Statistically Learning the Next SM from LHC data

1

André Lessa
(UFABC – São Paulo)

Wolfgang Waltenberger
(HEPHY - Vienna)

LPSC

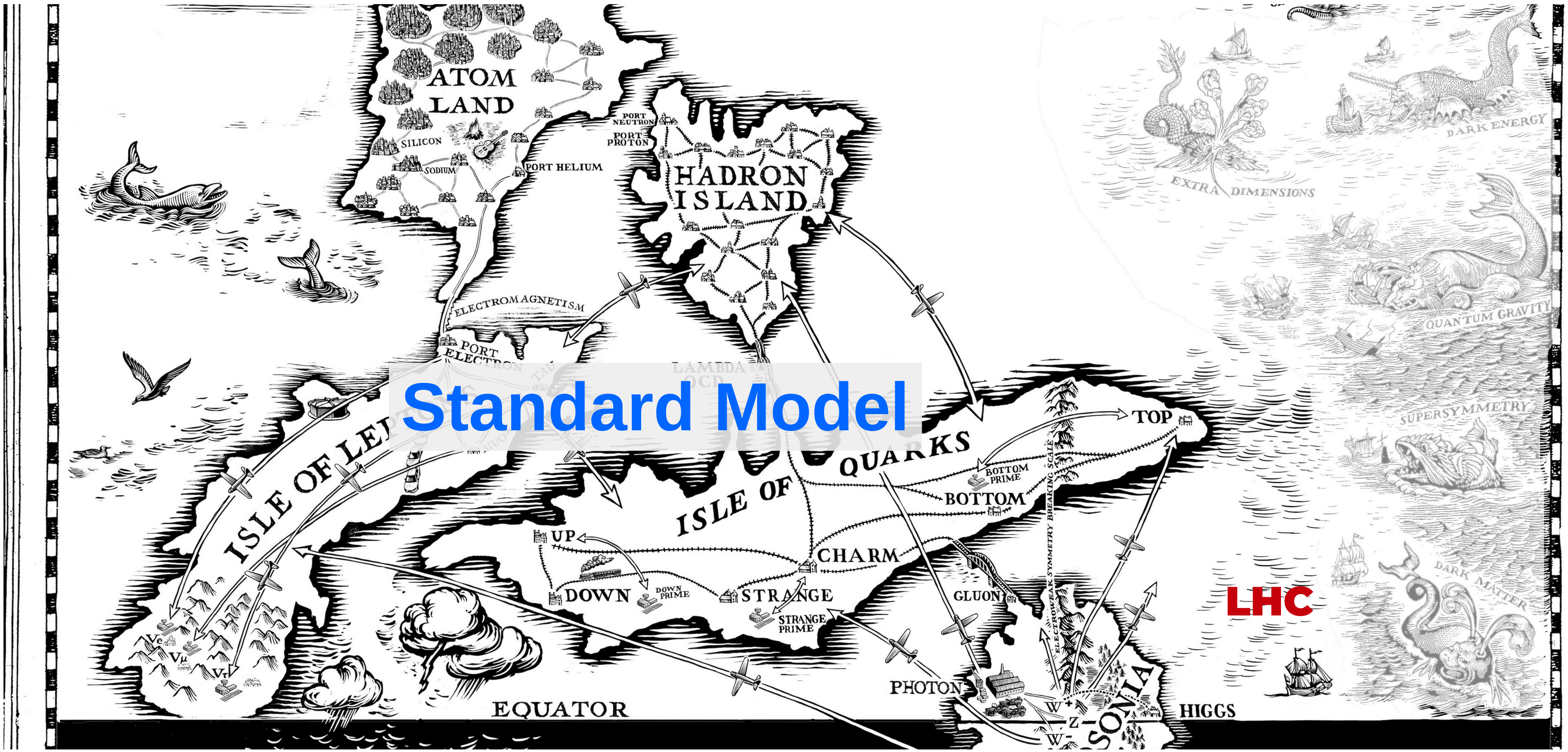
May 19, 2022

In collaboration with Sabine Kraml

JHEP 03 (2021) 207

Exploring the Unknown

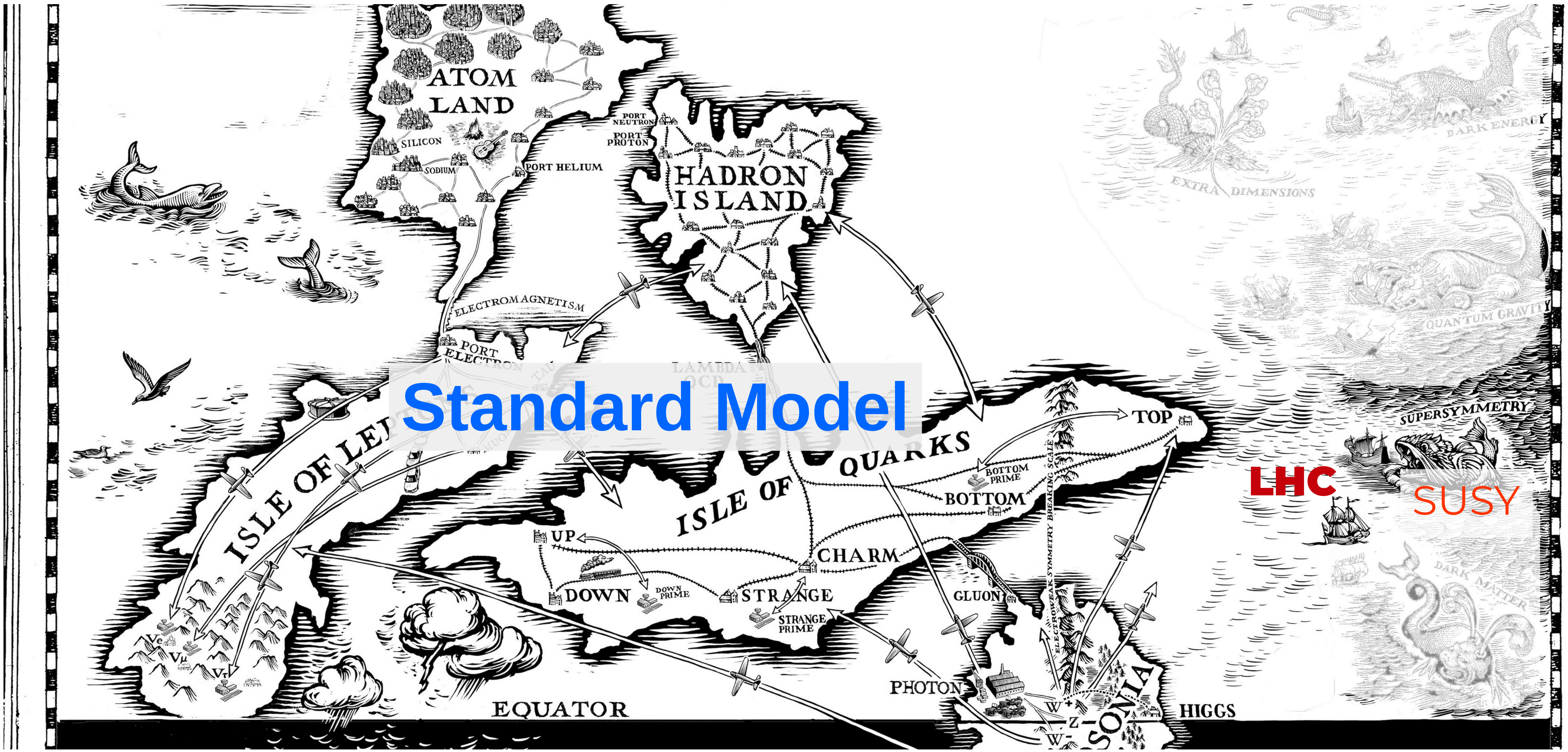
- The LHC is one of the main tools for exploring the frontier of our current understanding of the Universe



Adapted rom J. Butterworth's talk @ LPSC

Exploring the Unknown

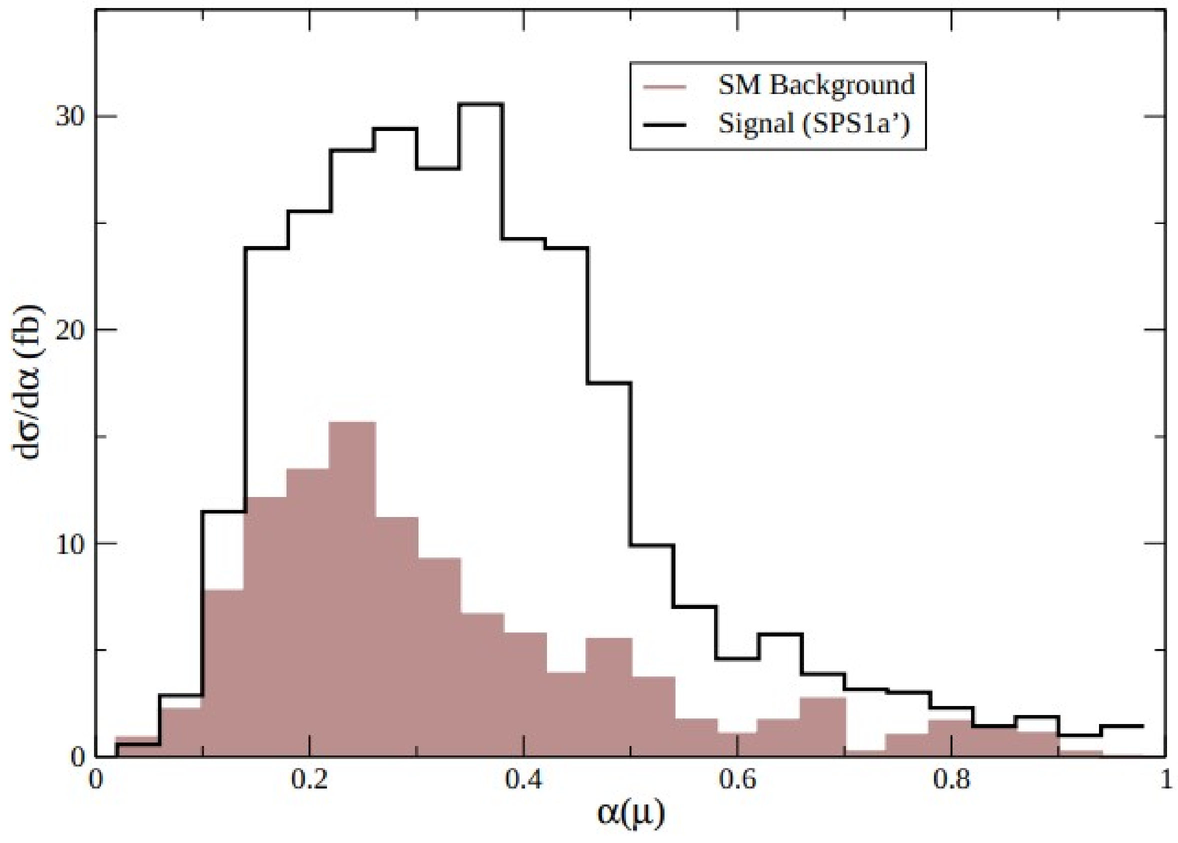
- Previous to LHC data, there was a strong bias in some of the HEP community:



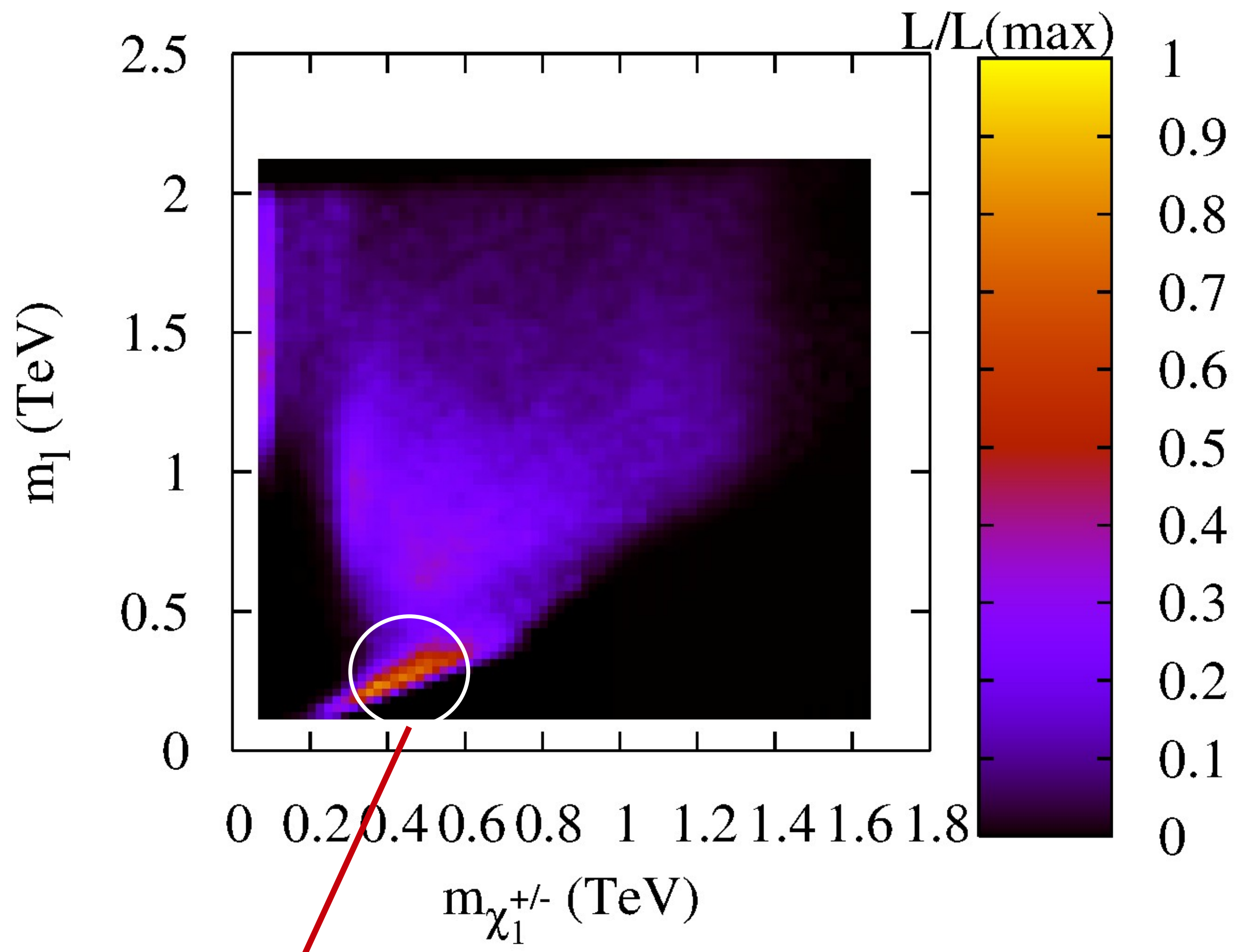
New Physics @ LHC

- It was expected that new physics would quickly emerge from LHC data

H. Baer, AL, H. Summy, Phys.Lett.B 674 (2009)



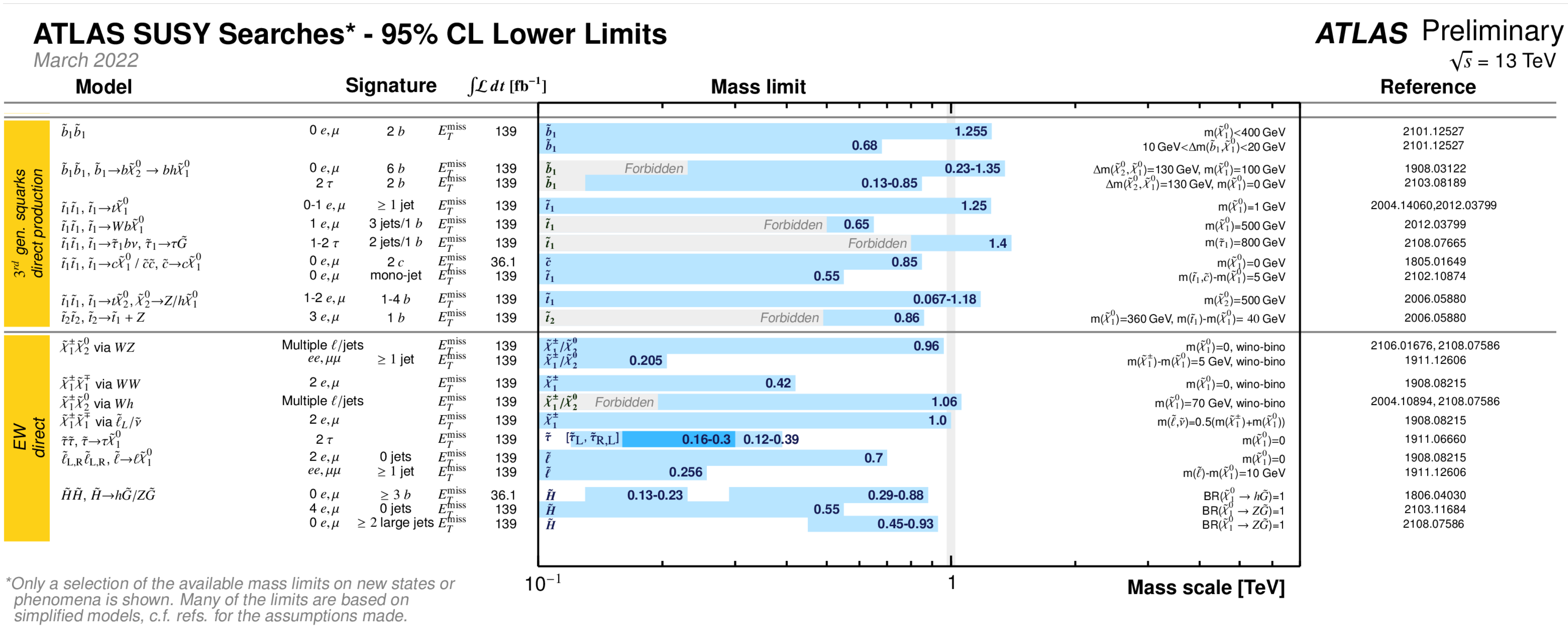
B.C. Allanach and C.G. Lester, Phys.Rev.D 73 (2006) 015013



