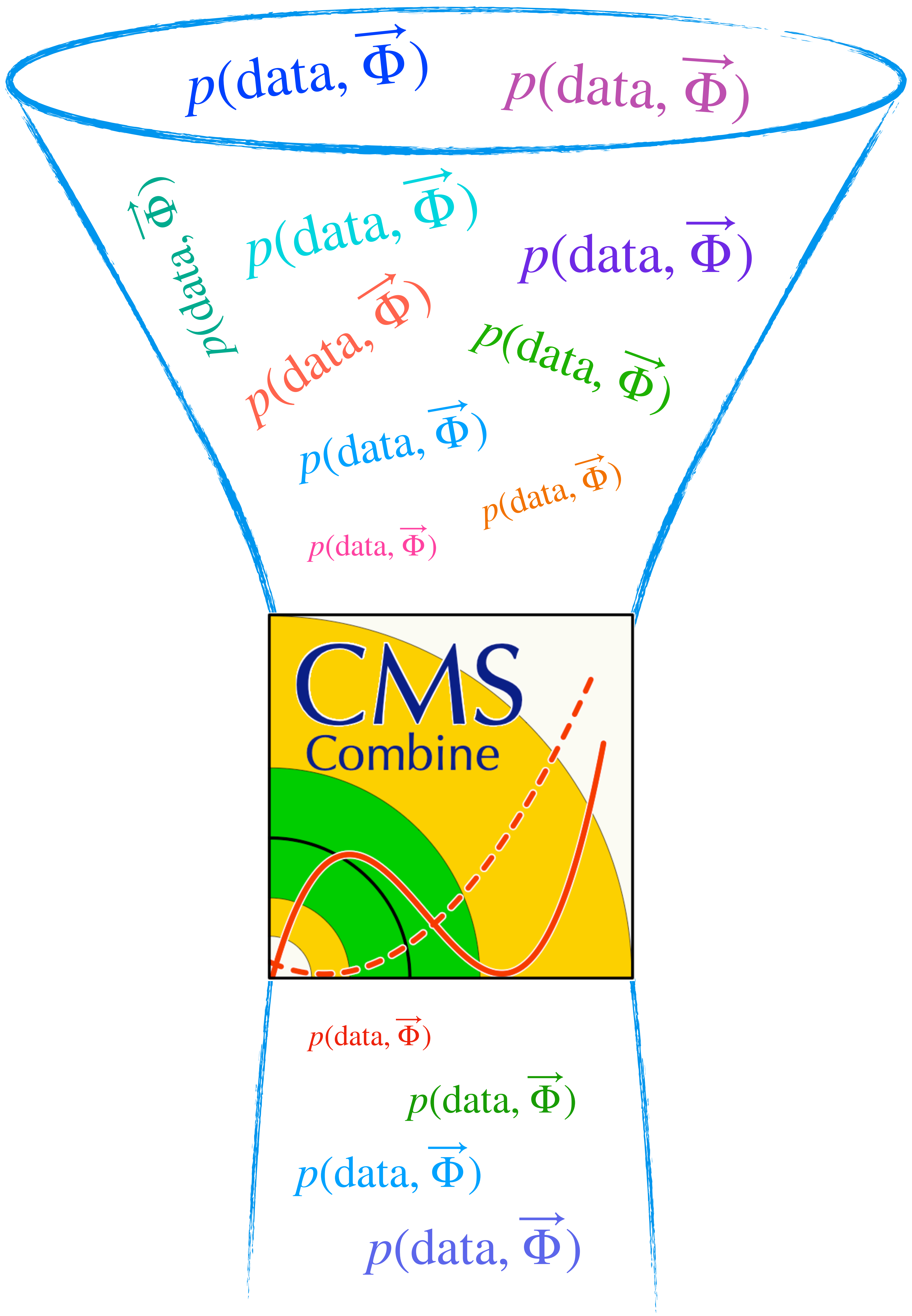


Publishing full statistical models of CMS physics analyses

Sezen Sekmen
Kyungpook National University
for the CMS Collaboration

ICHEP 2024, 17-24 July 2024, Prague

$p(\text{data}, \vec{\Phi})$

What is a statistical model?

Statistical model: The mathematical framework used to describe and make inferences about the underlying processes that generate observed data.

- Describes the probabilistic dependence of the observed quantities (i.e. data) on parameters of the model.
  - parameters of interest (POI): signal strength, resonance mass, …
  - nuisance parameters: not of direct interest, but required to explain data — uncertainties of experimental or theoretical origin: detector effects, background measurements, lumi calibration, cross-section calculation.
- Likelihood: Value of the statistical model at a given fixed set of data as a function of parameters.

Statistical model provides the complete mathematical description of an analysis and is the starting point of any interpretation.

$p(\text{data}, \vec{\Phi})$

Why publish statistical models?

Statistical models provide an excellent resource for the community.

Publishing them will help maximize the scientific impact of the analysis, and facilitate

- Preservation and documentation: the mathematical construction of the analysis in full detail.
- Combination of multiple analyses
- Reinterpretation and reuse (within and outside the collaborations):
- Education on statistics procedures
- Tool development: Statistical software updates can use real world examples to test and debug their recent developments.
- …

Unanimously accepted by stat gurus at the *"1st Workshop on Confidence Limits"*, 17-18 Jan 2000, CERN. See "panel discussion" in the Yellow Report CERN-2000-005.

Community report: *"Publishing statistical models: Getting the most out of particle physics experiments"*, SciPost Phys. 12, 037 (2022), arXiv:2109.04981 : Make the scientific case for statistical model publication, and discuss technical developments.

CMS starts publishing statistical models

$p(\text{data}, \vec{\Phi})$

In December 2023, CMS Collaboration took the decision to release statistical models for all forthcoming analyses by default.

- In accordance with open access policy of CERN and CMS.
- Must be well-documented and understandable.
- Publishing has always been highly desirable. The difficulty was technical implementation. Now we have a way.

Publish

- CMS statistical analysis package: ***Combine***
  - Containerized version + detailed publication
- Human-readable text files configuring the likelihood: ***Combine datacards***
  - Adhering to CMS-wide nuisance parameter naming conventions.

