Fast approximate statistical models ("bottom-up simplified likelihoods")

Nicolas Berger (LAPP)

Simplified Statistical Models

- Accurate reinterpretations and global combination need the full measurement PDF, or a very good approximation

- Currently, often rely on covariance matrices for the total uncertainty only – suboptimal in several ways
  - No description of non-Gaussian effects (important in tails, e.g. for EFT)
  - No uncertainty breakdowns ⇒ no way to correlate systematics across measurements

- Effort by experiments to publish the full measurement PDF (pyhf, combine) finally gathering steam:
  - ⊕ HistFactory description includes non-Gaussian effects
  - ⊕ Independent NPs for systematics: can in principle properly correlate (although not always trivial in practice)
  - ⊖ However models sometimes quite large: difficult to handle and long to evaluate
  - ⊖ More difficult to tackle unbinned models (need something like RooAbsPdf…)

- Simplified statistical models : retain key features of the full likelihood, going beyond covariance matrices

  → Several approaches: JHEP04 (2019) 064, CMS Note 2017/001, …

  → This talk: "bottom-up" approach starting from HistFactory workspaces

Starting point: the HistFactory description

Binned likelihood form, with parameters of interest (**μ**) and nuisance parameters (**θ**) :

$$L(\mu,\theta) = \prod_{c=1}^{n_{\text{channels}}} \prod_{b=b_c^{\text{first}}}^{b_c^{\text{last}}} P\left(n_b^{\text{obs}}; \sum_{s=1}^{n_{\text{samples}}} N_{s,b}^{\text{exp}}(\mu,\theta)\right) \prod_{p=1}^{n_{\text{non-free NPs}}} C(a_p;\theta_p)$$

Product over channels (independent measurement regions)

Product over channel bins

Poisson PDF in each bin

Observed bin yield

Expected bin yield, function of both POIs and NPs.

- Several possible forms: linear, exponential, etc.
- Implements correlations between bins

NP constraints (from auxiliary measurements a**)**

Complete description of binned measurements, but can be quite complex ⇒ how to simplify ?

Linear binned likelihoods

→ Consider only **Gaussian constraints**

→ Keep **full description** of **bin counting** (Poisson PDF) and **POIs** (μ)

→ Treat **NPs** (i.e. systematics, among others) at **linear order only**.

$$C(a_p; \theta_p) = G(a_p; \theta_p, \sigma_p)$$

$$N_{sb}^{\text{exp}}(\mu, \theta) = N_{sb}^{\text{nominal}}(\mu) \left(1 + \sum_p \delta_{sbp} \theta_p\right)$$

Exact dependence on the POI μ

NP dependence at linear order only

Linear impact coefficients describing all the systematic effects

Linear binned likelihoods

→ Consider only **Gaussian constraints**

→ Keep **full description** of **bin counting** (Poisson PDF) and **POIs** (μ)

→ Treat **NPs** (i.e. systematics, among others) at **linear order only**.

$$L(\mu,\theta) = \prod_{c=1}^{n_{\text{channels}}} \prod_{b=b_c^{\text{first}}}^{b_c^{\text{last}}} P\left(n_b^{\text{obs}}; \sum_{s=1}^{n_{\text{samples}}} N_{sb}^{\text{exp}}(\mu,\theta)\right) \prod_{p=1}^{n_{\text{non-free NPs}}} C(a_p;\theta_p)$$

$$L(\mu,\theta) = \prod_{c=1}^{n_{\text{channels}}} \prod_{b=b_c^{\text{first}}}^{b_c^{\text{last}}} P\left(n_b^{\text{obs}}; \sum_{s=1}^{n_{\text{samples}}} N_{sb}^{\text{nom}}(\mu)\left(1 + \sum_{p} \delta_{sbp}\,\theta_p\right)\right) \prod_{p=1}^{n_{\text{non-free NPs}}} G(\theta_p^{\text{obs}};\theta_p,\sigma_p)$$

Linear binned likelihoods

→ Consider only **Gaussian constraints**

→ Keep **full description** of **bin counting** (Poisson PDF) and **POIs** (μ)

→ Treat **NPs** (i.e. systematics, among others) at **linear order only**.

$$L(\boldsymbol{\mu},\boldsymbol{\theta}) = \prod_{c=1}^{n_{\text{channels}}} \prod_{b=b_c^{\text{first}}}^{b_c^{\text{last}}} \boldsymbol{P}\left(n_b^{\text{obs}} ; \sum_{s=1}^{n_{\text{samples}}} N_{sb}^{\text{nom}}(\boldsymbol{\mu}) \left(1 + \sum_{p} \delta_{sbp}\, \boldsymbol{\theta}_p \right) \right) \prod_{p=1}^{n_{\text{non-free NPs}}} G\left(\theta_p^{\text{obs}} ; \boldsymbol{\theta}_p, \sigma_p \right)$$