BSM masses ~ 300-500 GeV

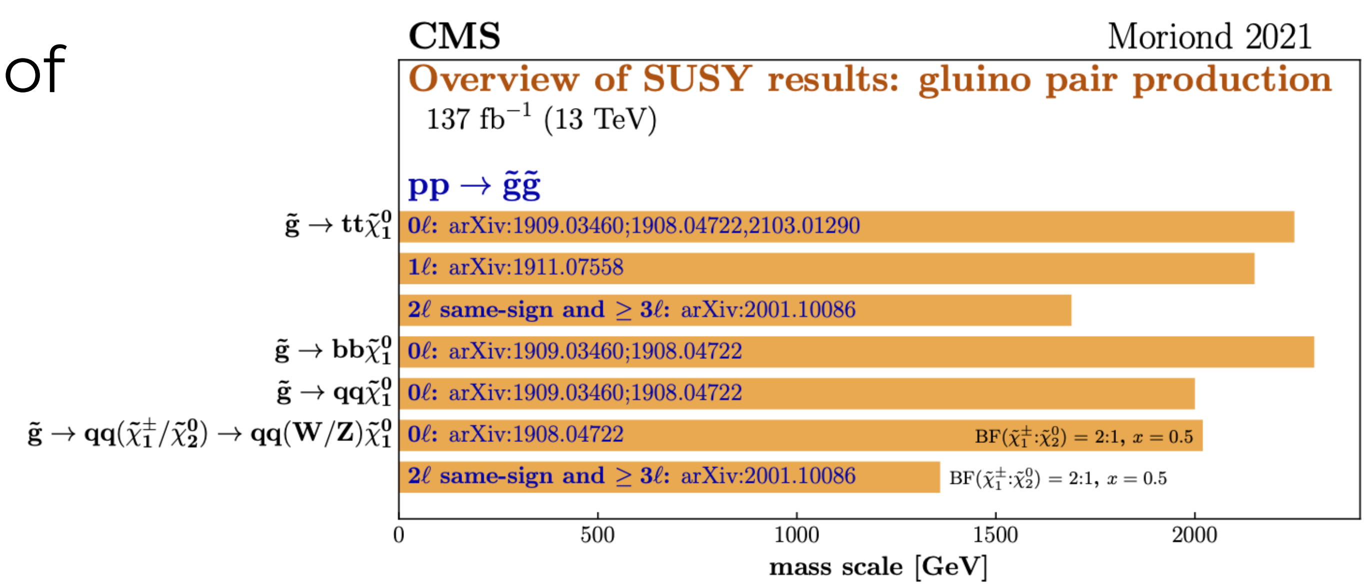
New Physics @ LHC

- After a decade of data taking…



No clear evidence of new physics

Many caveats!



Selection of observed limits at 95% C.L. (theory uncertainties are not included). Probe **up to** the quoted mass limit for light LSPs unless stated otherwise. The quantities ΔM and x represent the absolute mass difference between the primary sparticle and the LSP, and the difference between the intermediate sparticle and the LSP relative to ΔM, respectively, unless indicated otherwise.

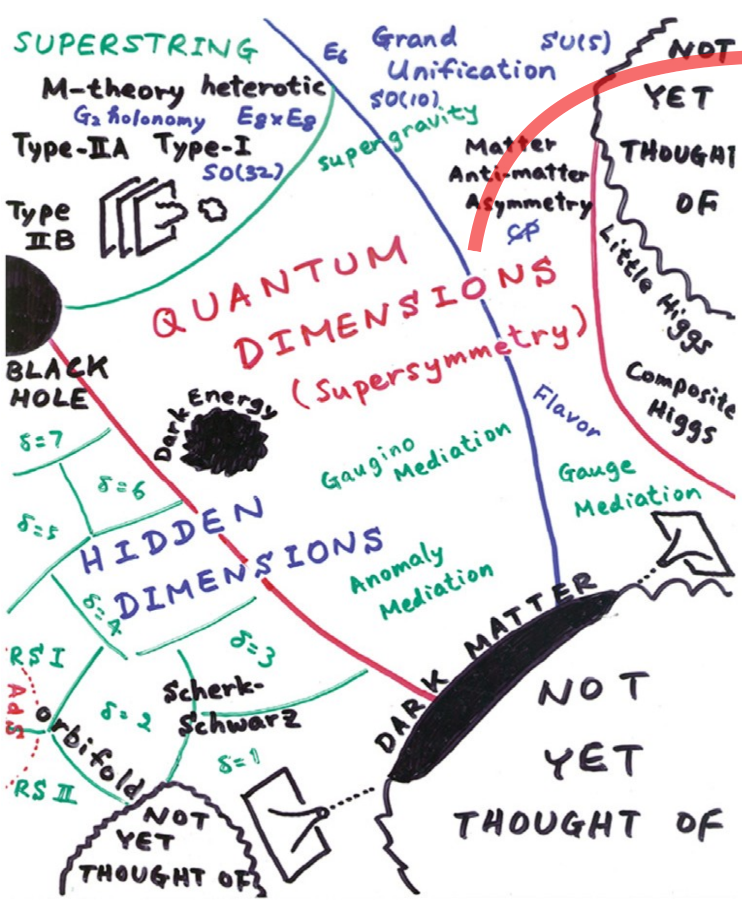
New Physics @ LHC

- Are we missing something?
- Could new physics be hiding in the data?
- Should we revisit our approach for searching for NP?

New Physics @ LHC

- Are we missing something?
- Could new physics be hiding in the data?
- Should we revisit our approach for searching for NP?

  - "Theory-oriented" (top-down) approach:



Choose model

Fix model parameters

Identify promising signatures

Simulate signal

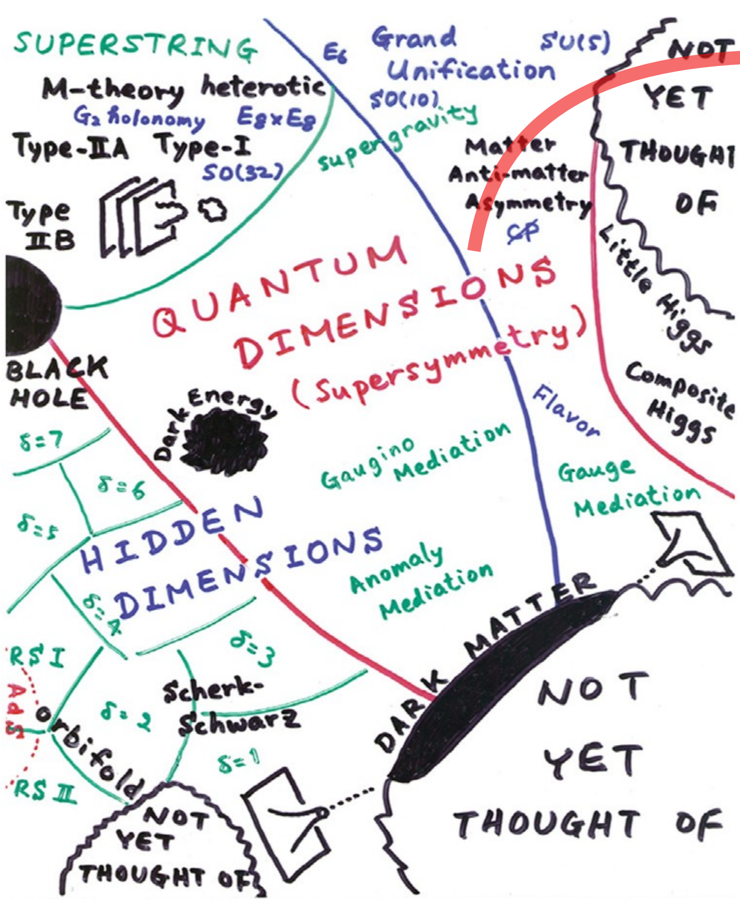
Compare with relevant LHC data

*it works well if we have a single well-motivated model candidate

New Physics @ LHC

- Are we missing something?
- Could new physics be hiding in the data?
- Should we revisit our approach for searching for NP?

- “Theory-oriented” (top-down) approach:



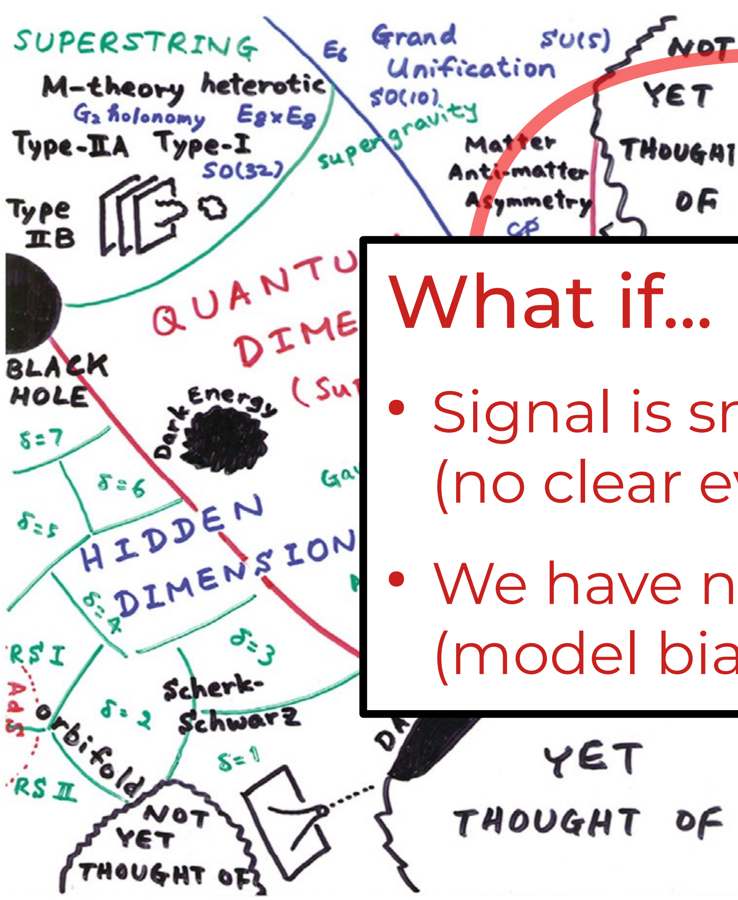
Choose model	Large number of models!
Fix model parameters	Infinite set of parameters!
Identify promising signatures	Non-obvious signatures
Simulate signal	CPU expensive
Compare with relevant LHC data	Large number of searches!

*it works well if we have a single well-motivated model candidate

New Physics @ LHC

- Are we missing something?
- Could new physics be hiding in the data?
- Should we revisit our approach for searching for NP?

- “Theory-oriented” (top-down) approach:



Choose model — Large number of models!

Fix model parameters — Infinite set of parameters!

es

Compare with relevant LHC data — Large number of searches!

What if…

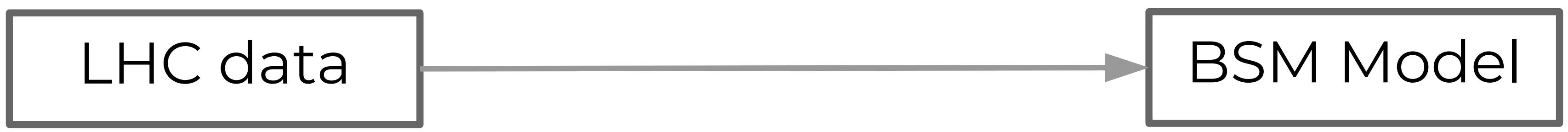
- Signal is small and dispersed over many signatures? (no clear evidence from individual searches)
- We have not yet considered the correct model? (model bias)

*it works well if we have a single well-motivated model candidate

Looking for the Next SM

- Our proposal: “Data-oriented” approach

Let the data guide the model building!



(LHC inverse problem)

Looking for the Next SM

- Our proposal: “Data-oriented” approach

Let the data guide the model building!

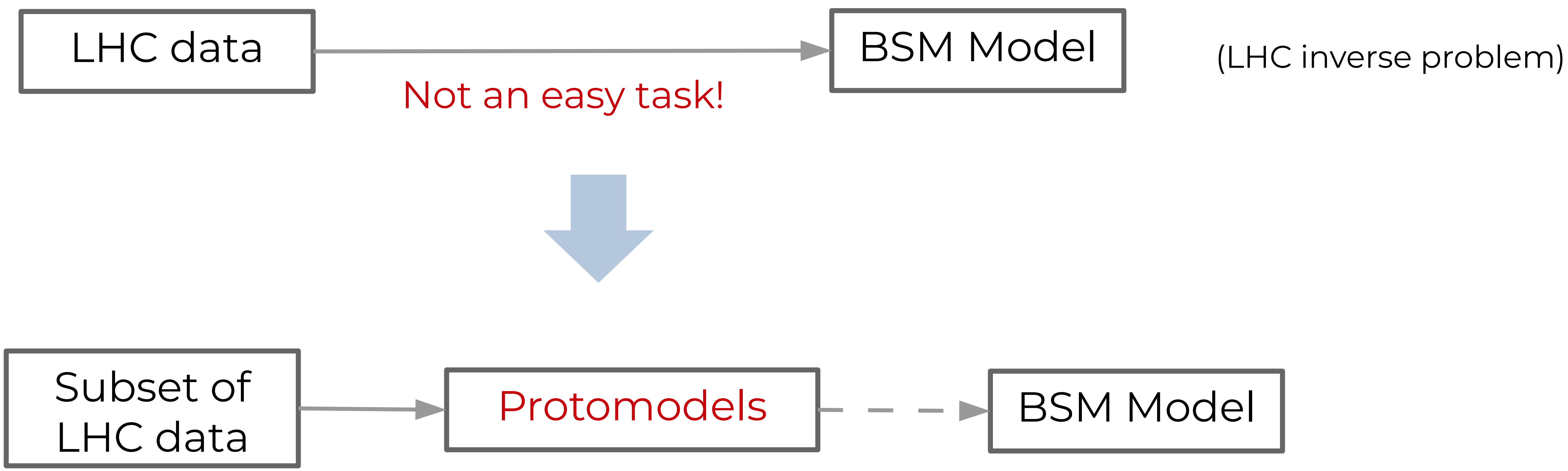
LHC data → BSM Model (LHC inverse problem)

Not an easy task!

Looking for the Next SM

- Our proposal: “Data-oriented” approach

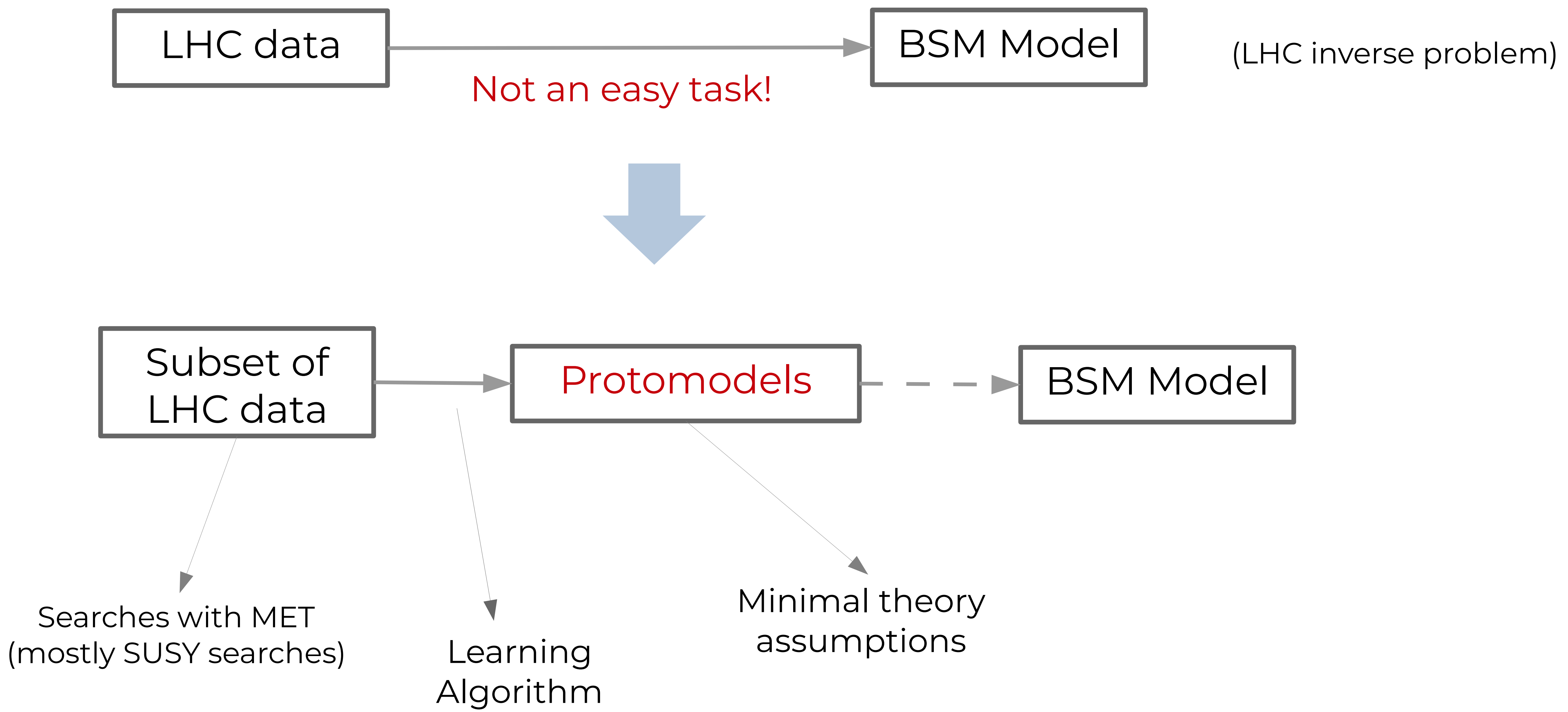
Let the data guide the model building!



