Artificial Proto-Modelling:
Building Precursors of a Next Standard Model from Simplified Model Results

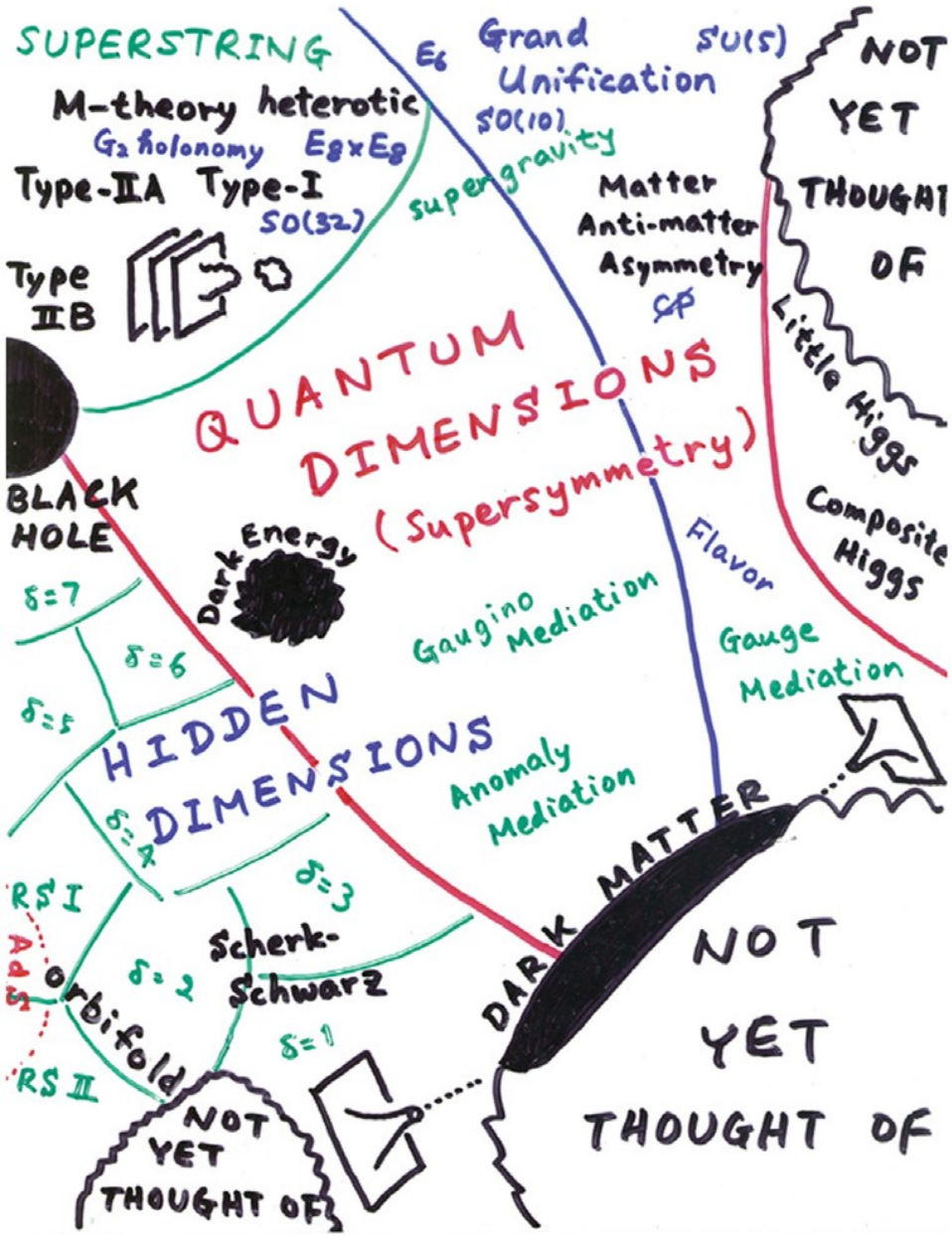
https://arxiv.org/abs/2012.12246

image courtesy of Jon Butterworth, Chris Wormell

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

ACAT 2021 @ virtual (South Korea),
Nov/Dec 2021

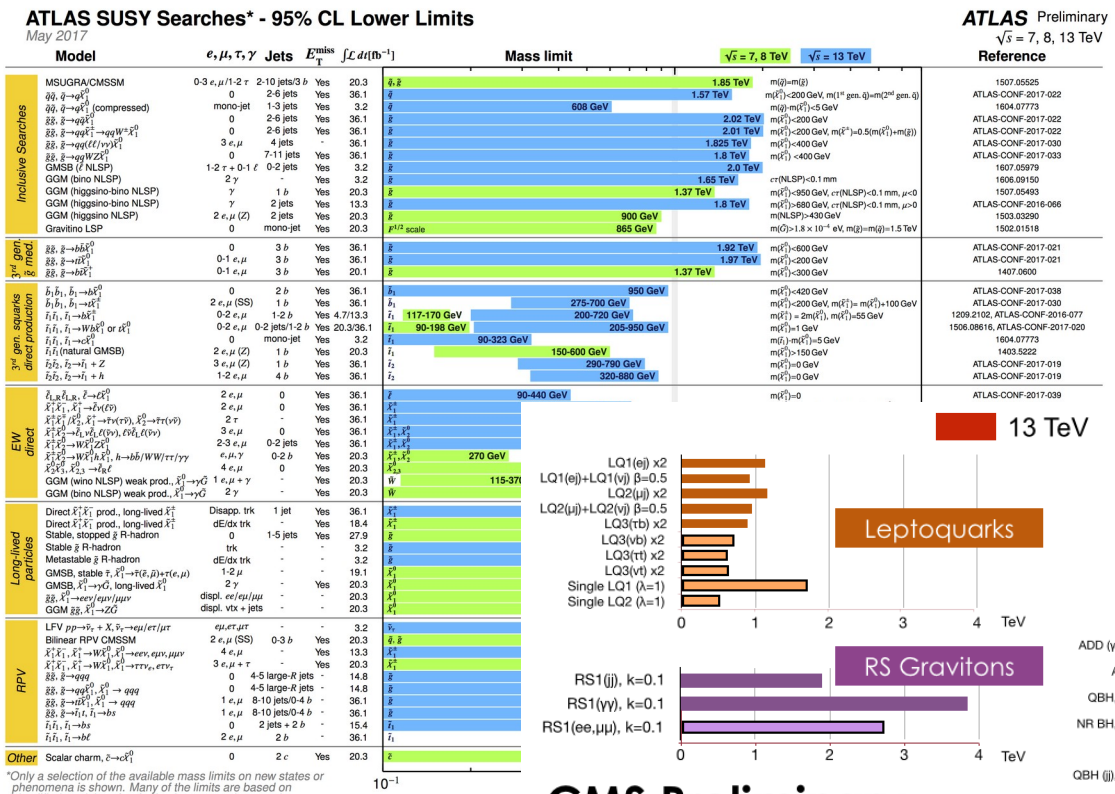
Problem Statement

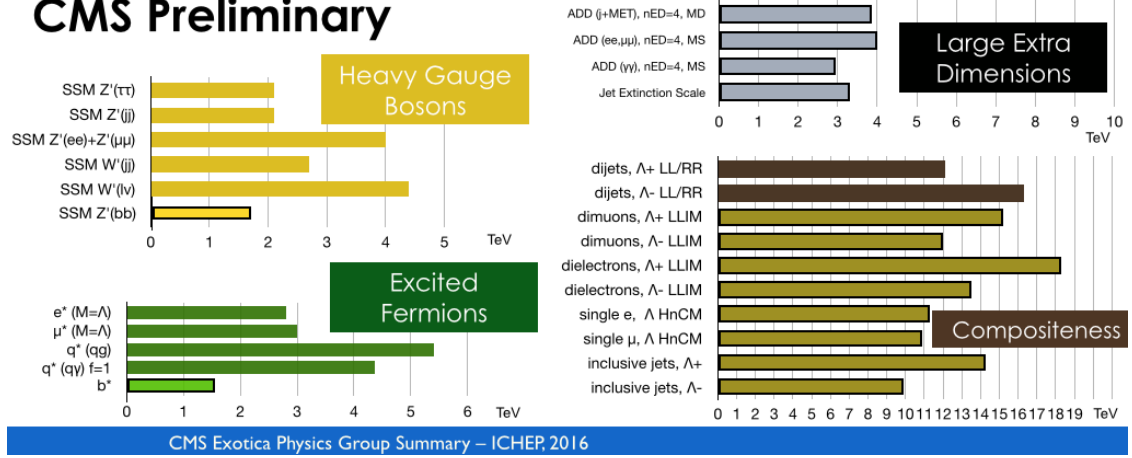


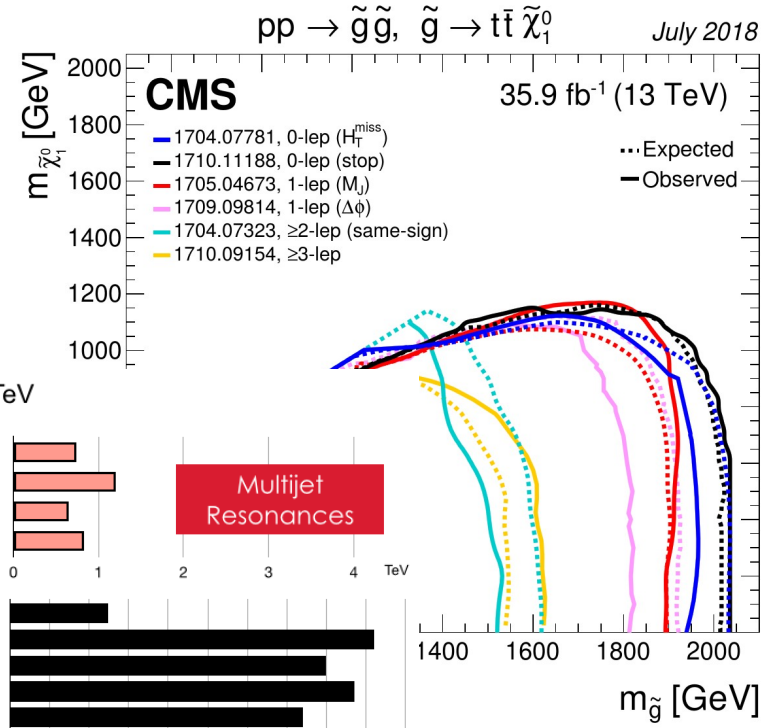