Fast profiling: with these assumptions, **can profile θ in closed form** using linear algebra (least squares) :

$$\hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}) = \left[\Gamma + P(\boldsymbol{\mu})\right]^{-1} \left[\Gamma\, \boldsymbol{\theta}^{\text{obs}} - \boldsymbol{Q}(\boldsymbol{\mu})\right]$$

$$P(\boldsymbol{\mu}) = \sum_{b=1}^{n_{\text{bins}}} n_b^{\text{obs}} \sum_{s s'} \frac{N_{sb}^{\text{nom}}(\boldsymbol{\mu}) N_{s'b}^{\text{nom}}(\boldsymbol{\mu})}{\sum_{s''} N_{s''b}^{\text{nom}}(\boldsymbol{\mu})^2} \boldsymbol{\Delta}_{sb} \otimes \boldsymbol{\Delta}_{s'b}$$

$$\boldsymbol{Q}(\boldsymbol{\mu}) = \sum_{b=1}^{n_{\text{bins}}} \sum_{s} N_{sb}^{\text{nom}} \left(1 - \frac{n_b^{\text{obs}}}{\sum_{s''} N_{s''b}^{\text{nom}}(\boldsymbol{\mu})} \right) \boldsymbol{\Delta}_{sb}$$

$$\Gamma = \text{diag}_p\left(1/\sigma_p^2\right) \qquad \left(\boldsymbol{\Delta}_{sb}\right)_p = \delta_{sbp}$$

Linear binned likelihoods

→ Consider only **Gaussian constraints**

→ Keep **full description** of **bin counting** (Poisson PDF) and **POIs** (μ)

→ Treat **NPs** (i.e. systematics, among others) at **linear order only**.

Profiling steps:

- Invert a matrix of size ($n_{NPs} \times n_{NPs}$)
- Perform matrix/vector multiplications (*np.einsum*)

→ Significantly faster than non-linear minimization

$$\hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}) = [\Gamma + P(\boldsymbol{\mu})]^{-1}[\Gamma\ \boldsymbol{\theta}^{\mathbf{obs}} - \mathbf{Q}(\boldsymbol{\mu})]$$

→ Non-linear minimization still performed in POI-space, since POIs treated exactly, but typically smaller dimension.

→ All NPs retained: systematics all included, can correlate across analyses as done within experiments.

→ Linear approximation: good for small systematics (e.g. searches) and systematics that are naturally linear (examples later)

→ Similar to approaches in JHEP04 (2019) 064, CMS Note 2017/001, but preserves sample and NP structure

Example: ATLAS Search for trilepton resonances

Analysis published in 2020, HEPData record include the full statistical model in pyhf JSON format

Search for trilepton resonances from chargino and neutralino pair production in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector

G. Aad *et al.*[*]
(ATLAS Collaboration)

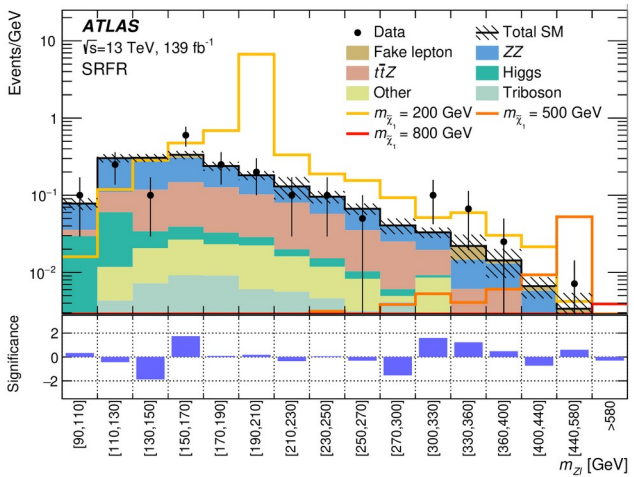
(Received 23 November 2020; accepted 23 April 2021; published 7 June 2021)