Looking for the Next SM

- Our proposal: “Data-oriented” approach

Let the data guide the model building!



Protomodels

- Minimizing the theory bias → protomodels

Protomodels

- Minimizing the theory bias → protomodels
- Protomodels are defined by:
  - Particle content
  - Masses
  - Branching ratios
  - Production cross-sections

(no considerations about symmetries, lagrangian, vertices, ...)

Protomodels

- Minimizing the theory bias → protomodels

- Protomodels are defined by:
  - Particle content
  - Masses
  - Branching ratios
  - Production cross-sections

(no considerations about symmetries, lagrangian, vertices, ...)

- For now we assume a Z_2-like symmetry
- The lightest particle must be stable and neutral (X_z)
- Protomodel ~ consistent set of simplified models

Protomodels

- Minimizing the theory bias → protomodels
- Protomodels are defined by:
  - Particle content
  - Masses
  - Branching ratios
  - Production cross-sections

(no considerations about symmetries, lagrangian, vertices, …)

- For now we assume a Z_2-like symmetry
- The lightest particle must be stable and neutral (X_z)
- Protomodel ~ consistent set of simplified models

- Protomodel Builder:
  - randomly selects particles and properties from a pool of available possibilities

masses,
BRs,
xsecs

particle	decay channels	particle	decay channels
X_q	$qX_Z^j,\ q'X_W^i,\ qX_g$	X_W^1	WX_Z^j
X_t^1	$tX_Z^j,\ bX_W^i,\ WX_b^1,\ tX_g$	X_W^2	$WX_Z^j,\ ZX_W^1,\ hX_W^1$
X_b^1	$bX_Z^j,\ tX_W^i,\ WX_t^1,\ bX_g$	$X_Z^{j\neq 1}$	$WX_W^i,\ ZX_Z^k,\ hX_Z^k$
X_t^2	$tX_Z^j,\ bX_W^i,\ ZX_t^1,\ WX_b^1,\ tX_g$	X_ℓ	$\ell X_Z^j,\ \nu_\ell X_W^i$
X_b^2	$bX_Z^j,\ tX_W^i,\ ZX_b^1,\ WX_t^1,\ bX_g$	X_{ν_ℓ}	$\nu_\ell X_Z^j,\ \ell X_W^i$
X_g	$q\bar{q}X_Z^i,\ q\bar{q}'X_W^i,\ b\bar{b}X_Z^i,\ t\bar{t}X_Z^j,\ btX_W^i,\ qX_q,\ bX_b,\ tX_t$		

Building the Next SM

- Overview:

Building the Next SM

- Overview:

PROTOMODEL BUILDER

Random changes to the protomodel

Building the Next SM

- Overview:

PROTOMODEL BUILDER → Random changes to the protomodel → Test against database (SModelS)

Building the Next SM

- Overview:

PROTOMODEL BUILDER

Random changes to the protomodel → Test against database (SModelS) → Model score (compatibility with data)

Building the Next SM

- Overview:

PROTOMODEL BUILDER

Random changes to the protomodel

Test against database (SModelS)

Model score
(compatibility with data)

Feedback for building new models

Building the Next SM

- Overview:

PROTOMODEL BUILDER

Random changes to the protomodel

Test against database (SModelS)

Model score
(compatibility with data)

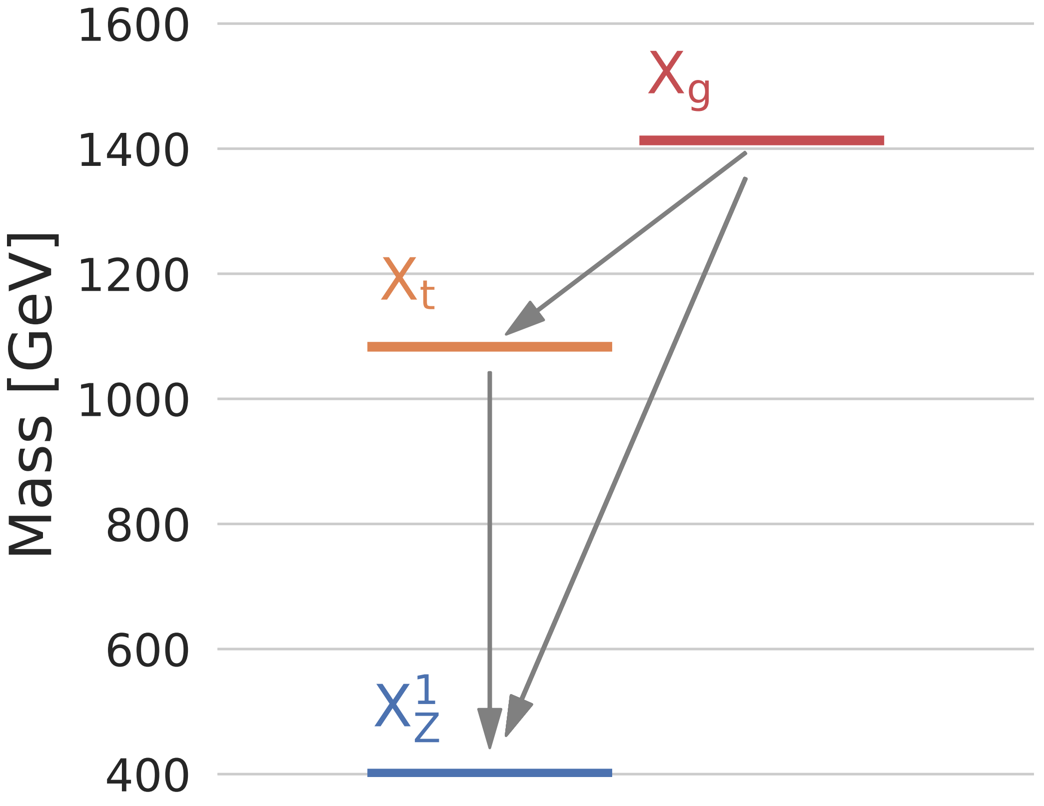
Feedback for building new models

➡ "MCMC-type walk" over model+parameter space

➡ After many iterations/steps, the builder "learns" the best BSM model

Walker Algorithm

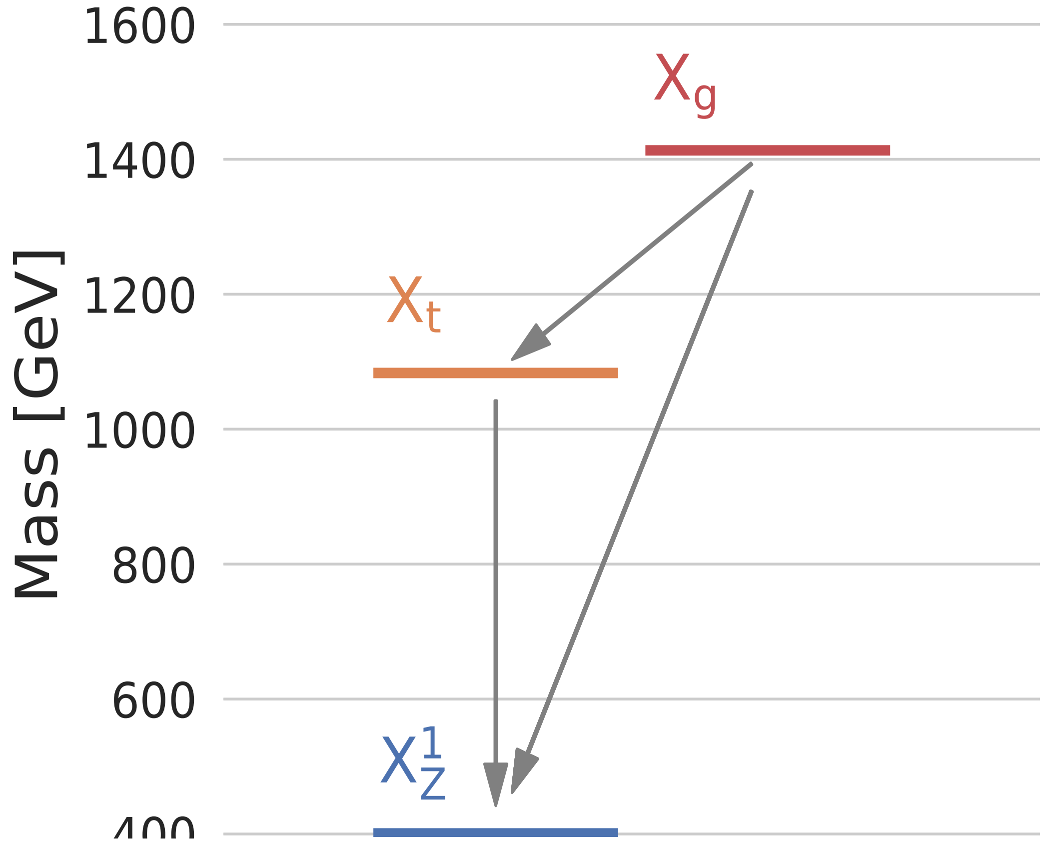
- At each step:
  - randomly add or remove a particle
  - randomly change branching ratios and masses
  - randomly change a production cross section



Walker Algorithm

- At each step:
  - randomly add or remove a particle
  - randomly change branching ratios and masses
  - randomly change a production cross section

- After each step:
  - compute new protomodel score “K”



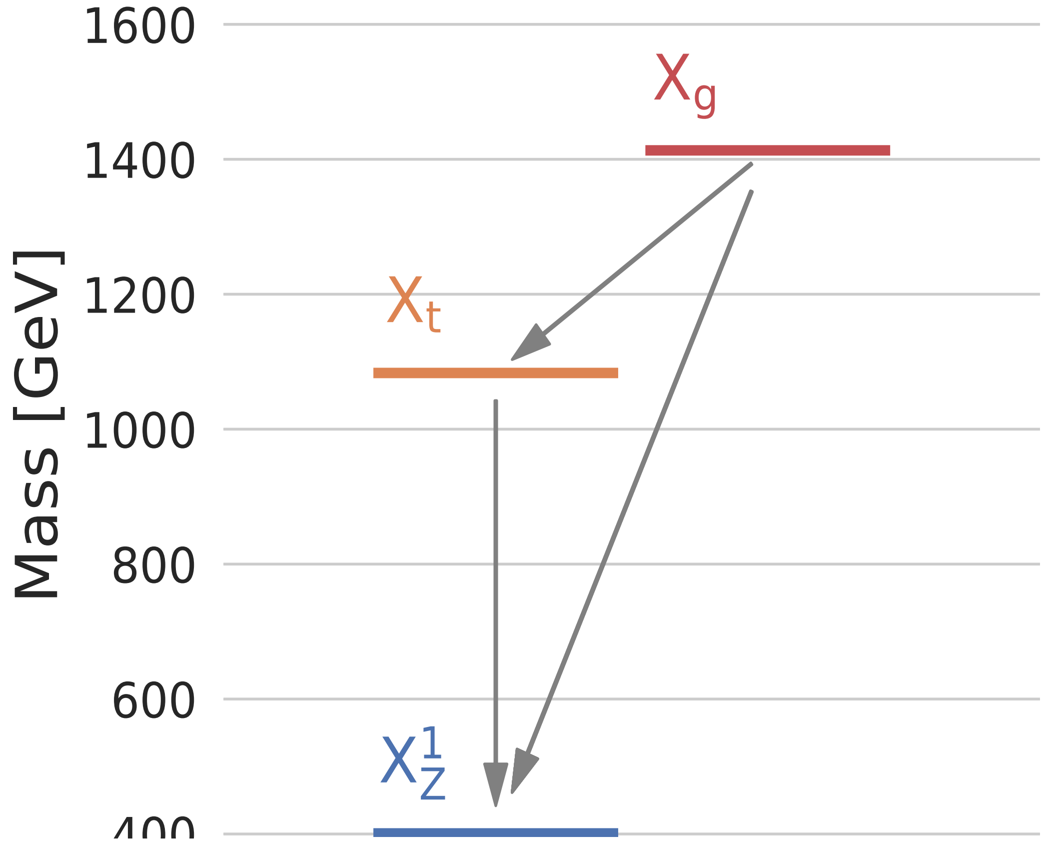
Walker Algorithm

- At each step:
  - randomly add or remove a particle
  - randomly change branching ratios and masses
  - randomly change a production cross section

- After each step:
  - compute new protomodel score “K”

- K got much worse?
  → revert to old protomodel

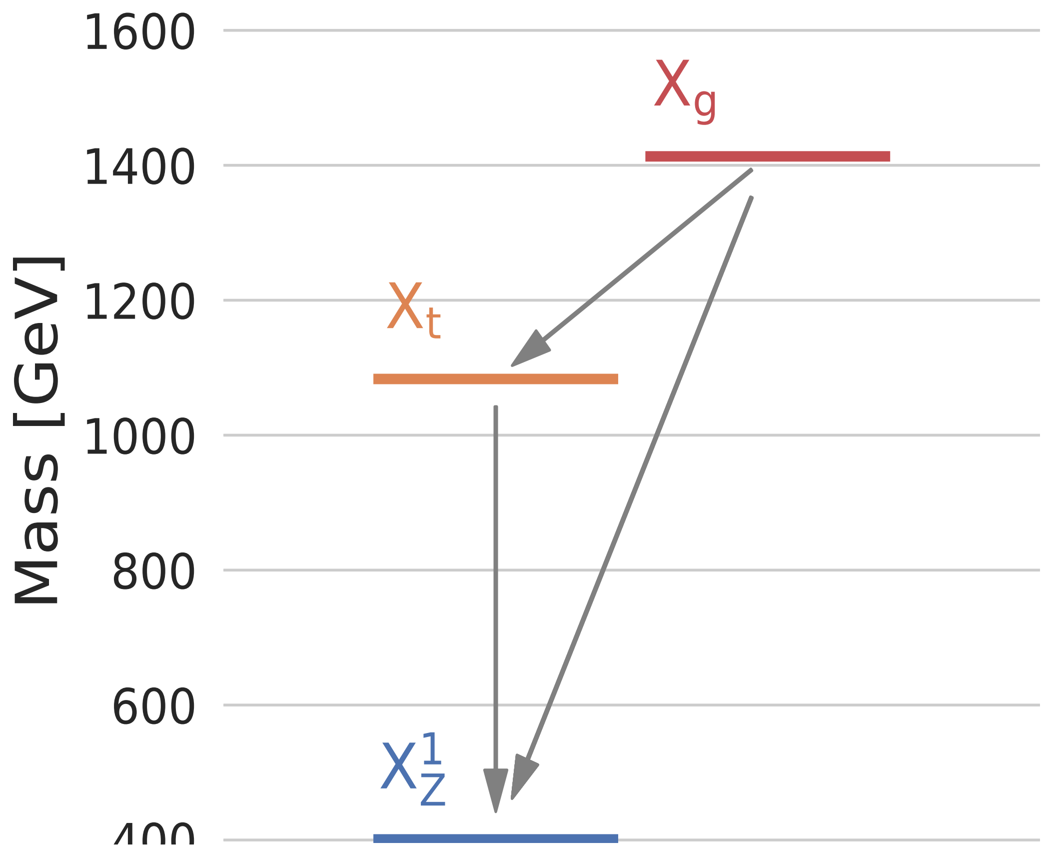
- K stayed the same or got better?
  → keep new protomodel



Walker Algorithm

- At each step:
  - randomly add or remove a particle
  - randomly change branching ratios and masses
  - randomly change a production cross section

- After each step:
  - compute new protomodel score "K"



- K got much worse?
  → revert to old protomodel

- K stayed the same or got better?
  → keep new protomodel

➡ The walker is driven by the protomodel score K

Test Statistic

The **test statistic 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).

Test Statistic

The **test statistic Kᶜ** 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).

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

Test Statistic

The **test statistic Kᶜ** 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).

Quantifies violation of Standard Model hypothesis

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

Test Statistic

The **test statistic Kᶜ** 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).

Quantifies violation of Standard Model hypothesis

Penalizes for model complexity

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

Test Statistic

The **test statistic 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).

Quantifies violation of Standard Model hypothesis

Penalizes 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. If a result can be added, it must be added.

Test Statistic

The **test statistic 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).