- Construction of a Next Standard Model (NSM) is arguably one of the big challenges of our time
- Many candidates theories exist
- Most of them have many free parameters. Example: n(MSSM) > 100

Hitoshi Murayama's "landscape of theories" – the number of "great ideas" is vast!

Problem Statement

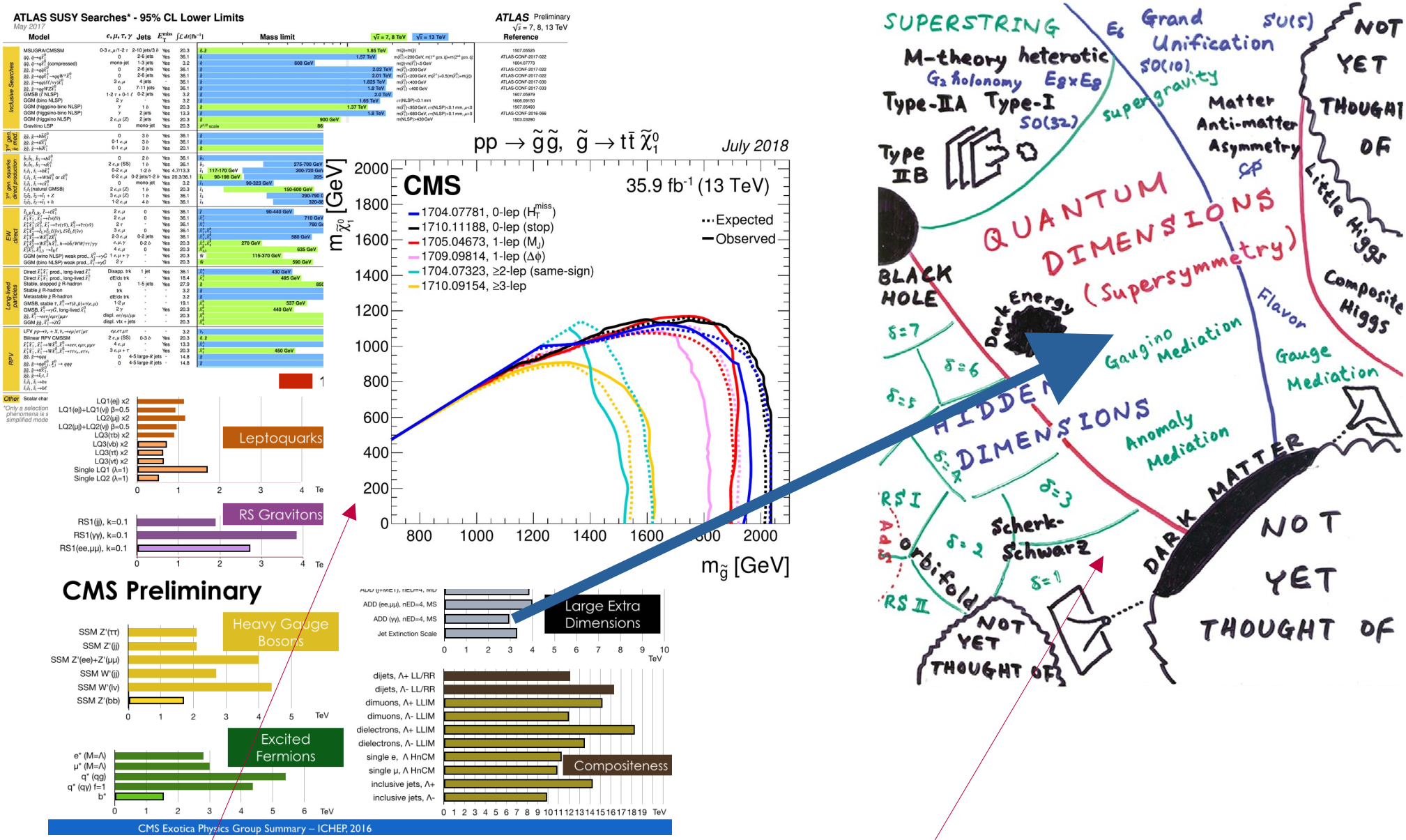






Not a handful of experimental signatures. Hundreds of publications with a wide range of signatures!