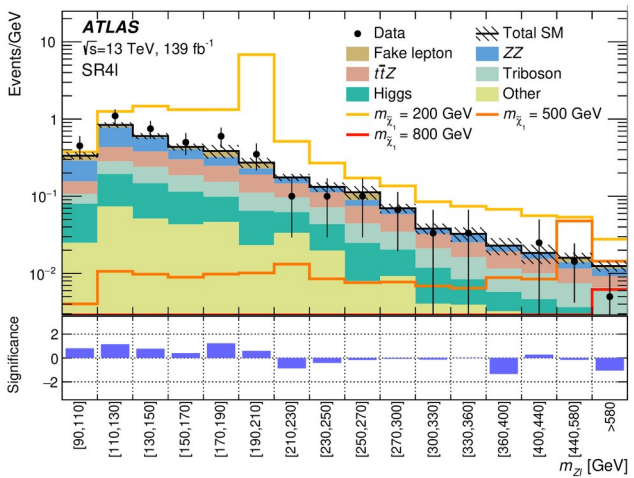
A search is performed for the electroweak pair production of charginos and associated production of a chargino and neutralino, each of which decays through an R-parity-violating coupling into a lepton and a W, Z, or Higgs boson. The trilepton invariant-mass spectrum is constructed from events with three or more leptons, targeting chargino decays that include an electron or muon and a leptonically decaying Z boson. The analyzed dataset corresponds to an integrated luminosity of 139 fb^{-1} of proton-proton collision data produced by the Large Hadron Collider at a center-of-mass energy of $\sqrt{s} = 13$ TeV and collected by the ATLAS experiment between 2015 and 2018. The data are found to be consistent with predictions from the Standard Model. The results are interpreted as limits at 95% confidence level on model-independent cross sections for processes beyond the Standard Model. Limits are also set on the production of charginos and neutralinos for a minimal supersymmetric Standard Model with an approximate $B - L$ symmetry. Charginos and neutralinos with masses between 100 and 1100 GeV are excluded depending on the assumed decay branching fractions into a lepton (electron, muon, or τ lepton) plus a boson (W, Z, or Higgs).

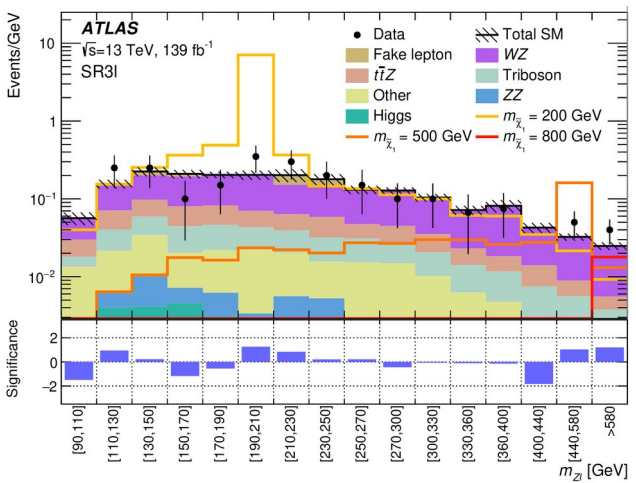
DOI: 10.1103/PhysRevD.103.112003

Analysis:

- 3 signal regions with 16 bins each, 3 control regions ⇒ 51 bins total
- 582 NPs: 3 free background normalizations, 579 systematics NPs
- Consider a 500 GeV signal, measure μ_{sig} ⇒ 1 POI (μ_{sig})