CMS Combine is public now

“The CMS statistical analysis and combination tool: COMBINE”, CMS-CAT-23-001, arXiv:2404.06614

- Detailed description and examples of statistical models constructed in combine
- Description of common statistical analysis routines
- Command-line examples to run the commonly used methods.
  - Calculation of maximum likelihood estimates, confidence / credible intervals, goodness-of-fit tests, diagnostics.
- Accompanied by pre-compiled containerized Combine release v9.2.0 [docs].

```
docker run [--platform linux/amd64] --name combine -it
gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.0
```


EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

CERN-EP-2024-078
2024/04/11

CMS-CAT-23-001

The CMS statistical analysis and combination tool: COMBINE

The CMS Collaboration*

Abstract

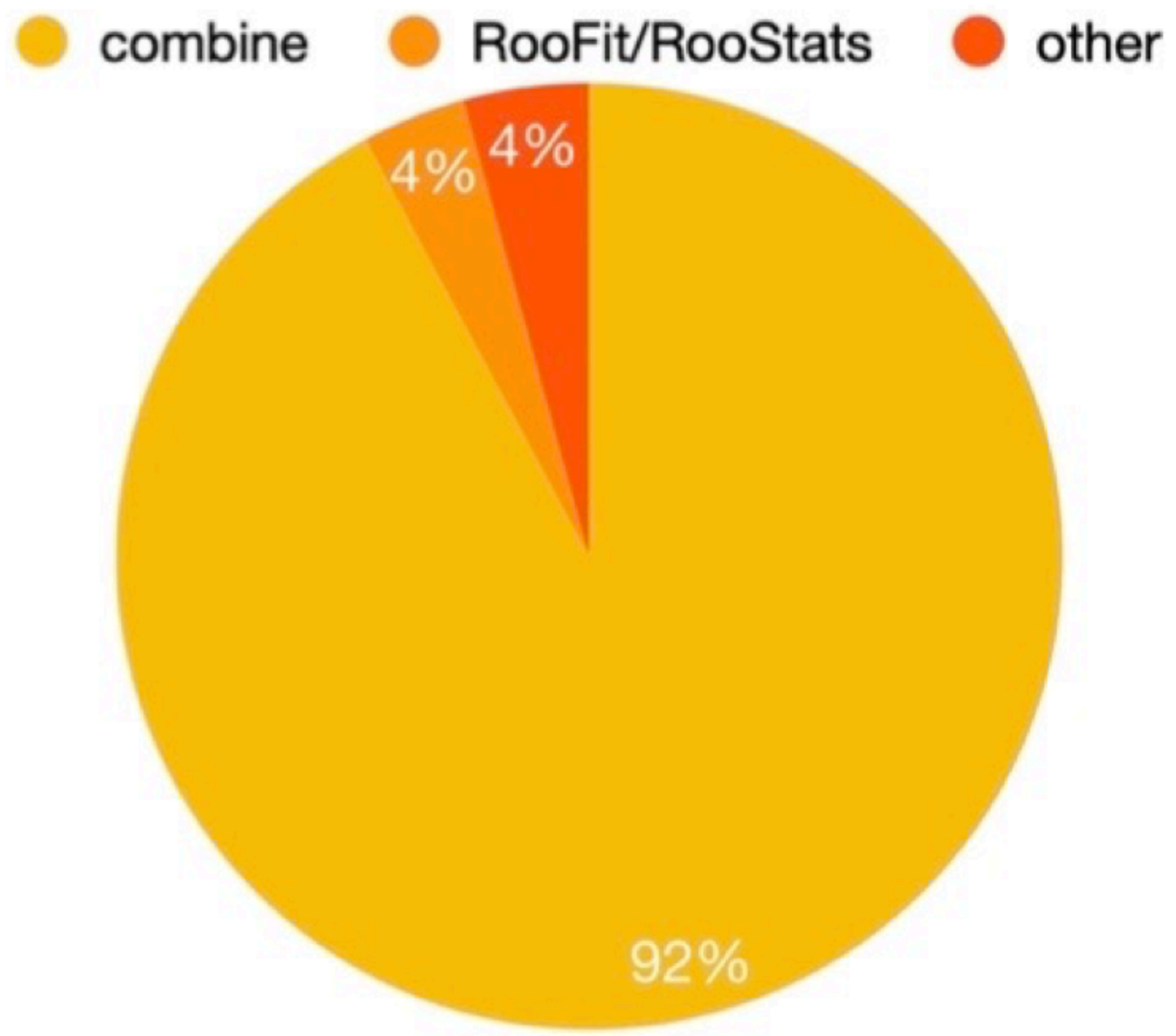
This paper describes the COMBINE software package used for statistical analyses by the CMS Collaboration. The package, originally designed to perform searches for a Higgs boson and the combined analysis of those searches, has evolved to become the statistical analysis tool presently used in the majority of measurements and searches performed by the CMS Collaboration. It is not specific to the CMS experiment, and this paper is intended to serve as a reference for users outside of the CMS Collaboration, providing an outline of the most salient features and capabilities. Readers are provided with the possibility to run COMBINE and reproduce examples provided in this paper using a publicly available container image. Since the package is constantly evolving to meet the demands of ever-increasing data sets and analysis sophistication, this paper cannot cover all details of COMBINE. However, the online documentation referenced within this paper provides an up-to-date and complete user guide.

Submitted to Computing and Software for Big Science

Introduction to Combine

Combine is the statistical analysis software used in CMS, built around ROOT, RooFit and RooStats:

- Command-line interface to several common workflows used in HEP statistical analysis (workflows recommended by CMS Statistics Committee).
- Encapsulate the statistical model in a human-readable configuration file, called the ***datacard***.
- Builds pre-defined statistical models: counting experiment, parametric shape (unbinned and binned), template-based shape
- Allows building custom stat models, e.g. with multiple signals.
- Powerful for combinations, scales well with model complexity
- Provides an extensive toolset for validation



From Statistics Committee Questionnaires 2021-2022

CMS' choice for statistical analysis!

but not specific to CMS

Statistical model built with Combine

Produce the statistical model $p(\text{data}, \vec{\Phi})$, where $\vec{\Phi}$ are the model parameters.

