Systematics in Searches

Event Selection, Limits, Discovery

Lukas Heinrich, CERN - PHYSTAT 2021

Goals for the talk

I will try to summarize

- how we build the statistical model at the LHC
  esp. how we incorporate nuisance parameters
- describe the current practice for statistical tests
- discuss common diagnostics and checks & point out some assumptions
- focus on frequentist viewpoint as its dominant at LHC

Caveat: slight bias towards ATLAS & Supersymmetry (SUSY) searches

LHC Stats largely has a fairly routine & well-established lived practice
- meeting like this great way to review our approach

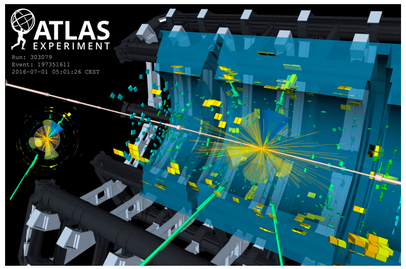
Highlevel View

Particle Physics is likelihood-free: $p(\text{data}\,|\,\theta)$ not available in closed form

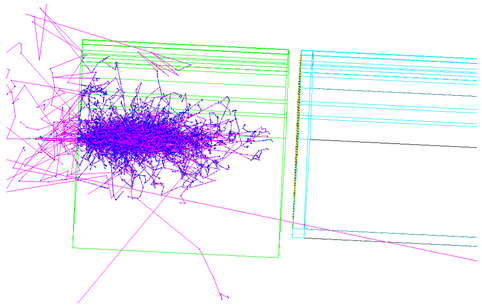
Deep & large graphical model :

- **very large latent space to be integrated over and huge observation space**
- **Inference on raw data space intractable**

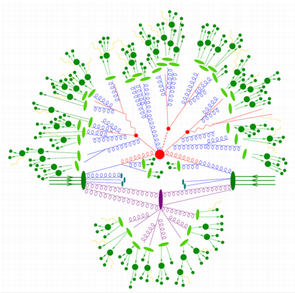
$$p(\text{data}\,|\,\theta) = \iiint \mathrm{d}z_d \mathrm{d}z_h \mathrm{d}z_p \; p(\text{data}\,|\,z_p) \;\; p(z_d|z_h)\, p(z_h|z_p)\, p(z_p|\theta)$$



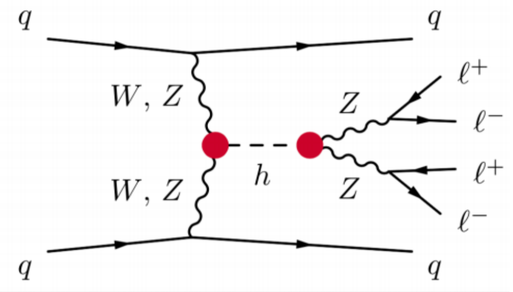
detector readout



detector interaction

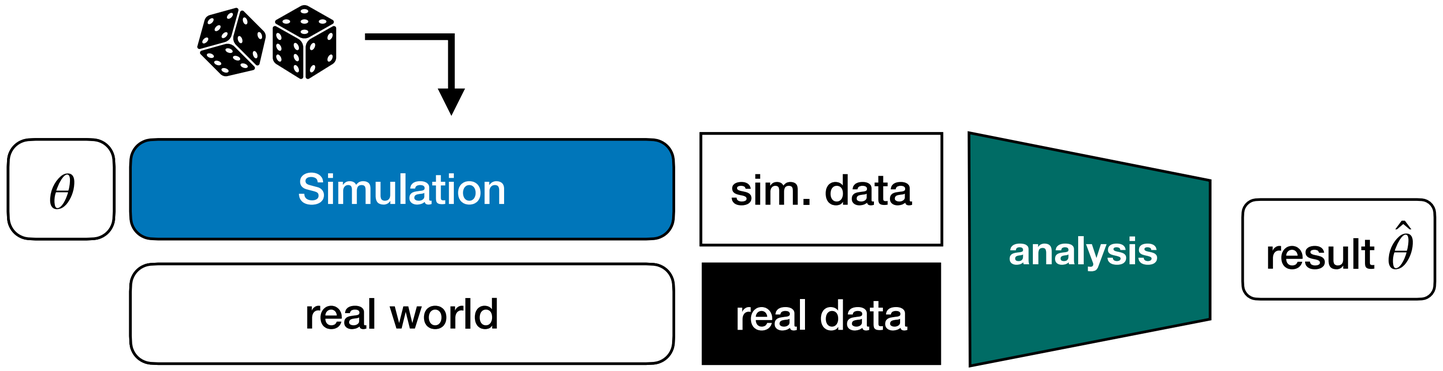


stochastic evolution



hard process

Highlevel View



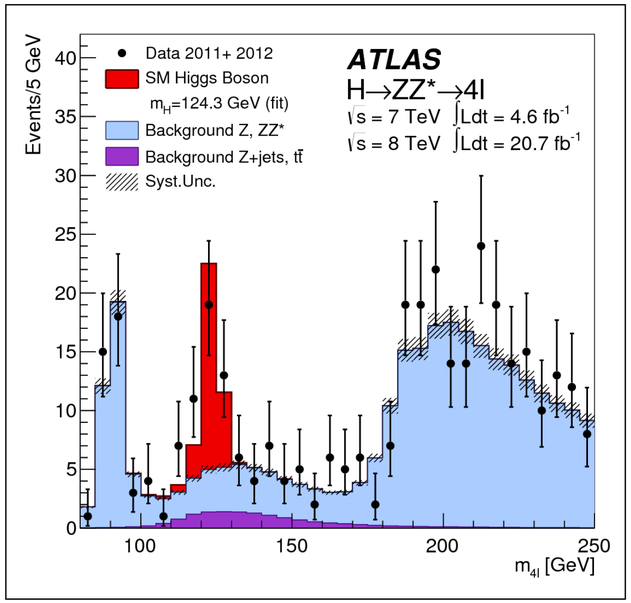
Two Solutions:

- **Simulation**: at least we sample from $p(\text{data}\,|\,\theta)$
  (lots of software: matrix elements, event generators, detector simulation)

  - can build simulation-driven density estimates from samples
  - lots of tunable (nuisance) parameters
  - computationally expensive O(days per theory point)
  - Quality of simulation not the same for all data
    - need to focus on subspace of data
    - or develop data-driven estimates

← Systematics!

Highlevel View



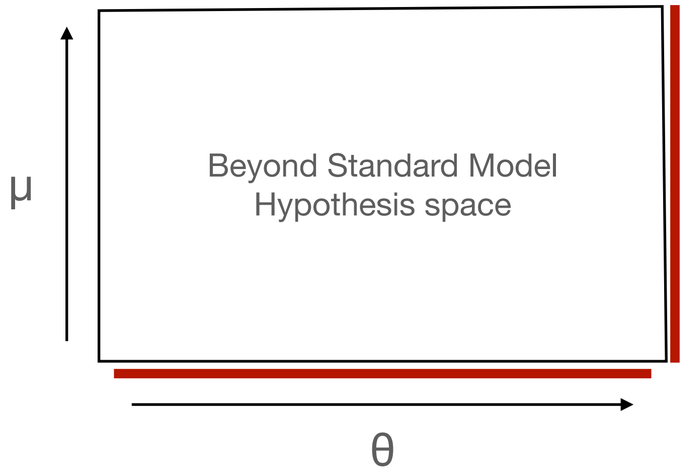
- **Analysis via summary statistics:** $p(x = t(\text{data}), \theta)$
  (even more software: reconstruction, analysis, ...)
  - summaries (observables) reduce dimensionality
    - often physics motivated: e.g. invariant masses (often 1-D)
  - $t(\text{data})$ is itself has many more nuisance parameters ← Systematics!
  - Curse: less sensitivity (data processing inequality
  - Blessing: also allows you tune likelihood for desiderata
    - e.g. robustness against NPs, shape of inference results

Nested Models

Want to study data with respect to (continuous set of) BSM Hypotheses

- **parametrized by new physics (NP) parameters**
- **often overall rate ("signal strength") a parameter**

$$\mathrm{BSM} = \mathrm{SM}(\nu) + \mu\mathrm{NP}(\theta, \nu)$$



Example:

- **pMSSM space (19 masses / couplings, etc)**
- **simplified models parametrized by masses**

The SM is part of this extended SM space

- **nested models make e.g. Wilk's theorem applicable**
- **discussion point: SM is not a point, but extended subset of the BSM space**
  - **in the absence of discoveries inference often on boundary**

Most of the effort: Building a model of the (summary) data

In our model we split parameters into $p(x \mid \theta) = p(x \mid \mu, \nu)$

- parameters of interest μ
- nuisance parameters ν (i.e. "systematics")