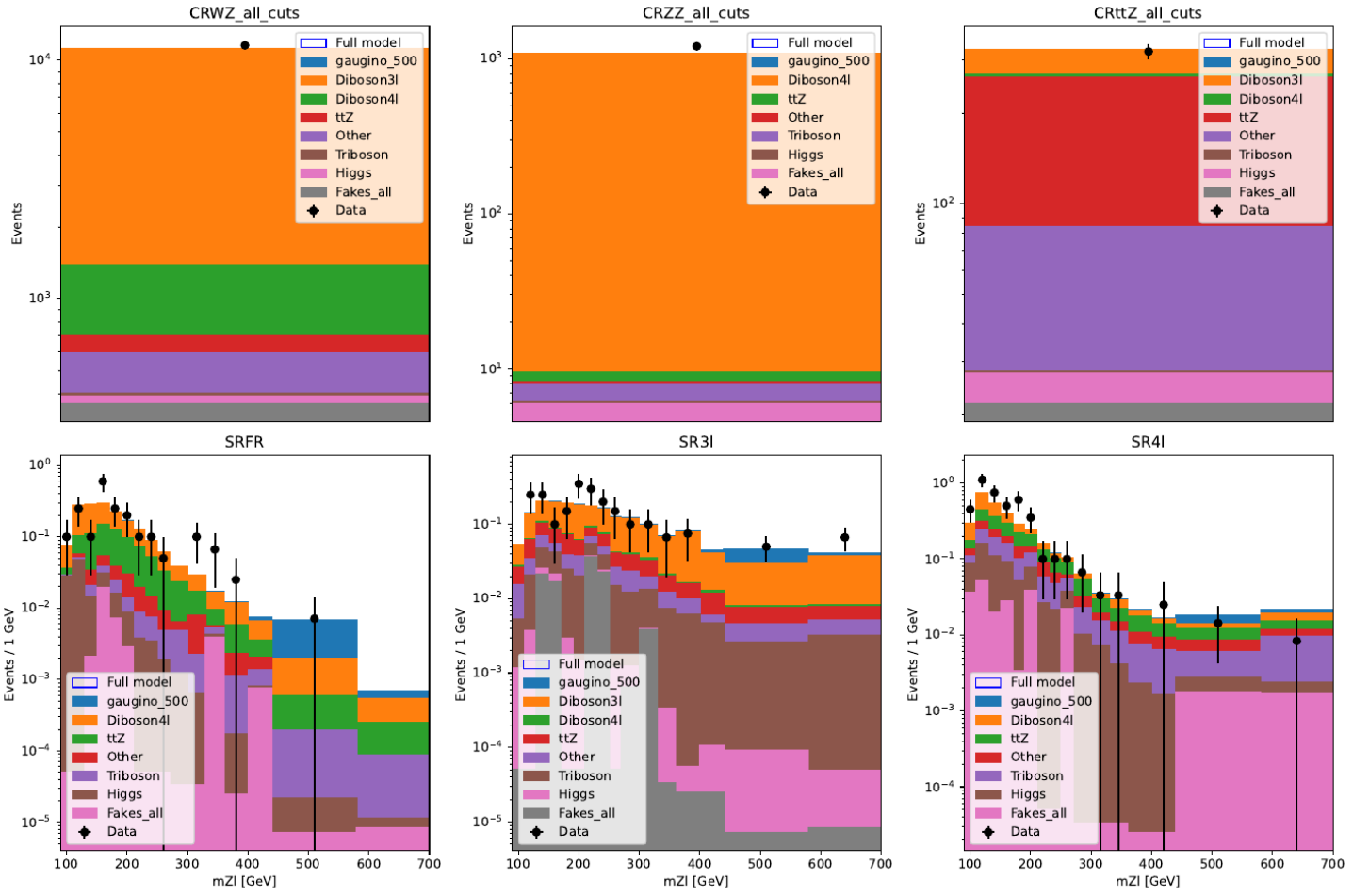




Fast profiling implementation

Automatically convert model from pyhf format (linearize pyhf *modifiers*)

Distributions similar to the ones in the published paper:



Fast profiling implementation

Perform profile-likelihood scans for μ_{sig} with

1) pyhf (pyhf.infer.mle.fixed_poi_fit) with numpy backend

2) fast profiling code, also using numpy

Fit time:

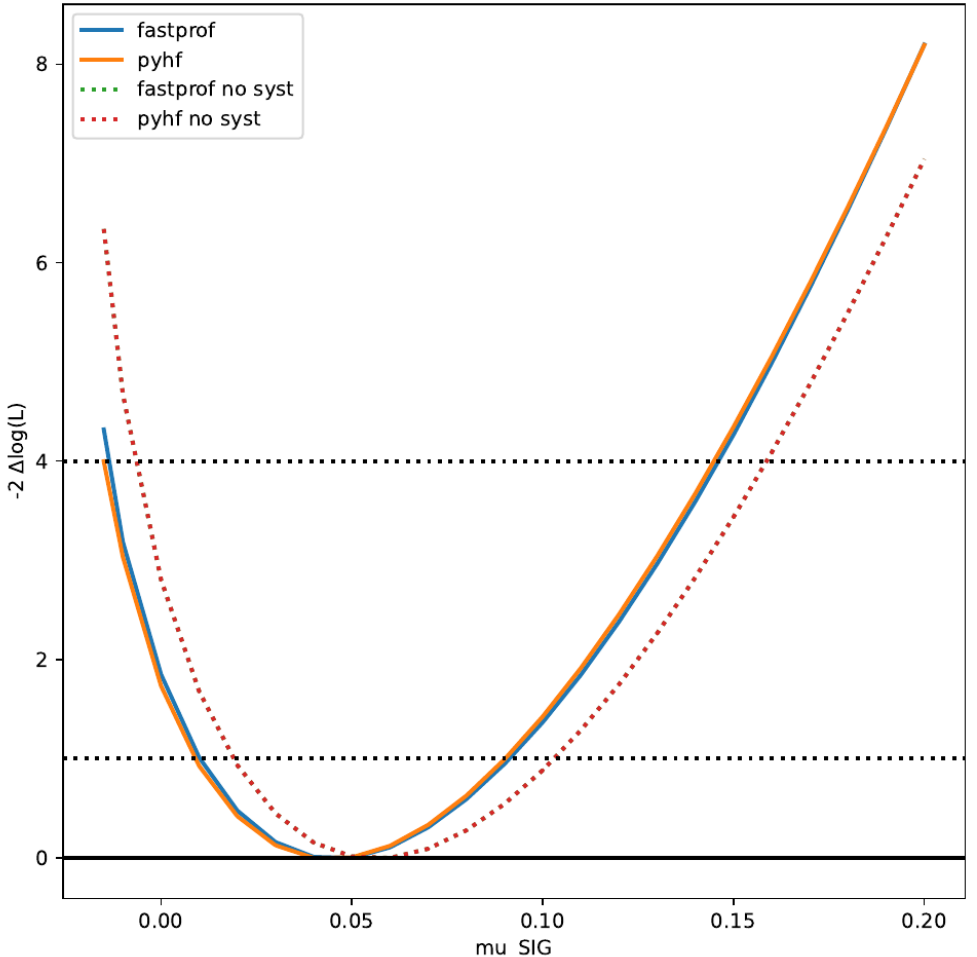
- 4 min using pyhf,
- ~0.025 sec using fastprof