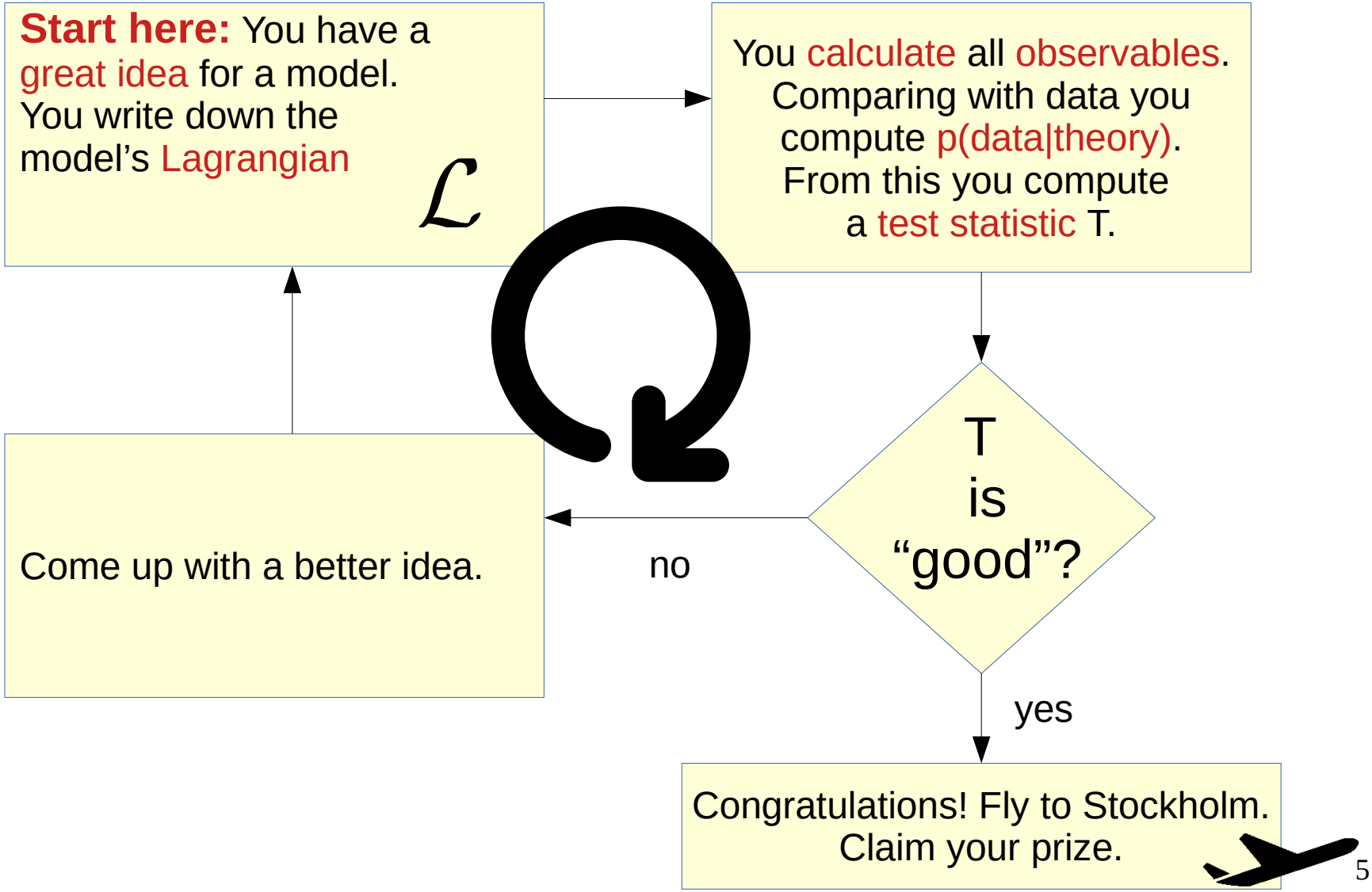
The Inverse Problem



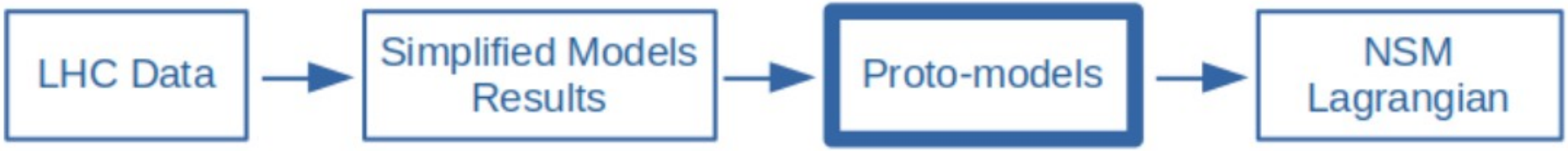
From these how do we construct which of these is the NSM (if any)?

The default approach: Top-down

Top-Down:



Our Approach: Bottom-Up

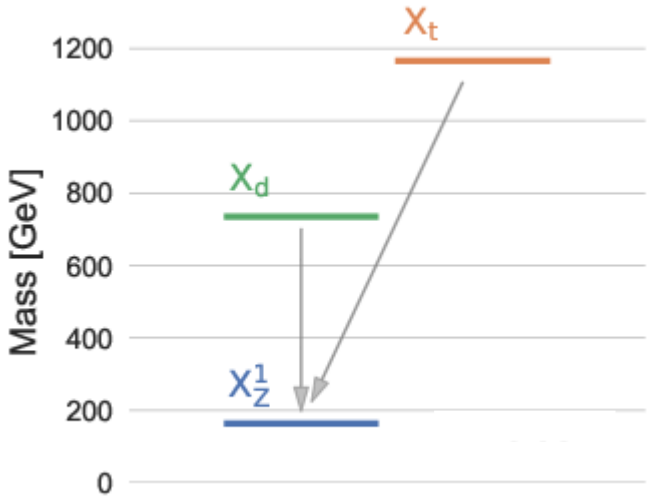


Instead of testing BSM scenarios one-by-one against the experimental data:

- identify potential dispersed signals in the slew of published LHC analyses
- build candidate "proto-models" (consistent sets of simplified models) from them.

MCMC-like random walk
through "proto-model" space of:

- particle content
- masses
- signal strengths [!]
- branching ratios

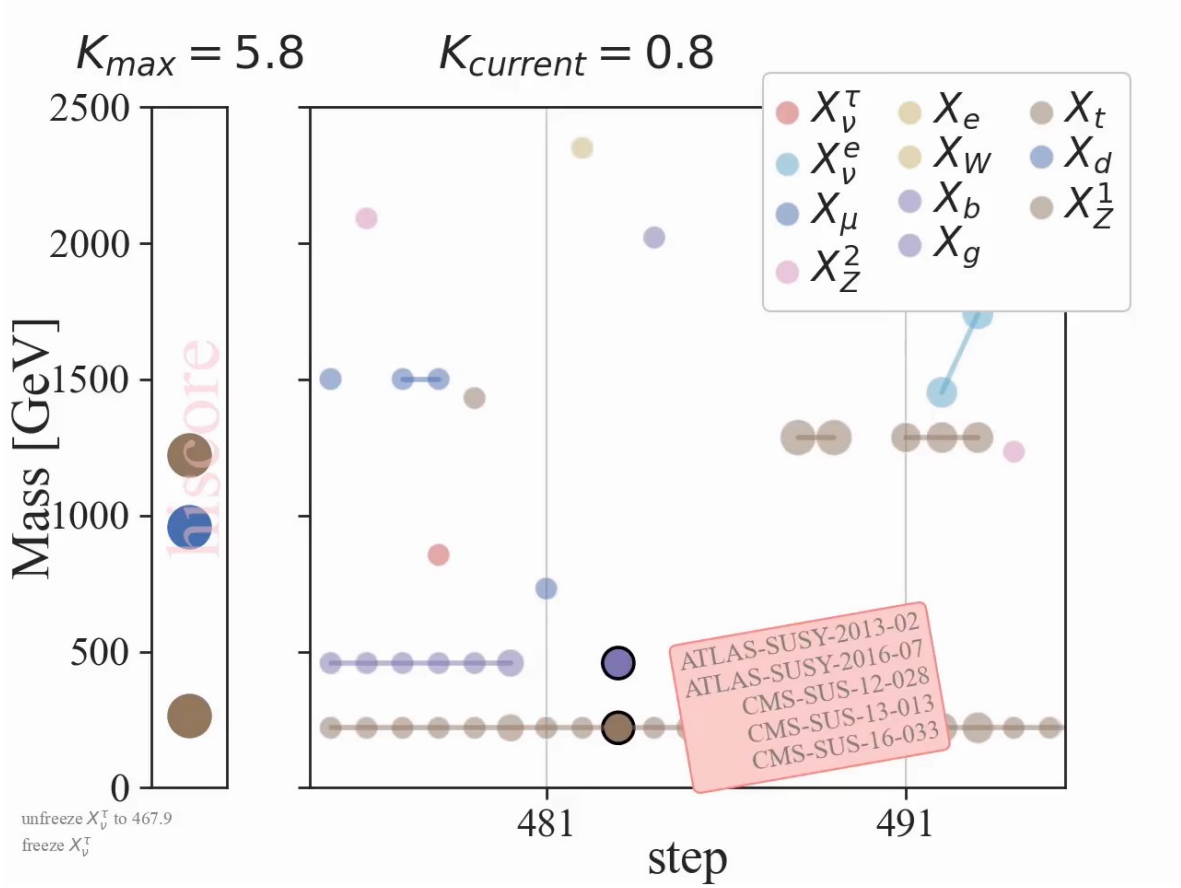


NSM: "Next Standard Model", BSM: "Beyond the Standard Model"

Our Approach

a test statistic

Particle spectra



A hiscore protomodel

Random modifications

potential dispersed signals

an MCMC-like random 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" test statistic

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.
- Assume simplified statistical models "behind" the data → simplified likelihoods

#	ID	Short Description	Type	$\mathcal{L}$ [fb^{-1}]
1	CMS-PAS-EXO-16-036	hscp search	ul, eff	12.9
2	CMS-PAS-SUS-16-052	soft l, <= 2 jets	ul, eff	35.9
3	CMS-SUS-16-009	multijets + $\not{E}_T$, top tagging	ul	2.3
4	CMS-SUS-16-032	Sbottom and compressed stop	ul	35.9
5	CMS-SUS-16-033	0ℓ + jets + $\not{E}_T$	ul, eff	35.9
6	CMS-SUS-16-034	2 OSSF l's	ul	35.9
7	CMS-SUS-16-035	2 SS l's	ul	35.9
8	CMS-SUS-16-036	0ℓ + jets + $\not{E}_T$	ul	35.9
9	CMS-SUS-16-037	1ℓ + jets + $\not{E}_T$ with MJ	ul	35.9
10	CMS-SUS-16-039	multi-l EWK searches	ul	35.9
11	CMS-SUS-16-041	multi-ls + jets + $\not{E}_T$	ul	35.9
12	CMS-SUS-16-042	1ℓ + jets + $\not{E}_T$	ul	35.9
13	CMS-SUS-16-043	EWK WH	ul	35.9
14	CMS-SUS-16-045	Sbottom to bHbH and H $\to \gamma\gamma$	ul	35.9
15	CMS-SUS-16-046	$\gamma + \not{E}_T$	ul	35.9
16	CMS-SUS-16-047	γ + HT	ul	35.9
17	CMS-SUS-16-049	All hadronic stop	ul	35.9
18	CMS-SUS-16-050	0ℓ + top tag	ul	35.9
19	CMS-SUS-16-051	1ℓ stop	ul	35.9
20	CMS-SUS-17-001	Stop search in dil + jets + $\not{E}_T$	ul	35.9
21	CMS-SUS-17-003	2 taus + $\not{E}_T$	ul	35.9
22	CMS-SUS-17-004	EW-ino combination	ul	35.9
23	CMS-SUS-17-005	1ℓ + multijets + $\not{E}_T$, top tagging	ul	35.9
24	CMS-SUS-17-006	jets + boosted H(bb) + $\not{E}_T$	ul	35.9
25	CMS-SUS-17-009	SFOS l's + $\not{E}_T$	ul	35.9
26	CMS-SUS-17-010	2L stop	ul	35.9
27	CMS-SUS-18-002	γ, jets, b-jets+ $\not{E}_T$, top tagging	ul	35.9
28	CMS-SUS-19-006	0ℓ + jets, MHT	ul	137.0
17	CMS-SUS-14-010	b-jets + 4 Ws	ul	19.5
18	CMS-SUS-14-021	soft l's, low n_{jets}, high $\not{E}_T$	ul	19.7

#	ID	Short Description	Type	$\mathcal{L}$ [fb^{-1}]
1	ATLAS-SUSY-2015-01	2 b-jets + $\not{E}_T$	ul	3.2
2	ATLAS-SUSY-2015-02	single l stop	ul, eff	3.2
3	ATLAS-SUSY-2015-06	0 l's + 2-6 jets + $\not{E}_T$	eff	3.2
4	ATLAS-SUSY-2015-09	jets + 2 SS l's or >=3 l's	ul	3.2
5	ATLAS-SUSY-2016-07	0ℓ + jets + $\not{E}_T$	ul, eff	36.1
6	ATLAS-SUSY-2016-14	2 SS or 3 l's + jets + $\not{E}_T$	ul	36.1
7	ATLAS-SUSY-2016-15	0ℓ stop	ul	36.1
8	ATLAS-SUSY-2016-16	1ℓ stop	ul, eff	36.1
9	ATLAS-SUSY-2016-17	2 opposite sign l's + $\not{E}_T$	ul	36.1
10	ATLAS-SUSY-2016-19	stops to staus	ul	36.1
11	ATLAS-SUSY-2016-24	2-3 l's + $\not{E}_T$, EWino	ul, eff	36.1
12	ATLAS-SUSY-2016-26	>=2 c jets + $\not{E}_T$	ul	36.1
13	ATLAS-SUSY-2016-27	jets + $\gamma + \not{E}_T$	ul, eff	36.1
14	ATLAS-SUSY-2016-28	2 b-jets + $\not{E}_T$	ul	36.1
15	ATLAS-SUSY-2016-33	2 OSSF l's + $\not{E}_T$	ul	36.1
16	ATLAS-SUSY-2017-01	EWK WH(bb) + $\not{E}_T$	ul	36.1
17	ATLAS-SUSY-2017-02	0ℓ + jets + $\not{E}_T$	ul	36.1
18	ATLAS-SUSY-2017-03	multi-l EWK searches	ul	36.1
19	ATLAS-SUSY-2018-04	2 hadronic taus	ul, eff	139.0
20	ATLAS-SUSY-2018-06	3 l's EW-ino	ul	139.0
21	ATLAS-SUSY-2018-31	2b + 2H(bb) + $\not{E}_T$	ul, eff	139.0
22	ATLAS-SUSY-2018-32	2 OS l's + $\not{E}_T$	ul	139.0
23	ATLAS-SUSY-2019-08	1ℓ + higgs + $\not{E}_T$	ul, eff	139.0
14	ATLAS-SUSY-2013-19	2 OS l's + (b-)jets + $\not{E}_T$	ul	20.3
15	ATLAS-SUSY-2013-21	monojet or c-jet + $\not{E}_T$	eff	20.3
16	ATLAS-SUSY-2013-23	1ℓ + 2 b-jets (or 2 γs) + $\not{E}_T$	ul	20.3
17	ATLAS-SUSY-2014-03	>= 2(c-)jets + $\not{E}_T$	eff	20.3