Often (but not always!): some external infomation w.r.t. nuisance parameters

- someone else did analysis of disjoint data more sensitive to value of ν

Goal 1: need strategy to incorporate external information into our analysis

Nuisance Parameters: not primary target of inference

Goal 2: want strategies to excise NPs from our inferences

Incorporating Information

Consistent Bayesian & Frequentist treatment of external information

- **make source manifest as its own probability model and underlying data**

Bayesian: resolve prior into "subsidiary imeasurement" and "ur-prior"

$$p(\mu,\nu\,|\,x,a) \sim p_{\text{main}}(x\,|\,\mu,\nu)\,p_{\text{subs}}(\nu\,|\,a)\,p(\mu) \sim p_{\text{main}}(x\,|\,\mu,\nu)\,p_{\text{subs}}(a\,|\,\nu)\,p(\nu)\,p(\mu)$$

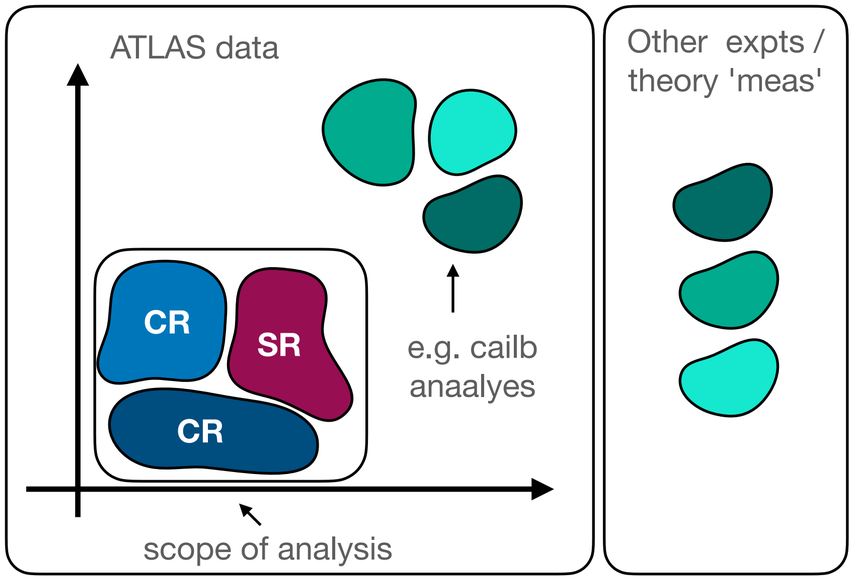
Frequentist: model is combination of main and subsidiary measurements

$$p_{\text{main}}(x,a\,|\,\mu,\nu) = p_{\text{main}}(x\,|\,\theta,\nu)\,p_{\text{subs}}(a\,|\,\nu)$$

Modelling $p_{\text{main}}(x,a\,|\,\mu,\nu)$ accomodates both inference styles

Actual Subsidiary Measurements vs Proxy Model

When modelling $p_{\text{main}}(x, a \,|\, \mu, \nu) = p_{\text{main}}(x \,|\, \theta, \nu) p_{\text{subs}}(a \,|\, \nu)$ need to decide how to handle subs. measurement



- **"Near-field" systematics (analysis specific NPs)**
  - **in-situ measurement $p_{\text{subs}}(a \,|\, \nu)$ as part of the analysis ("control regions")**
  - **new category of "excellent" in Sinervo rubric?**

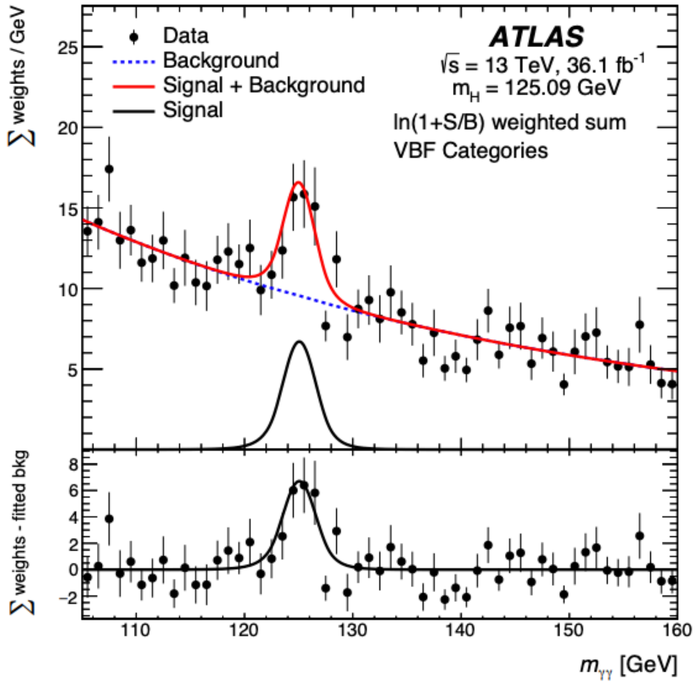
- **"Far-field" systematics (experiment-global, theory, ...)**
  - **$p_{\text{subs}}(a \,|\, \nu)$ either too complicated or not available**
  - **model with "simplified" $\hat{p}_{\text{subs}}(a \,|\, \nu)$: analogous with "easy" priors**
  - **common assumption all subsidiary measurements independent**

Probability Model as the user sees it

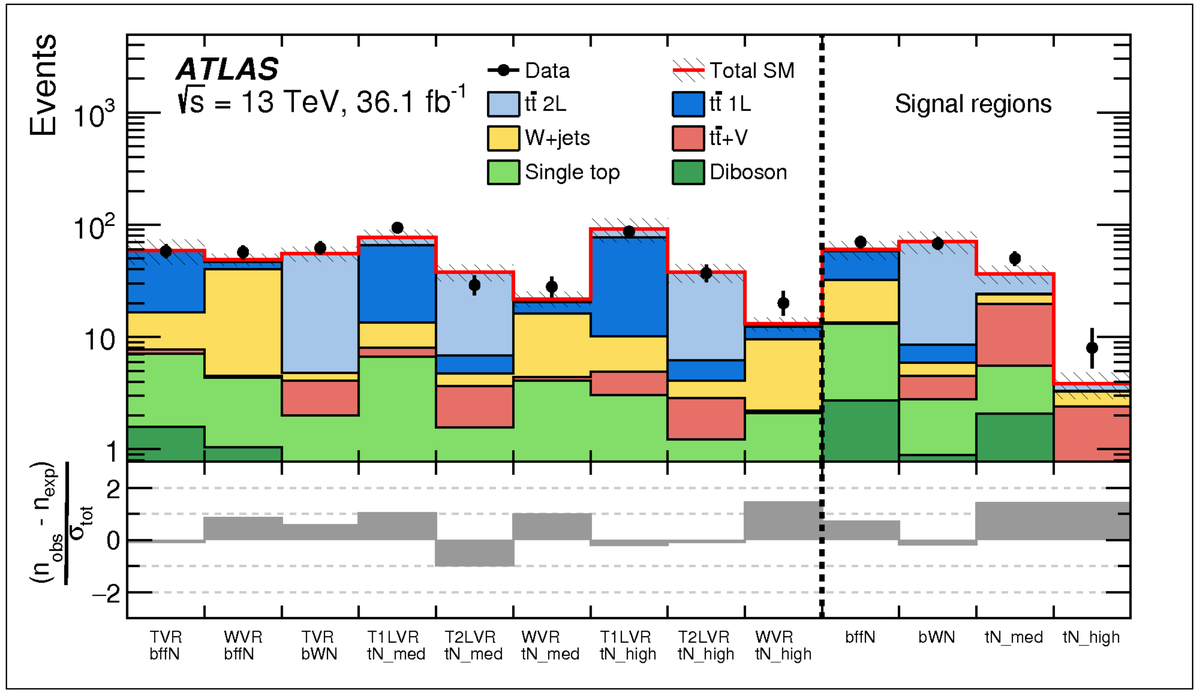
So focus on building a model of low-dimensional summary x: $p(x \mid \theta)$

Two broad styles:

will focus on this



Unbinned with (functional) form
binning just for viz



Binned with Simulation Templates

A successful template for binned analyses:

HistFactory summarizes the approach well:

$$f(\boldsymbol{n}, \boldsymbol{a} \mid \boldsymbol{\eta}, \boldsymbol{\chi}) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}\left(n_{cb} \mid \nu_{cb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{Simultaneous measurement of multiple channels}} \underbrace{\prod_{\chi \in \boldsymbol{\chi}} c_\chi(a_\chi \mid \chi)}_{\text{constraint terms for “auxiliary measurements”}},$$

$$\overline{\nu_{cb}\left(\boldsymbol{\phi}\right)} = \sum_{s \in \text{samples}} \nu_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) = \sum_{s \in \text{samples}} \underbrace{\left(\prod_{\kappa \in \boldsymbol{\kappa}} \kappa_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{multiplicative modifiers}} \underbrace{\left(\nu^0_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) + \sum_{\Delta \in \boldsymbol{\Delta}} \Delta_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{additive modifiers}}$$

nominal rates

Total estimated rate / yield func of POIs & NPs

sum over components in mixture mode

"systematics"

[pyhf docs]

Serves vast majority of e.g. ATLAS SUSY results. Used as well in e.g. Belle-II, EIC, pheno BSM studies, ...