Absolute times not relevant (pyhf is much faster using better backends than numpy!)

Key point is the speedup (>1000) using linearization.

Excellent agreement(*) in the scans!

- No systs: exact by construction (POI effects only)
- With systs: shows small effect of linear approx.



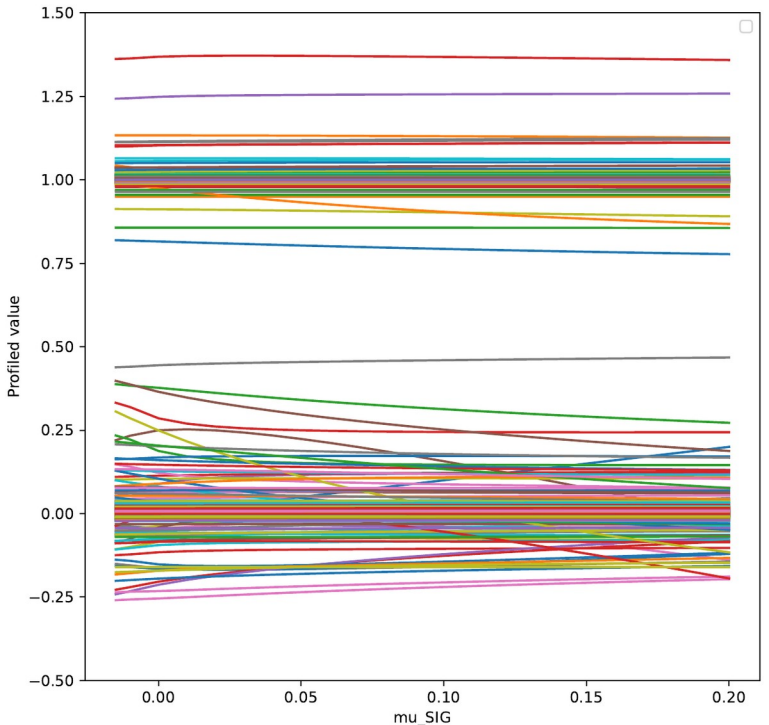
(*) 8 NPs (of 582) pruned: either very large impacts in some bins (> 100%) or ± 1σ impacts large and same-sign

Fast profiling implementation

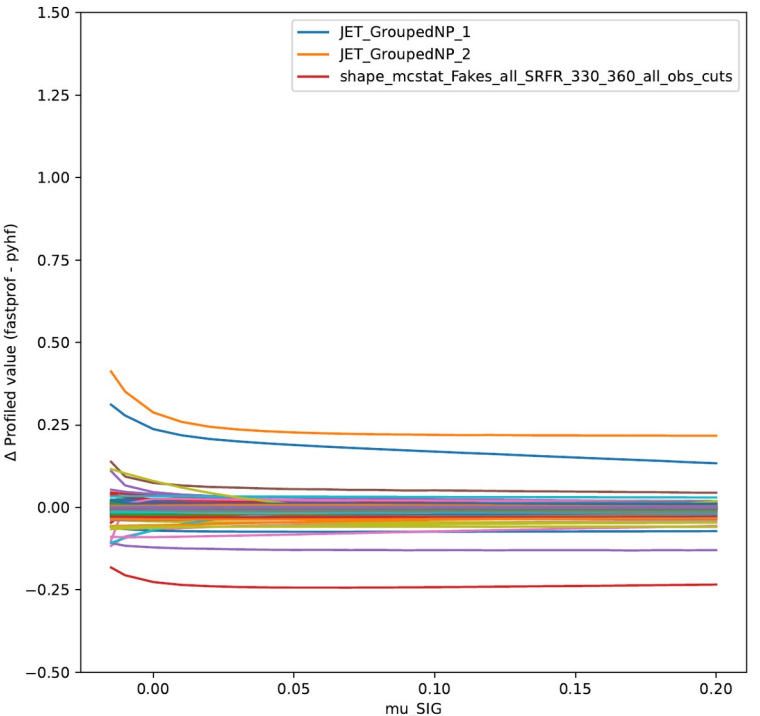
What is the impact of linearization on the low-level profiling mechanics ?