https://smodels.github.io/docs/ListOfAnalyses124

The Combiner

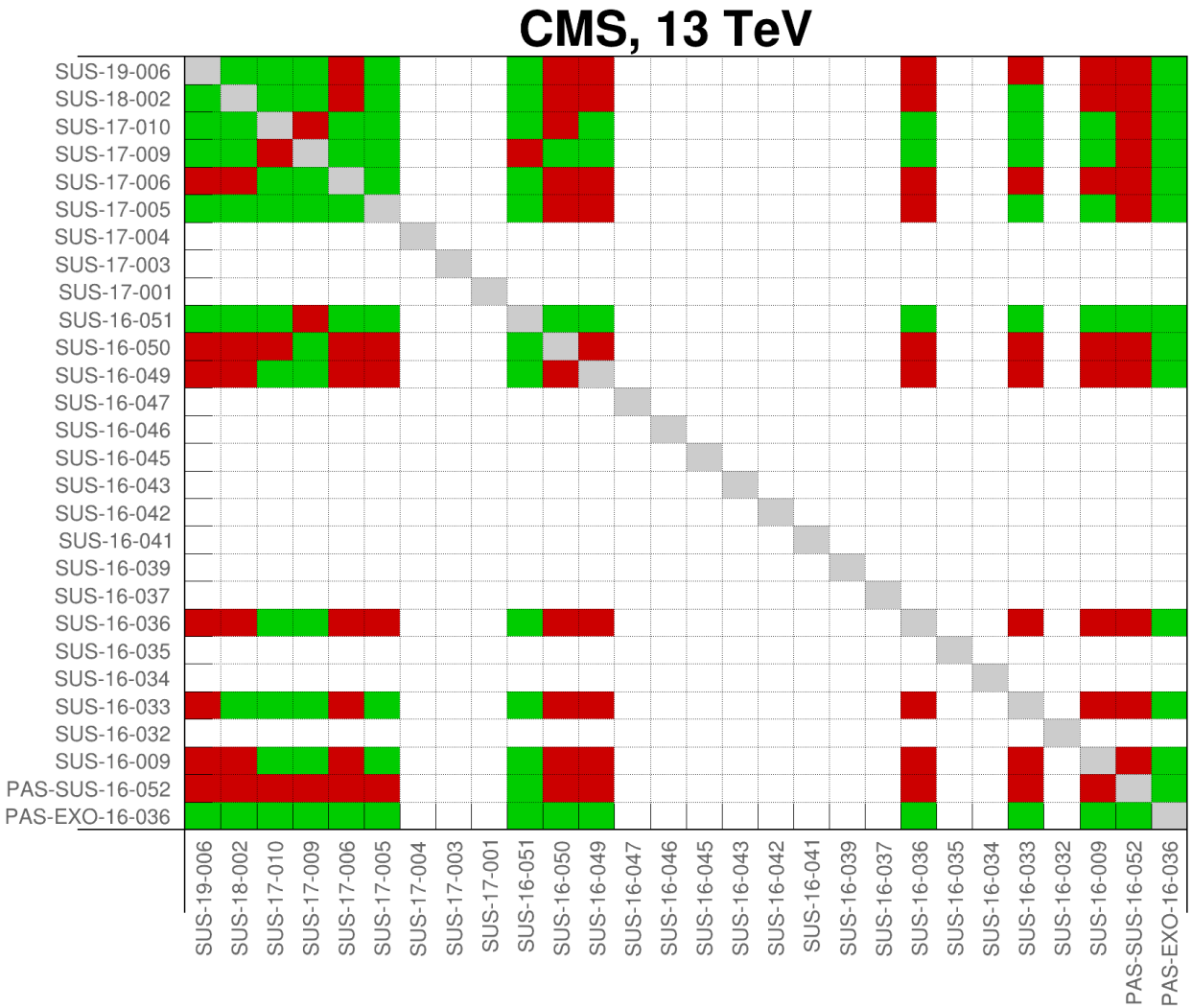
As we are chasing dispersed signals, we need to allow the machine to combine likelihoods. Simplified, binaric “inter-analyses correlations matrix”:

green: approximately uncorrelated → combinable

red: correlated, not combinable

White: cannot construct a likelihood

Signal regions within each analysis: correlated



Les Houches effort:

https://arxiv.org/abs/2002.12220

Current version: “educated guesses” from description of signatures in signal regions. Ongoing effort to determine this matrix automatically with recasting tools.

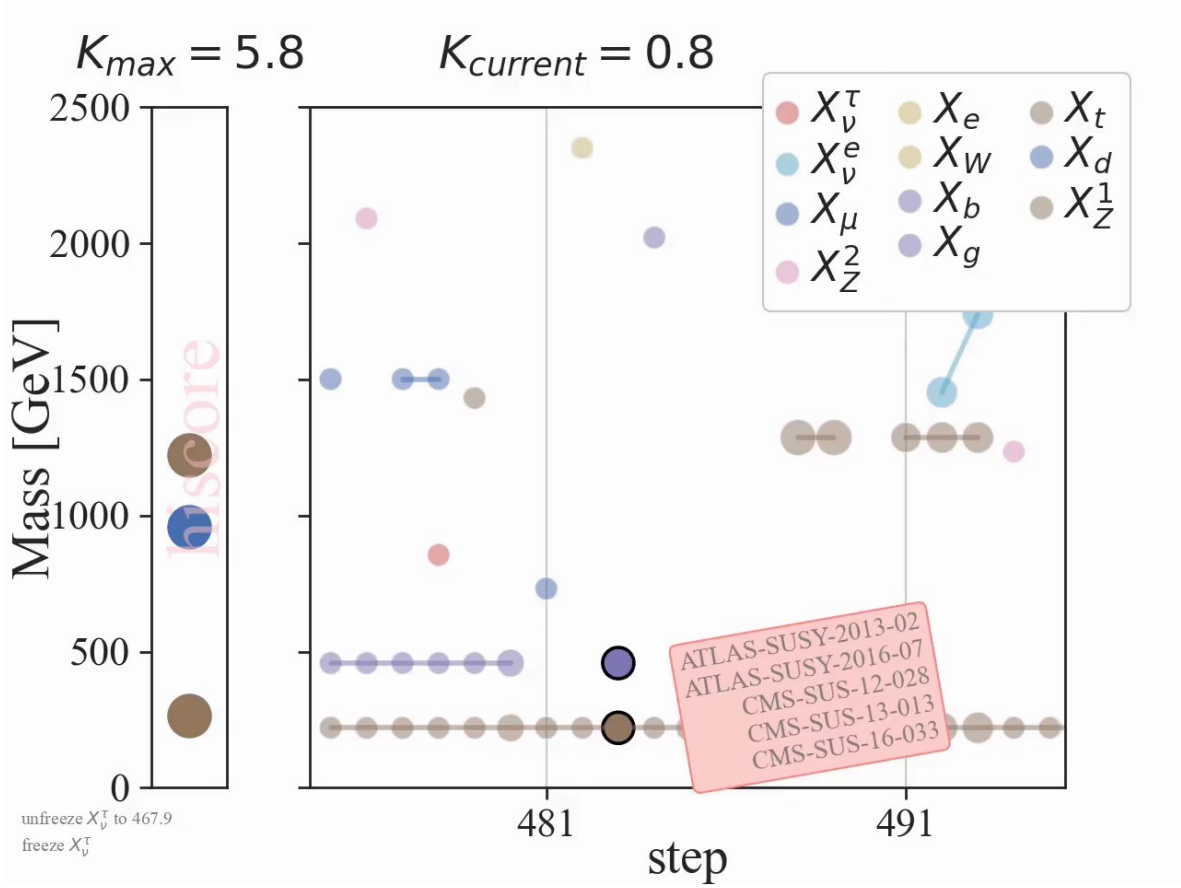
The Test Statistic

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

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

- $\hat{\mu}$ is the signal strength of the model that maximizes the likelihood.
- Its support is limited such as to ensure that all negative results in the SModelS database are respected.