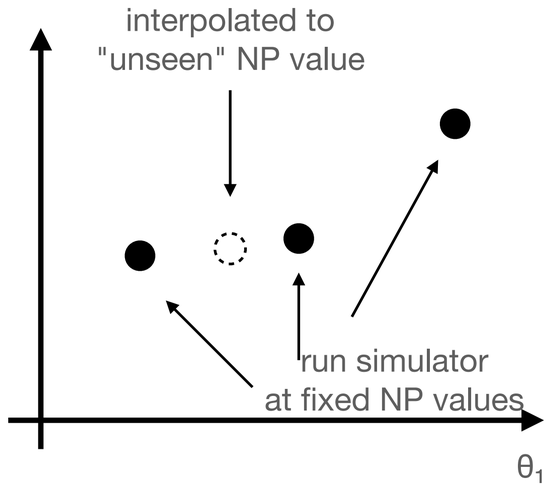
Modelling the actual measurements

Main measurement & non-simplified contraints: need to model $p(x \mid \mu, \nu)$

Ideal case: can run simulator at (μ, ν)

But only feasible for limited set of points:

- **cannot afford for each evaluation during inference**



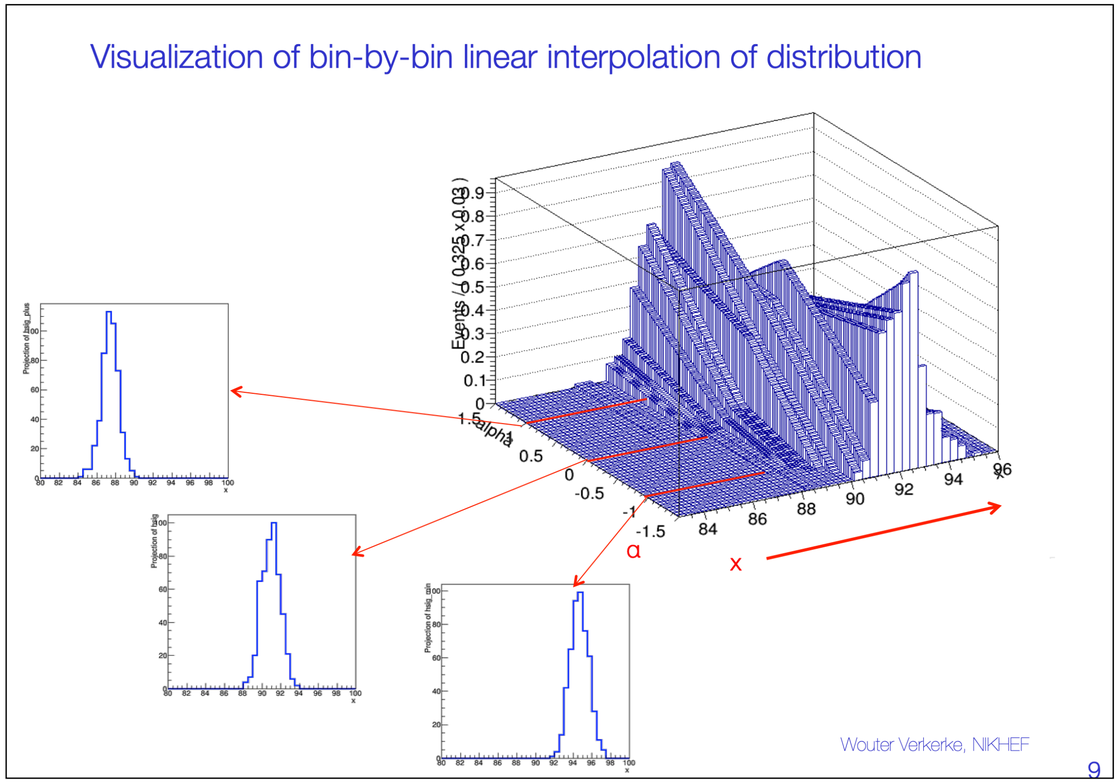
Resulting model has several layers or approximation

- **finite size of N_i events simulation at $\theta_i \neq \theta$** [issue: how to pick θ_i]
- **density estimate of statistic/observable $p(x \mid \theta_i)$** [issue: finite sample size, how to pick N_i / binning]
- **interpolate from set of $p(x \mid \theta_i) \rightarrow p(x \mid \theta)$** [issue: reliability of interpolation / unertainty?]

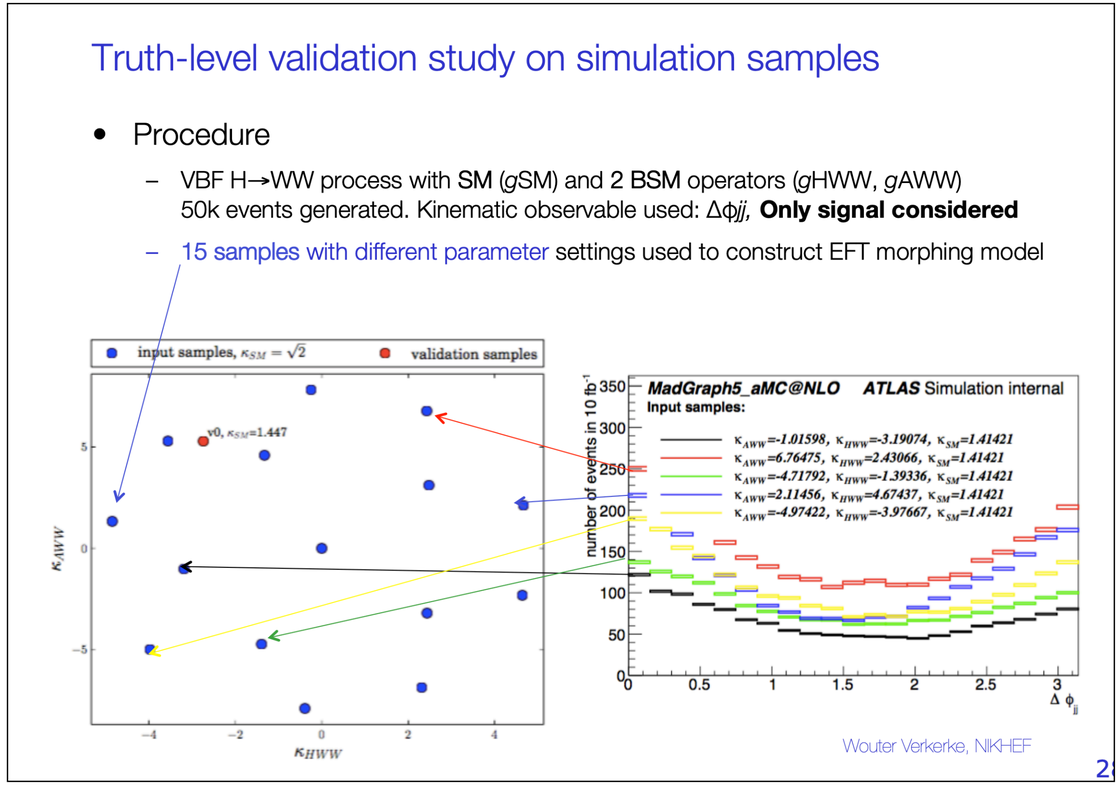
Example Interpolations

Widely used default: Vertical Interpolation
(used in HistFactory)

For signal (e.g. EFT) more complex examples exist
but usually not used for NP (also e.g. moment morphing etc)
(but not as relevant for searches)



one-dimensional interpolation
based on temlates

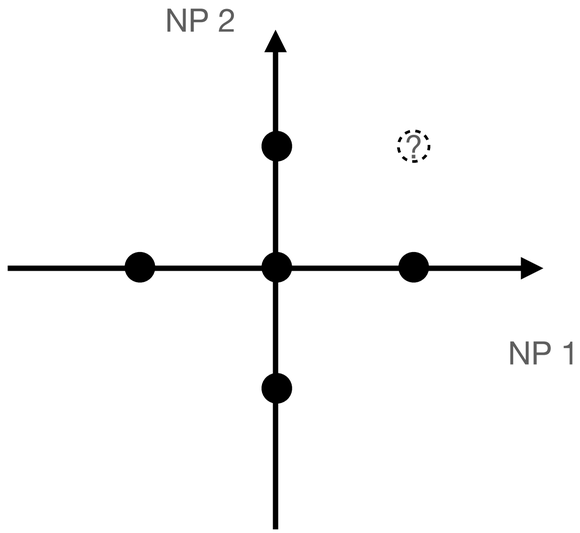


Two-dimensional interpolation [W. Verkerke]

Discussion: on-axis only evaluation

Possible technical issue:

Simulators are practically not always evaluatable at arbitrary points in nuisance parameter space