- For numerical efficiency, factorize into probabilities for
  - primary component: POI $\vec{\mu}$, primary observables $\vec{x}$
  - auxiliary component: nuisance parameters $\vec{\nu}$, auxiliary observables $\vec{y}$ that constrain $\vec{\nu}$

$$p(\vec{x}, \vec{y}; \vec{\Phi}) = p(\vec{x}; \vec{\mu}, \vec{\nu}) \prod_k p_k(\vec{y}_k; \vec{\nu}_k)$$

- Likelihood function is constructed by evaluating $p(\vec{x}, \vec{y}; \vec{\Phi})$ on a dataset

$$\mathcal{L}(\vec{\Phi}) = \prod_d p(\vec{x}_d; \vec{\mu}, \vec{\nu}) \prod_k p_k(\vec{y}_k; \vec{\nu}_k)$$

d runs over all entries in data

- Combine implements a (RooFit based) custom class to build the likelihood.
- Likelihood used in both frequentist and Bayesian calculations.

Supported models: Counting experiment, template shape

$$p(\vec{x}, \vec{y}; \vec{\Phi}) = p(\vec{x}; \vec{\mu}, \vec{\nu}) \prod_k p_k(\vec{y}_k; \vec{\nu}_k)$$

focus here

Counting analysis:

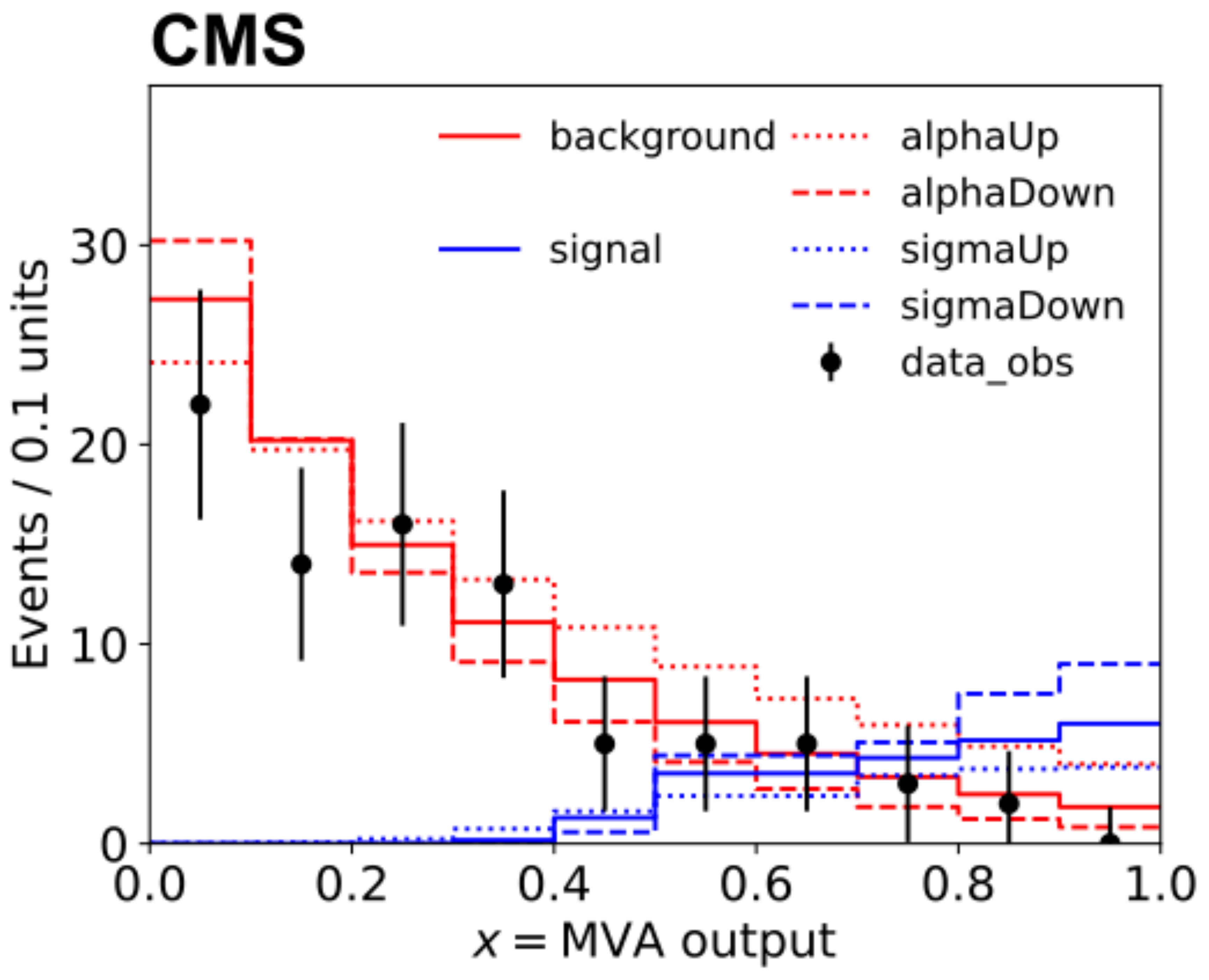
- only one primary observable, the total event count in a single channel.
- Poisson probability: $p(n; \lambda(\vec{\mu}, \vec{\nu})) = \lambda^n \frac{e^{-\lambda}}{n!}$

Template shape analysis:

- Observable in each channel partitioned into N_B bins.

$$p(x; \vec{\mu}, \vec{\nu}) = \prod_{b=1}^{N_B} P(n_b; \lambda(\vec{\mu}, \vec{\nu}))$$

Poisson probability



sigma, alpha: systematic uncertainties

- Input in histograms: data, central expectations, uncertainties as variations on expectations.
- Model most used in CMS.

Supported models: Parametric shape

$$p(\vec{x},\vec{y};\vec{\Phi}) = p(\vec{x};\vec{\mu},\vec{\nu}) \prod_k p_k(\vec{y}_k;\vec{\nu}_k)$$