Quantifies violation of Standard Model hypothesis

Penalizes for model complexity

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

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

Joint likelihoods: combining “complete” sets of results that are assumed to be approximately uncorrelated. If a result can be added, it must be added.

$$\pi(\mathrm{M}) \approx \exp[-c\cdot \mathrm{n}^{\mathrm{BSM}}_{\mathrm{particles}}]$$

→ “AIC-like” test statistic

Test Statistic

The **test statistic 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).

Quantifies violation of Standard Model hypothesis

Penalizes for model complexity

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

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

$$\pi(\mathrm{M}) \approx \exp[-c \cdot \mathrm{n}^{\mathrm{BSM}}_{\mathrm{particles}}]$$

→ "AIC-like" test statistic

Joint likelihoods: combining "complete" sets of results that are assumed to be approximately uncorrelated. If a result can be added, it must be added.

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

Test Statistic

- We choose that combination of signal regions that maximally violates the SM hypothesis (“anomaly hunt”)

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

Test Statistic

- We choose that combination of signal regions that maximally violates the SM hypothesis ("anomaly hunt")

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

By restricting the support of the parameter of interest we guarantee compatibility with all negative results in the entirety of the SModelS database

Test Statistic

- We choose that combination of signal regions that maximally violates the SM hypothesis ("anomaly hunt")

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

By restricting the support of the parameter of interest we guarantee compatibility with all negative results in the entirety of the SModelS database

- The test statistic is based on likelihoods
- The likelihood is computed using simplified models results in SModelS database

Input Data

- SModelS Database:
  - Searches for production of new particles with missing energy (DM-inspired)
  - Around 50 CMS and 50 ATLAS publications
  - Simplified statistical models for the data → simplified likelihoods

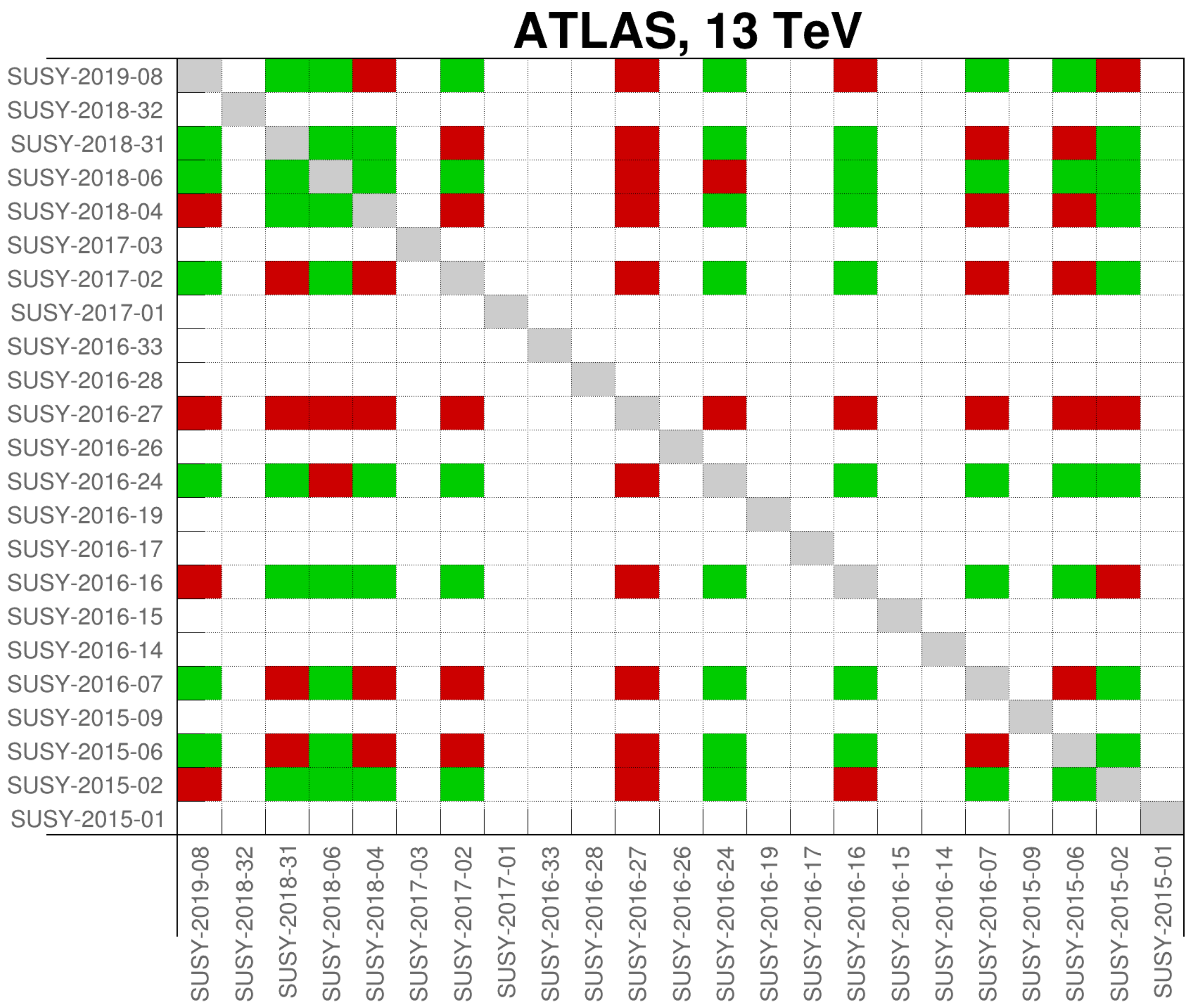
ID	Short Description	$\mathcal{L}$ [fb^{-1}]	UL$_{obs}$	UL$_{exp}$	EM	comb.
ATLAS-SUSY-2015-01 [67]	2 b-jets	3.2	✓			
ATLAS-SUSY-2015-02 [68]	1ℓ stop	3.2	✓		✓	
ATLAS-SUSY-2015-06 [69]	0ℓ + 2–6 jets	3.2			✓	
ATLAS-SUSY-2015-09 [70]	jets + 2 SS or $\geq 3\,\ell$	3.2	✓			
ATLAS-SUSY-2016-07 [55]	0ℓ + jets	36.1	✓		✓	
ATLAS-SUSY-2016-14 [71]	jets + 2 SS or $\geq 3\,\ell$	36.1	✓			
ATLAS-SUSY-2016-15 [72]	0ℓ stop	36.1	✓			
ATLAS-SUSY-2016-16 [48]	1ℓ stop	36.1	✓		✓	
ATLAS-SUSY-2016-17 [73]	2 OS leptons	36.1	✓			
ATLAS-SUSY-2016-19 [74]	2 b-jets + τ's	36.1	✓			
ATLAS-SUSY-2016-24 [53]	2–3 ℓ, 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-2016-33 [78]	2 OSSF ℓ	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 [81]	2–3 leptons, EWino	36.1	✓			
ATLAS-SUSY-2018-04 [38]	2 hadronic taus	139.0	✓		✓	JSON
ATLAS-SUSY-2018-06 [82]	3 leptons, EWino	139.0	✓	✓		
ATLAS-SUSY-2018-31 [39]	2b + 2H(bb)	139.0	✓		✓	JSON
ATLAS-SUSY-2018-32 [83]	2 OS leptons	139.0	✓			
ATLAS-SUSY-2019-08 [40]	1ℓ + higgs	139.0	✓		✓	JSON

ID	Short Description	$\mathcal{L}$ [fb^{-1}]	UL$_{obs}$	UL$_{exp}$	EM	comb.
CMS-PAS-EXO-16-036 [84]	HSCP	12.9	✓		✓	
CMS-PAS-SUS-16-052 [34]	ISR jet + soft ℓ	35.9	✓		✓	Cov.
CMS-SUS-16-009 [85]	0ℓ + jets, top tagging	2.3	✓	✓		
CMS-SUS-16-032 [86]	2 b- or 2 c-jets	35.9	✓			
CMS-SUS-16-033 [57]	0ℓ + jets	35.9	✓	✓	✓	
CMS-SUS-16-034 [87]	2 OSSF leptons	35.9	✓			
CMS-SUS-16-035 [88]	2 SS leptons	35.9	✓			
CMS-SUS-16-036 [58]	0ℓ + jets	35.9	✓	✓		
CMS-SUS-16-037 [89]	1ℓ + jets with MJ	35.9	✓			
CMS-SUS-16-039 [90]	2–3ℓ, EWino	35.9	✓			
CMS-SUS-16-041 [91]	jets + $\geq 3\ell$	35.9	✓			
CMS-SUS-16-042 [92]	1ℓ + jets	35.9	✓			
CMS-SUS-16-043 [93]	WH(bb), EWino	35.9	✓			
CMS-SUS-16-045 [94]	jets + H $\rightarrow \gamma\gamma$	35.9	✓			
CMS-SUS-16-046 [95]	high-p_T γ	35.9	✓			
CMS-SUS-16-047 [96]	γ + jets, high H_T	35.9	✓			
CMS-SUS-16-049 [97]	0ℓ stop	35.9	✓	✓		
CMS-SUS-16-050 [49]	0ℓ stop, m_{T2}	35.9	✓	✓		
CMS-SUS-16-051 [59]	1ℓ stop	35.9	✓	✓		
CMS-SUS-17-001 [98]	2ℓ stop	35.9	✓			
CMS-SUS-17-003 [99]	2 taus	35.9	✓			
CMS-SUS-17-004 [43]	EWino combination	35.9	✓			
CMS-SUS-17-005 [100]	1ℓ stop, soft	35.9	✓	✓		
CMS-SUS-17-006 [101]	jets + boosted H(bb)	35.9	✓	✓		
CMS-SUS-17-009 [52]	2 OSSF leptons	35.9	✓	✓		
CMS-SUS-17-010 [102]	2ℓ EWino, stop	35.9	✓	✓		
CMS-SUS-18-002 [103]	γ + (b-)jets	35.9	✓	✓		
CMS-SUS-19-006 [15]	0ℓ + jets, MHT	137.0	✓	✓		

...

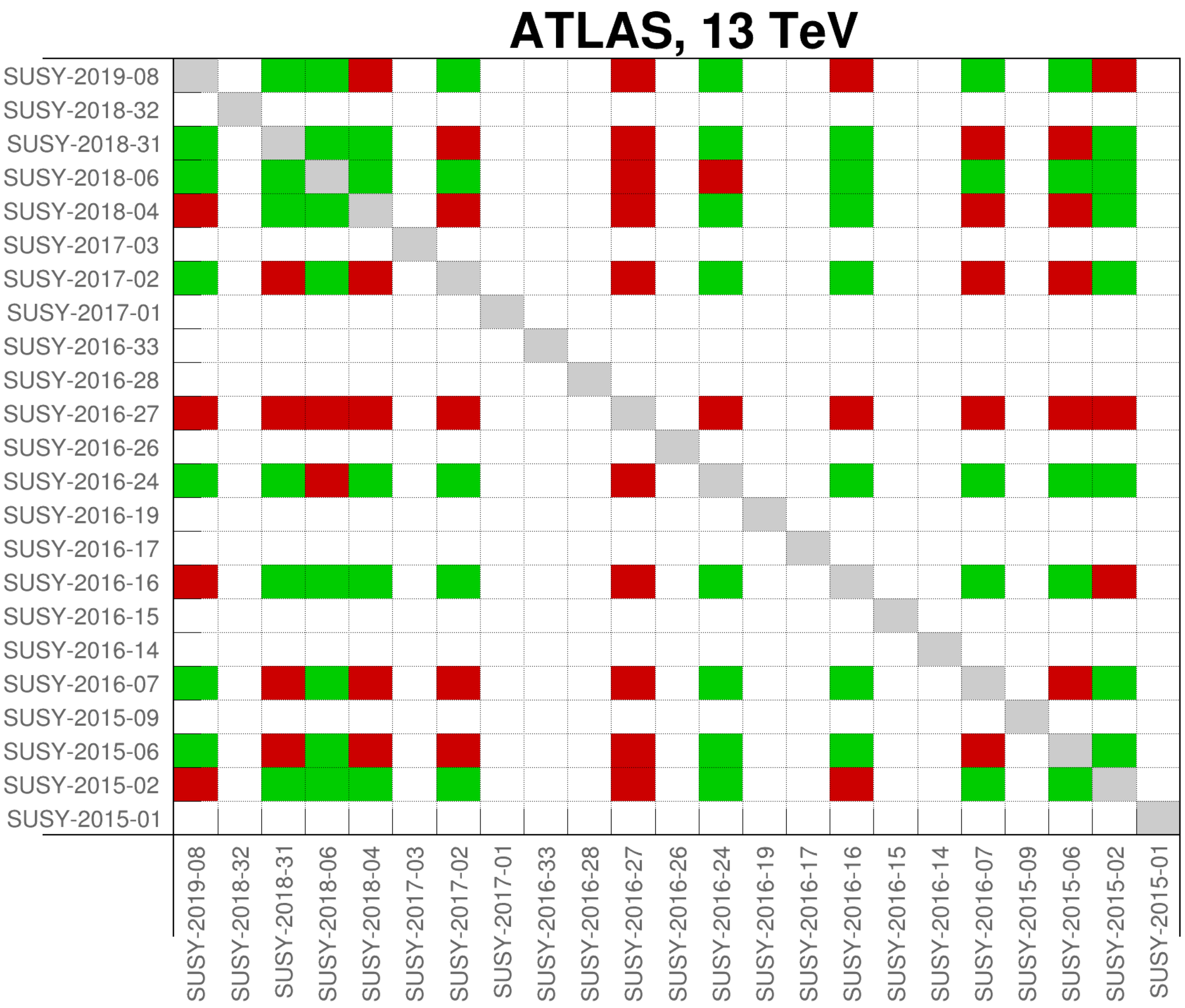
Combining Data

- As we are chasing dispersed signals, we need to combine likelihoods. We assume a simplified, binary “inter-analyses correlations matrix”:



Combining Data

- As we are chasing dispersed signals, we need to combine likelihoods. We assume a simplified, binary "inter-analyses correlations matrix":



Green → approximately uncorrelated → combinable

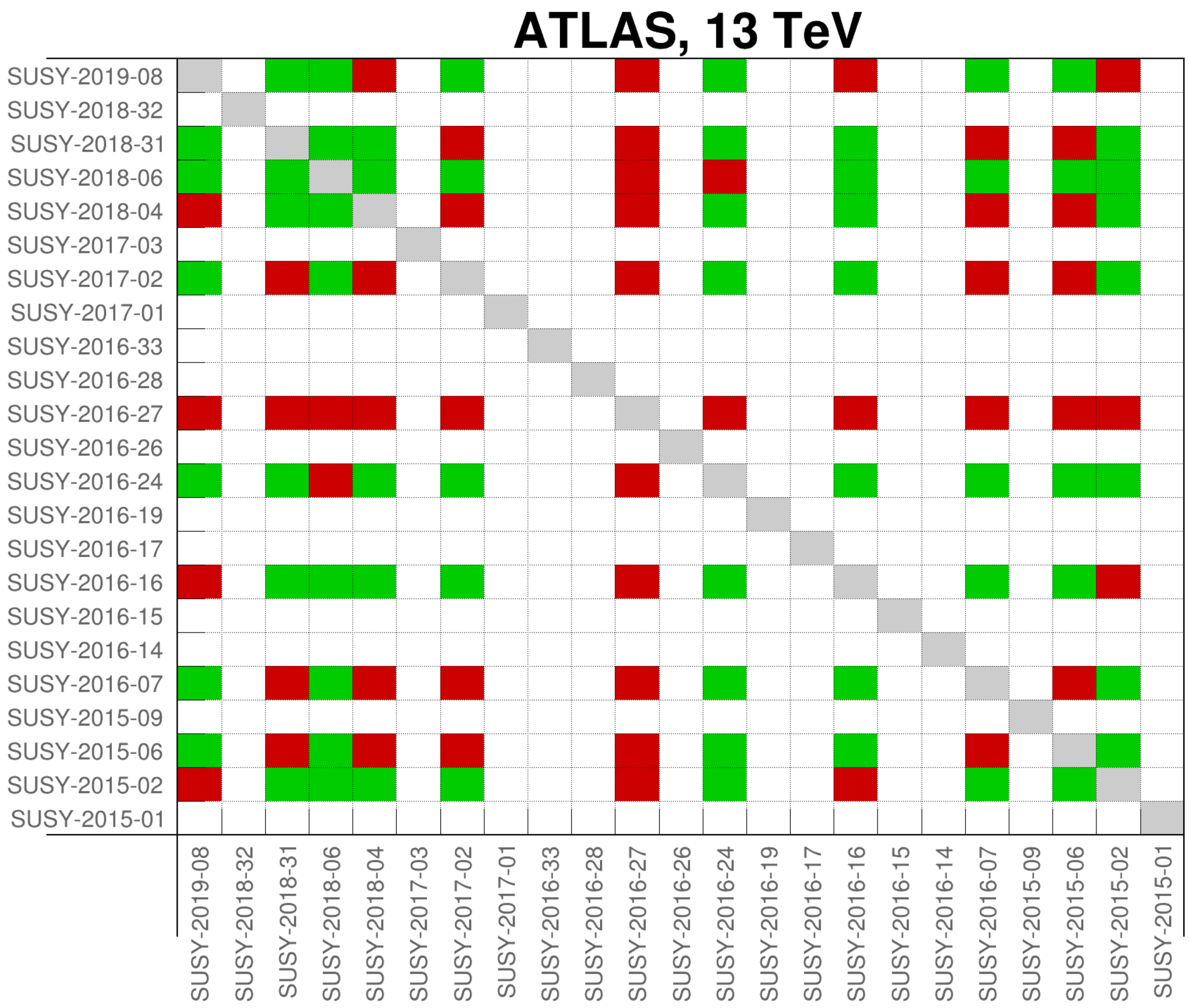
Red → correlated → not combinable

White → cannot construct a likelihood

Signal regions within each analysis → correlated

Combining Data

- As we are chasing dispersed signals, we need to combine likelihoods. We assume a simplified, binary “inter-analyses correlations matrix”:



Green → approximately uncorrelated → combinable

Red → correlated → not combinable

White → cannot construct a likelihood

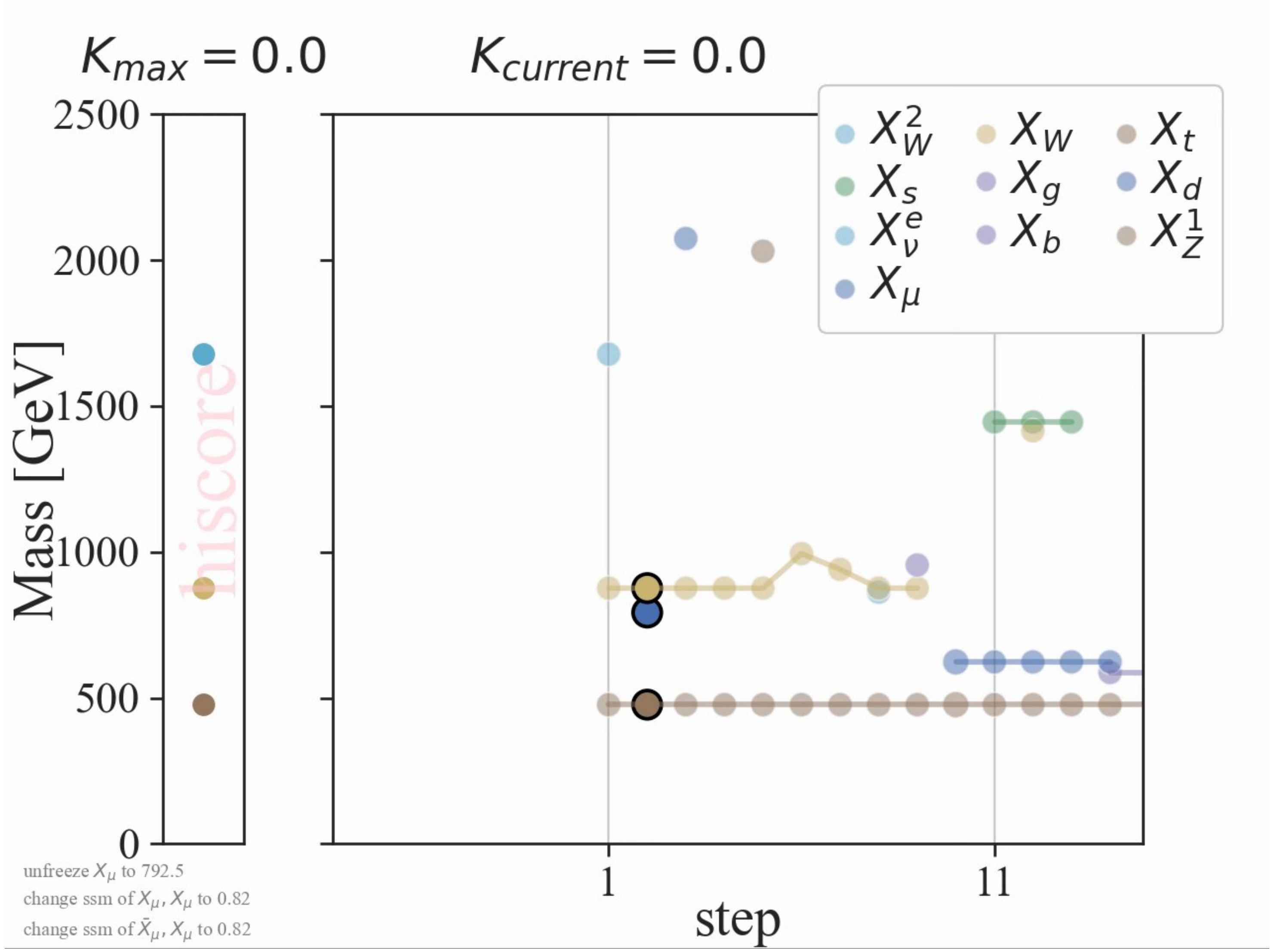
Signal regions within each analysis → correlated

Les Houches effort:

https://arxiv.org/abs/2002.12220

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

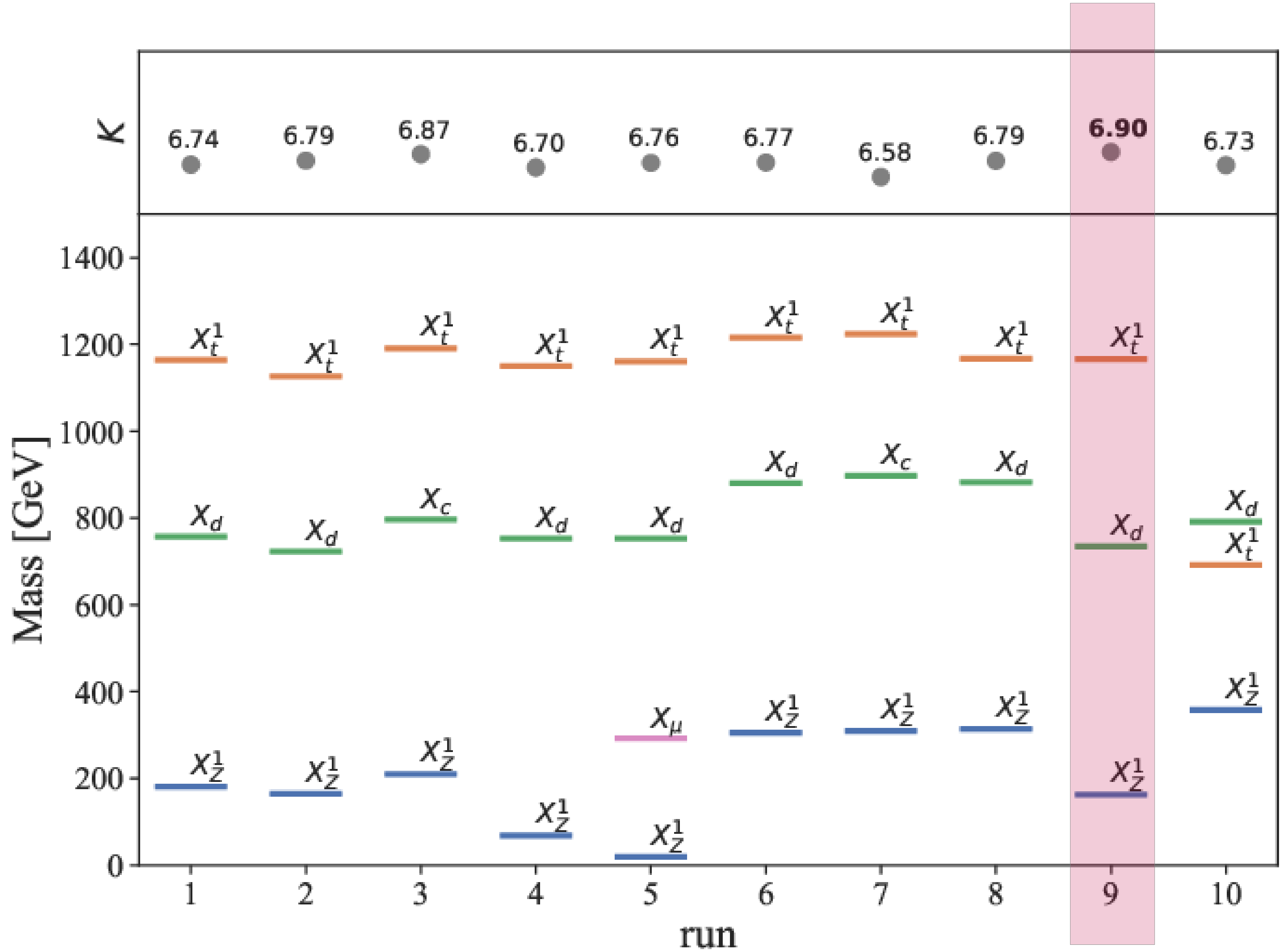
Running the algorithm...



- We defined a "run" as 50 parallel walkers, making 1,000 steps each.
- We performed 10 such runs on the SModelS database.
- We validated with simulated versions of the SModelS database, synthesized from our statistical models.
- Total computing resources spent: ~ 1,000,000 CPU hours

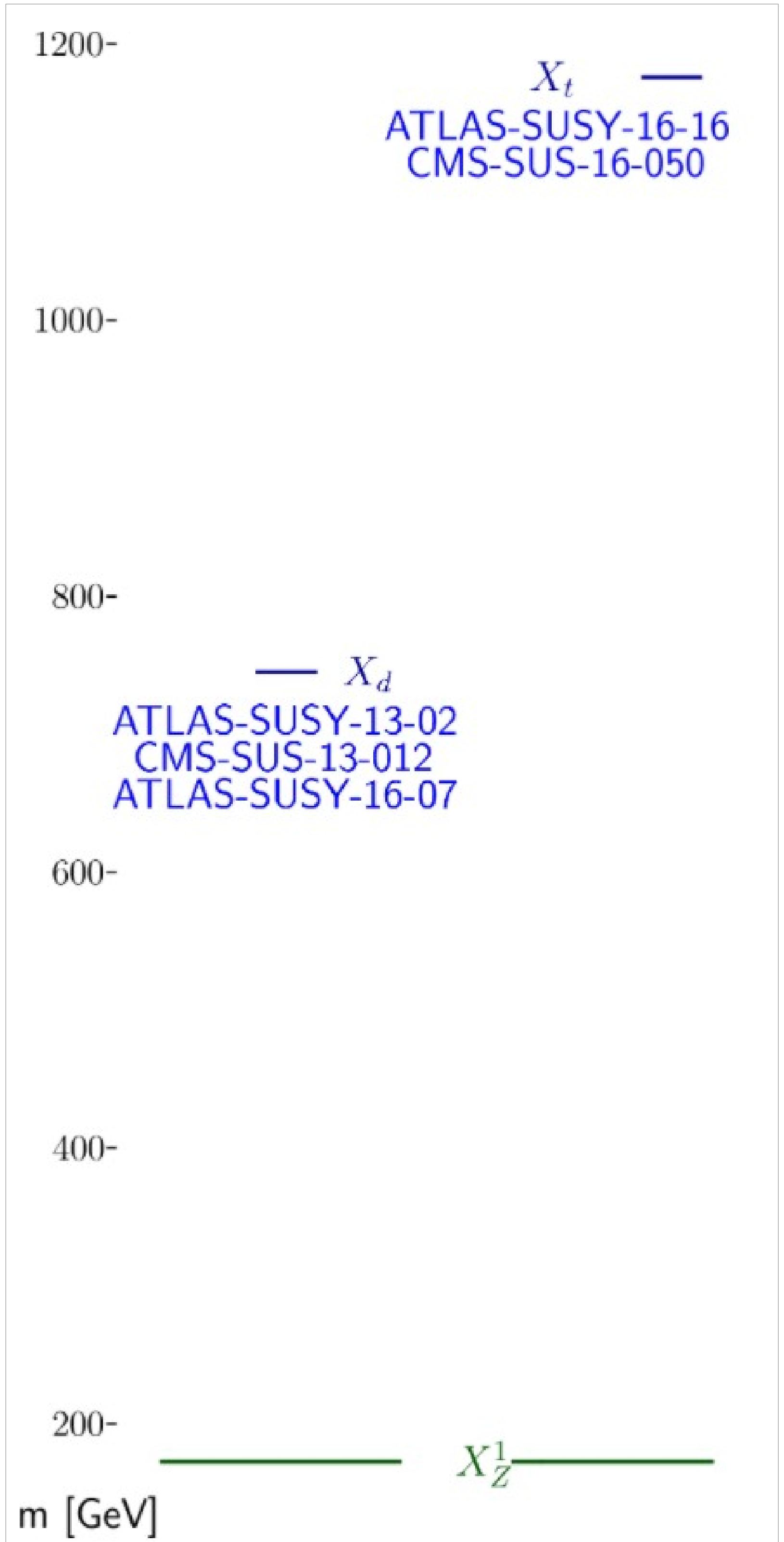
Walking over the SModelS Database

- 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 are compatible with values expected from the MSSM. The best test statistic was K=6.9.

The High Score Protomodel



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

negative results in the database

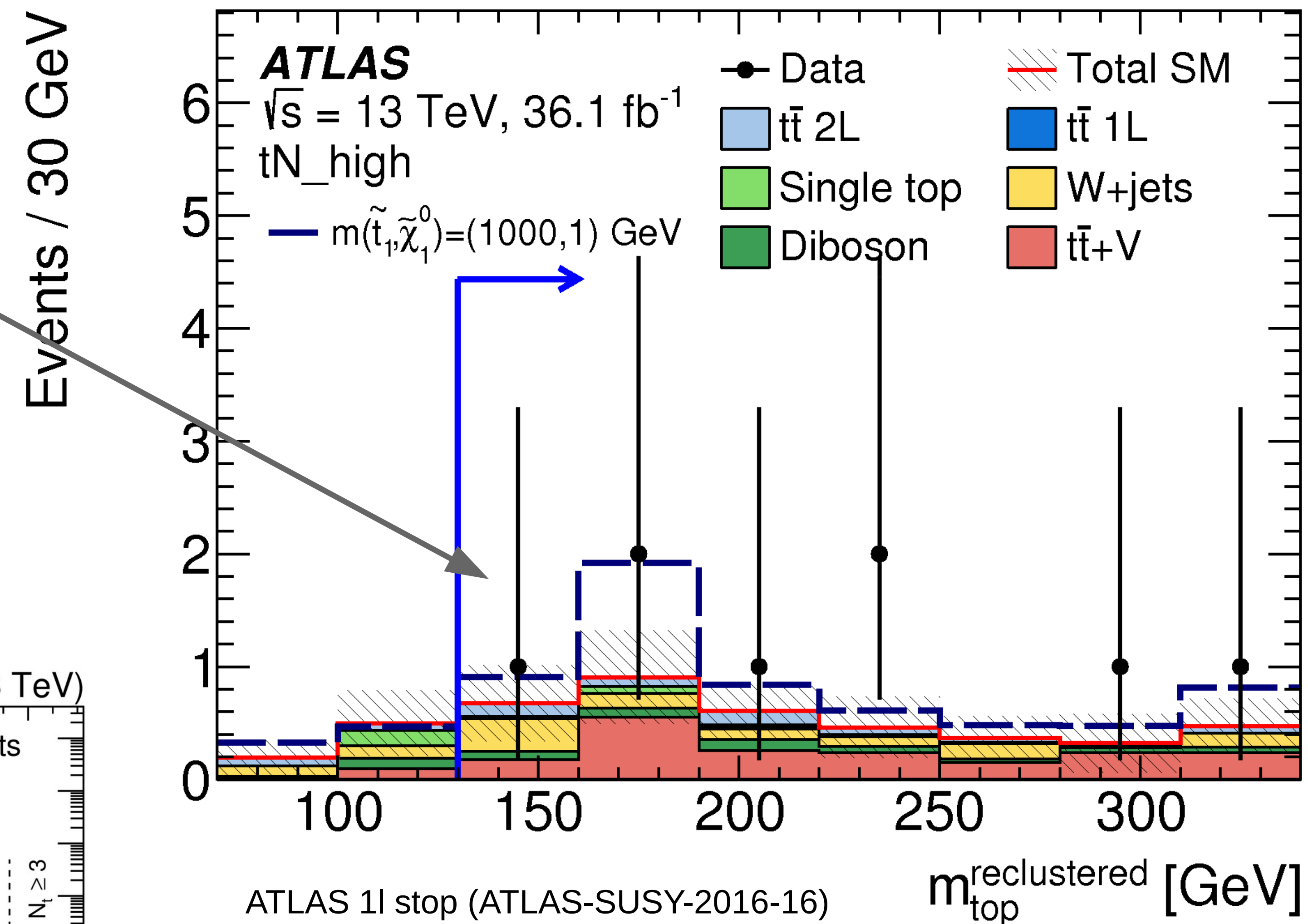
Signal strength multipliers: $(\bar{X}_t, X_t) = 1.2; (\bar{X}_d, X_d), (X_d, X_Z^1), (\bar{X}_d, X_Z^1) = 0.49$