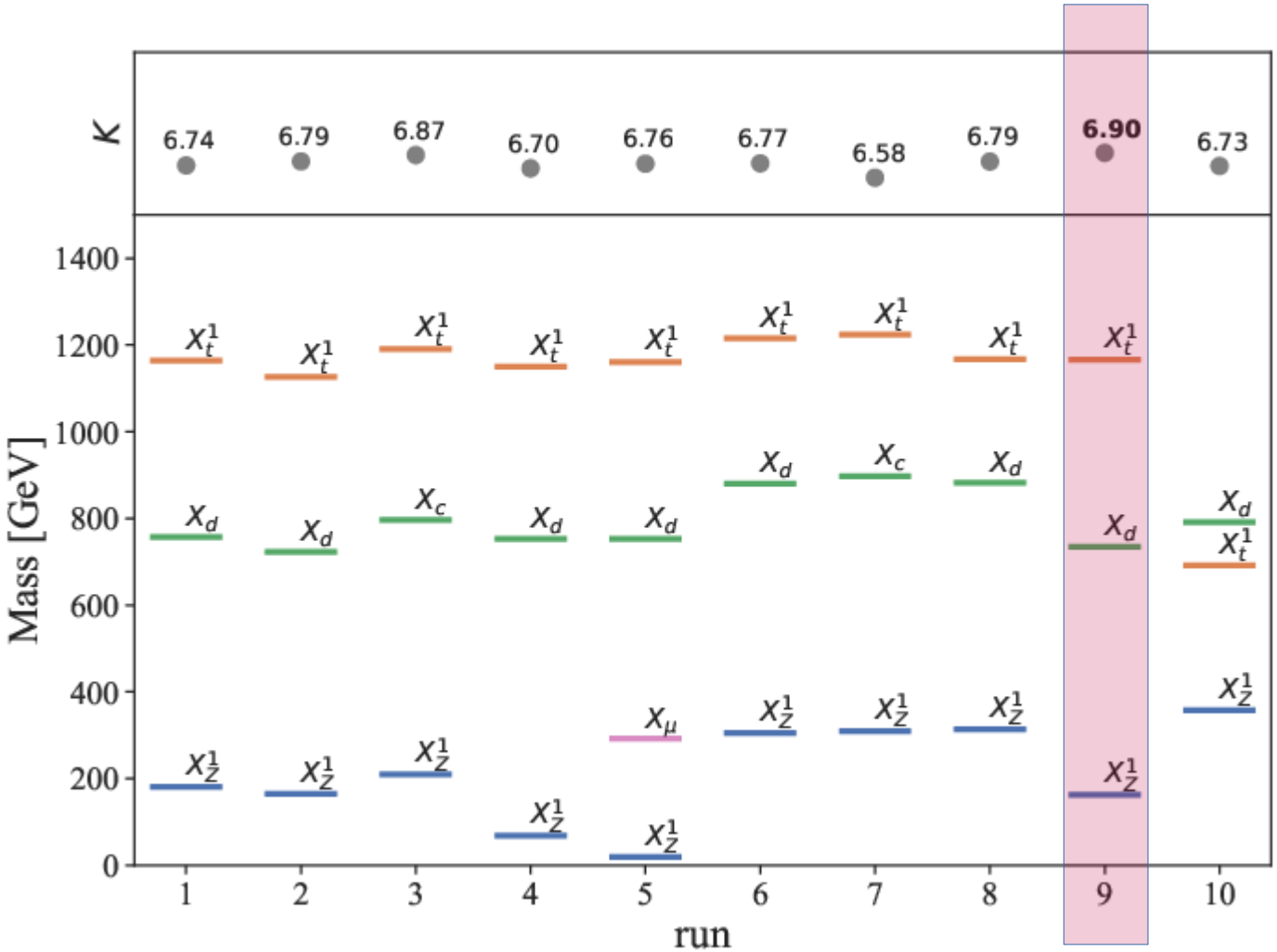
And Then We Ran the Algorithm ...



We defined a “run” as 50 parallel walkers, making 1,000 steps each.
We performed 10 such runs on the SModelS database.
Total computing resources spent: ~ 1,000,000 CPU hours

Walking Over the SModelS Database

We performed 10 such runs on the SModelS database:

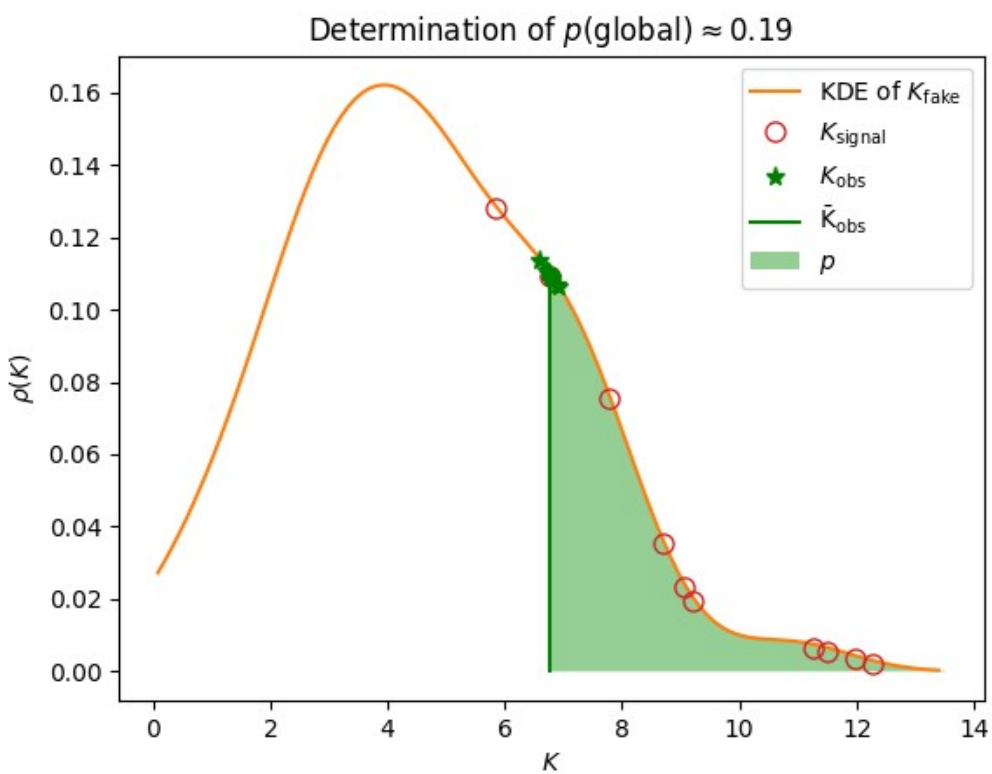


All 10 runs introduced a top partner as well as a light quark partner. The cross sections found by the algorithm are compatible with values expected from the MSSM. The highest test statistic was K=6.9.

Global *p*-Value

- Because we have statistical models of the search results, we can synthesize statistically correct databases of results that are “typical”, if no new physics is in the data.

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



By construction, no Look-Elsewhere Effect applies.