focus here

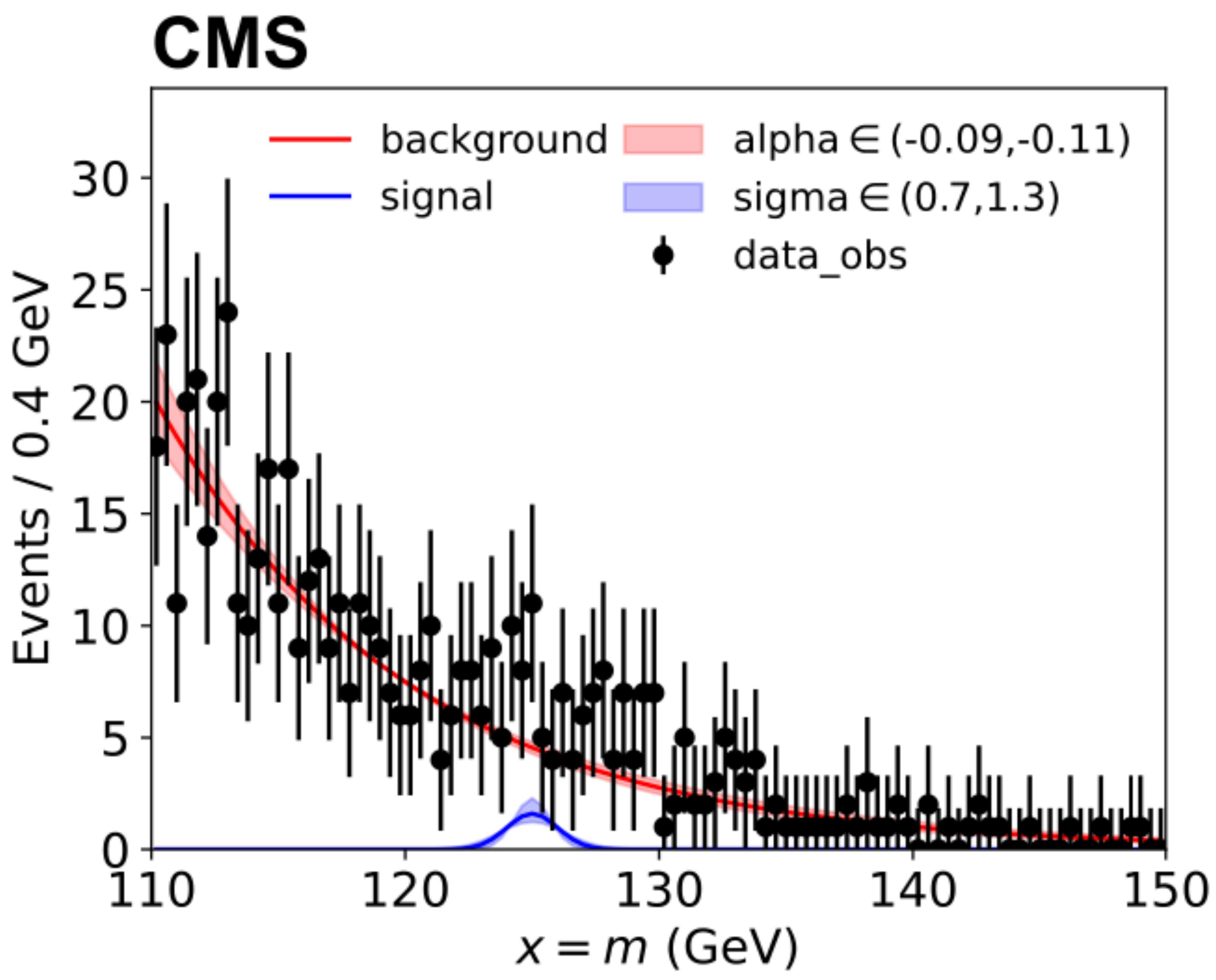
Parametric shape:

- Model uses analytic functions rather than histograms.
- e.g. Higgs -> γγ invariant mass fit.
- Data can be binned or unbinned.
- Uncertainties on the expected distributions are uncertainties on the analytic function parameters.

$$p(x;\vec{\mu},\vec{\nu}) = \sum_p \frac{\lambda_p(\vec{\mu},\vec{\nu}) f_p(x;\vec{\mu},\vec{\nu})}{\sum_p \lambda_p(\vec{\mu},\vec{\nu})}$$

pdfs for each process p

process



sigma, alpha: uncertainties on parameters of the analytic function.

Combine datacard structure

Example for a counting experiment:

```
imax 1        number of channels,
jmax 2        number of backgrounds,
kmax 3        number of nuisance parameters
# A single channel - ch1 - in which 0 events are observed in data
bin                ch1
observation        0
# ----------
bin                ch1    ch1    ch1
process            ppX    WW     tt
process            0      1      2
rate               1.47   0.64   0.22
# ----------
lumi     lnN       1.11   1.11    1.11
xs       lnN       1.20    -       -
nWW      gmN 4      -     0.16     -
```

Nuisance - constraint - effect on processes

Primary and auxiliary likelihood components.

$$\mathcal{L}(\vec{\Phi}) = \prod_d p(\vec{x}_d; \vec{\mu}, \vec{\nu}) \prod_k p_k(\vec{y}_k; \vec{\nu}_k)$$

Dedicated formats exist for template-shape, parametric-shape, multi-signal, multiplicative scale factors, etc.

Extracting results with Combine - I

```
combine <datacard.[txt|root]> -M <method>
```

- HybridNew: compute modified frequentist limits with pseudo-data, p-values, significance and confidence intervals with several options, `--LHCmode LHC-limits` is the recommended one.
- AsymptoticLimits: limits calculated according to the asymptotic formulas in arxiv:1007.1727, valid for large event counts.

LHC-style test statistic, defined using a ratio of profile likelihoods.

$$\tilde{q}_{\mathrm{LHC}}(\mu) = \begin{cases} -2\ln\left(\dfrac{\mathcal{L}(\mu,\hat{\hat{\vec{\nu}}}(\mu))}{\mathcal{L}(\hat{\mu},\hat{\vec{\nu}})}\right) & \text{if } 0 \leq \hat{\mu} \leq \mu, \\ -2\ln\left(\dfrac{\mathcal{L}(\mu,\hat{\hat{\vec{\nu}}}(\mu))}{\mathcal{L}(0,\hat{\hat{\vec{\nu}}}(0))}\right) & \text{if } \hat{\mu} < 0, \\ 0 & \text{if } \hat{\mu} > \mu, \end{cases}$$

$\hat{\mu}$: Maximum likelihood estimator.