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

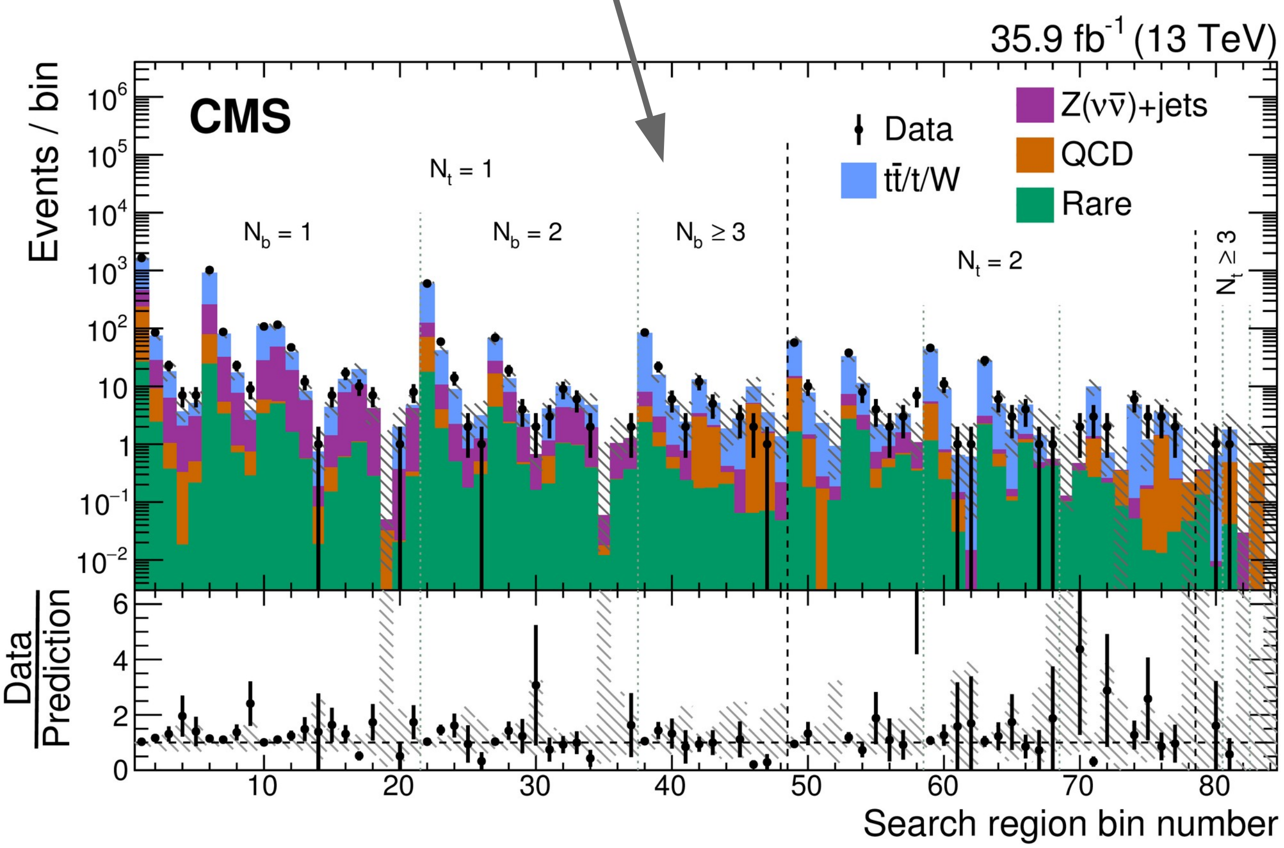
https://smodels.github.io/protomodels/2020_PioneerStudy/real9/index.html

Data driving the protomodel

Nothing unusual is visible in the individual results



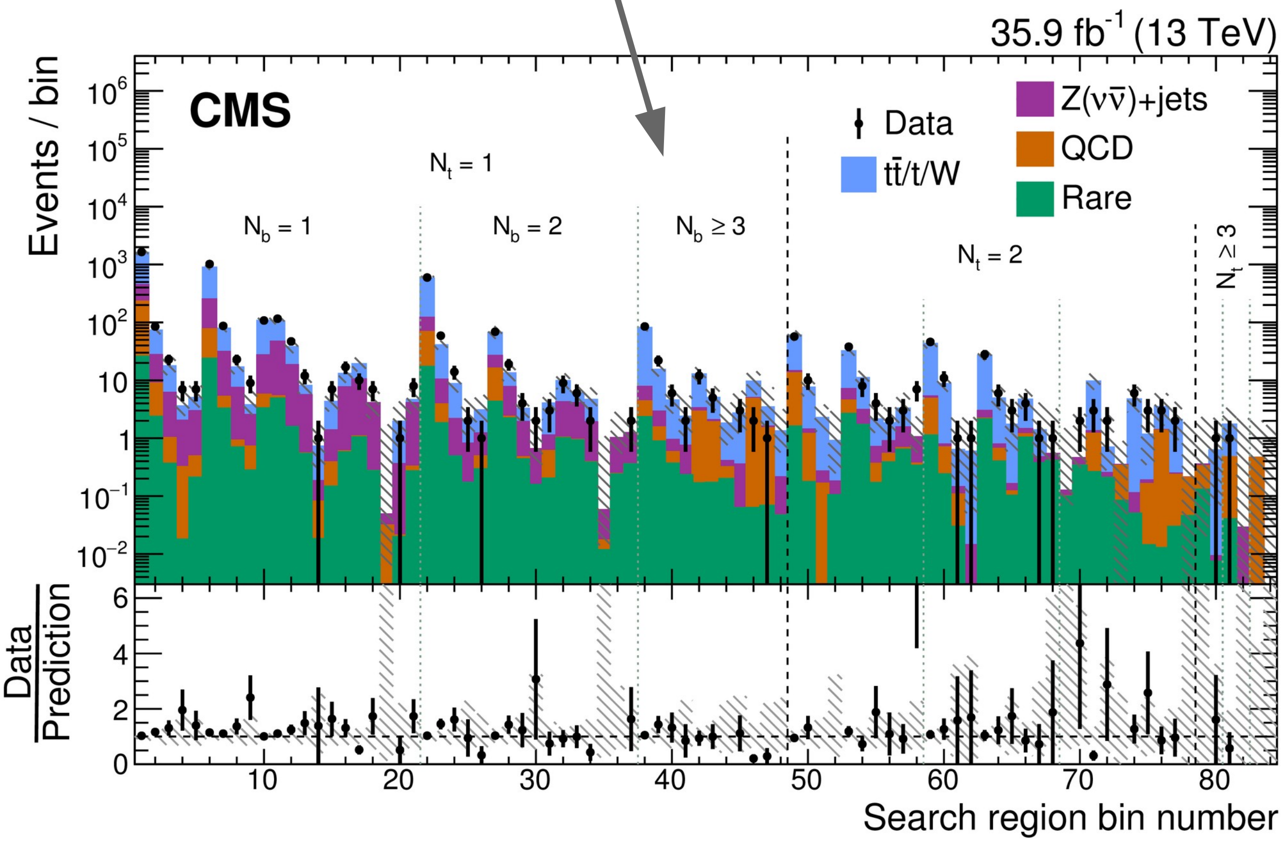
ATLAS 1l stop (ATLAS-SUSY-2016-16)



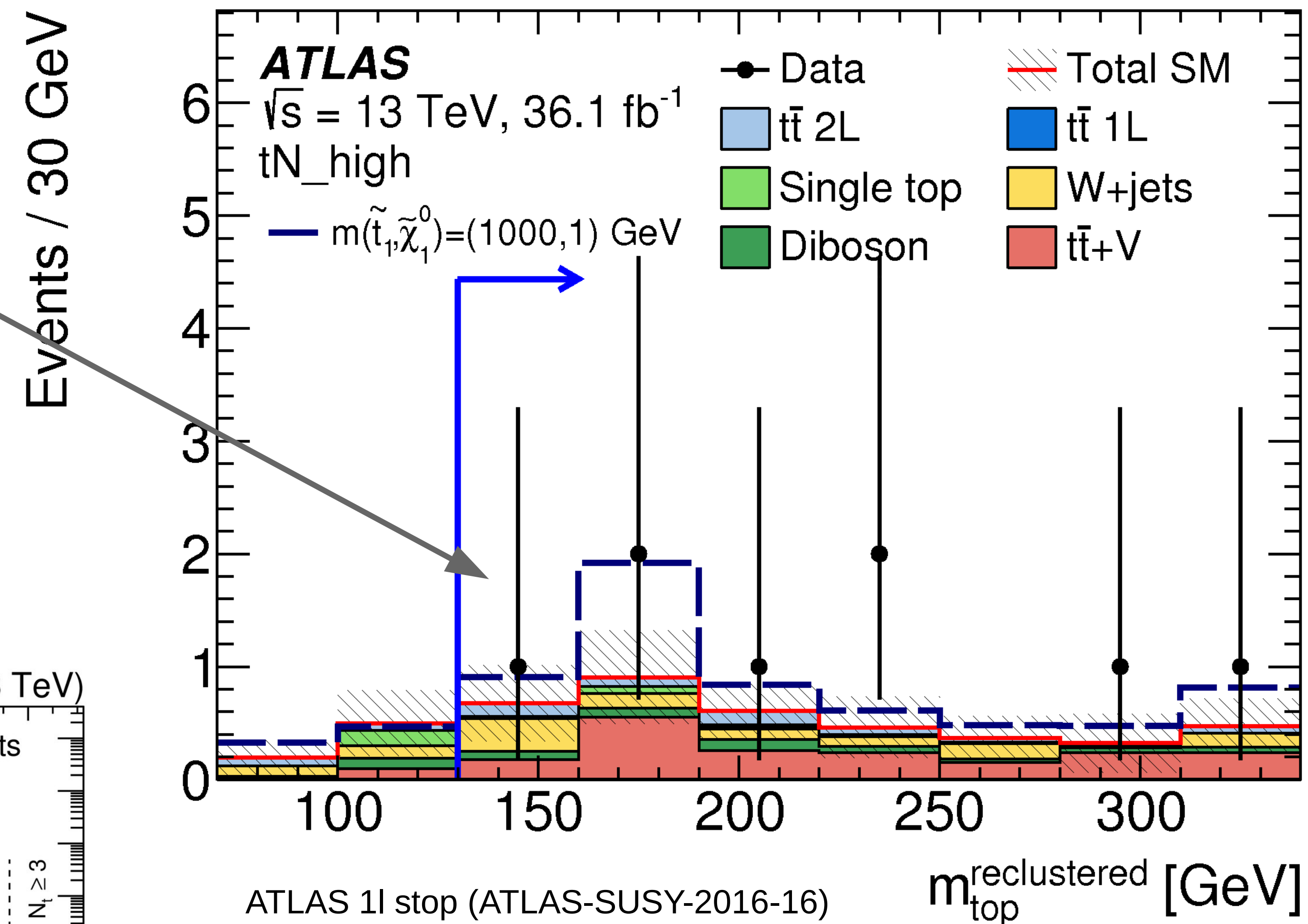
CMS 0l stop (CMS-SUS-16-050)

Data driving the protomodel

Nothing unusual is visible in the individual results



CMS 0l stop (CMS-SUS-16-050)



ATLAS 1l stop (ATLAS-SUSY-2016-16)

But when interpreted as a correlated signal from the protomodel, things may seem different!

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.

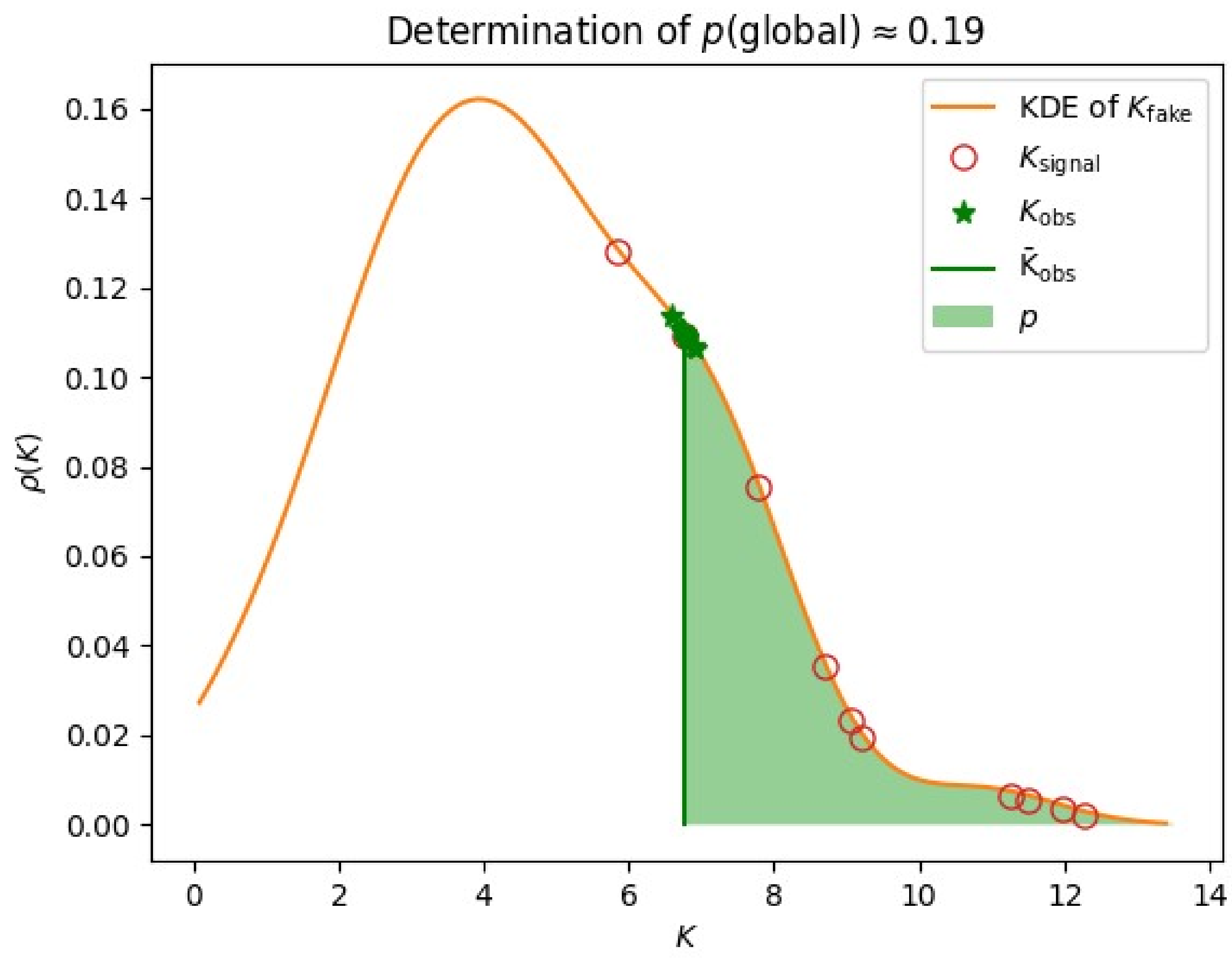
- From this we can compute a p-value for the Standard Model hypothesis: that is the chances that – under the SM hypothesis – we would obtain a result as extreme as ours or more extreme.

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.

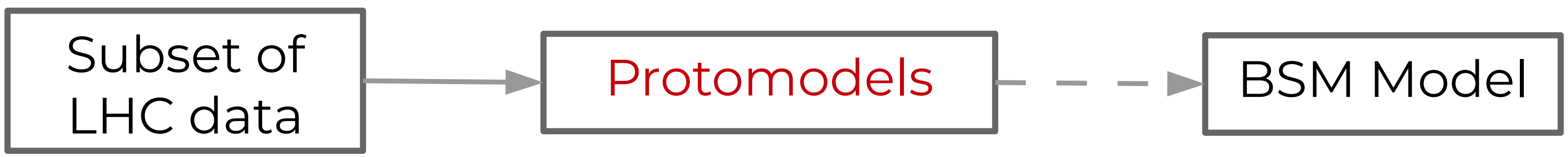
smaller p-value → SM "more excluded"



By construction, no Look-Elsewhere Effect applies.

