Analysis Combinations and Proto-modelling

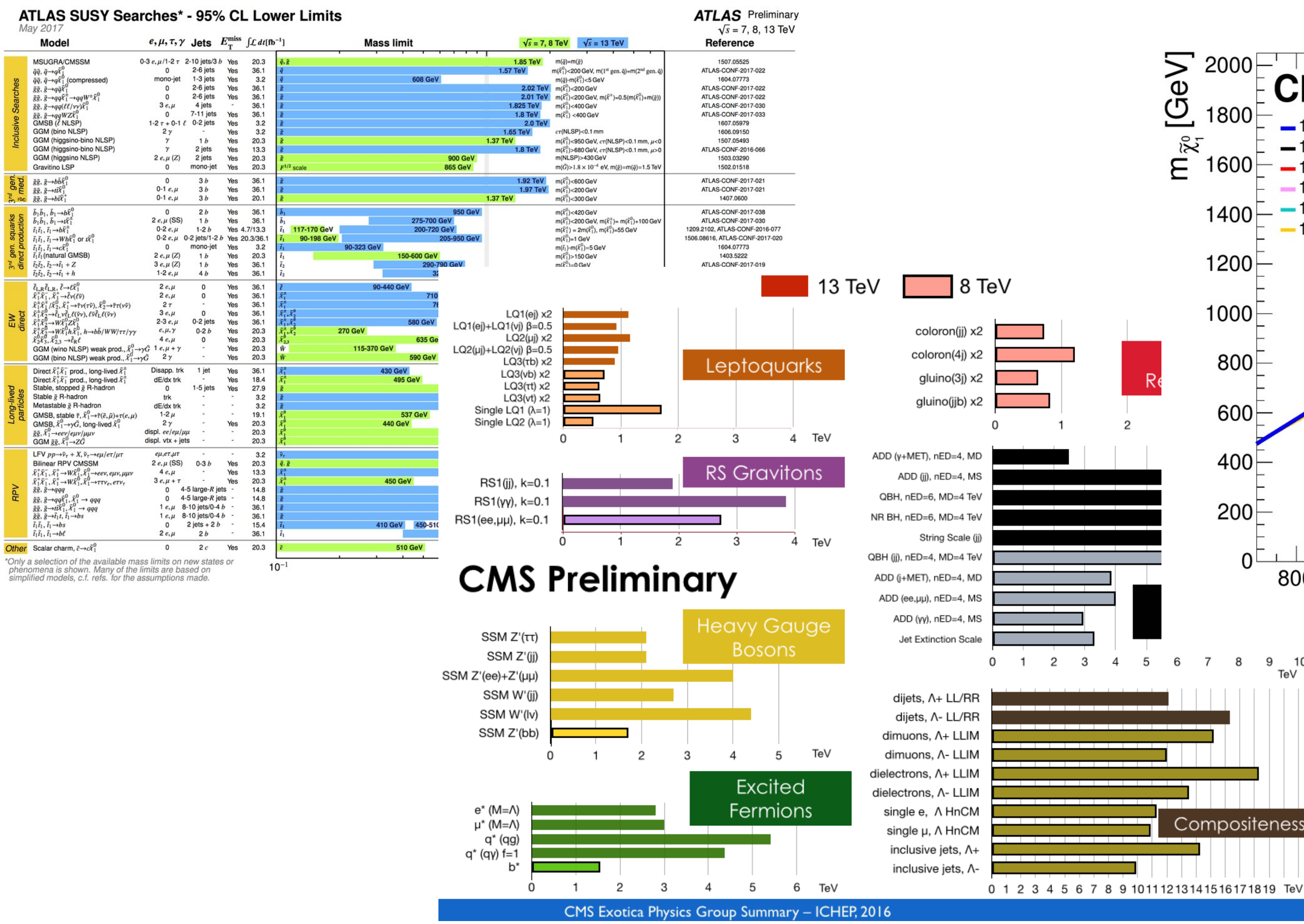
PROTOMODEL BUILDER

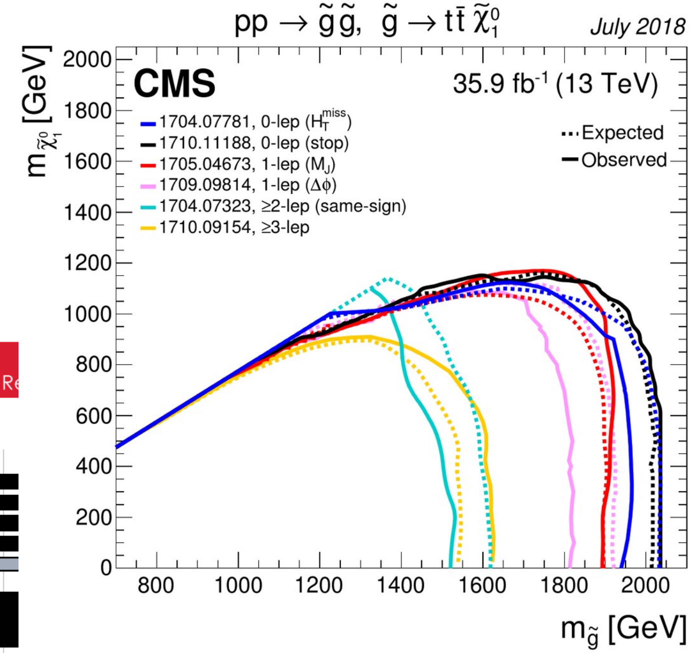
https://arxiv.org/abs/2012.12246

Sabine Kraml (LPSC), Andre Lessa (UFABC),
Wolfgang Waltenberger (ÖAW, Uni Wien)

Publication of statistical models, hands-on workshop, Nov 2021

Searching For Dispersed Signals



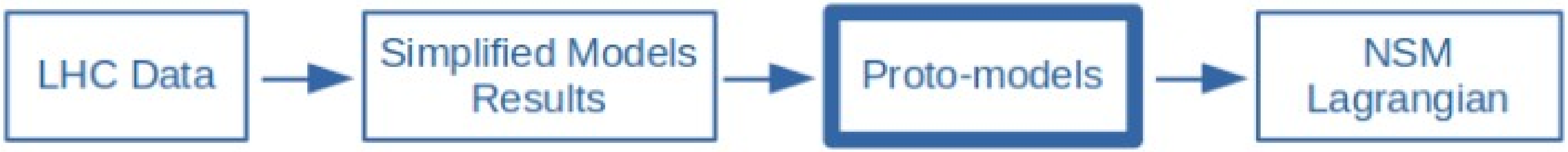


- Huge number of experimental results
- What if new physics is already slowly seeping into our analyses, but in a very dispersed manner?

Our aim:

Hunting For Dispersed Signals in LHC's Published BSM Search Results

Hunting For Dispersed Signals in LHC's Published BSM Search Results



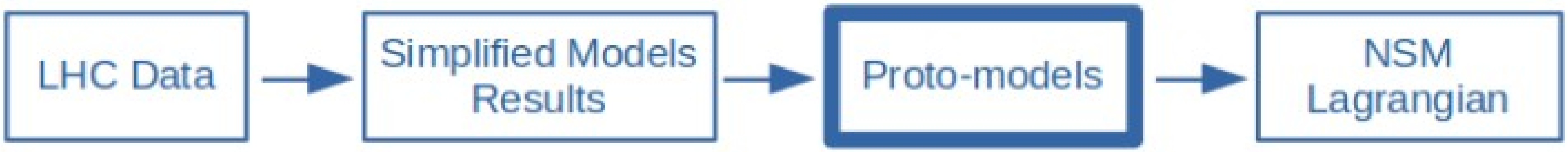
The enormity of LHC data of the BSM searches ….