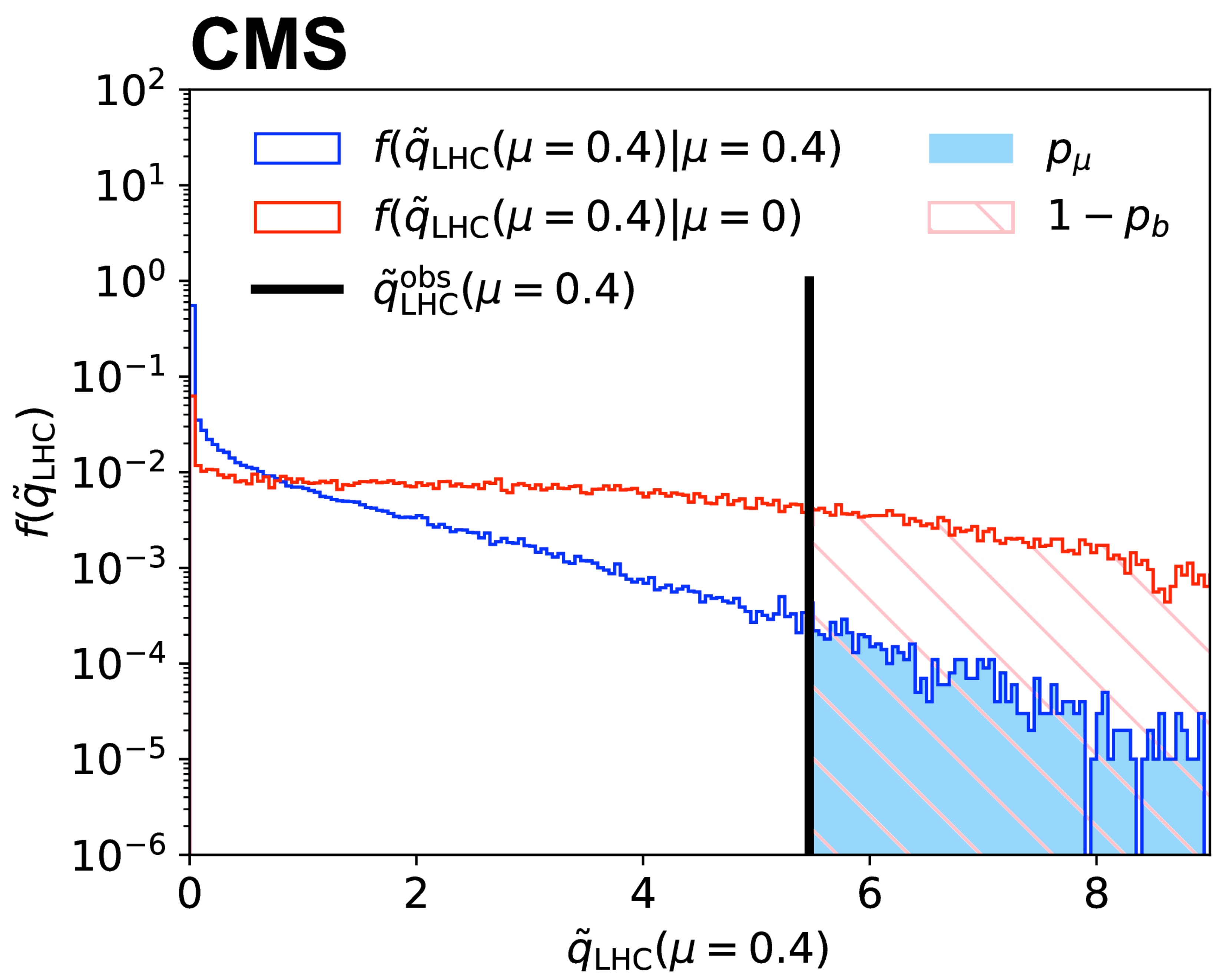
$\hat{\hat{\vec{\nu}}}(\mu)$, $\hat{\vec{\nu}}$ Values of nuisance parameters that maximize the likelihood for a specific value of μ and for $\mu = \hat{\mu}$

Extracting results with Combine - II

```
combine <datacard.[txt|root]> -M <method>
```

- HybridNew: compute modified frequentist limits with pseudo-data, p-values, significance and confidence intervals with several options, `--LHCmode LHC-limits` is the recommended one.
- AsymptoticLimits: limits calculated according to the asymptotic formulas in arxiv:1007.1727, valid for large event counts.

Test statistic distributions f and p-values for a template-based shape model:



$$p_\mu = \int_{q_x^{obs}(\mu)}^{\infty} f(q_x(\mu)\,|\,\mu)dq \qquad p_b = \int_0^{q_x^{obs}(\mu)} f(q_x(\mu)\,|\,0)dq$$

$$\mathrm{CL_s} = p_\mu/(1-p_b)$$

Value of μ at $\mathrm{CL_s} = \alpha$

—> upper limit on μ at $100(1-\alpha)\,\%$

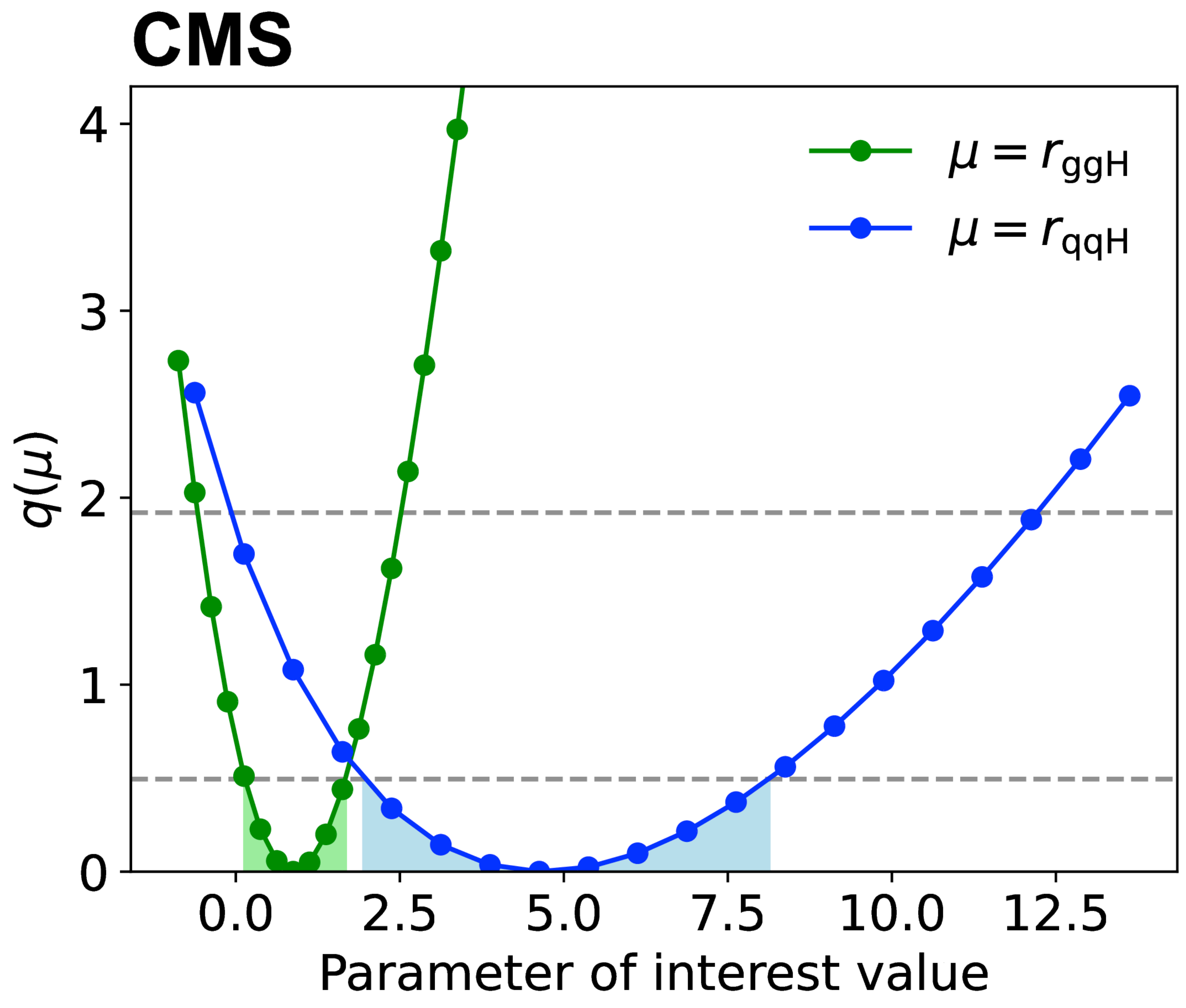
Extracting results with Combine - III

```
combine <datacard.[txt|root]> -M <method>
```