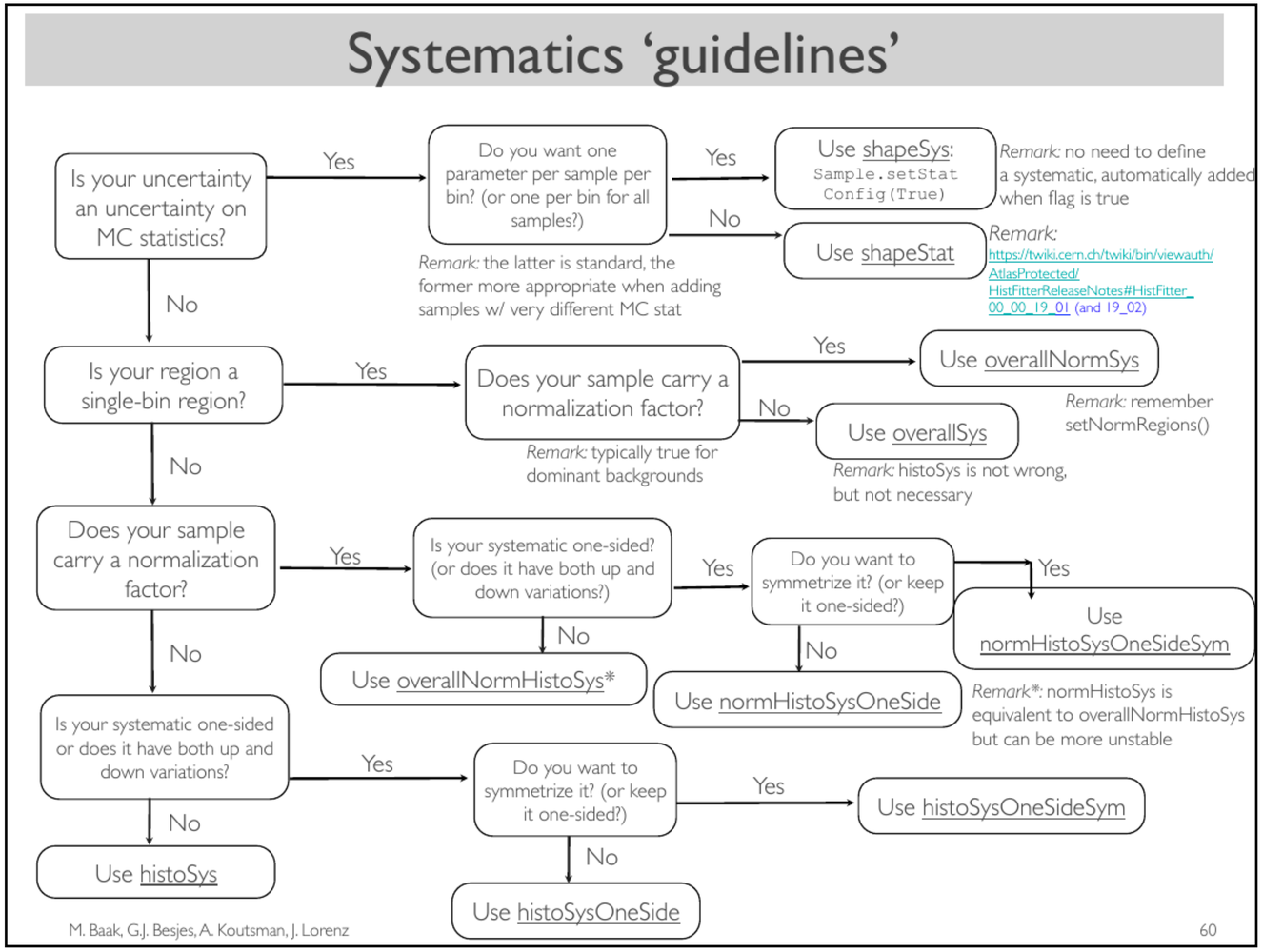
- **Lots lot of tooling just supports "on-axis" evaluation (only single non-nominal setting)**
  - **Per axis only have fixed evaluation points (cannot run ATLAS reco at e.g. +0.7σ nominal JES)**

Corrolary:

- **often unclear whether parametrized interpolation would match the true response of the simulator at that point**
- **possible solutin: Gaussian Processes (backup)**

How to incorporate Systematic uncertainties in a models

Once you know which evaluations you can get we provide guidance how to build interpolations and appropriate subs. terms



[Source: HistFitter Twiki]

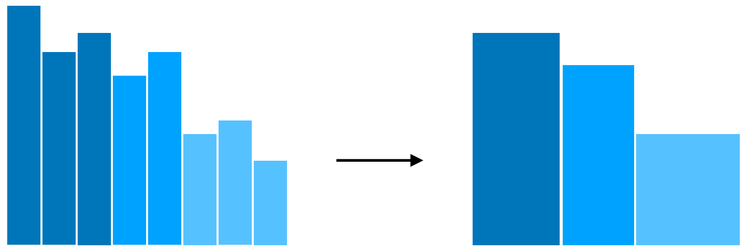
Stat uncertainty from finite-sample stats

Every template derived from simulation is affected by finite sample variance

- **exact treatment would lead to explosion of Nuisance Pars**

Solution: Light-weight Barlow-Beeston

- **only allocate single NP per bin**
- **overall σ_{MC}: quad.-sum of sample $\sigma_{i,\mathrm{MC}}$**



Additionally:

Smoothing of templates to reduce stat. fluctuation in templates

- **Targeted rebinning to reduce inter-bin variation**
- **Discussion:**
  - **constraints on smoothing alg to avoid under-estimating stat. uncert**
  - **principled way to find / learn "optimal binning"?**

Systematics Pruning

Modern LHC analyses can have hundreds of nuisance parameters

- not actively added by analyzers but determined by central groups
  not always relevant for studied phasespace
- can lead to numerical instabililty in fits / minimization

Idea: remove systematics that have negligible effect on inference

Example: drop normalization effects < 0.5%

Discussion:

- operating on symptoms? Should we give up modelling or improve tools?
- better to merge systematics than drop?
- principled way to assess drop tiny systematics?

[Example; autodiff]

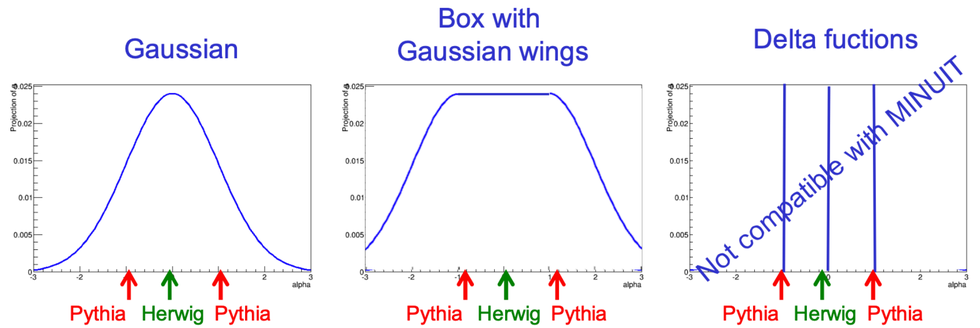
2-point systematics

Potential source of uncertainty: simuator itself, rather than its configuration
Example: Pythia vs Sherpa (LHC Lingo: "2-point systematics")

- not a continuous set of sensible NP values where we choose ref. points (like e.g. energy scale)
- fundamentally discontinuous (?)
- more model selection than nuisance parameter

Specific issues with theory uncertainties

- Subsidiary measurement of a theoretical 2-point uncertainty effectively quantifies the ‘knowledge’ on these models
  - *Extra difficult to make meaningful statement about this*, since meaning of parameter is not well embedded in underlying theory model
  - But again, all procedures need to assume some distribution… Profiling requires you to spell it out
- Some options and their effects