Summary, Outlook

- In light of no clear evidence for new physics in the individual channels/results, a more global attempt at finding new physics seems appropriate

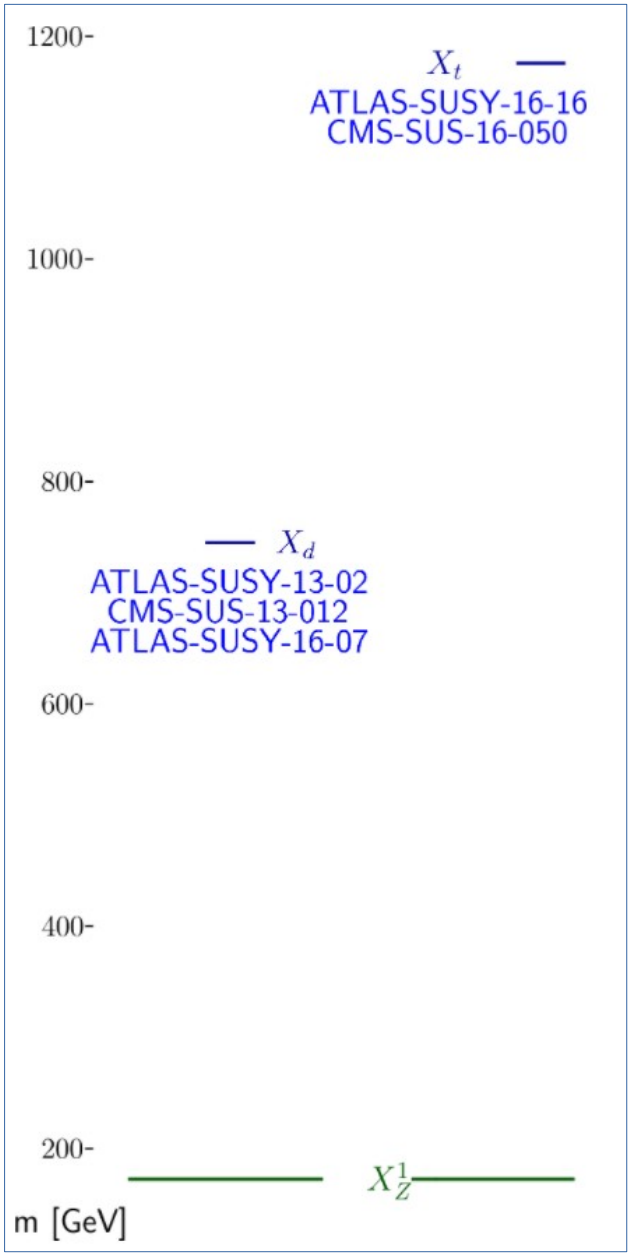
- First prototype run of protomodels builder with results from ~ 100 analyses resulted in p-value of SM hypothesis of ~ 0.2: a very small tension with the Standard Model hypothesis (but also some tension between some results)

- Working on next iteration with more results, better likelihoods, improved algorithm, covering more signatures, larger protomodels space

- Can we combine information from searches with information from measurements?

Backup

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)	$\sigma^{\rm UL}_{\rm obs}$ (fb)	$\sigma^{\rm UL}_{\rm exp}$ (fb)	$r_{\rm obs}$
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, $M_{\rm T2}$, 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: $X_t : K_{\rm without} = 2.59(59\%), X_d : K_{\rm without} = 3.90(41\%)$
Last updated: Mon Dec 14 20:08:06 2020

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

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

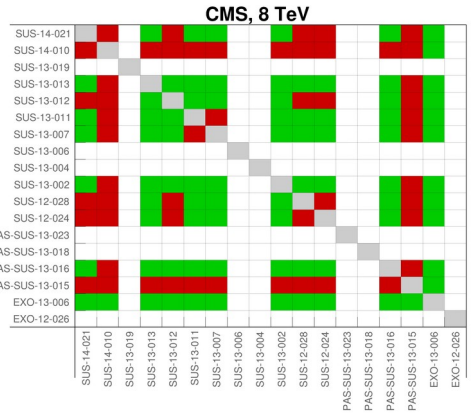
- **Efficiency maps + full likelihoods**
  full realism, correct statistical model

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.

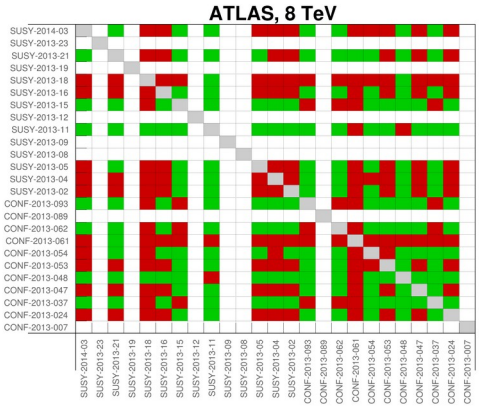
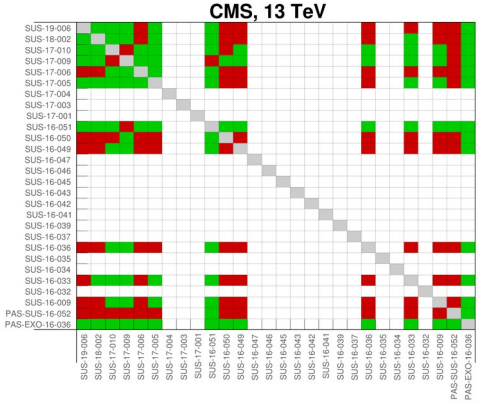


Fig. 2

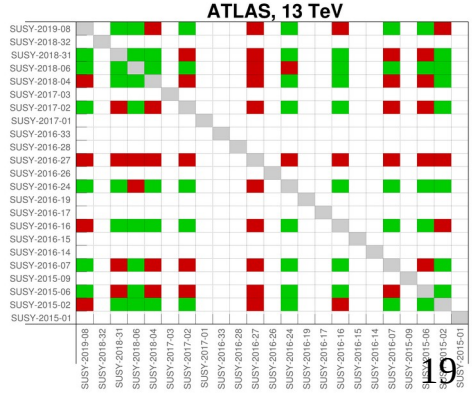
Approximately uncorrelated are analyses that are:

- 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

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

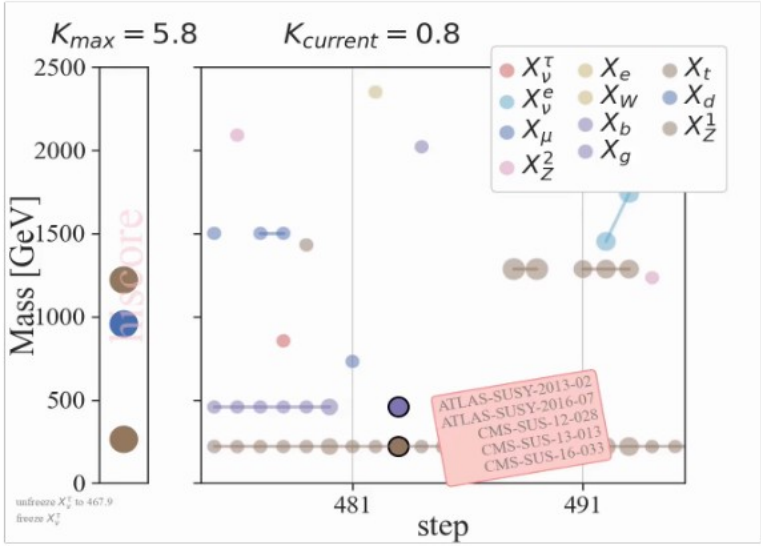
It is maximized in the denominator, but its support is confined such that no limits in the SModelS database are violated (the "critic"),