- HybridNew: compute modified frequentist limits with pseudo-data, p-values, significance and confidence intervals with several options, `--LHCmode LHC-limits` is the recommended one.
- AsymptoticLimits: limits calculated according to the asymptotic formulas in arxiv:1007.1727, valid for large event counts.
- Significance: simple profile likelihood approximation for calculating significances.
- BayesianSimple and MarkovChainMC compute Bayesian upper limits and credible intervals for simple and arbitrary models.
- MultiDimFit: perform maximum likelihood fit, with multiple POIs, estimate CI from likelihood scans.

Profile likelihood ratio $q(\vec{\mu}) = -\ln\left(\frac{\mathcal{L}(\vec{\mu}, \hat{\hat{\vec{\nu}}}(\vec{\mu}))}{\mathcal{L}(\hat{\vec{\mu}}, \hat{\vec{\nu}})}\right)$

for a multisignal model, and 68% CL intervals for each POI (ggH and qqH signal strength)

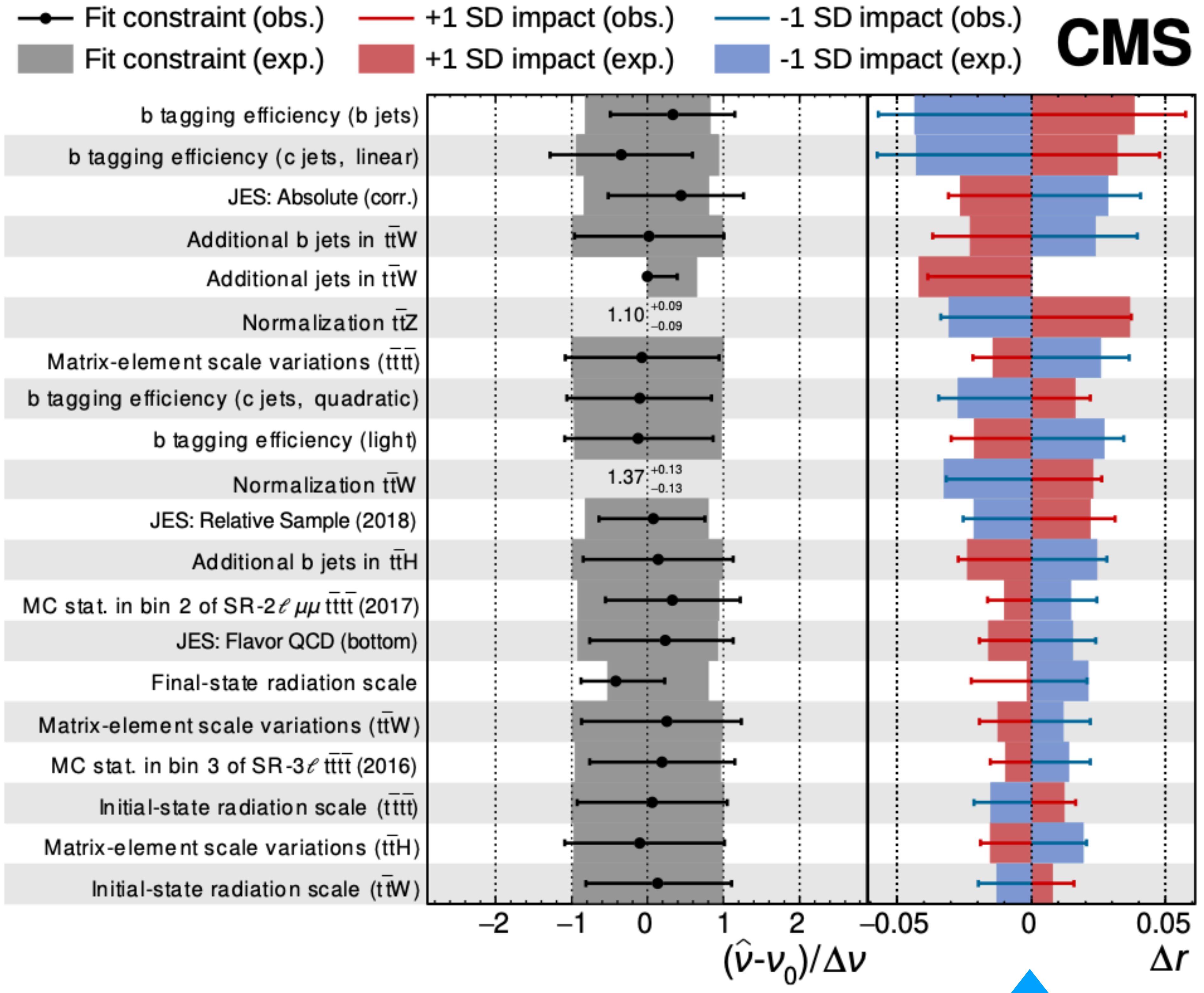


Combine diagnostic tools

```
combine <datacard.[txt|root]> -M <method>
```

- GoodnessOfFit: perform a goodness of fit test for models including shape information using several GoF estimators (saturated, Kolmogorov-Smirnov, Anderson-Darling)
- Impacts: evaluate the shift in POI from $\pm\sigma_{\text{postfit}}$ variation for each nuisance parameter.
- ChannelCompatibiltyCheck: check how consistent are the individual channels of a combination are
- GenerateOnly: generate random or Asimov pseudo-datasets for use as input to other methods