Prefers Herwig at 1σ

All predictions ‘between’ Herwig and Pythia equally probable

Only ‘pure’ Herwig and Pythia exist

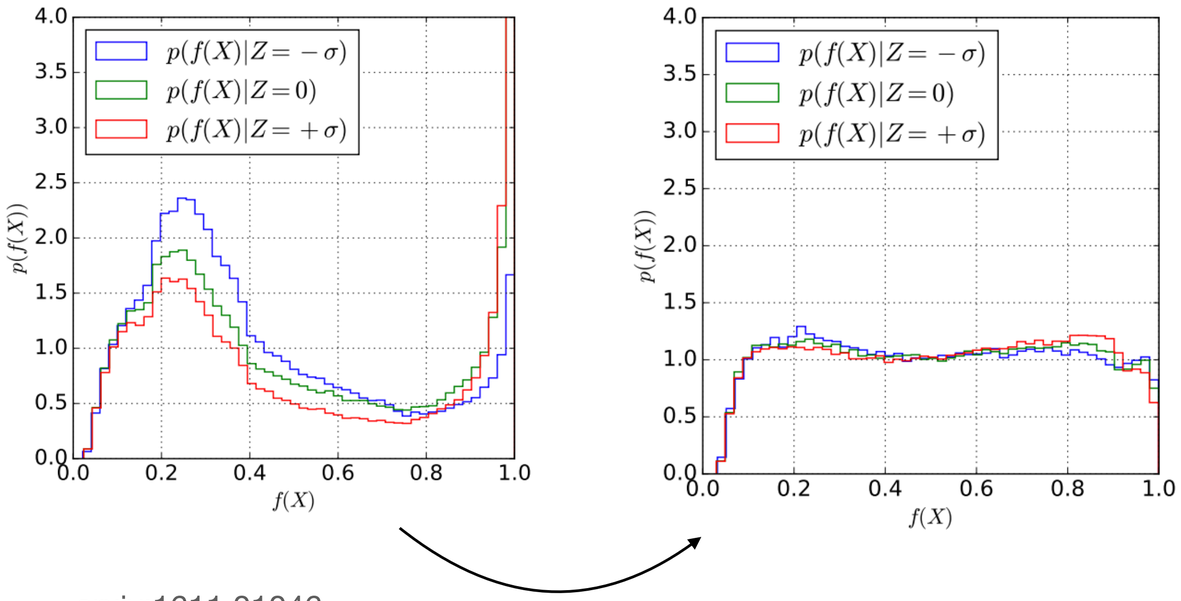
Wouter Verkerke, NIKHEF

Machine-Learning to shape systematics response

building a model on low-dimensional data gives as leeway to process the data in a robust way: shaping of the likelihood

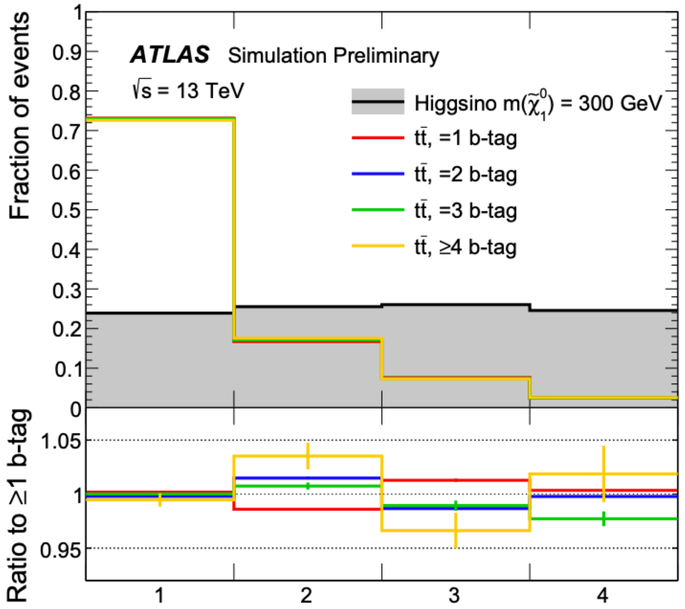
Goal: representations of the data that

- **good at** separating sig v. bkg
- **bad at** measuring NPs (i.e. invariant)



arxiv:1611.01046

adversarial training



arxiv:2106.09609

arxiv:2001.05310 distance correlation penalty

Inference

Once you have the parametrized model $p(x \mid \mu, \nu)$ you can do inference

At LHC: predominantly frequentist.

Broad Strokes:

Point Estimates: Maximum Likelihood Estimates
Interval Estimates (Upper Limits)

- **Inversion of series of hypothesis tests (Neyman construction)**
- **hypothesis tests with profile-likelihood-based test statistic and modified p-values (CLs method)**
- **heavy use of asymptotic test-statistic shapes (Wilks & Wald)**

Test Statistics

Use of default test statistics based on profile likelihood

- **profiling: conditional maximization of nuisance parameter**
- **compare to LEP, Tevatron, Neutrino**

$$\frac{L(\mu, \hat{\hat{\nu}}(\mu))}{L(\hat{\mu}, \hat{\nu})}$$

Profile Likelihood Ratio (e.g. LHC)

$$\frac{L(\mu, \nu)}{L(\mu = 0, \nu)}$$

simple Likelihood Ratio of fixed Hypotheses (e.g. LEP)

$$\frac{L(\mu, \hat{\hat{\nu}}(\mu))}{L(\hat{\mu}, \hat{\hat{\nu}}(0))}$$

Ratio of profile Likelihoods (e.g. Tevatron)

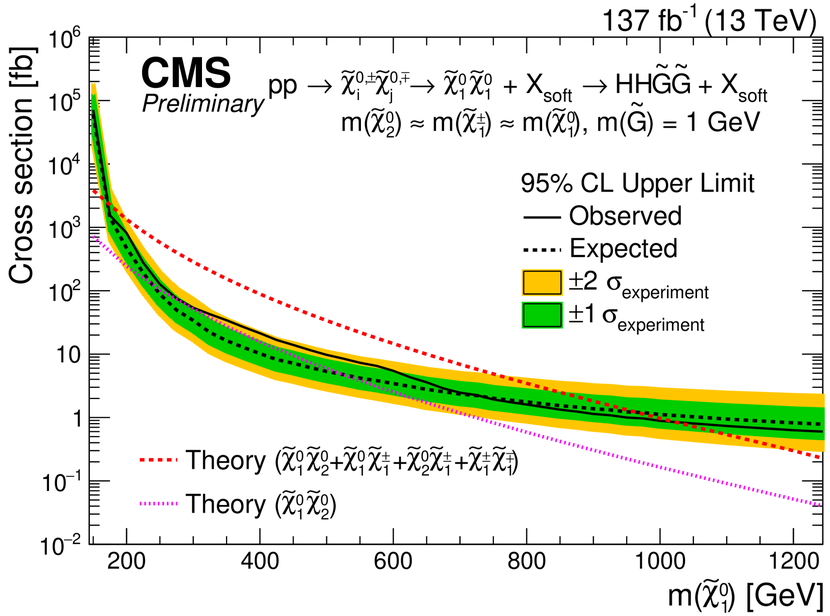
$$\frac{L(\mu, \nu)}{L(\hat{\mu}, \nu)}$$

Feldman-Cousins (MicroBooNE?)

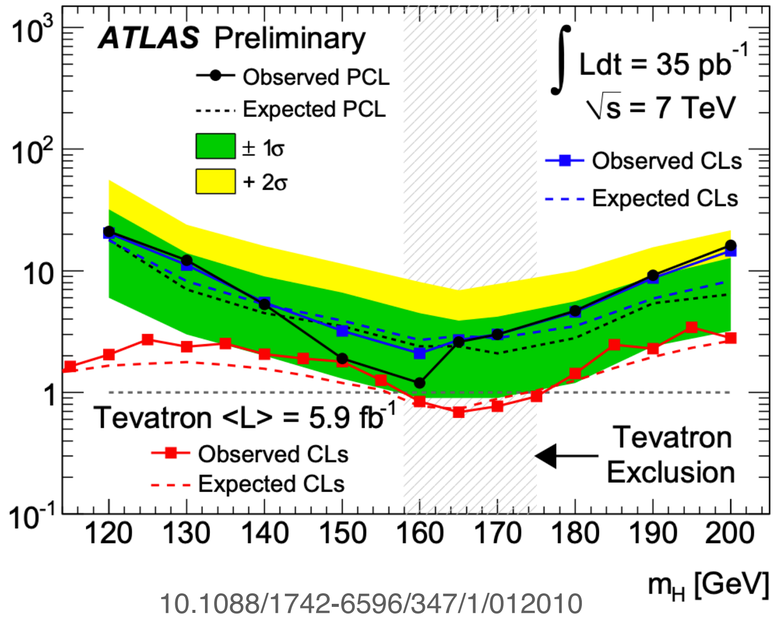
Names differ in communities e.g. last week MicroBooNE quoted Feldman-Cousins but other papers mention Profile-Likelihood Ratio for MicroBooNE (see backup)

CLs: p_{s+b}/p_b

Overwhelming use of CLs to avoid rejecting hypothesis w/o sensitivity: deliberately overcovers intervals (but unreported how much)