… has been summarized in simplifed models results ….

… from which we construct precursor theories, identifying potential dispersed signals ….

…. only later will we worry about fundamental BSM Lagrangians.

Hunting For Dispersed Signals in LHC's Published BSM Search Results



The enormity of LHC data of the BSM searches ….

… has been summarized in simplifed models results ….

… from which we construct precursor theories, identifying potential dispersed signals ….

…. only later will we worry about fundamental BSM Lagrangians.

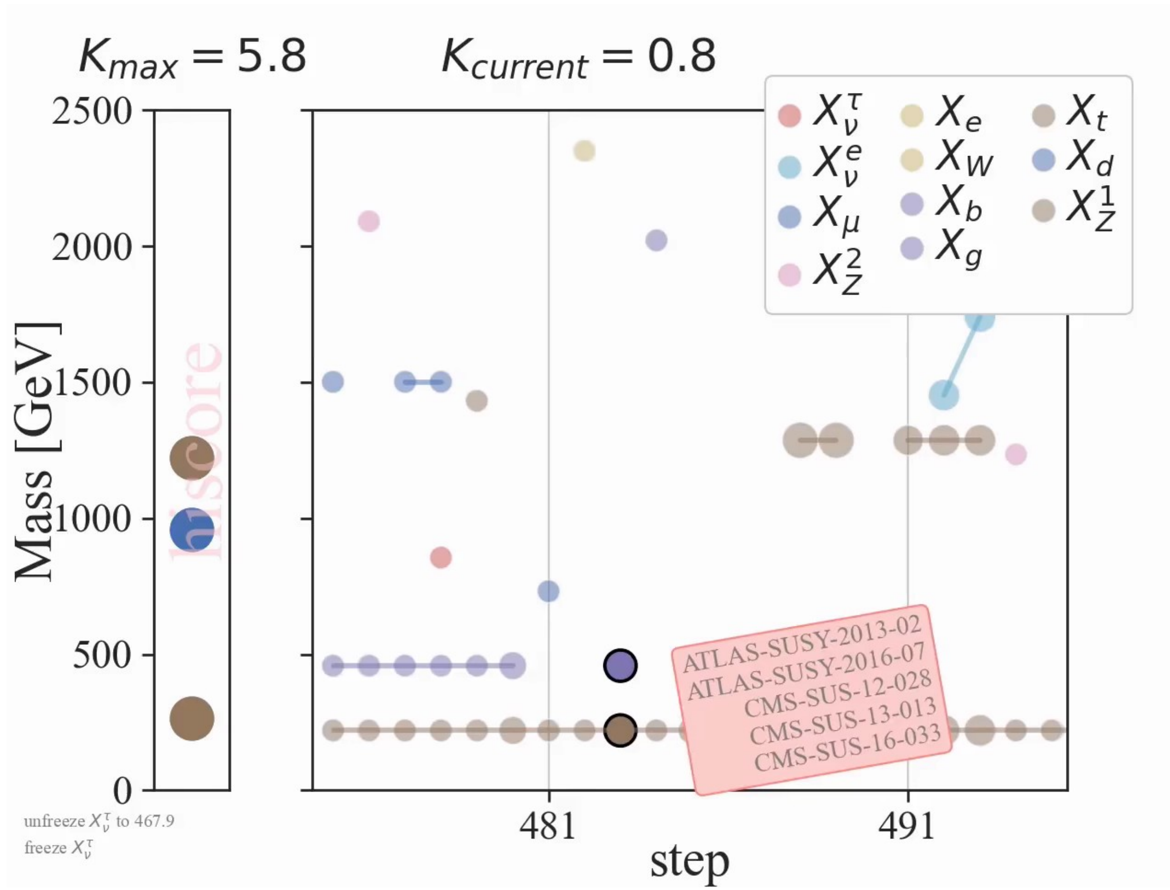
Two important aspects:

- need to combine likelihoods for all possible combinations of results
- build up precursor theories “proto-models” (:= consistent sets of simplified models) for context

Artificial Proto-modelling

a test statistic

Particle spectra



A hiscore protomodel

potential dispersed signals

Random modifications

an MCMC-like walk

https://smodels.github.io/protomodels/videos

The Test Statistic

The test statistic $\mathbf{K}^c$ is a likelihood-ratio test that quantifies how much better the proto-model describes the data than the Standard-Model (plus a penalty for model complexity).

$$K^c := -2\ln \frac{\mathrm{L}^c_{\mathbf{SM}} \cdot \pi(\mathbf{SM})}{\mathrm{L}^c_{\mathbf{BSM}}(\hat{\mu}) \cdot \pi(\mathbf{BSM})}$$

Joint likelihoods: combining “complete” sets of results that are assumed to be approximately uncorrelated.

Priors of the models used to penalize for model complexity, they are of the form

$$\pi(\mathrm{M}) \approx \exp[-c \cdot \mathrm{ndf}]$$

“AIC-like”

We search for proto-models and combinations of results / likelihoods that maximize K^c while remaining compatible with all negative results in our database.

INPUT DATA

The test statistic is based on likelihoods.

- likelihood computation based on simplified models results in SModelS database
- vast number efficiency and upper limit maps from ~ 50 CMS and ~ 50 ATLAS publications.

ID	Short Description	$\mathcal{L}$ [fb^{-1}]	UL$_{\rm obs}$	UL$_{\rm exp}$	EM	comb.
CMS-PAS-EXO-16-036 [83]	HSCP	12.9	✓			
CMS-PAS-SUS-16-052 [84]	ISR jet + soft ℓ	35.9	✓		✓	Cov.
CMS-SUS-16-009 [85]	0ℓ + jets, top tag	2.3	✓	✓		
CMS-SUS-16-032 [86]	2 b- or 2 c-jets	35.9	✓			
CMS-SUS-16-033 [87]	0ℓ + jets	35.9	✓	✓	✓	
CMS-SUS-16-034 [88]	2 OSSF leptons	35.9	✓			
CMS-SUS-16-035 [89]	2 SS leptons	35.9	✓			
CMS-SUS-16-036 [90]	0ℓ + jets	35.9	✓	✓		
CMS-SUS-16-037 [91]	1ℓ + jets with MJ	35.9	✓			
CMS-SUS-16-039 [92]	multi-ℓ, EWino	35.9	✓			
CMS-SUS-16-041 [93]	multi-ℓ + jets	35.9	✓			
CMS-SUS-16-042 [94]	1ℓ + jets	35.9	✓			
CMS-SUS-16-043 [95]	$WH(bb)$, EWino	35.9	✓			
CMS-SUS-16-045 [96]	2 b + 2 $H(\gamma\gamma)$	35.9	✓			
CMS-SUS-16-046 [97]	high-p_T γ	35.9	✓			
CMS-SUS-16-047 [98]	γ + jets, high H_T	35.9	✓			
CMS-SUS-16-049 [99]	0ℓ stop	35.9	✓	✓		
CMS-SUS-16-050 [100]	0ℓ + top tag	35.9	✓	✓		
CMS-SUS-16-051 [101]	1ℓ stop	35.9	✓	✓		
CMS-SUS-17-001 [102]	2ℓ stop	35.9	✓			
CMS-SUS-17-003 [103]	2 taus	35.9	✓			
CMS-SUS-17-004 [58]	EWino combination	35.9	✓			
CMS-SUS-17-005 [104]	1ℓ + jets, top tag	35.9	✓	✓		
CMS-SUS-17-006 [105]	jets + boosted $H(bb)$	35.9	✓	✓		
CMS-SUS-17-009 [106]	SFOS leptons	35.9	✓	✓		
CMS-SUS-17-010 [107]	2ℓ stop	35.9	✓	✓		
CMS-SUS-18-002 [108]	γ + (b-)jets, top tag	35.9	✓	✓		
CMS-SUS-19-006 [109]	0ℓ + jets, MHT	137.0	✓	✓		
CMS-SUS-19-009 [110]	1ℓ + jets, MHT	137.0	✓			
CMS-EXO-19-001 [111]	non-prompt jets	137.0			✓	
CMS-EXO-19-010 [10]	disappearing tracks	101.0			✓	