Nuisance parameter uncertainties and impacts for the observation of four top quark production:



$$\Delta\mu^{\pm} = \hat{\mu}(\nu_k \pm \Delta^{\pm}\nu_k) - \hat{\mu}$$

Combine web documentation [link]

Combine v10.0.X Search

Introduction

These pages document the RooStats / RooFit - based software tool used for statistical analysis within the CMS experiment - Combine. Note that while this tool was originally developed in the Higgs Physics Analysis Group (PAG), its usage is now widespread within CMS.

Combine provides a command-line interface to many different statistical techniques, available inside RooFit/RooStats, that are used widely inside CMS.

The package exists on GitHub under https://github.com/cms-analysis/HiggsAnalysis-CombinedLimit

For more information about Git, GitHub and its usage in CMS, see http://cms-sw.github.io/cmssw/faq.html

The code can be checked out from GitHub and compiled on top of a CMSSW release that includes a recent RooFit/RooStats, or via standalone compilation without CMSSW dependencies. See the

- Installation (w/ CMSSW, standalone, using LCG, conda, latest release) [link]
- Setting up the analysis (counting, template based, parametric) [link]
- Running Combine [link]
- Underlying statistics
  - Likelihood definition [link]
  - How the fits are performed (profiling, marginalization, confidence intervals) [link]
  - Statistical tests (test statistic, goodness-of-fit) [link]
- Tutorials: main features (template-based model), parametric, unfolding

First statistical model: Higgs boson observation - I

CDS | Search records... | Communities | My dashboard | Log in

CMS statistical models

Published April 15, 2024 | Version v1.0   Model   Open

CMS Higgs boson observation statistical model

CMS Collaboration

Introduction

This resource contains the full statistical model from the Higgs Run-1 combination, which led to the Higgs boson discovery, in the format of `Combine` datacards. The instructions below include a few basic examples on how to extract the significance and signal strength measurements, for more details please consult the `Combine` documentation.

Datacards

Datacards for the combination (and per-decay channel sub-combinations) leading to the Higgs-boson discovery at CMS are in the `125.5` folder. The nuisance parameters corresponding to different sources of systematic uncertainties are described in the `*.html` files located in that folder.

For the full combination of decay channels, the relevant datacard is `125.5/comb.txt`. The individual datacards for each of the analyses in CMS targeting the main Higgs boson decay modes are also in the `125.5` folder.

Software instructions

General installation instructions for `Combine` can be found in the `Combine` documentation.

A container image is provided to ensure reproducible results. The results in this README are obtained using `v9.2.1`:

```
docker run --name combine -it gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1
```

A slim version of the container image is also available at `gitlab-registry.cern.ch/cms-cloud/combine-standalone:v9.2.1-slim`. Versions of packages in the slim container image do not match exactly with the ones in the default container, so small differences in the output of commands with respect to the ones shown below are to be expected.

You can copy files (such as the datacards and other inputs for `combine`) using `docker cp` as documented here.

For the commands below, you may require running `ulimit -s unlimited; ulimit -u unlimited` to avoid memory issues.

Significance Calculation

2K VIEWS | 164 DOWNLOADS

Show more details

Versions

Version v1.0 — Apr 15, 2024
10.17181/c2948-e8875

Cite all versions? You can cite all versions by using the DOI 10.17181/2cp5k-ggn24. This DOI represents all versions, and will always resolve to the latest one. Read more.

Communities

CMS statistical models

Details

DOI *(Cite this version - v1.0)*
DOI 10.17181/c2948-e8875

DOI *(Cite all versions)*
DOI 10.17181/2cp5k-ggn24

Resource type
Model

Publisher
CERN

CMS Higgs Run 1 combination of 5 main Higgs channels CMS-HIG-12-028.

Public statistical model in CERN Document Server [link]

Table snippet describing nuisance parameters

	class	description
BR_hzz	branching_ratios	uncertainty on the branching ratio of higgs to Z bosons
CMS_zz4l_bkgMELA	custom	shape uncertainties jet energy scale and resolution modifying the background shape
CMS_hzz4mu_Zjets	custom	uncertainty on irreducible Z+jets background split in different channels
CMS_hzz4e_Zjets	custom	uncertainty on irreducible Z+jets background split in different channels
CMS_hzz2e2mu_Zjets	custom	uncertainty on irreducible Z+jets background split in different channels
pdf_hzz4l_accept	custom	acceptance uncertinty derived in h->4l analysis

First statistical model: Higgs boson observation - II

1. Combine channels

```
combineCards.py 125.5/comb_hgg.txt 125.5/comb_hzz.txt > 125.5/comb_hgg_hzz.txt
combine 125.5/comb_hgg_hzz.txt --mass 125.5 -M Significance
```

2. Calculate the significance

```
combine 125.5/comb.txt --mass 125.5 -M Significance
```

The output will be:

```
 <<< Combine >>>
 <<< v9.2.1 >>>
>>> Random number generator seed is 123456
>>> Method used is Significance

 -- Significance --
Significance: 4.87557
Done in 1.76 min (cpu), 1.76 min (real)
```

3.Measure the signal strength

We can measure the signal strength of the Higgs boson (`r`) and its uncertainty using `Combine`,

```
combine 125.5/comb.txt -m 125.5 -M MultiDimFit --algo singles --setParameterRanges r=0.2,1.5
```

and the output will be:

```
 <<< Combine >>>
 <<< v9.2.1 >>>
>>> Random number generator seed is 123456
>>> Method used is MultiDimFit
Set Range of Parameter r To : (0.2,1.5)
Doing initial fit:

 --- MultiDimFit ---
best fit parameter values and profile-likelihood uncertainties:
   r :    +0.785   -0.204/+0.218 (68%)
Done in 1.86 min (cpu), 1.86 min (real)
```

4. Build a model as H-vector boson, H-fermion coupling modifiers as POIs: [HiggsCouplings_ICHEP12:cVcF]

```
text2workspace.py -P HiggsAnalysis.CombinedLimit.HiggsCouplings_ICHEP12:cVcF 125.5/comb.txt -m 125.5 -o comb_kVkF.root
```

Note that since the discovery, the Physics Model `HiggsCouplings_ICHEP12:cVcF` has evolved to use make use of higher precision theoretical calculations, but for the discovery analysis, this is the model that was used.