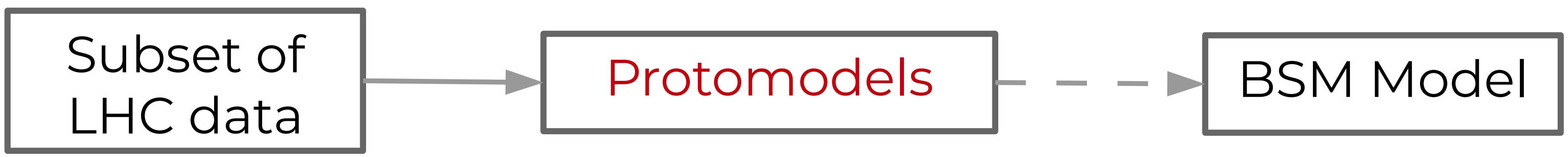
Building the UV Model

- The protomodels are an intermediate step:

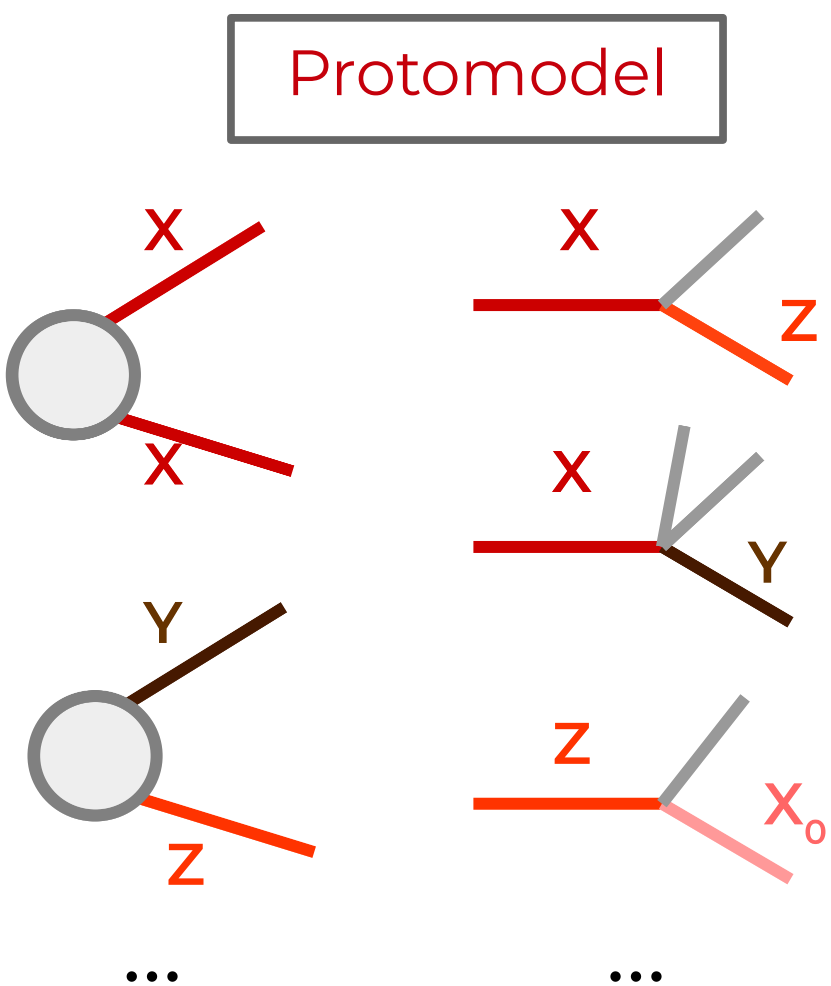


Building the UV Model

- The protomodels are an intermediate step:



- Taking the second step (future development):

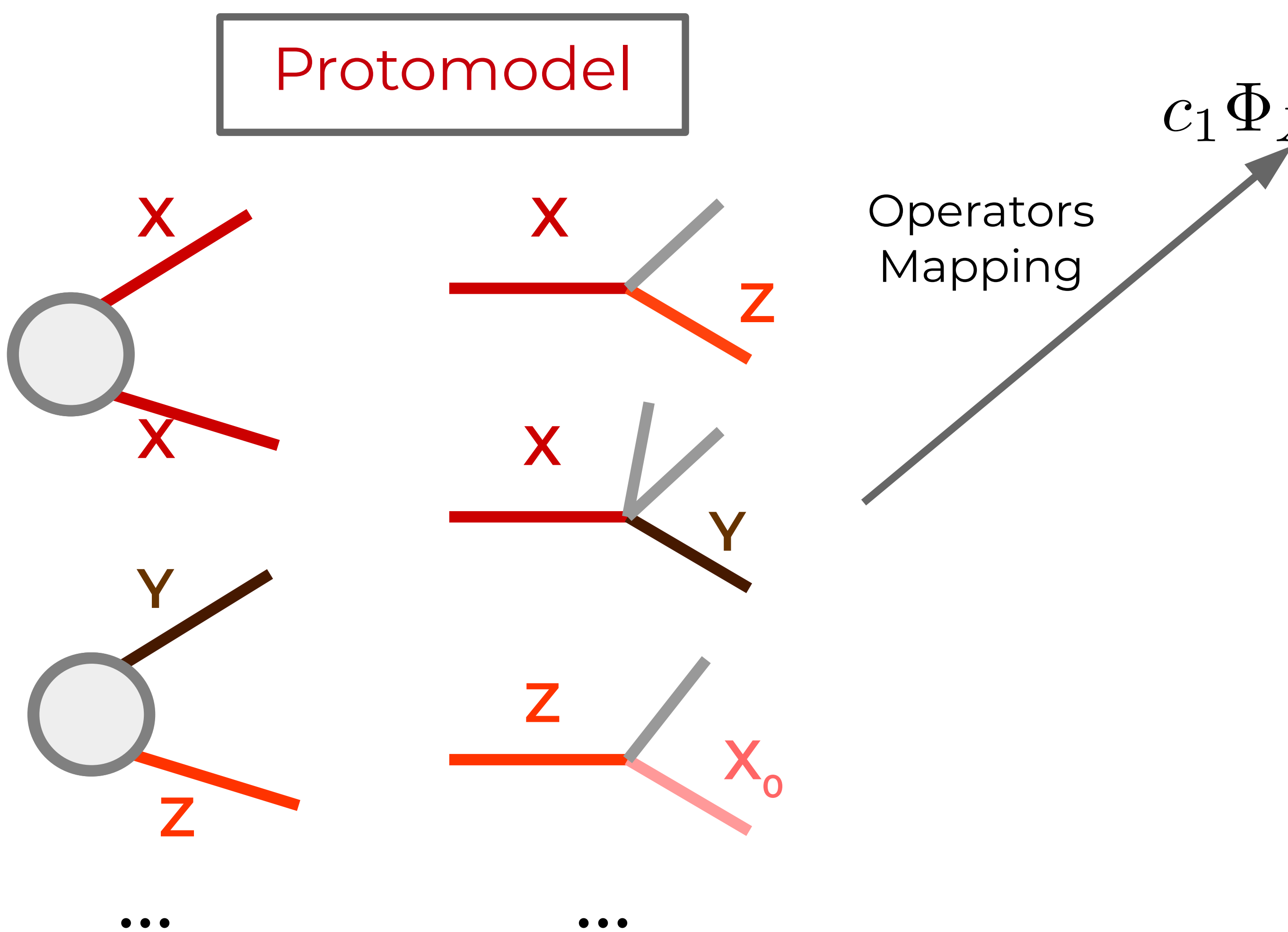


Building the UV Model

- The protomodels are an intermediate step:

Subset of LHC data → Protomodels ⇢ BSM Model

- Taking the second step (future development):



$$c_1 \Phi_X \Phi_Y u,\ c_2 \Phi_X \Phi_Y h^2,\ c_3 \Phi_X \Phi_Z h,\ \ldots$$

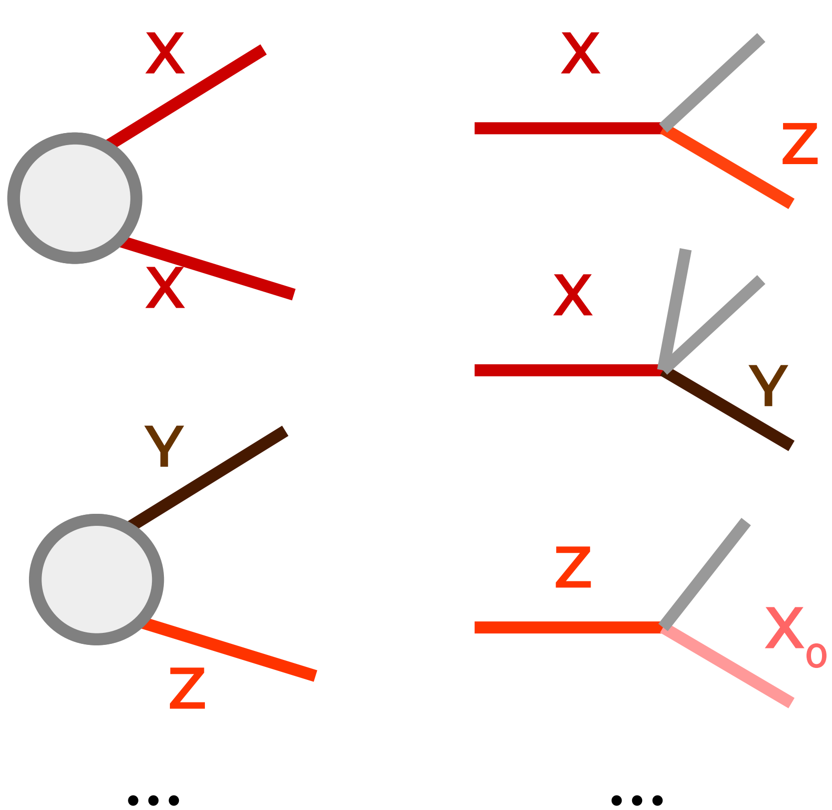
Building the UV Model

- The protomodels are an intermediate step:

Subset of LHC data → Protomodels ⇢ BSM Model

- Taking the second step (future development):

Protomodel



Operators Mapping

$$c_1\Phi_X\Phi_Y u,\ c_2\Phi_X\Phi_Y h^2,\ c_3\Phi_X\Phi_Z h,\ \ldots$$

Consistent lagrangian

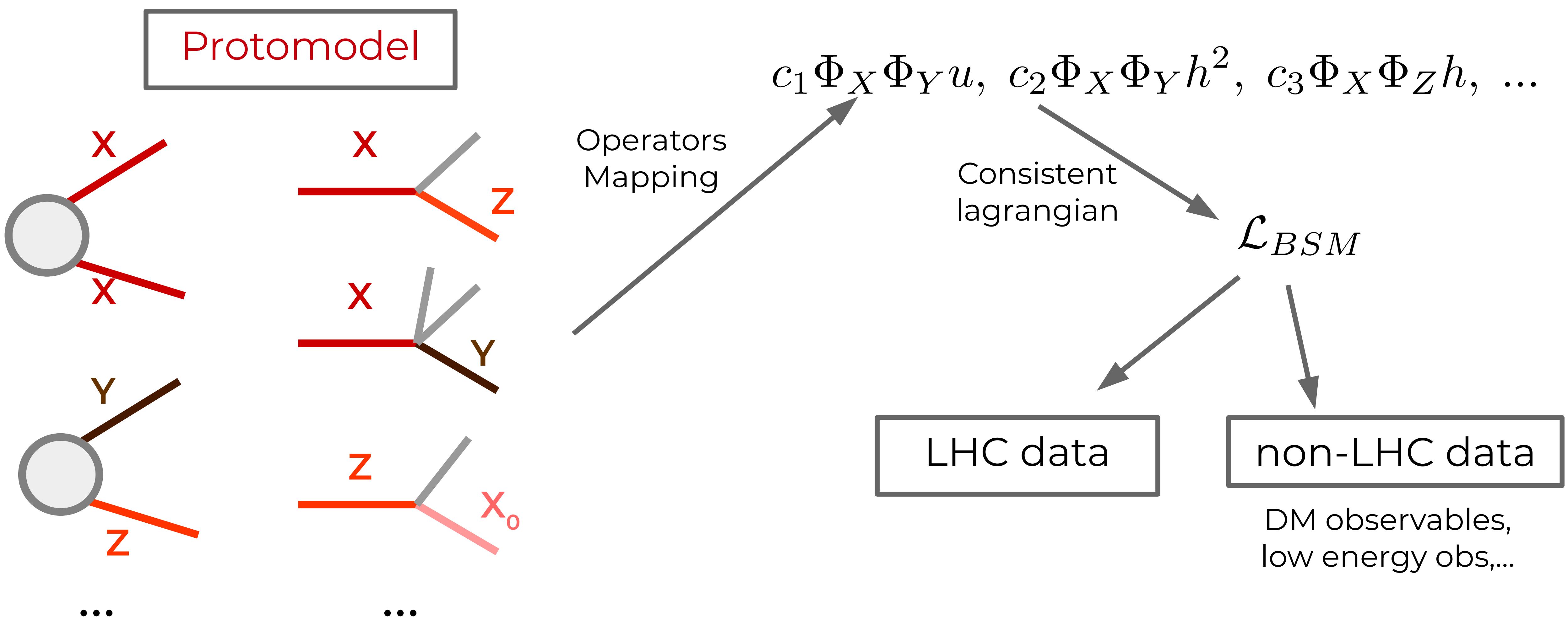
$$\mathcal{L}_{BSM}$$

Building the UV Model

- The protomodels are an intermediate step:

Subset of LHC data → Protomodels → BSM Model

- Taking the second step (future development):



Other Future Developments

- SModelS framework:
  - Extend to generalized signals (without the Z_2 restriction) using graphs
  - Include non-SUSY searches
- SModelS database:
  - Add latest full run-2 CMS and ATLAS publications
  - Learn the database
- Statistical calculation:
  - Move to more complete statistical models (e.g. via pyhf)
  - Improve analyses combination matrix
  - Learn likelihoods

Thank you!

Work funded by:

- joint French-Austrian fund
  FWF - I 5767 47045 and ANR-21-CE31-0023
- IN2P3 master project "Théorie – BSMGA"
- AL's "invited professor" by CPTGA
- WW's "invited professor" by UGA, Enigmass

Backup

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 Test Statistic

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

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

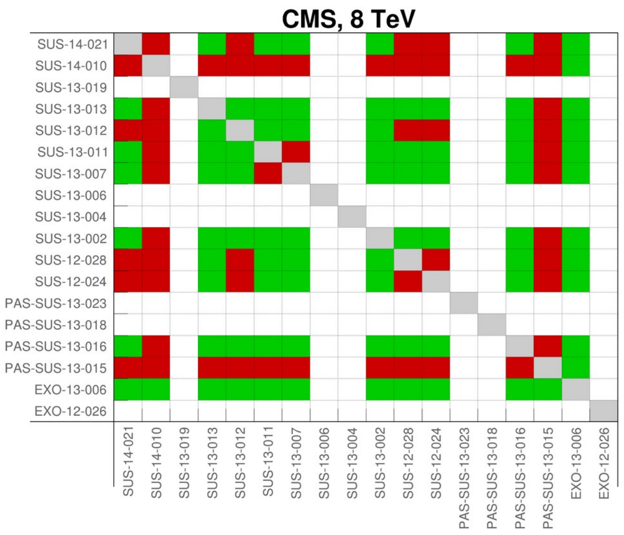
π(BSM) is the prior of the BSM model. We use it to “regularize” the model, i.e. impose the *law of parsimony*:

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

Resulting in a test statistic that resembles an “Akaike information criterion”:

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

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



The Combiner

we allow the machine
to combine likelihooods.

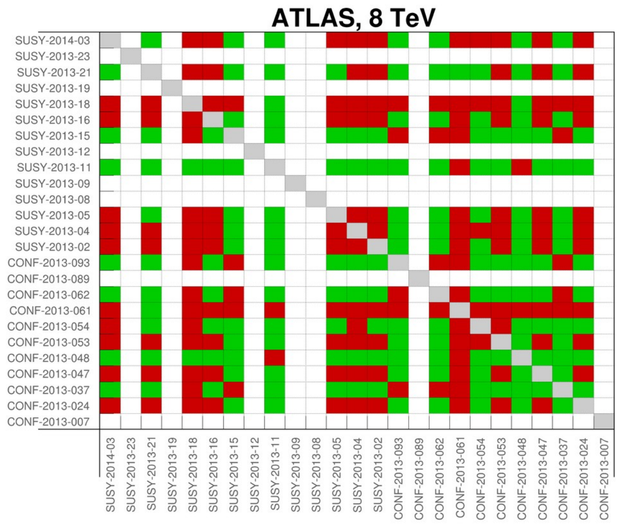
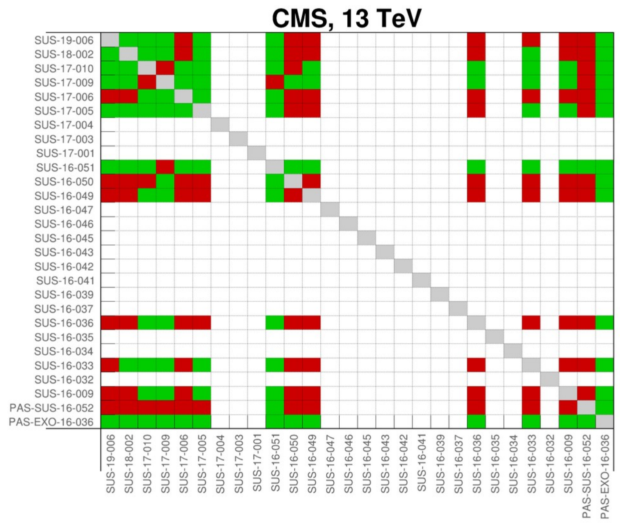


Fig. 2

Approximately uncorrelated are analyses that are:

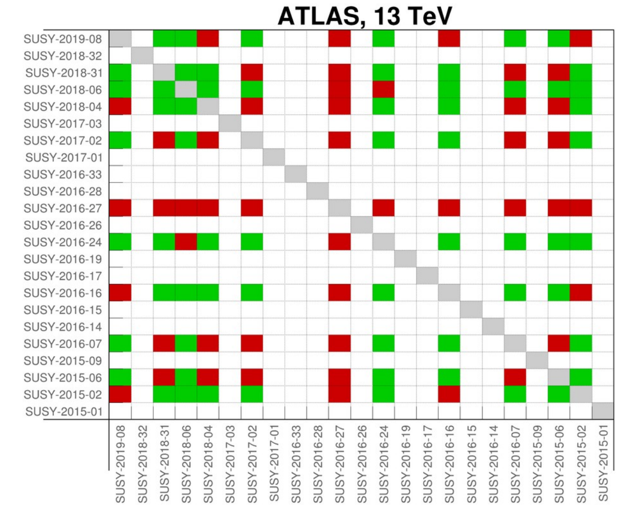
- from different runs, and/or
- from different experiments, and/or
- looking for (clearly) different signatures



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 r nost 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