ID					
CMS-E					
CMS-E					
CMS-P					
CMS-P					
CMS-P					
CMS-P					
CMS-S					
CMS-S					
CMS-S					
CMS-S					
CMS-S					
CMS-S					
CMS-S					
CMS-S					
CMS-SUS-13-013	2 SS l's + (b-)jets + $\not\!\!E_T$	19.5	✓	✓	✓
CMS-SUS-13-019	>= 2 jets + $\not\!\!E_T$, MT2	19.5	✓		
CMS-SUS-14-010	b-jets + 4 Ws	19.5	✓	✓	
CMS-SUS-14-021	soft l's, low n$_{jets}$, high $\not\!\!E_T$	19.7	✓	✓	✓

ID	Short Description	$\mathcal{L}$ [fb^{-1}]	UL$_{\rm obs}$	UL$_{\rm exp}$	EM	comb.
ATLAS-SUSY-2015-01 [62]	2 b-jets	3.2	✓			
ATLAS-SUSY-2015-02 [63]	1ℓ stop	3.2	✓		✓	
ATLAS-SUSY-2015-06 [64]	0ℓ + 2–6 jets	3.2			✓	
ATLAS-SUSY-2015-09 [65]	jets + 2 SS or $\geq 3\ell$	3.2	✓			
ATLAS-SUSY-2016-06 [66]	disappearing tracks	36.1			✓	
ATLAS-SUSY-2016-07 [67]	0ℓ + jets	36.1	✓		✓	
ATLAS-SUSY-2016-08 [68]	displaced vertices	32.8	✓			
ATLAS-SUSY-2016-14 [69]	2 SS or 3 ℓ's + jets	36.1	✓			
ATLAS-SUSY-2016-15 [70]	0ℓ stop	36.1	✓			
ATLAS-SUSY-2016-16 [71]	1ℓ stop	36.1	✓		✓	
ATLAS-SUSY-2016-17 [72]	2 OS leptons	36.1	✓			
ATLAS-SUSY-2016-19 [73]	2 b-jets + τ's	36.1	✓			
ATLAS-SUSY-2016-24 [74]	2–3 ℓ's, EWino	36.1	✓		✓	
ATLAS-SUSY-2016-26 [75]	$\geq 2\,c$-jets	36.1	✓			
ATLAS-SUSY-2016-27 [76]	jets + γ	36.1	✓		✓	
ATLAS-SUSY-2016-28 [77]	2 b-jets	36.1	✓			
ATLAS-SUSY-2016-32 [44]	HSCP	31.6	✓	✓	✓	
ATLAS-SUSY-2016-33 [78]	2 OSSF ℓ's	36.1	✓			
ATLAS-SUSY-2017-01 [79]	$WH(bb)$, EWino	36.1	✓			
ATLAS-SUSY-2017-02 [80]	0ℓ + jets	36.1	✓	✓		
ATLAS-SUSY-2017-03 [21]	multi-ℓ EWino	36.1	✓		✓	
ATLAS-SUSY-2018-04 [81]	2 hadronic taus	139.0	✓		✓	JSON
ATLAS-SUSY-2018-06 [22]	3 leptons, EWino	139.0	✓	✓	✓	
ATLAS-SUSY-2018-10 [17]	1ℓ + jets	139.0	✓		✓	
ATLAS-SUSY-2018-12 [19]	0ℓ + jets	139.0	✓	✓	✓	
ATLAS-SUSY-2018-14 [15]	displaced leptons	139.0			✓	JSON
ATLAS-SUSY-2018-22 [18]	multi-jets	139.0	✓		✓	
ATLAS-SUSY-2018-23 [20]	$WH(\gamma\gamma)$, EWino	139.0	✓	✓		
ATLAS-SUSY-2018-31 [82]	$2b + 2H(bb)$	139.0	✓		✓	JSON
ATLAS-SUSY-2018-32 [59]	2 OS leptons	139.0	✓			
ATLAS-SUSY-2019-08 [60]	1ℓ + $H(bb)$, EWino	139.0	✓		✓	JSON

ID					xp	EM
ATLAS-SUSY-2013-21	monojet or c-jet + $\not\!\!E_T$	20.3				✓
ATLAS-SUSY-2013-23	1ℓ + 2 b-jets (or 2 γs) + $\not\!\!E_T$	20.3		✓		
ATLAS-SUSY-2014-03	>= 2(c-)jets + $\not\!\!E_T$	20.3				✓