$\hat{\hat{\theta}}(\mu)$

fastprof profiled values for each NP



Δ(fastprof-pyhf) profiled values for each NP



Still some somewhat problematic NPs, but good agreement overall parameter-by-parameter

How good is the linear approximation ?

Dependence on POIs is exact, so all depends on effect of NPs. Some possible cases:

- **Normalization factor**
  → Already linear

$$N(\boldsymbol{\mu}, \boldsymbol{\theta}) = \theta_{\text{norm}} \; N^{\text{nom}}(\boldsymbol{\mu})$$

- **Symmetric Gaussian systematic**
  → Already linear

$$N(\boldsymbol{\mu}, \boldsymbol{\theta}) = N^{\text{nom}}(\boldsymbol{\mu}) \left(1 + \delta_{\text{syst}} \theta \right)$$

- **Asymmetric Gaussian systematic**
  → Non-linear if $|\theta_+ - \theta_-|$ large.

$$N(\boldsymbol{\mu}, \boldsymbol{\theta}) = N^{\text{nom}}(\boldsymbol{\mu}) \left(1 + \delta_+ \max(\theta, 0) + \delta_- \min(\theta, 0) \right)$$

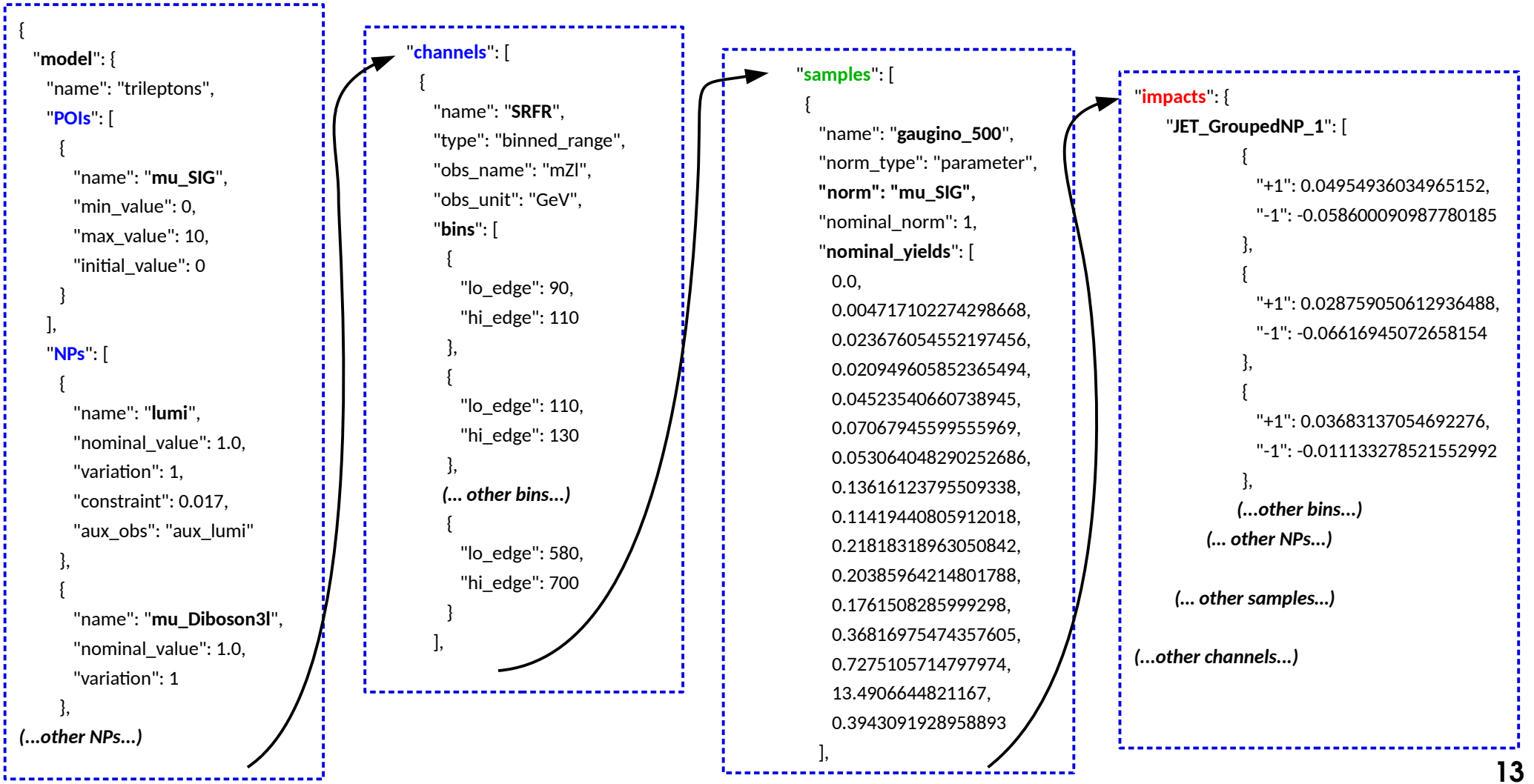
- **Log-normal systematic**
  → Non-linear, approximate linearity for $\delta_{\text{syst}} \ll 1$