We can measure these parameters and their uncertainties with,

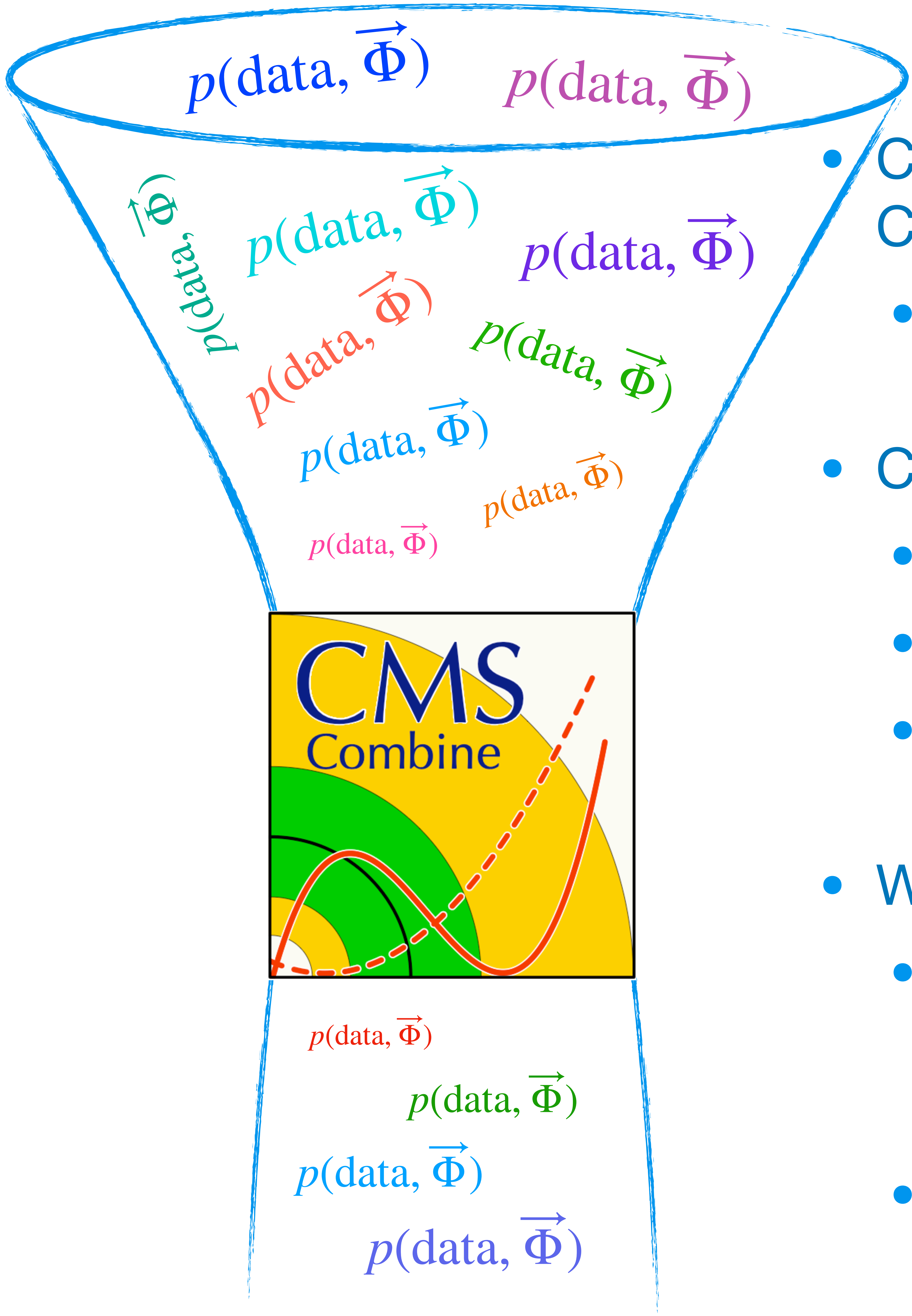
```
combine comb_kVkF.root -m 125.5 -M MultiDimFit --algo singles
```

```
 --- MultiDimFit ---
best fit parameter values and profile-likelihood uncertainties:
   CV :    +0.946   -0.120/+0.113 (68%)
   CF :    +0.497   -0.170/+0.203 (68%)
Done in 3.09 min (cpu), 3.09 min (real)
```

Upcoming: Models for BSM searches

- Datacards for several BSM searches are on their way to publication.
  - Coming very soon: SUSY disappearing track search (CMS-SUS-21-006).
- Particularly interesting for BSM reinterpretation studies.
- Challenge for BSM models: multiple BSM model parameters, large scans.
  Each datacard has BSM model-dependent signal systematics.
  - How to avoid publishing thousands of datacards?
    Solution: Provide a single template datacard + interpolated functions of rates and signal systematics per region versus physics model parameters (e.g. SUSY particle masses) for each physics model
    - Interpolation via RooSplineND for counting-style analyses, or automated via keyword input for shape analyses.
  - Open question: How to treat systematics when there is a brand new signal model?

Summary, outlook



- CMS released the first statistical model implemented in Combine and in process of releasing more statistical models.
  - Upcoming CMS analyses are adapting a nuisance parameter naming convention to facilitate publication.
- Combine tool is public: documentation + standalone container
  - Self-documenting statistical model building
  - Extensive toolset for statistical inference
  - Constantly improving documentation, ensuring compatibility with the latest ROOT version
- Working towards compatibility with other formats:
  - Combine <—> pyHF conversion tool was developed and extensively validated in the context of ATLAS + CMS tttt EFT combination [github].
  - Work started to implement HS3 (HEP Statistics Serialization Standard) [link].