https://smodels.github.io/docs/ListOfAnalyses210

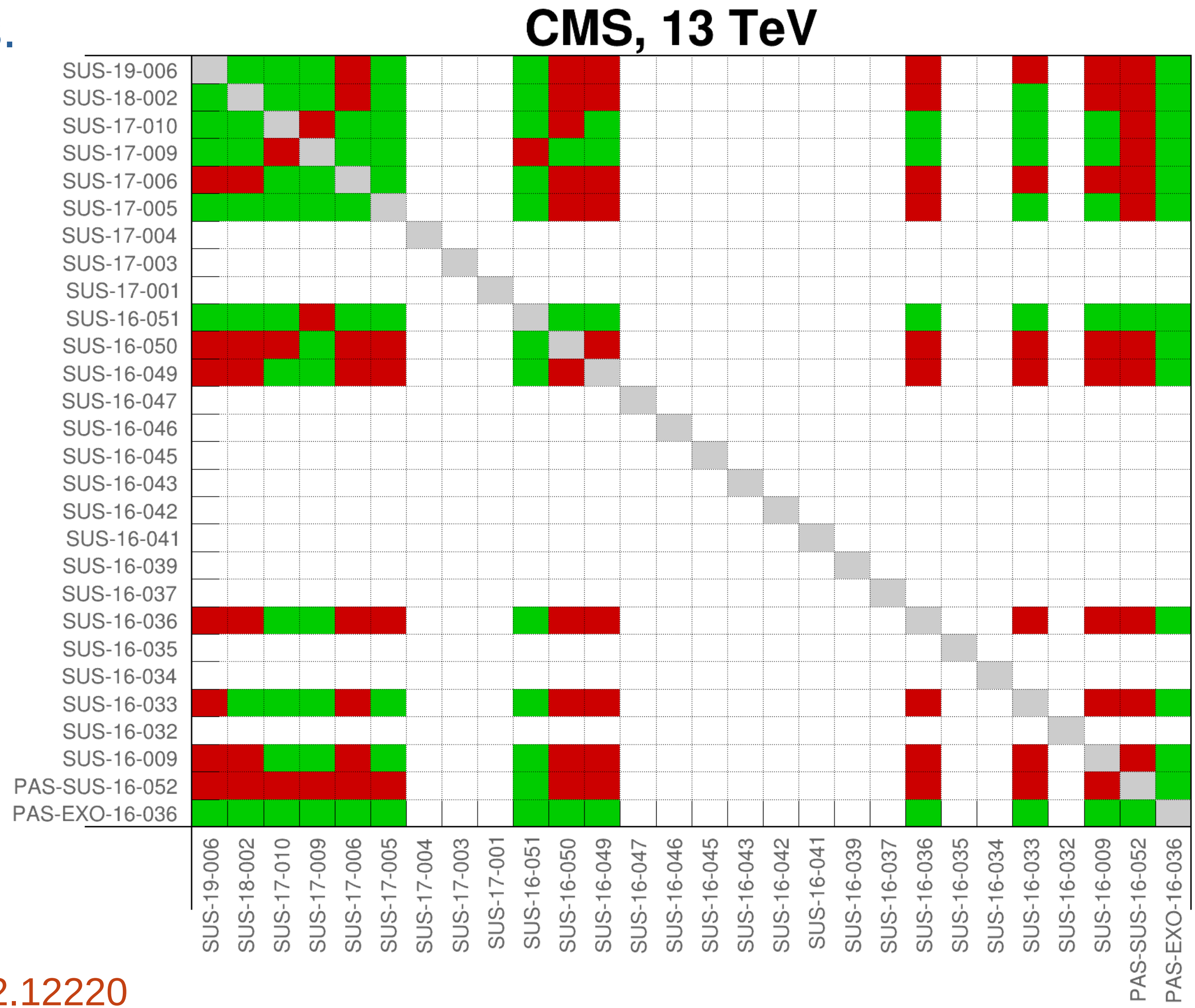
The Combiner

As we are chasing dispersed signals, we need to allow the machine to combine likelihoods.

green: approximately uncorrelated → combinable

red: correlated, not combinable

White: cannot construct a likelihood



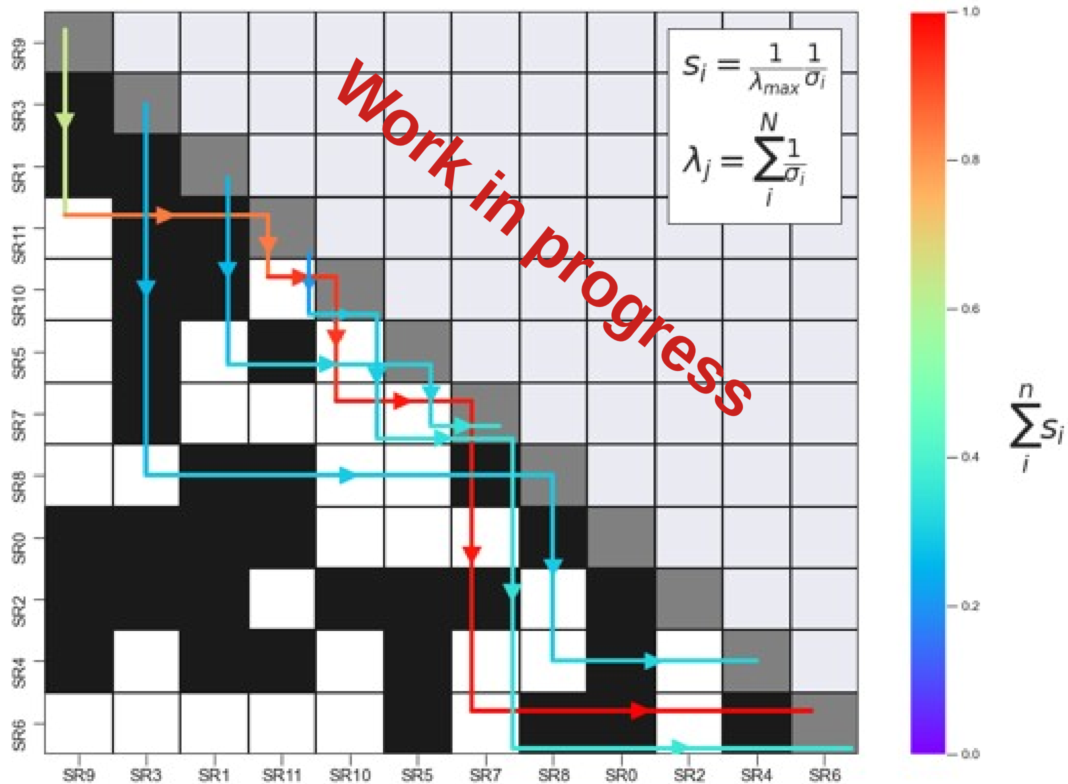
Les Houches effort:

https://arxiv.org/abs/2002.12220

Study of inter-analyses correlations

Ongoing effort: TACO collaboration

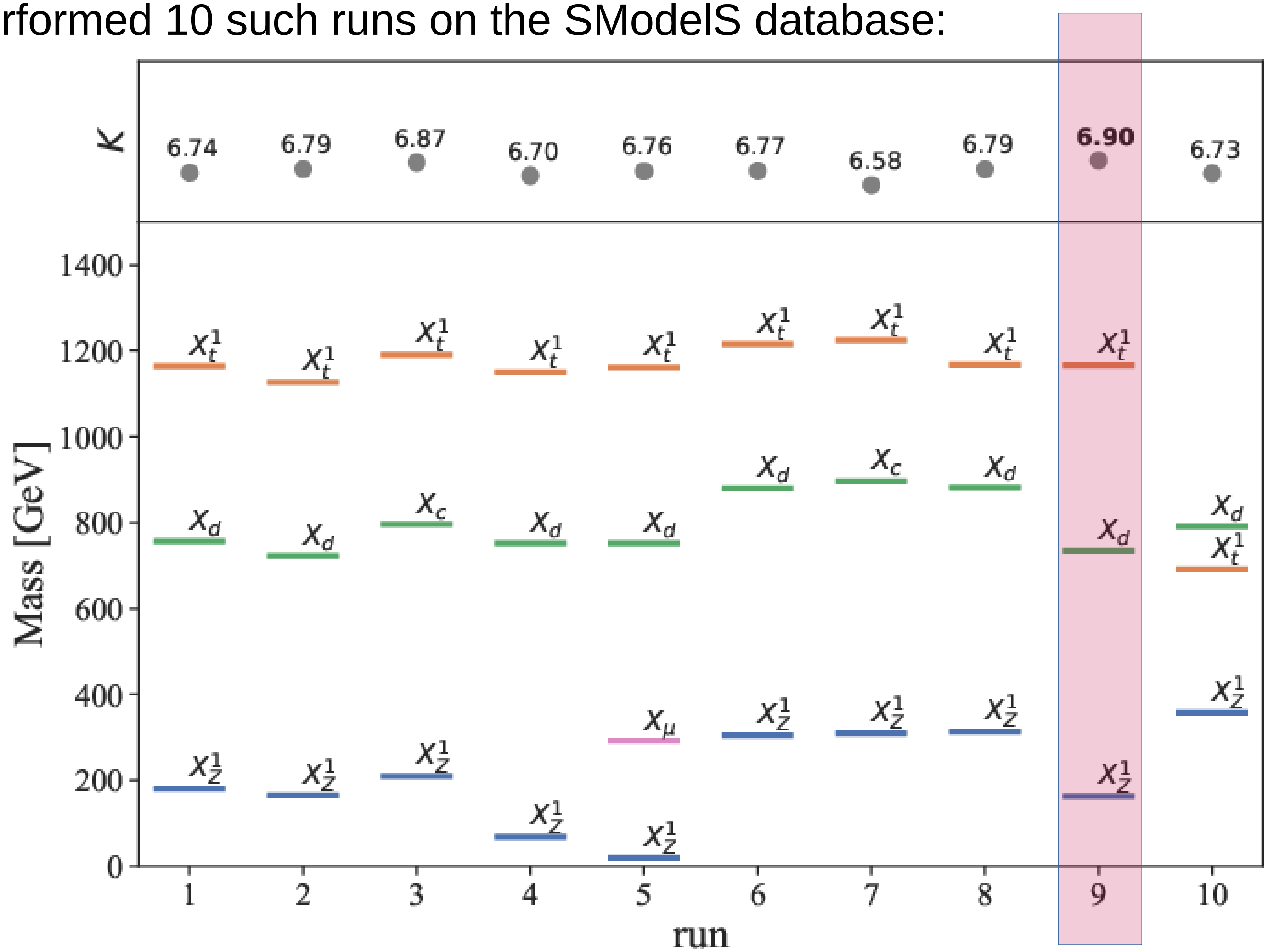
- Small collaboration: members of MA5, SModelS, rivet, gambit
- Builds on Les Houches effort (see previous slide)
- Aim: systematically study "overlaps" between signal regions, develop smart combination algorithm
- Best possible effort in case of simplified likelihoods
- Can go further with full likelihoods (and standardized naming conventions for nuisances)