$$N(\boldsymbol{\mu}, \boldsymbol{\theta}) = N^{\text{nom}}(\boldsymbol{\mu}) \left(1 + \delta_{\text{syst}} \right)^{\theta}$$

Approximation should be good for **small systematics**, **symmetric systematics** and cases where the **systematics are already linear** (e.g. normalization factors, BLUE, …)

JSON Model description

YAML also supported, more compact format



Approximating Unbinned Models

Binned models: just linearize NP impact in each bin

→ Automatic conversion tool for pyhf models

Unbinned models: can defined a binned approximation, for a given fine binning:

- Integrate unbinned distribution in each bin.
- Compute linearized impact of each NP on the integral

For bin width $\lesssim$ experimental resolution, should provide a good approximation.

Same number of POIs and NPs as the unbinned model.

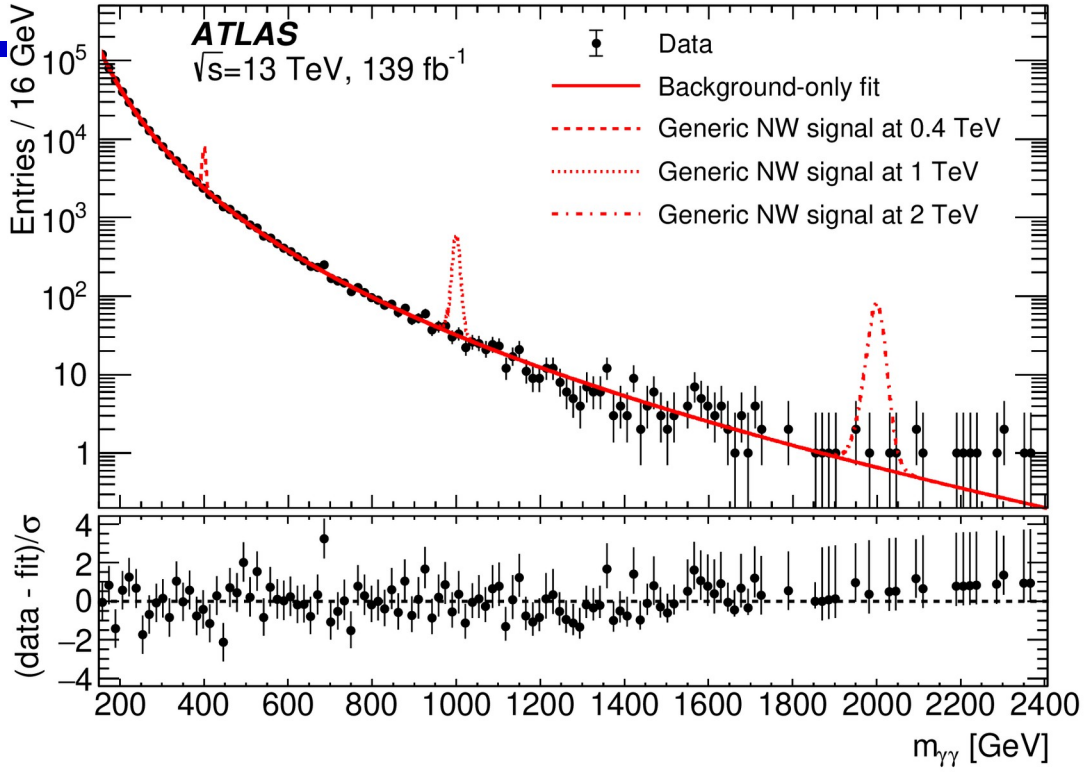
Many numbers! Need to define ($n_{bins} \times n_{samples} \times n_{NPs}$) impacts.

X→γγ Search

Unbinned search for narrow γγ resonances over a wide mass range.

Asymptotic approximation not valid in the tail
⇒ Need to use toys for limit computations
(not only expected, but also observed!)

Use binned approximation to the likelihood:
250 bins in log($m_{\gamma\gamma}$), 11 NPs, 1 POI
Need O(100k) toys for each signal mass point



⇒ Tested 500 mass points to get the observed limit, compute limits in toy datasets at O(10-100 Hz)
→ Good agreement with asymptotics at low masses, expected deviations at higher mass.