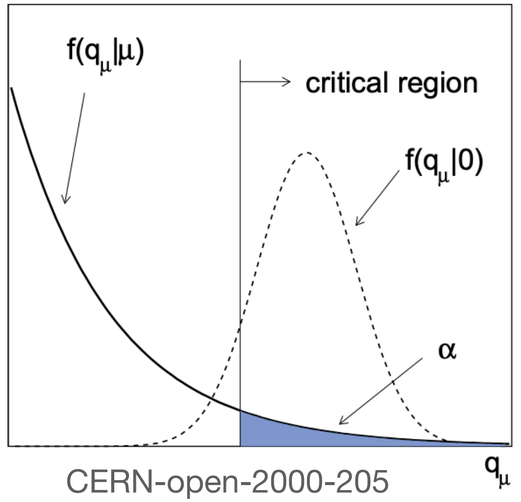


SUS-20-004



10.1088/1742-6596/347/1/012010

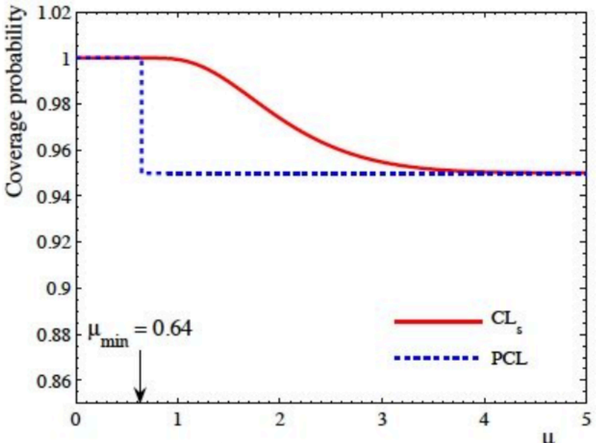
CLs: ratio of p-values



CERN-open-2000-205

arxiv:1007.1727

PC: manifest cut-off



arxiv:1105.3166

Early LHC: Power-Constrained Limits (PCL) but CLs prevails

- **settled question? PHYSTAT-DM now recommends PCL (2105.00599)**

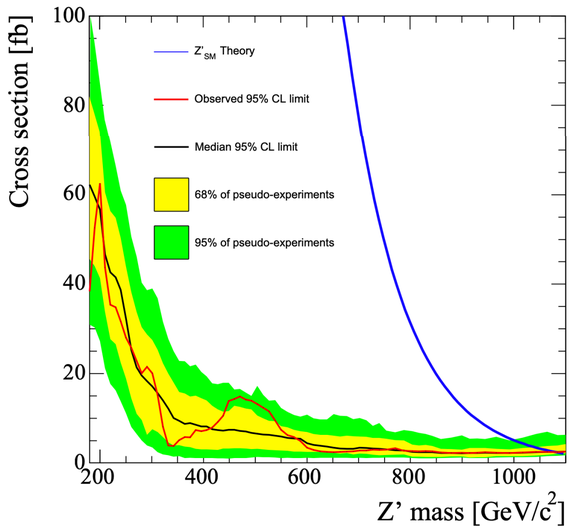
Discussion: Space in which you do interval construction?

Many results are presented in the space of multiple "theory parameters"

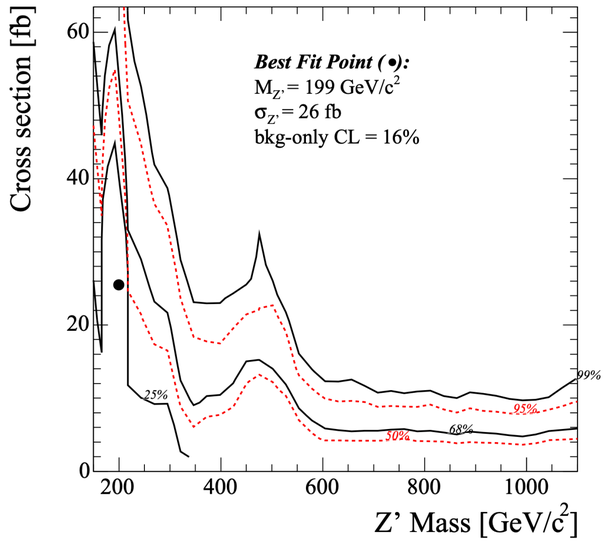
- **classic example: 2D SUSY Contour Plots**

But hypothesis test most often done with POI being signal strength

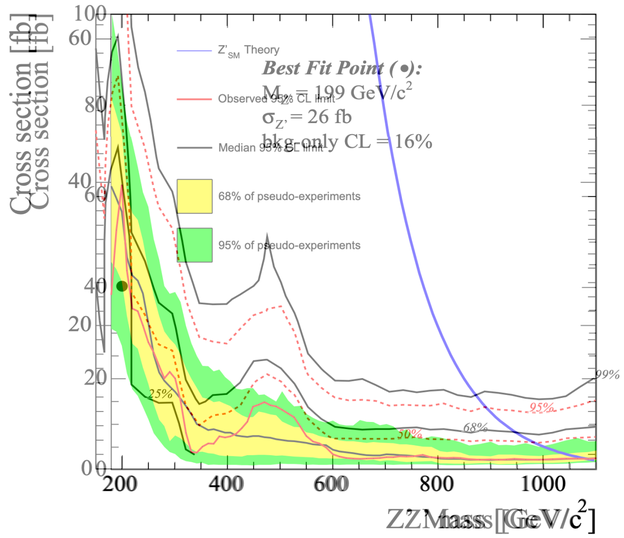
- **those most often are "hyperparameters" not actual floating in profile likelihood computation**



Raster Scan



2-D Neyman Construction

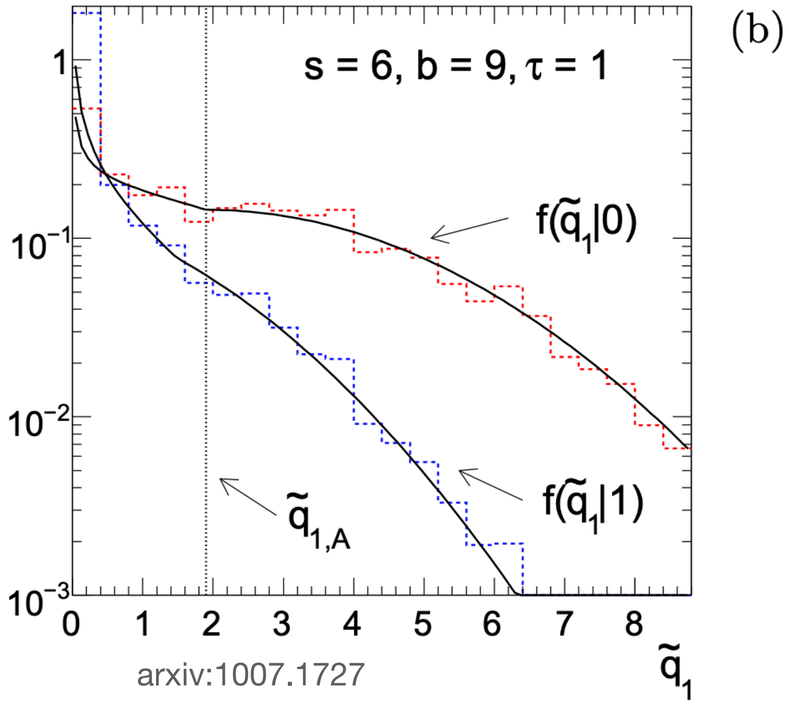


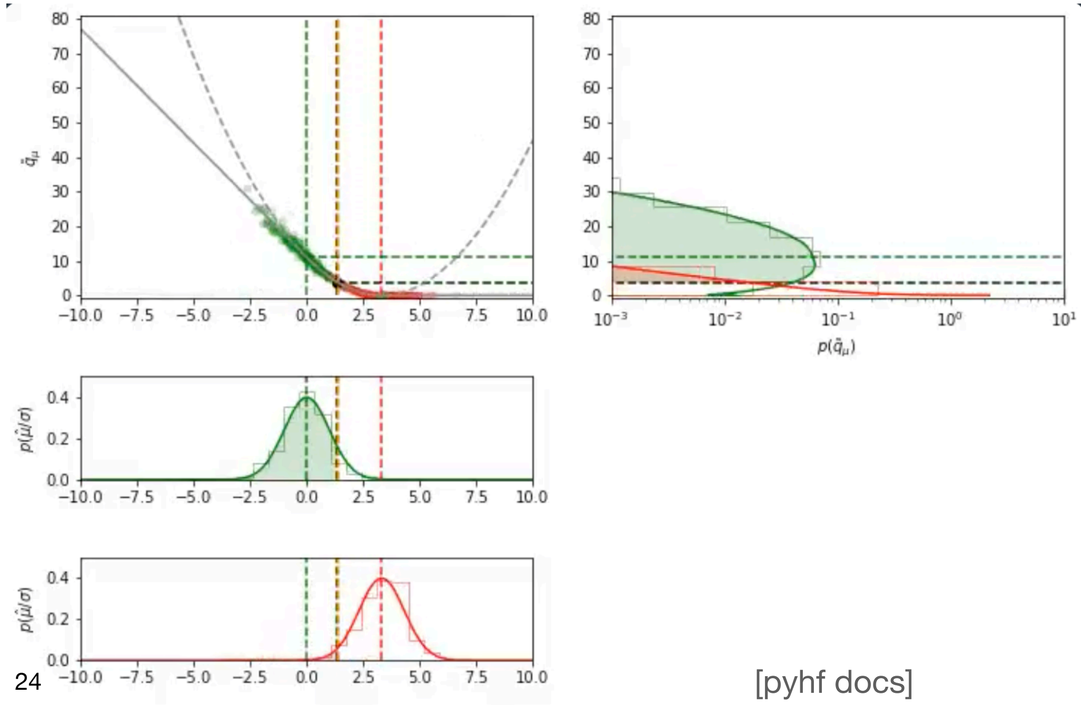
Overlay

arXiv: 1101.4578

Asymptotic Calculations

- Many searches rely heavily on asymptotic shapes of test stat distributions
- Assumption: Wilks & Wald hold:
  - could be checked more directly (e.g. $\hat{\mu}$ distribution)
  - often regularity assumptions (e.g. $\sigma_{\hat{\mu},s+b} = \sigma_{\hat{\mu},\mathrm{b-only}}$)