Andy Buckley, Benjamin Fuks, Humberto Reyes-González, WW, Sophie Williamson, Jamie Yellen

ProtoModeller: Walking Over the SModelS Database

We defined a “run” as 50 parallel walkers, making 1,000 steps each.
We performed 10 such runs on the SModelS database:

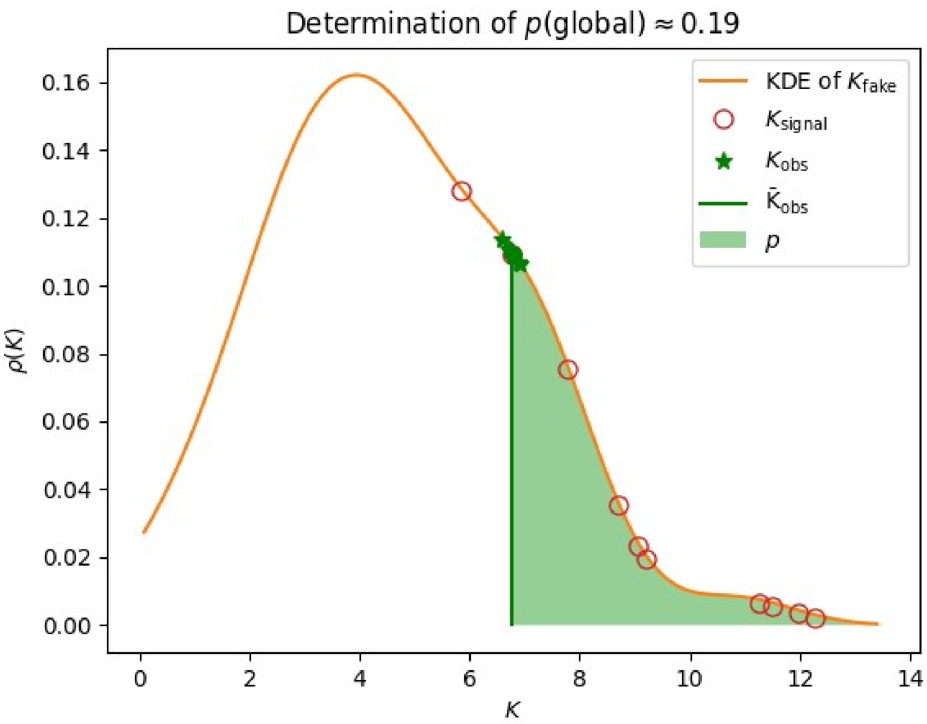


All 10 runs introduced a top partner as well as a light quark partner. The cross sections are compatible with values expected from the MSSM. The best test statistic was *K*=6.9.

Global p-Value

With our **simplified statistical models** of the search results, we can synthesize databases of results that are “typical”, if no new physics is in the data.

From this we can compute an approximate p-value for the Standard Model hypothesis: that is the chances that – under the SM hypothesis – we would obtain a results as extreme as ours or more extreme.



Long term goal: full statistical models, proper sampling also of shared systematics.

So .... what's our Verdict?

- After 1,000,000 CPU hours of running the procedure, it is still unclear! The result is not entirely as expected, under the Standard Model hypothesis.

- But there are actually tensions between some of the experimental results.

- Proof-of-concept has been given. Now rerun with more results, larger number of “theory parameters”, improved methodology, more accurate statistical models.

https://arxiv.org/abs/2012.12246

Backup

The Test Statistic

For every legal combination, we define a test statistic *K*

$$K^c := -2 \ln \frac{\mathrm{L}^c_{\mathbf{SM}} \cdot \pi(\mathbf{SM})}{\mathrm{L}^c_{\mathbf{BSM}}(\hat{\mu}) \cdot \pi(\mathbf{BSM})}$$ Eq. 6

π(BSM) is the prior of the BSM model. We use it to “regularize” the model, i.e. impose the *law of parsimony*:

$$\pi(M) = \exp\left[-\left(\frac{n_{\mathrm{particles}}}{a_1} + \frac{n_{\mathrm{BRs}}}{a_2} + \frac{n_{\mathrm{productionmodes}}}{a_3}\right)\right]$$ Eq. 9

That way, one new particle with one non-trivial branching ratio and two production modes is similar to one degree of freedom in Akaike’s information criterion (the sign is however flipped, and it’s a likelihood ratio), i.e. the test statistic is roughly equivalent to

$$K \approx \Delta\chi^2 - 2n_{\mathrm{particles}}$$

An additional particle will have to increase the “(delta-)chi-square” by approximately two units.

Likelihoods

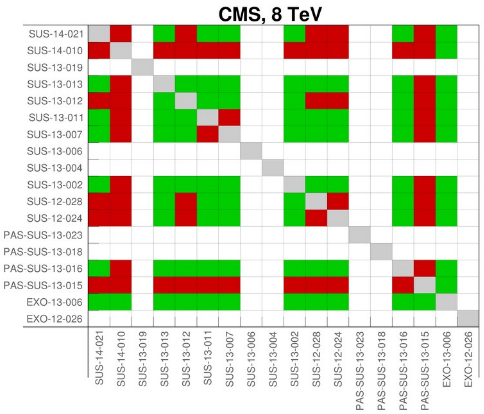
- **Only exclusion lines**
  If only exclusion lines are given, without upper limits, we can do nothing
- **Observed 95% CL upper limits only:**
  cannot construct likelihood, binary decision “excluded” / “not-excluded” only (“critic”)
- **Expected and observed 95% CL upper limits**
  can construct an approximate likelihood with truncated Gaussian,
  cannot combine topologies, very crude approximation
- **Efficiency maps**
  can construct a likelihood as Gaussian (for the nuisances) * Poissonian (for yields), can work per SR, and combine topologies in each SR [*]
- **Efficiency maps + correlation matrices**
  can combine signal regions via multivariate Gaussian * Poissonians
- **Efficiency maps + full likelihoods**
  full realism, correct statistical model

Combos

Likelihoods

BETTER

[*] if efficiency maps are not supplied, we can try to produce them with recasting frameworks



The Combiner

we allow the machine
to combine likelihooods.