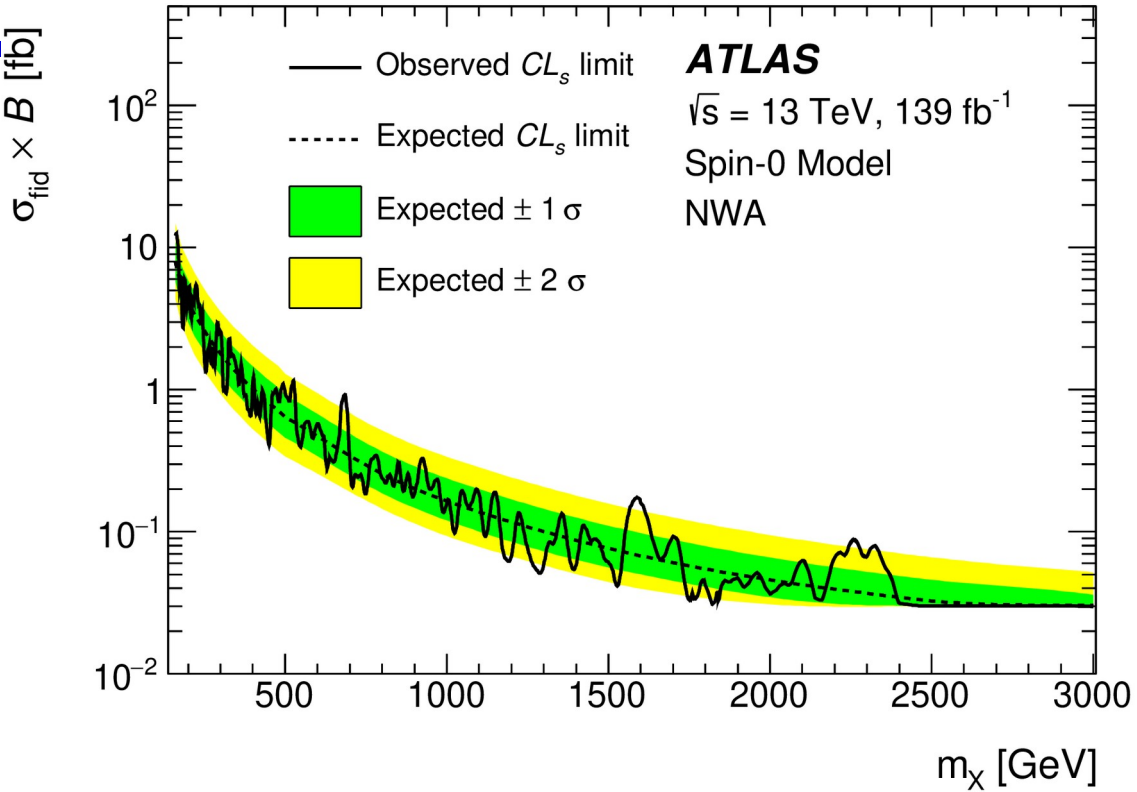
Same approach seems promising for simplified description of the unbinned H→γγ analysis, but model is very large (O(10000) bins × O(100) NPs × O(100) samples ⇒ 10^8 impact values)

X→γγ Search

Unbinned search for narrow γγ resonances over a wide mass range.

Asymptotic approximation not valid in the tail
⇒ Need to use toys for limit computations
(not only expected, but also observed!)

Use binned approximation to the likelihood:
250 bins in log($m_{\gamma\gamma}$), 11 NPs, 1 POI
Need O(100k) toys for each signal mass point



⇒ Tested 500 mass points to get the observed limit, compute limits in toy datasets at O(10-100 Hz)
→ Good agreement with asymptotics at low masses, expected deviations at higher mass.

Same approach seems promising for simplified description of the unbinned H→γγ analysis, but model is very large (O(10000) bins × O(100) NPs × O(100) samples ⇒ 10^8 impact values)

Outlook

- Fast linear profiling can provide very large speedups (× 10000) compared to non-linear minimization
- Linear NP impacts can provide a good approximation to full likelihoods, especially for small systematics (searches)
- Linear models can be created from HistFactory models (pyhf) in straightforward way
- Implements arbitrary number of POIs and POI dependence, can reparameterize to other POI sets (e.g. EFT ci)
- Can also be used for "brute-force" (small-bin) description of unbinned PDF, in spite of large n_{bins}.
- Complementary to other approaches (different simplified (or not) models can suite different needs…)

Code available on github (or gitlab for CERN-based users)

To install: pip install fastprof

Steps to reproduce the trilepton analysis shown here are available here