Assessing Impact, ranking systematics

An important question: ranking of systematic effects

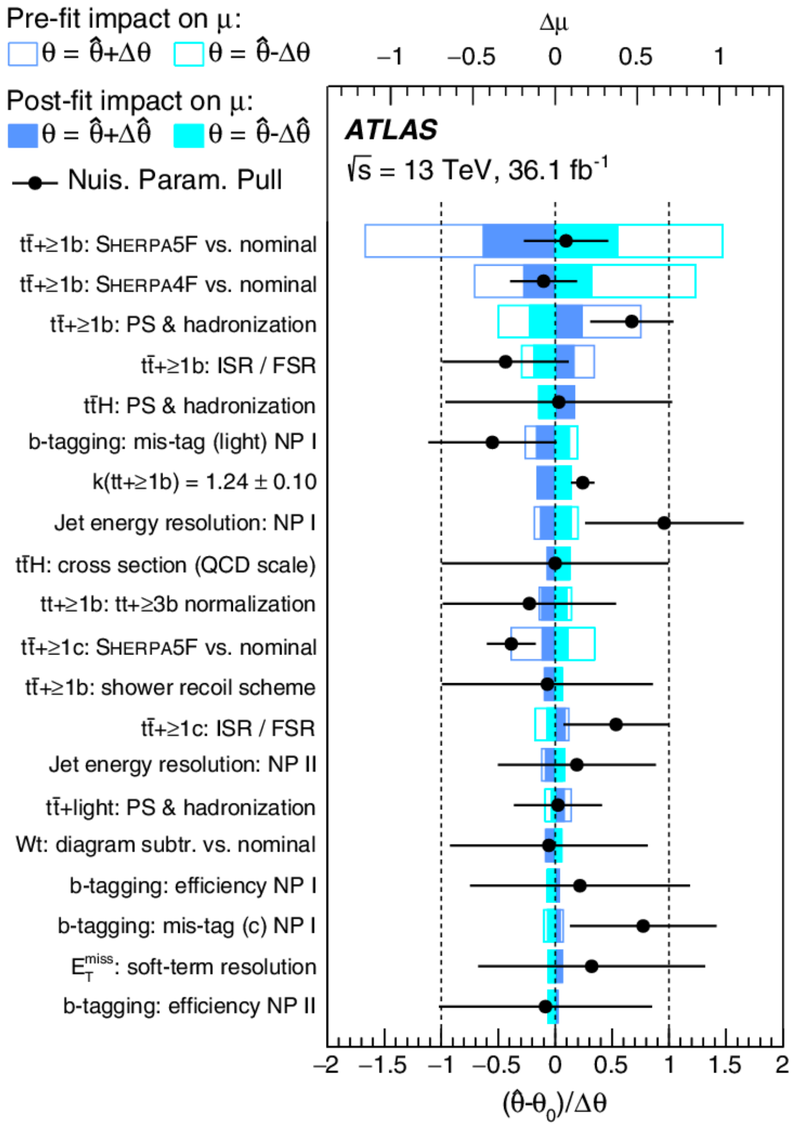
- feedback where to invest in modelling work
- perhaps convert into an in-situ measurement

Standard Diagnostic "The Ranking Plot"

Gives both: pre- and post-fit impact

Procedure:

- fix NP to value to $\hat{\theta} \pm \Delta\theta_{\text{pre/post}}$
- observe impact on POI



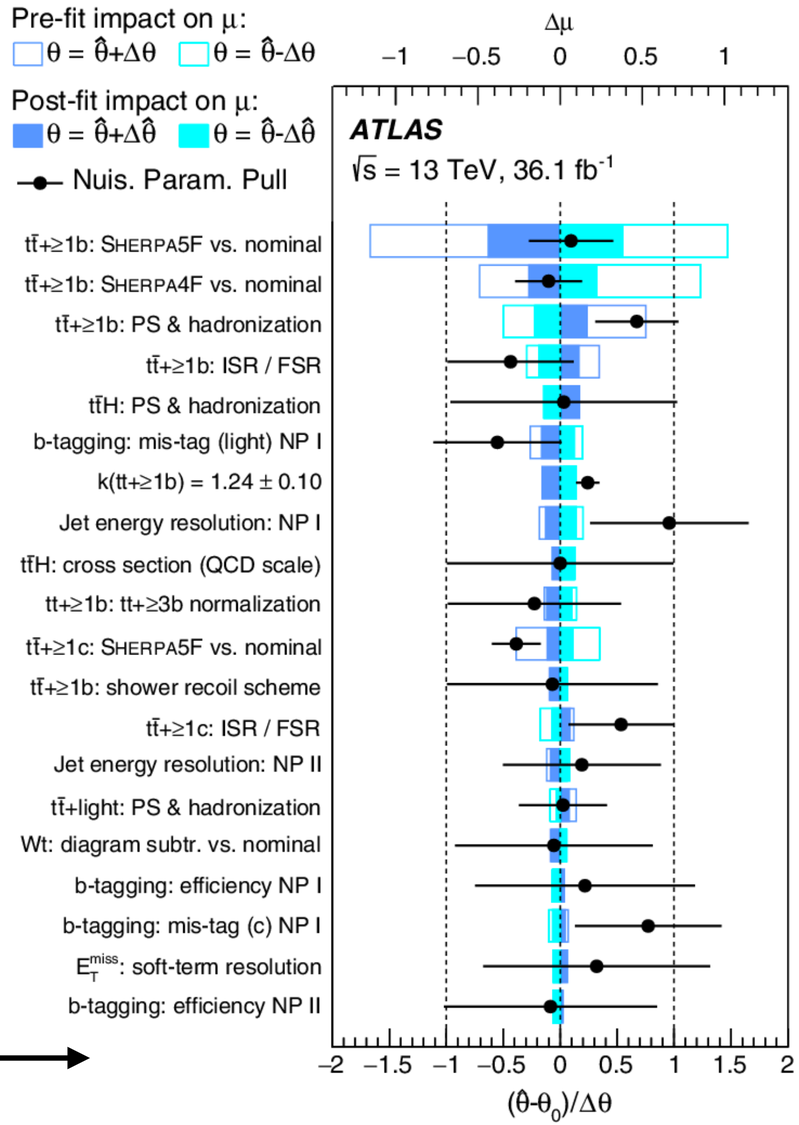
Assessing Validity of simplified constraints

Assumption of replacing full nuisance likelihood with simple term:

- present analysis **should not be more sensitive** to NP than dedicated subsidiary measurement

Standard Diagnostic: "The Pull Plot"

- "Pulls" show NP value after profiling
- "Pull Width" post-fit uncertainty
  - if it's smaller than pre-fit look closely
  - rarely it can be larger: often red flag



Definition of signal and background hypotheses

Calling something a signal or background hypothesis only fixes the POIs

- but we need to fix the NP as well → certain degree of ambiguity
- **affects e.g. expected limits** we quote, want "representative NP values"

Options:

1. pre-fit: choose some **predefined nominal value** $\nu = \nu_0$
2. blinded fit: **fit "near-field" measurements + constraint terms** $\hat{\hat{\nu}}(\text{subs. data})$
3. unblinded fit: **fit the full data** $\hat{\hat{\nu}}(\text{data})$ [NB: expected limits depend on observed data]

etc.