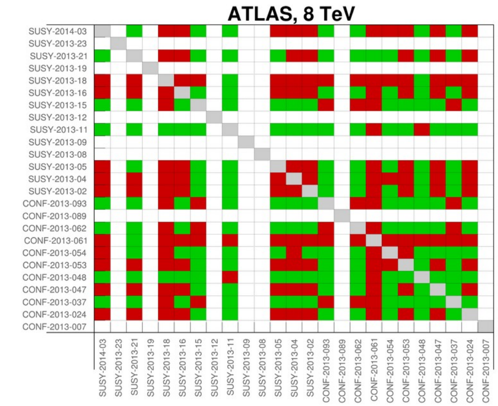
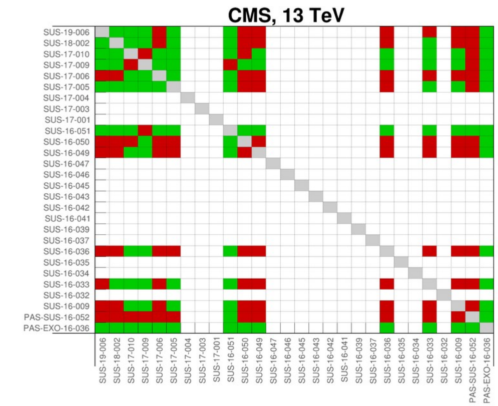


Fig. 2

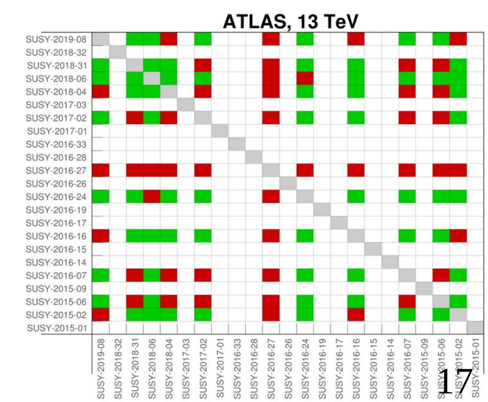
Approximately uncorrelated are analyses that are:

- from different runs, and/or
- from different experiments, and/or
- looking for (clearly) different signatures



A combination “c” of analyses is “legal” if the following conditions are met:

- all results are mutually uncorrelated (= ”combinable”)
- if a result can be added, it has to be added (any subset of a legal combination is not itself legal)
- combined likelihood: $L_c = \prod_{i \in c} L_i$



The Test Statistic

For every legal combination, we define a test statistic *K*

$$K^c := -2\ln\frac{\mathrm{L}^c_{\mathbf{SM}}\cdot\pi(\mathbf{SM})}{\mathrm{L}^c_{\mathbf{BSM}}(\hat{\mu})\cdot\pi(\mathbf{BSM})}$$ Eq. 6

(Remember, we have a database of results from ~ 100 CMS+ATLAS searches. We want to find the most interesting combinations of these results, i.e. the ones that maximally violate the SM hypothesis)

Of all “legal” combinations of experimental results, the builder chooses the one combination “c” that maximizes *K*:

$$K := \max_{\forall c\in C} K^c$$ Eq. 7

μ denotes an global signal strength multiplier – the production cross sections are free parameters

$$\forall i,j : \sigma\left(pp\to X_iX_j\right) = \mu\bar{\sigma}\left(pp\to X_iX_j\right)$$

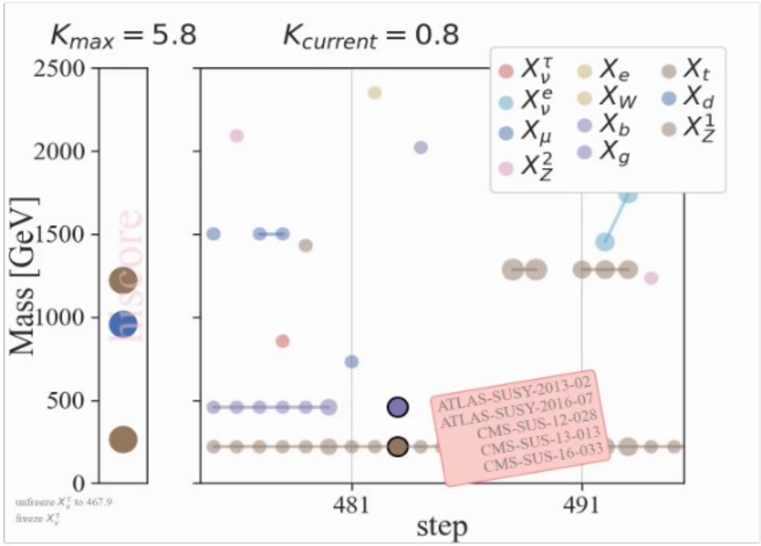
It is maximized in the denominator, but its support is confined such that no limits in the SModelS database are violated (the “critic”),

$$\hat{\mu}\in[0,\mu_{\mathrm{max}}]$$

The Walker

The Walker takes care of moving in the protomodel space with varying dimensionality by performing the following types of modifications to the protomodel:

- add or remove particles from the protomodel
- change the masses of particles
- change the signal strengths of production modes
- change decay channels and branching ratios



At each step the test statistic *K* is computed. An MCMC-like procedure[*] is then applied in the sense that the step is reverted with a probability of

$$\exp\left[\frac{1}{2}(K_i - K_{i-1})\right]$$

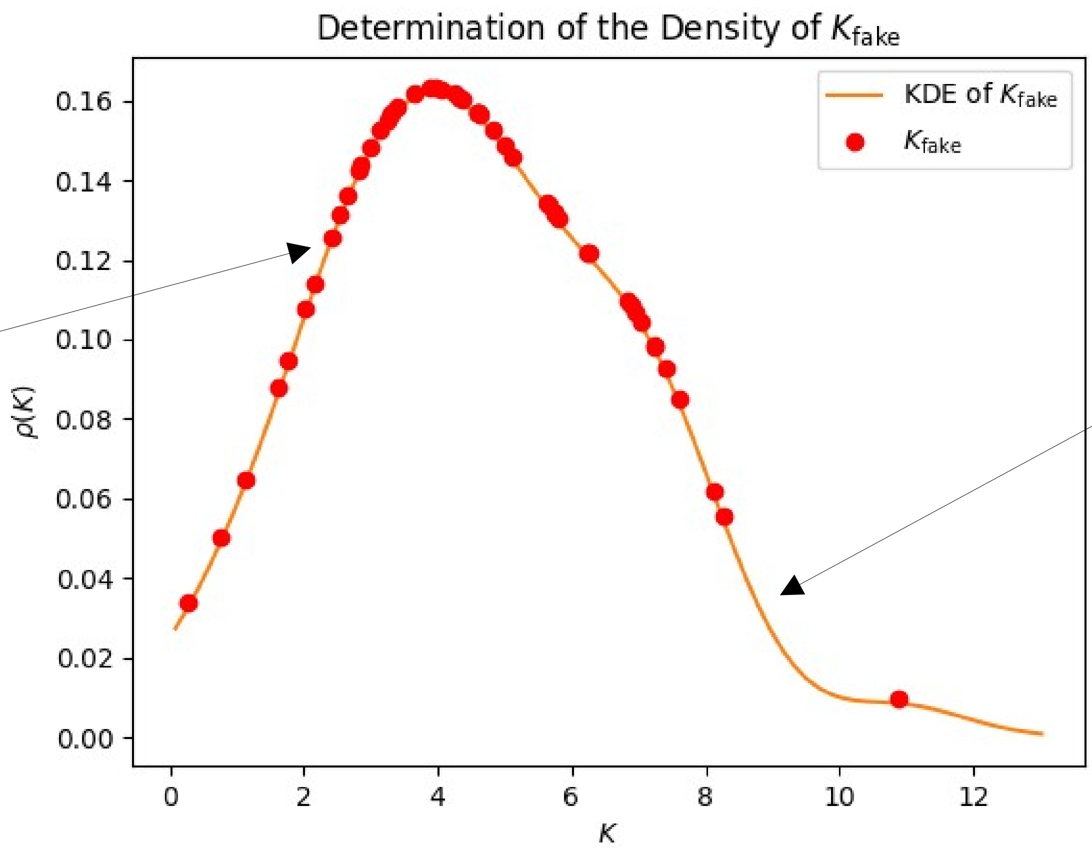
if and only if K_i is smaller than K_{i-1}

* (note however, instead of ratios of unnormalized posteriors we have ratios of ratios of unnormalized posteriors)

Walking Over Fake Standard Model Databases

- Produced 50 “fake” SModelS databases by sampling background models
- Corresponds to typical LHC results if no new physics is in data
- Determine 50 “fake” K values by running 50 walkers on each of the 50 databases (50 x 50 walkers in total) → density of K under null SM-only hypothesis

K for one “fake” background-only database.