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

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

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

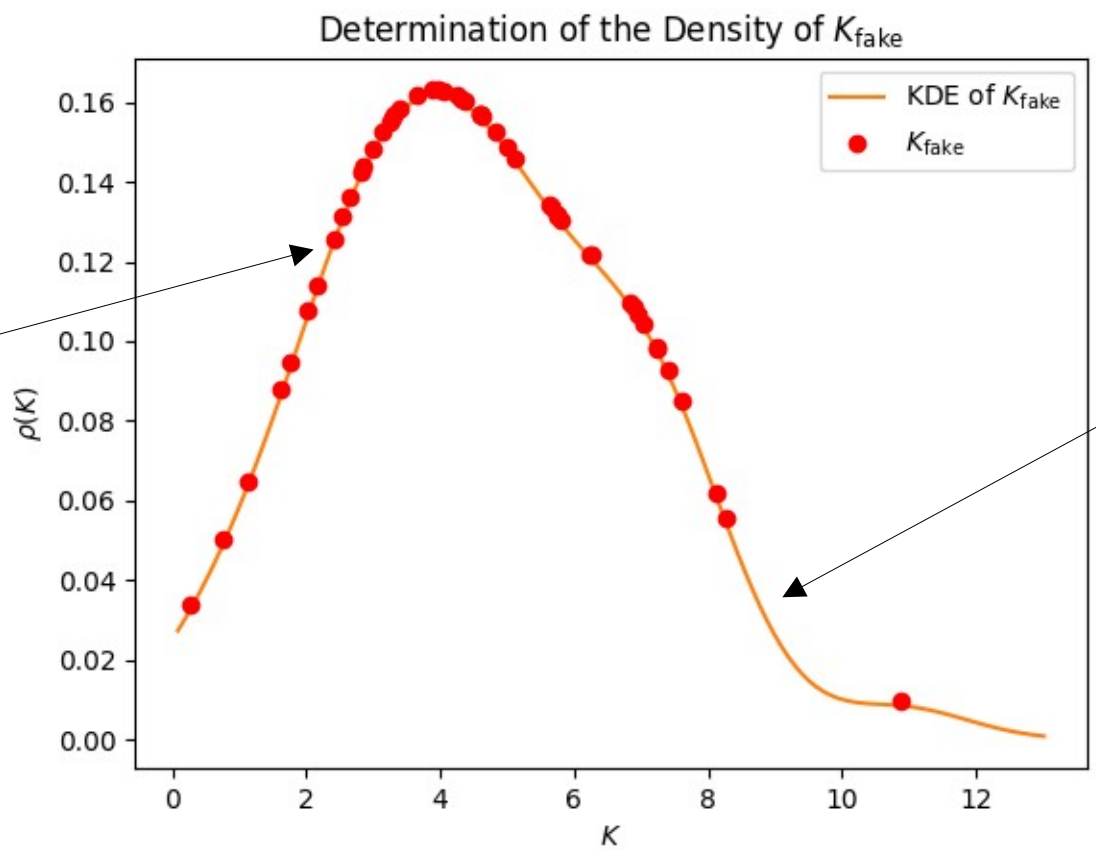
if and only if K_i is smaller than K_{i-1}

* (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

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

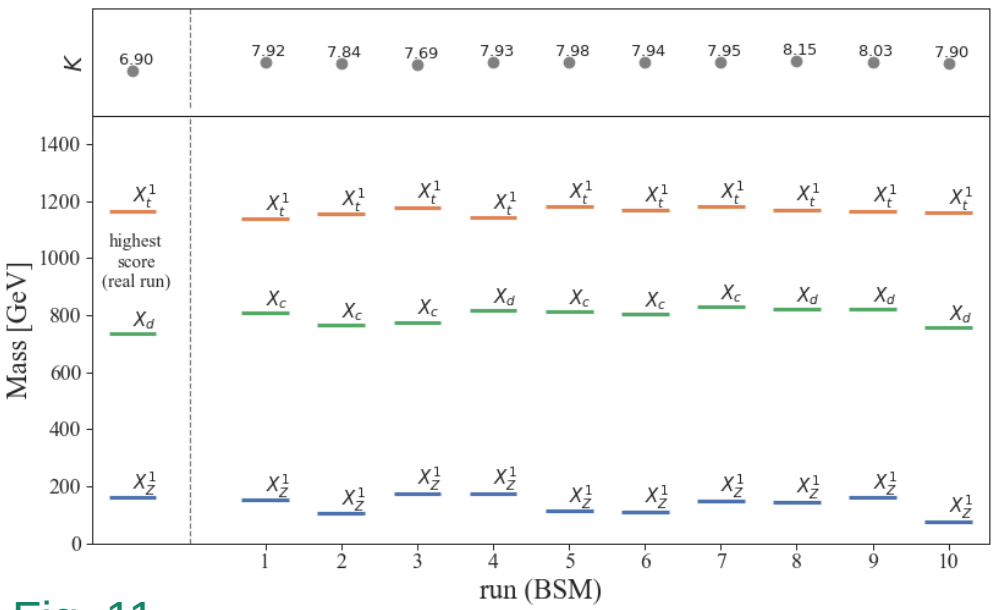


Fig. 11

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

Physics closure test

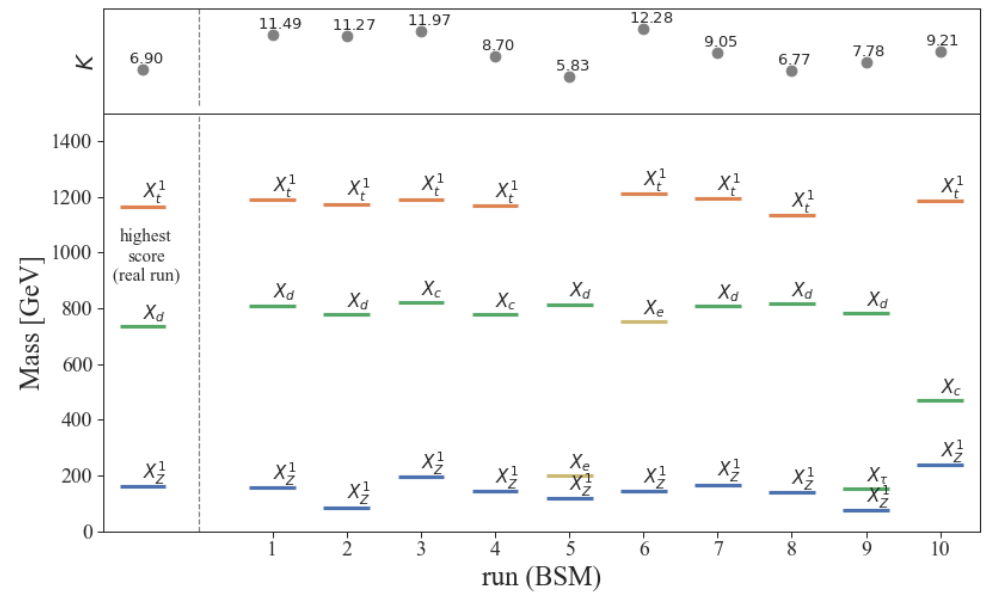


Fig. 10

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:

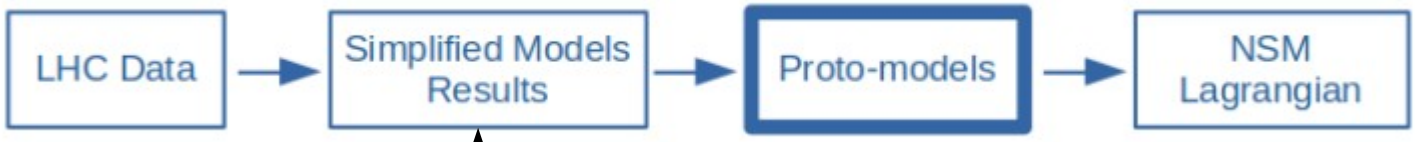
- learn the SModelS database
- make everything differentiable

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!