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

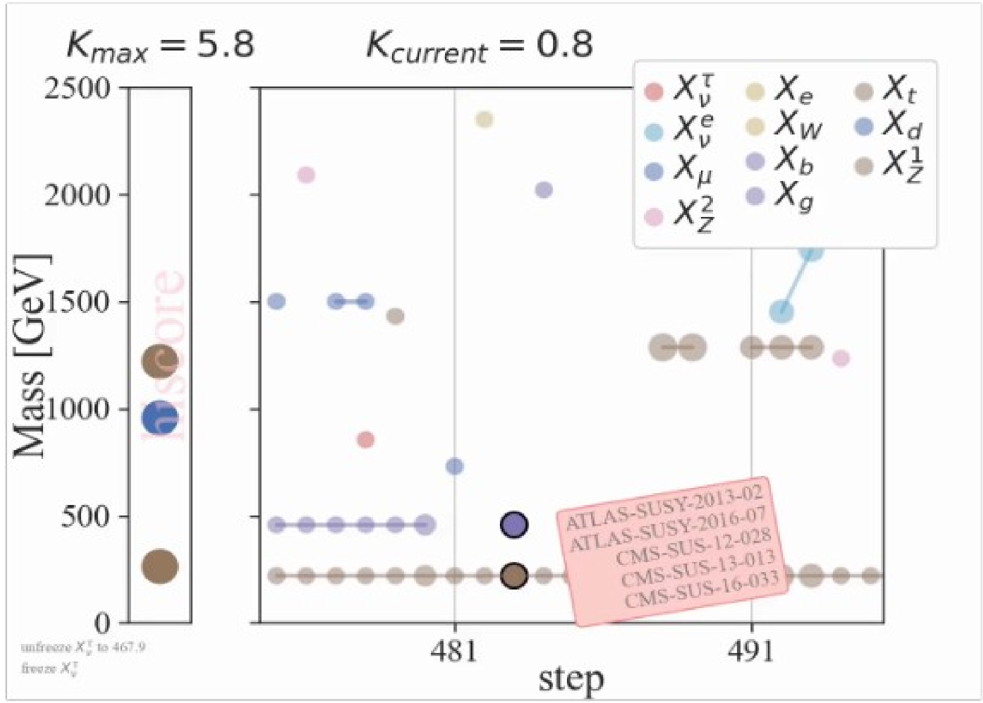
It is maximized in the denominator, but its support is confined such that no limits in the SModelS database are violated (the “critic”),

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

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 M … ed in the sense that the step is reverted with a probability of

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

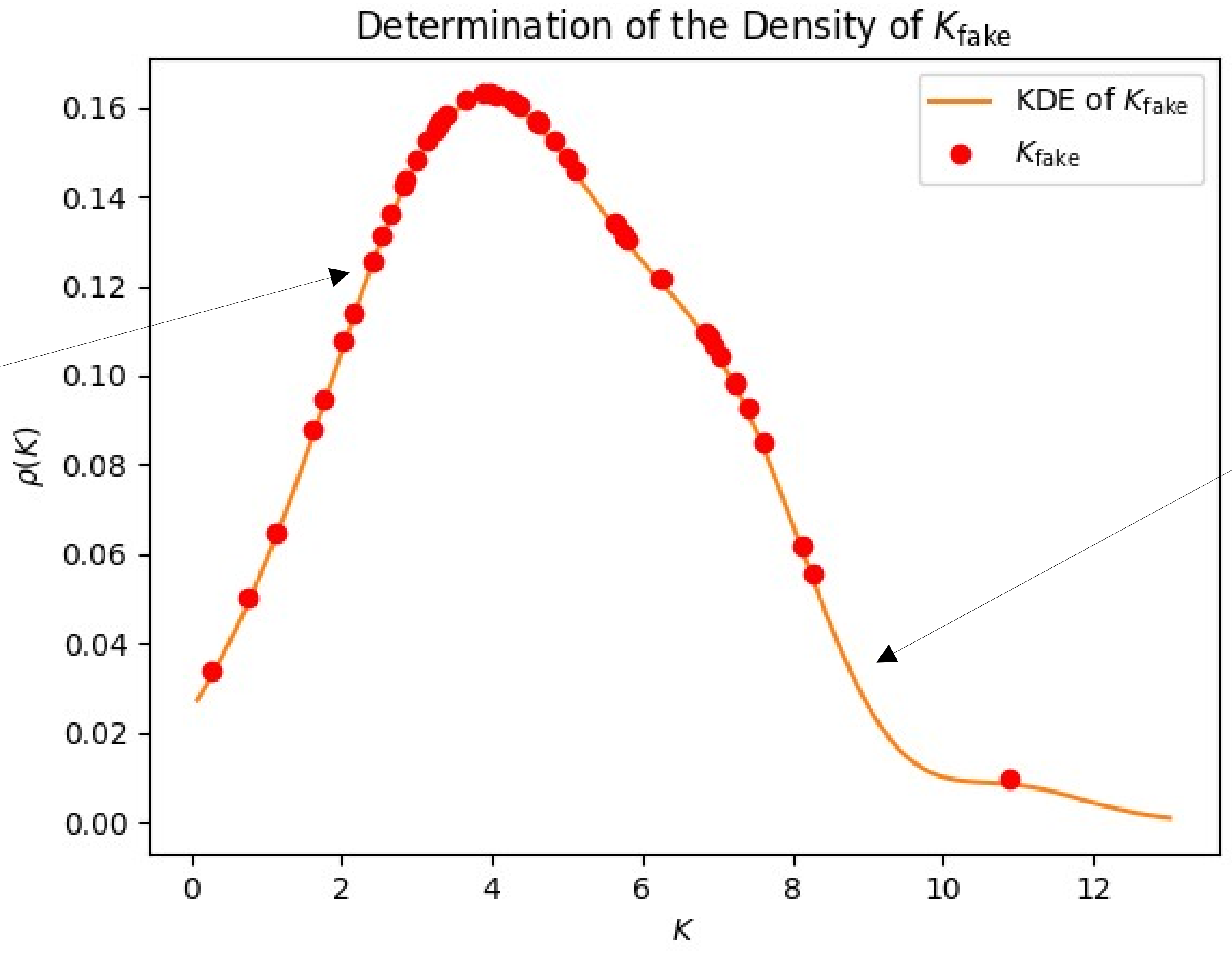
if and only if K_i is smaller than K_{i-1}

* (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

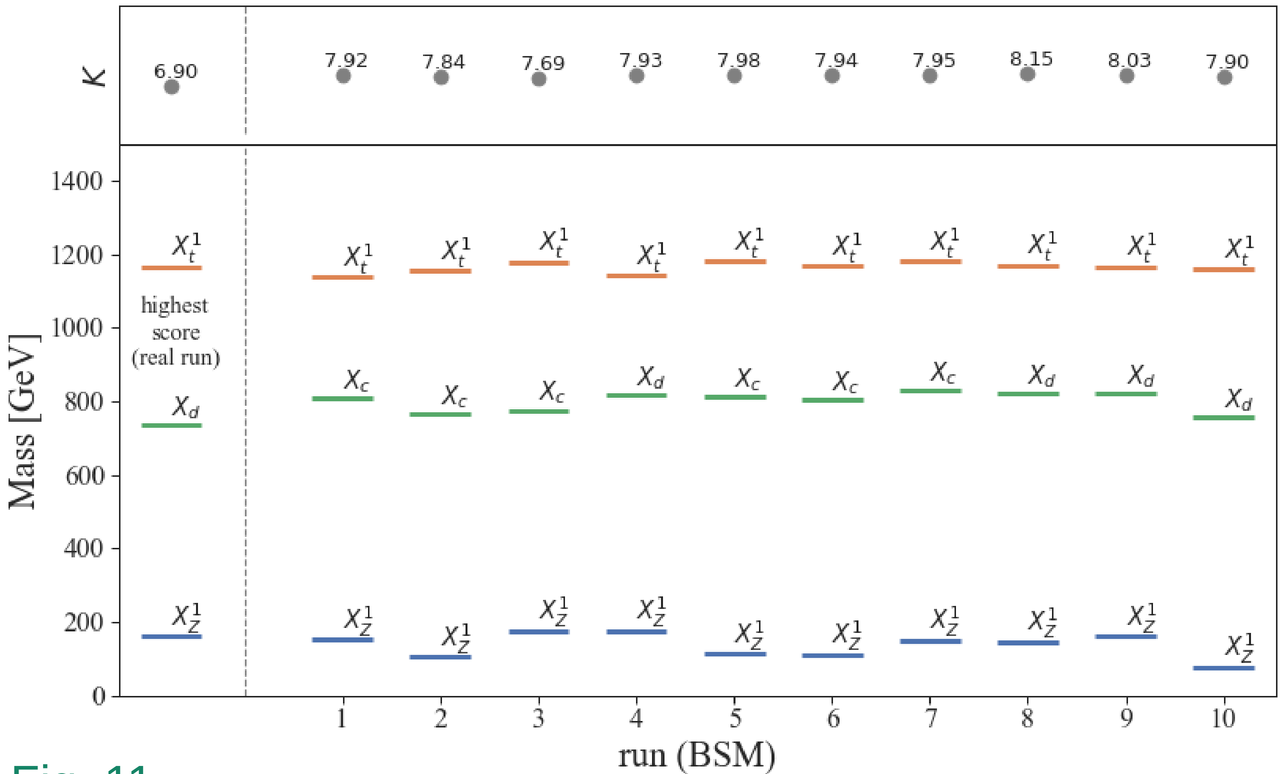
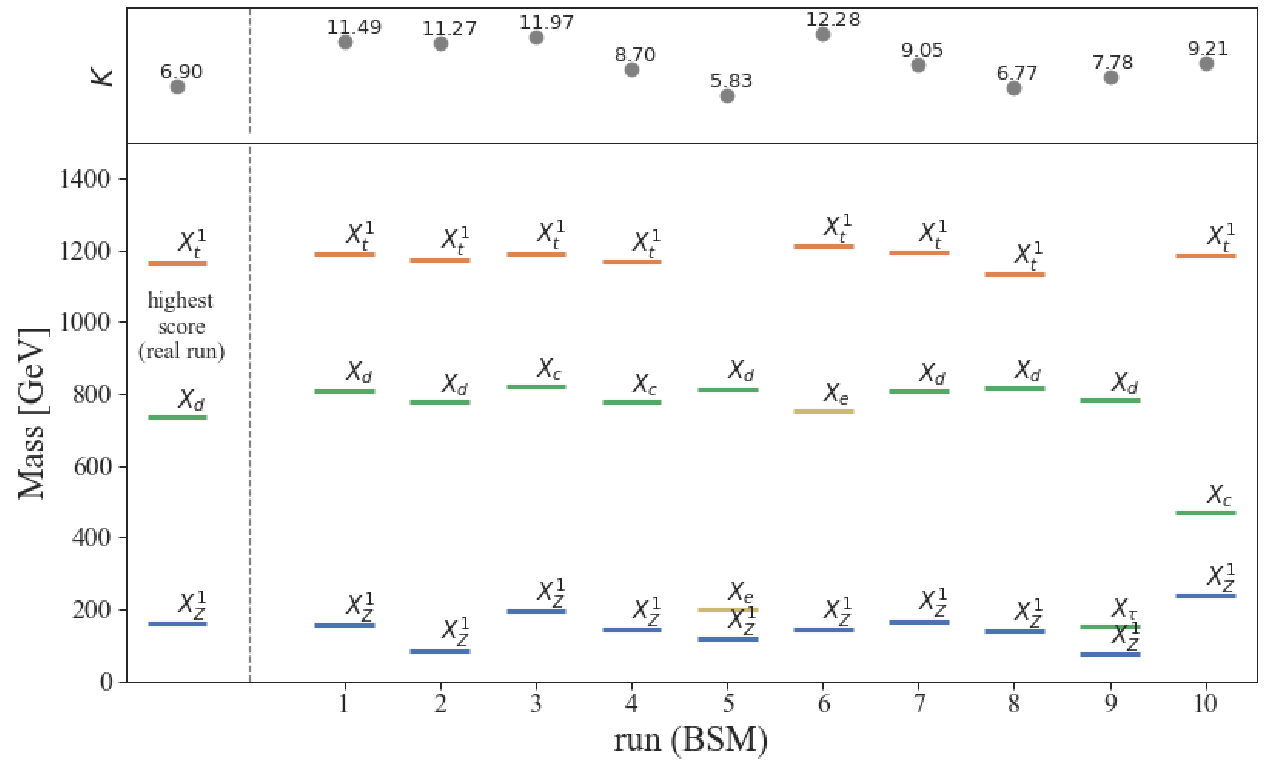


Fig. 11

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

Physics closure test

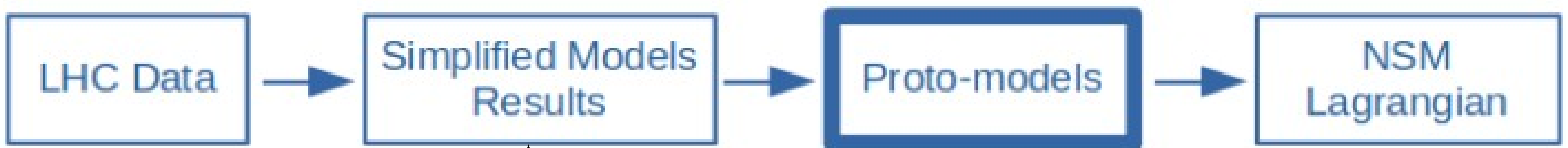


Sampling turned on

Fig. 10

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" [*]:

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

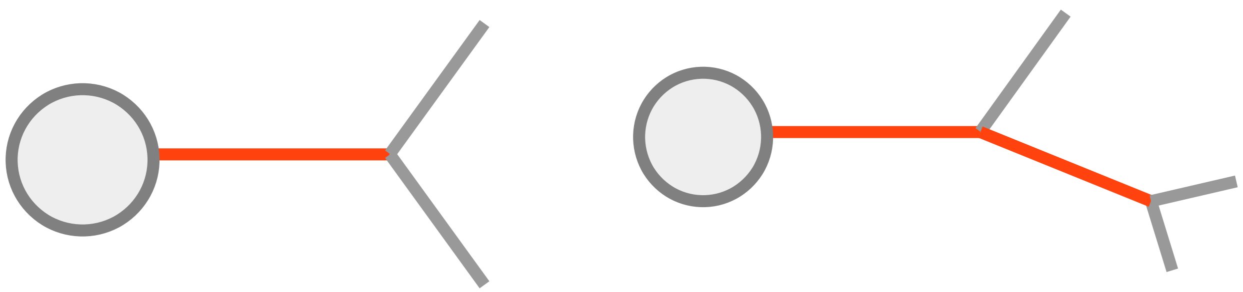
Needless to say, the data pipeline sketched above is not the only feasible one. Differentiability however would be a helpful tool for all possible data pipelines. A similar rationale would apply also to EFTs, Wilson coefficients and data from measurements.

https://arxiv.org/abs/1810.08312

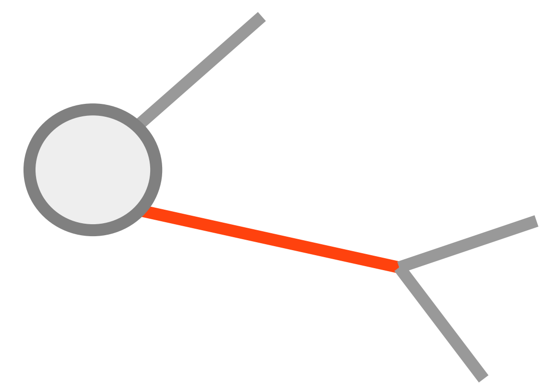
→ Differentiable Inductive Reasoning!

Generalized Topologies with Graphs

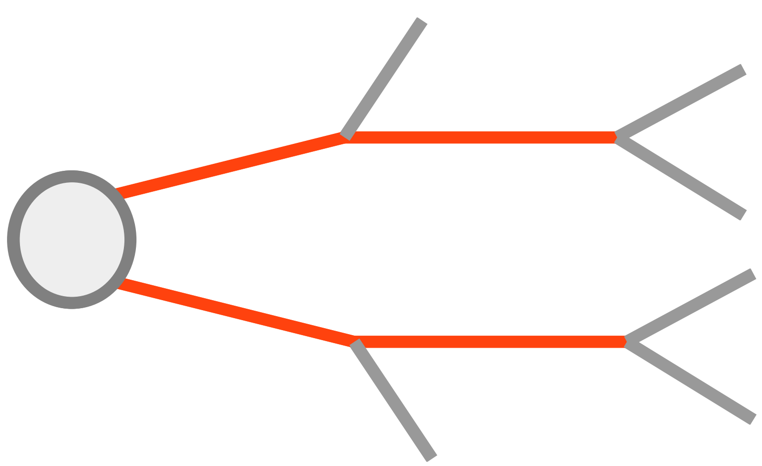
- Describe signal topologies with arbitrary shapes, such as:
  - Resonant production



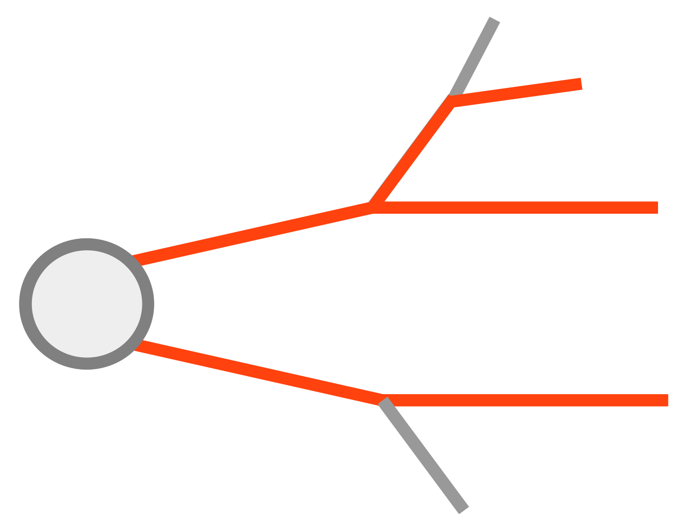
  - Associated Production



  - R-Parity Violating Decays



  - Non-Z_2 decays



...