Density of K estimated via a simple Kernel density estimator.

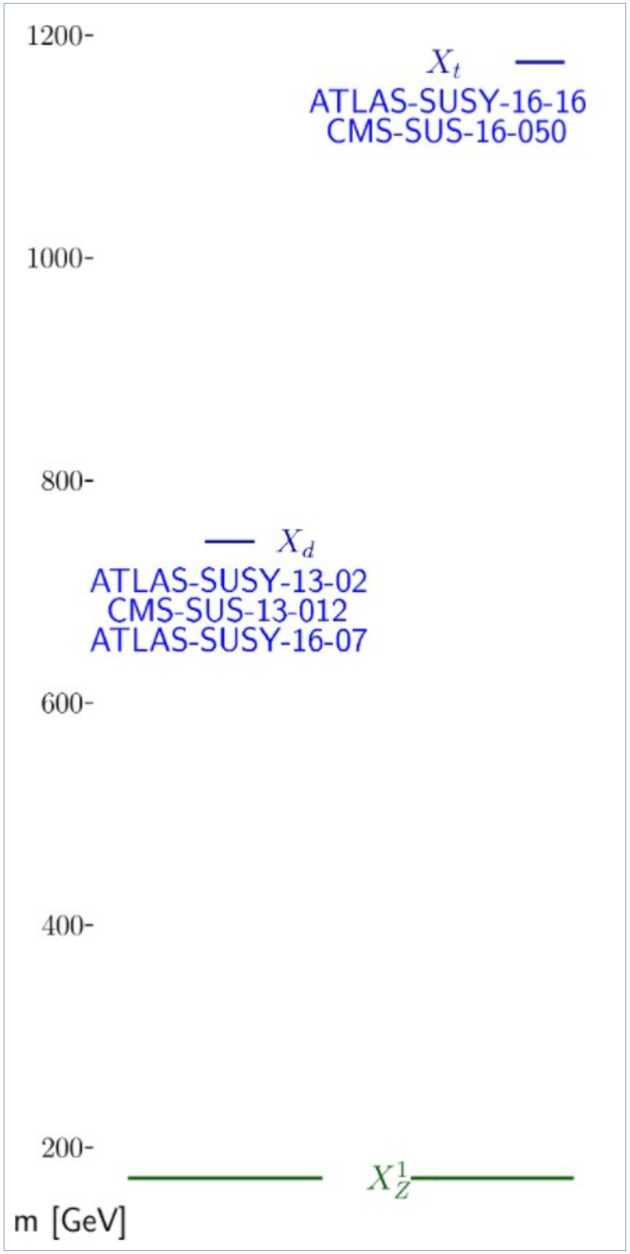
The Walks

We define a “run” as 50 parallel walks, each taking 1000 steps.

We performed

- 10 runs on the SModelS database (Sec. 5.2)
- 50 runs on fake “Standard Model-like” databases (Sec 5.1)
  to be able to determine a global *p*-value under the SM hypothesis
- 2x10 runs on fake “Signal-like” databases (Sec 5.3)
  to show closure of the method

The Hiscore Proto-Model



Analysis	Dataset	Obs	Exp	Z	P	Signal
ATL multijet, 8 TeV [54]	SR6jtp	6	4.9 ± 1.6	$0.4\ \sigma$	X_d	0.25
ATL multijet, 13 TeV [55]	2j_Me ...	611	526 ± 31	$2.2\ \sigma$	X_d	44.18
ATL 1ℓ stop, 13 TeV [48]	tN_high	8	3.8 ± 1	$1.9\ \sigma$	X_t	3.93
CMS multijet, 8 TeV [56]		30.8 fb	19.6 fb	$1.1\ \sigma$	X_d	2.66 fb
CMS 0ℓ stop, 13 TeV [49]		4.5 fb	2.5 fb	$1.6\ \sigma$	X_t	2.62 fb

Table 3: Analyses contributing to the K value of the highest score proto-model

Tension!

the dispersed excess

Analysis (all CMS 13 TeV)	Prod	σ_{XX} (fb)	$\sigma^{\mathrm{UL}}_{\mathrm{obs}}$ (fb)	$\sigma^{\mathrm{UL}}_{\mathrm{exp}}$ (fb)	r_{obs}
CMS multijet, M_{H_T}, 137 fb^{-1} [15]	$(\bar{X}_d, X_d)$	23.96	18.45	21.57	1.30
CMS multijet, M_{H_T}, 137 fb^{-1} [15]	$(\bar{X}_t, X_t)$	2.62	2.04	2.08	1.28
CMS multijet, M_{H_T}, 36 fb^{-1} [57]	$(\bar{X}_d, X_d)$	23.96	19.26	28.31	1.24
CMS multijet, M_{T2}, 36 fb^{-1} [58]	$(\bar{X}_d, X_d)$	23.96	26.02	31.79	0.92
CMS 1ℓ stop, 36 fb^{-1} [59]	$(\bar{X}_t, X_t)$	2.62	2.91	4.44	0.90

Table 4: List of the most constraining results for the highest score proto-model. The

what is driving the "critic"

Signal strength multipliers: $(\bar{X}_t, X_t) = 1.2; (\bar{X}_d, X_d), (X_d, X_Z^1), (\bar{X}_d, X_Z^1) = 0.49$

Contributions by particles: $X_t : K_{\mathrm{without}} = 2.59 (59\%), X_d : K_{\mathrm{without}} = 3.90 (41\%)$
Last updated: Mon Dec 14 20:08:06 2020

https://smodels.github.io/protomodels/2020_PioneerStudy/real9/index.html

Walking Over Databases With Fake Signals

To show closure of our method, we inject the winning protomodel as a signal in fake databases, and see if the algorithm can reconstruct the injected signal.

