$$\mathcal{L} \supseteq \lambda\, \phi q_i^c q_j^c + g \phi^* q_i^c \psi + m_\psi \psi\psi' + m_\phi^2 |\phi|^2 + g' \psi' N \tilde{N}$$

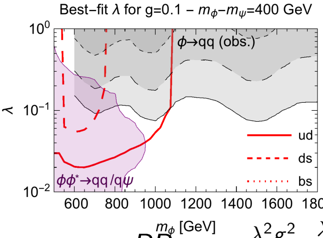
Excess would be due to couplings of order $0.1 - 1$ depending if it couples to light or heavy flavors.

Additional model signatures

A priori, initial parton flavors and the branching ratio into $q\psi$ are undefined:

- dijet resonance: $qq \to \phi \to qq$ (interestingly, a 2σ deviation in last CMS search near 1.2 TeV)
- pair production: $gg \to \phi\phi^*$: $2j + \not{E}_T$, $3j + \not{E}_T$, $(2j)(2j)$.





$$\sigma_\phi \times BR_{q\psi} \sim \frac{\lambda^2 g^2}{\lambda^2+g^2} \overset{\lambda \gg g}{\to} const.$$

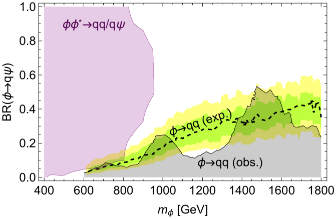
$$\mathcal{L} \supseteq \lambda\,\phi q_i^c q_j^c + g\phi^* q_i^c \psi + m_\psi \psi\psi' + m_\phi^2|\phi|^2 + g'\psi' N\tilde{N}$$

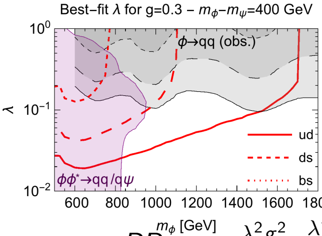
Excess would be due to couplings of order $0.1 - 1$ depending if it couples to light or heavy flavors.

Additional model signatures

A priori, initial parton flavors and the branching ratio into $q\psi$ are undefined:

- dijet resonance: $qq \to \phi \to qq$ (interestingly, a 2σ deviation in last CMS search near 1.2 TeV)
- pair production: $gg \to \phi\phi^*$: $2j + \not{E}_T$, $3j + \not{E}_T$, $(2j)(2j)$.





$$\sigma_\phi \times BR_{q\psi} \sim \frac{\lambda^2 g^2}{\lambda^2 + g^2} \overset{\lambda \gg g}{\longrightarrow} const.$$

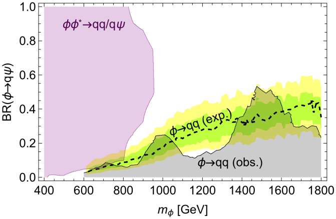
$$\mathcal{L} \supseteq \lambda\, \phi q_i^c q_j^c + g \phi^* q_i^c \psi + m_\psi \psi \psi' + m_\phi^2 |\phi|^2 + g' \psi' N \tilde{N}$$

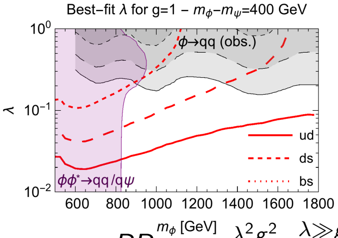
Excess would be due to couplings of order $0.1 - 1$ depending if it couples to light or heavy flavors.

Additional model signatures

A priori, initial parton flavors and the branching ratio into $q\psi$ are undefined:

- dijet resonance: $qq \to \phi \to qq$ (interestingly, a 2σ deviation in last CMS search near 1.2 TeV)
- pair production: $gg \to \phi\phi^*$: $2j + \not{E}_T$, $3j + \not{E}_T$, $(2j)(2j)$.





$$\sigma_\phi \times BR_{q\psi} \sim \frac{\lambda^2 g^2}{\lambda^2+g^2} \overset{\lambda \gg g}{\to} const.$$

$$\mathcal{L} \supseteq \lambda\,\phi q_i^c q_j^c + g\phi^* q_i^c \psi + m_\psi \psi\psi' + m_\phi^2 |\phi|^2 + g'\psi' N\tilde{N}$$

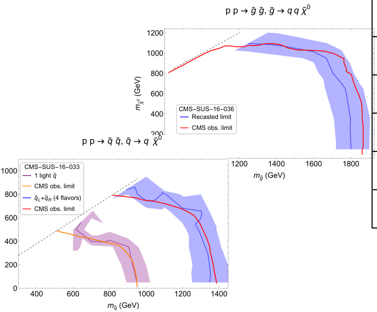
Excess would be due to couplings of order $0.1 - 1$ depending if it couples to light or heavy flavors.

Recasting Pipeline

Our pipeline:

- Madgraph5 2.4.3
- Pythia 8.219
- Delphes 3.4
- (py)ROOT 6
- iminuit

validated against official plots



ICHEP	Moriond
ATLAS-CONF-2016-037 (ATLAS SSL/3L + MET)	CMS-PAS-SUS-16-032 (b's+MET)
ATLAS-CONF-2016-052 (ATLAS multi-b+MET)	CMS-PAS-SUS-16-033 (jets+MHT)
ATLAS-CONF-2016-054 (ATLAS 1L + jets + MET)	CMS-PAS-SUS-16-036 (jets+MT2)
ATLAS-CONF-2016-057 (ATLAS multijet [RPV])	CMS-PAS-SUS-16-035 (SS2L)
ATLAS-CONF-2016-078 (ATLAS 2-6 jets+MET)	CMS-PAS-SUS-16-042 (1L+jets+MET - $\Delta\Phi$)
ATLAS-CONF-2016-094 (ATLAS 1L + many jets)	CMS-PAS-SUS-16-051 (stop 1L)
ATLAS-CONF-2016-095 (ATLAS 8-10 jets)	CMS-PAS-SUS-17-001 (stop 2L)
ATLAS-CONF-2016-077 (ATLAS stop 0L)	ATLAS-CONF-2017-020 (stop 0L)
ATLAS-CONF-2016-050 (ATLAS stop 1L)	ATLAS-CONF-2017-021 (b's + MET)
SUS-16-014-pas (CMS jets+MET)	ATLAS-CONF-2017-022 (2-6 jets + MET)
SUS-16-028-pas (CMS stop 1L)	ATLAS-CONF-2017-013 (1L+jets [RPV])
SUS-16-030-pas (CMS stop 0L boosted)	
ATLAS-CONF-2016-094 (ATLAS 1L + jets [RPV])	

– 50% "recast uncertainty"