ATLAS practice: use unblinded fit to report expected limits

Summary: Many things work but also lots of interesting questions

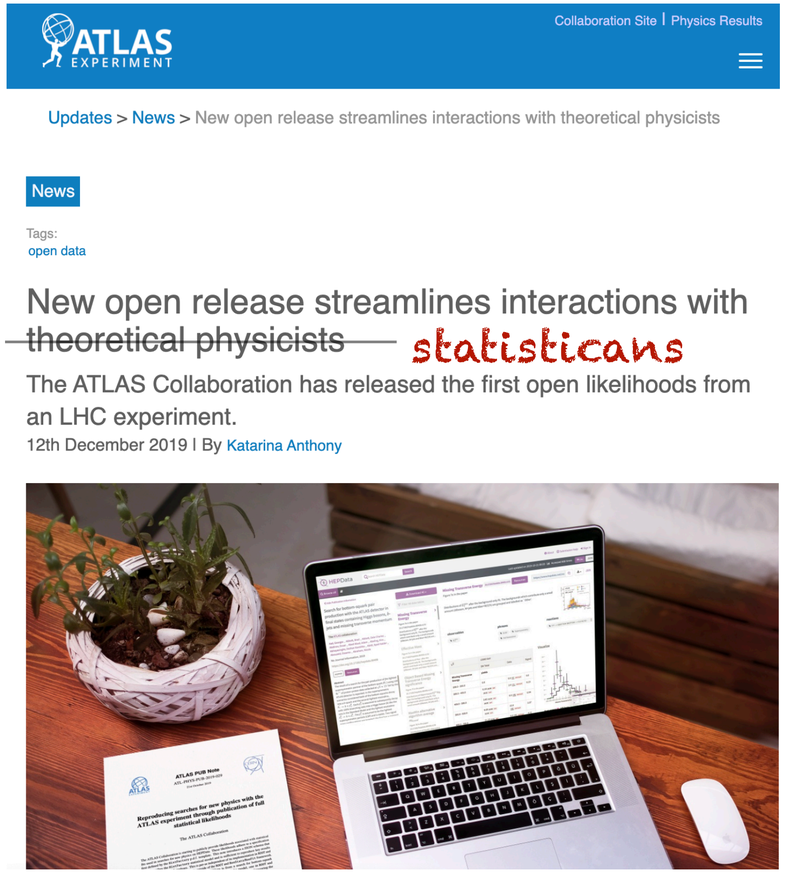
Exciting development: public real-world probability models

Not only opportunity to push new science but also new stats methodology / answer "what if"s

A new public workhorse / platform akin to e.g. on-off problem, $\mathrm{Pois}(n \mid \mu s + b)$, ...

Examples:

- **Bayesian Workflow on LHC models**
- **Coverage & Asymptotics Studies**
- **Signal interpolation strategies**
- **Reinterpretation Tools**
- **Model destillation ("simplified likelihoods")**



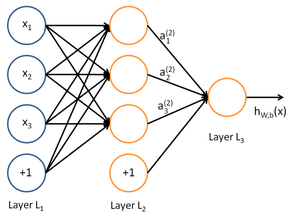
Backup

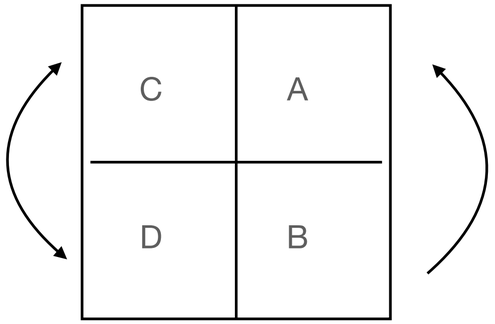
Beyond ABCD - a complicated example

Searches often push towards phase-spaces where simulation is unreliable

- data-driven estimates (i.e. in-situ subsidary measurements)
- classic example: ABCD

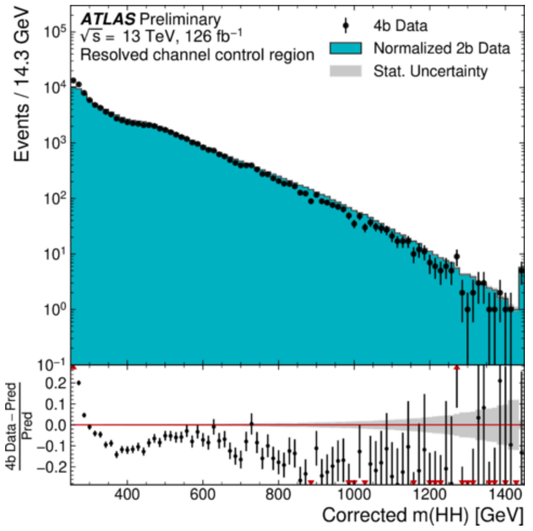
More advanced: event-by-event reweighting
Classifiers (e.g. NNs) ↔ density ratio estimators

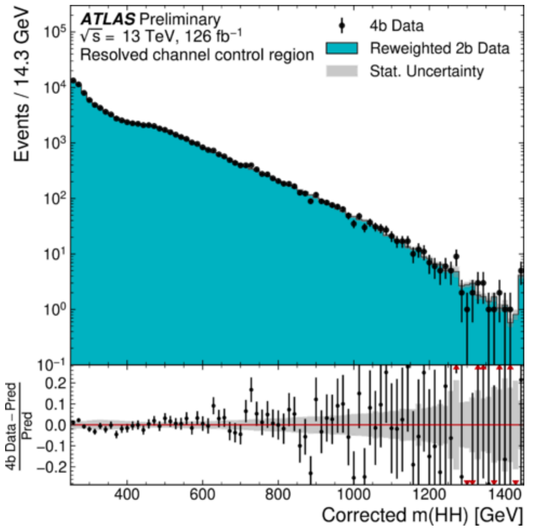




Train reweighting D→C, apply to B→A

- not straight forward to built into likelihood
- Solution:
  - ensemble of trainings (bootstrapping)
  - differences as uncertainty
  - vertical interpolation as usual



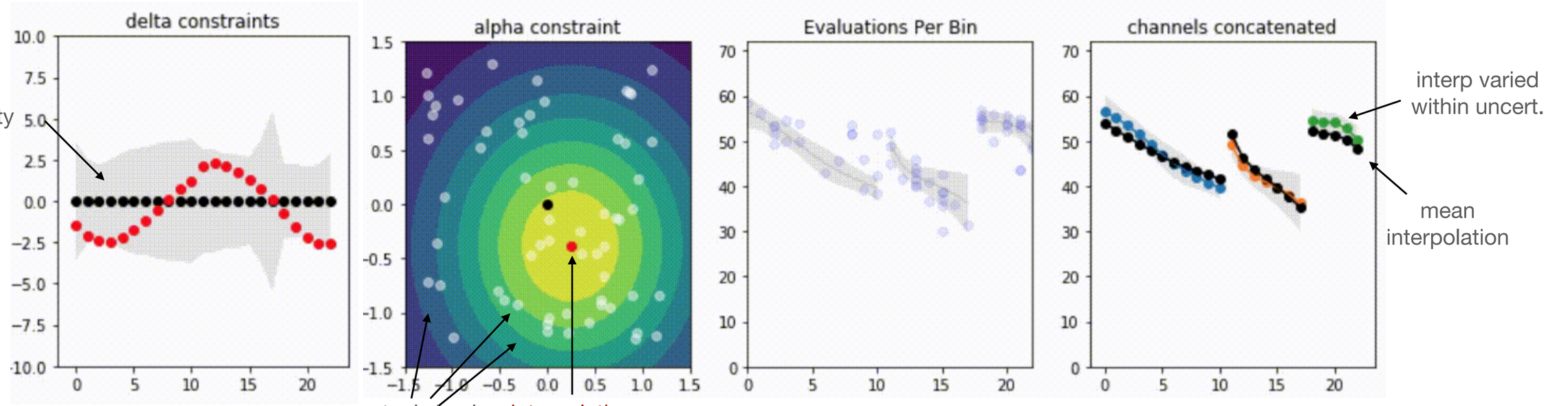


Example Interpolations

Possible solution for NPs: Gaussian Processes

- **model $p(\text{hist}(\theta) \,|\, \{\text{set of hist}(\theta_i)\})$ (more natural in Bayesian picture)**
- **either use MAP estimate as interpolation result**
- **but can also provide full prior/constraint term to model for uncertainty on the interpolation itself**

Example: Template Interpolation base on variable set of input templates



[[prototype code]]

Test Statistics for MicroBooNE

Table 2 Definition of the test statistics used for sterile-neutrino searches in the presence of nuisance parameters ($\boldsymbol{\eta}$). The null hypothesis is $\mathrm{H}_0 : \{\sin^2(2\theta), \Delta m^2 : \sin^2(2\theta) = X, \Delta m^2 = Y\}$ for all tests while the alternative hypothesis H_1 changes. The free parameters of interest in H_1 are shown in the second column. The name of the techniques based on each test statistics and a selection of experiments using them are listed in the last columns.

Test Statistic Computed for $\mathrm{H}_0 : \{\sin^2(2\theta), \Delta m^2 : \sin^2(2\theta) = X, \Delta m^2 = Y\}$	Free Parameters of Interest	Associated Names	Experiments
$\mathrm{T}_2 = -2\ln \dfrac{\sup\limits_{\boldsymbol{\eta}} \mathcal{L}\left(\sin^2(2\theta) = X, \Delta m^2 = Y, \boldsymbol{\eta} \vert N^{obs}\right)}{\sup\limits_{\sin^2(2\theta), \Delta m^2, \boldsymbol{\eta}} \mathcal{L}\left(\sin^2(2\theta), \Delta m^2, \boldsymbol{\eta} \vert N^{obs}\right)}$	$\sin^2(2\theta)$, Δm^2	2D Scan or global p-value	LSND, MiniBooNE, PROSPECT

arxiv:1906.11854