Sec 5.3

Technical closure test

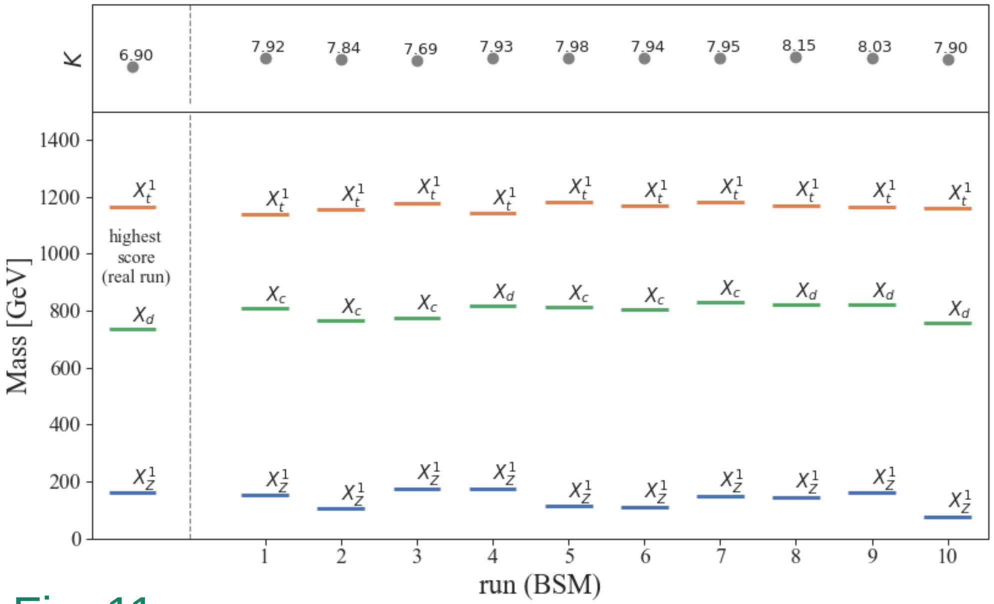


Fig. 11

No sampling of the models for the SRs, i.e. observed events := expected SM + expected signal events

Physics closure test

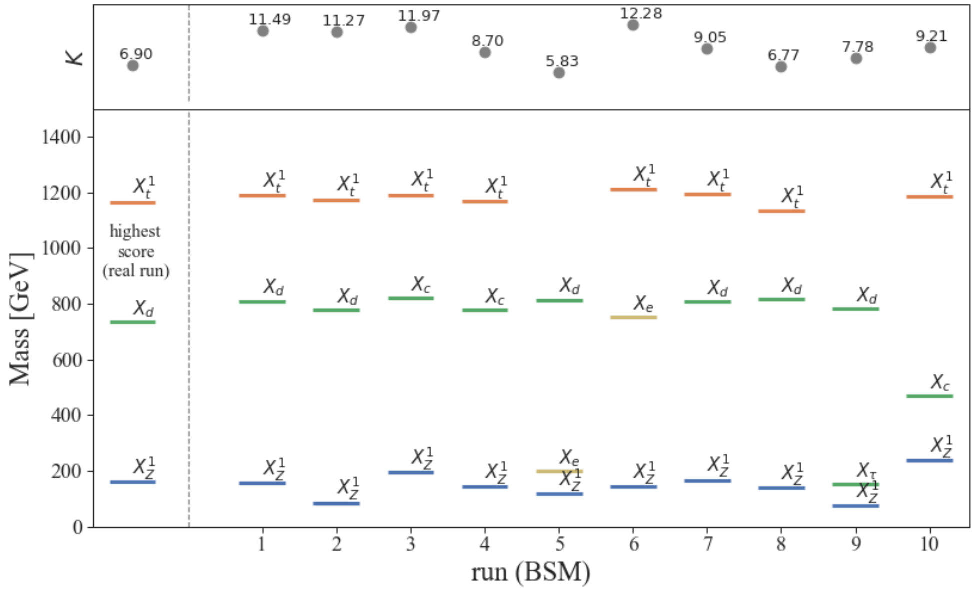


Fig. 10

Sampling turned on

Future Developments

Future Improvements

Improvements of the SModelS database:

- add latest full run-2 CMS and ATLAS publications (Moriond!)
- produce efficiency maps for existing results
- enlarge mass range of older efficiency maps

Improvements in speed:

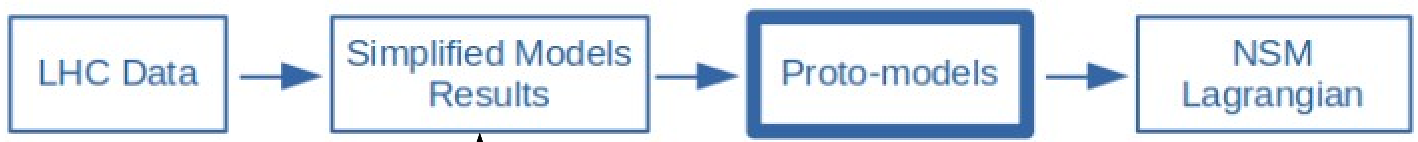
- learn the SModelS database
- make everything differentiable

Improvements in procedure:

- improve the “analyses correlation matrix”, automate the determination
- ponder relationship between proto-models and effective field theories
- connect proto-models with complete theories

Why Differentiable?

If we had gradients we could perform gradient descent to find the best model, and we could use e.g. the Fisher information to infer the error on its parameters (or, alternatively we can then MCMC-sample).



described as likelihoods L that are differentiable with respect to the yields y_i

we have started an effort to make SModelS differentiable w.r.t SMS parameters p_j, by learning our entire database:

that's just a sum of simplified models → differentiable!

for individual candidates we can make this differentiable w.r.t fundamental parameters Θ_l, via neural networks, with efforts similar to DeepXS, or "TheoryGANs" [*]:

$$\frac{\partial L}{\partial \theta_l} = \frac{\partial L}{\partial y_i} \cdot \frac{\partial y_i}{\partial p_j} \cdot \frac{\partial p_j}{\partial (m_k, \Gamma_k, \sigma_k)} \cdot \frac{\partial (m_k, \Gamma_k, \sigma_k)}{\partial \theta_l}$$

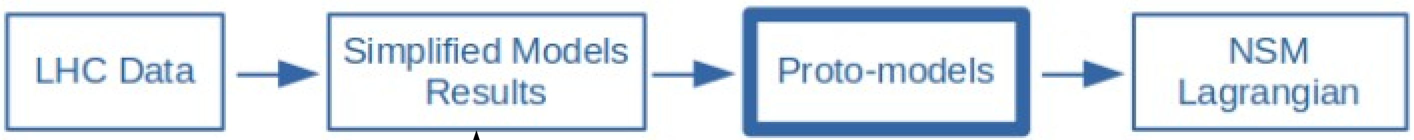
Needless to say, the data pipeline sketched above is not the only feasible one. Differentiability however would be a helpful tool for all possible data pipelines. A similar rationale would apply also to EFTs, Wilson coefficients and data from measurements.

https://arxiv.org/abs/1810.08312

→ Differentiable Inductive Reasoning!

Why Differentiable?

If we had gradients we could perform gradient descent to find the best model, and we could use e.g. the Fisher information to infer the error on its parameters (or, alternatively we can then MCMC-sample).



described as likelihoods L that are differentiable with respect to the yields y_i

we have started an effort to make SModelS differentiable w.r.t SMS parameters p_j, by learning our entire database:

that's just a sum of simplified models → differentiable!

for individual candidates we can make this differentiable w.r.t fundamental parameters Θ_l, via neural networks, with efforts similar to DeepXS, or "TheoryGANs" [*]:

$$\frac{\partial L}{\partial \theta_l} = \frac{\partial L}{\partial y_i} \cdot \frac{\partial y_i}{\partial p_j} \cdot \frac{\partial p_j}{\partial (m_k, \Gamma_k, \sigma_k)} \cdot \frac{\partial (m_k, \Gamma_k, \sigma_k)}{\partial \theta_l}$$

Needless to say, the data pipeline sketched above is not the only feasible one. Differentiability however would be a helpful tool for all possible data pipelines. A similar rationale would apply also to EFTs, Wilson coefficients and data from measurements.

https://arxiv.org/abs/1810.08312

→ Differentiable Inductive Reasoning!