Aspects of HEP – (Re-)interpretation

Thiago R. F. P. Tomei

SPRACE-Unesp

Disclaimer

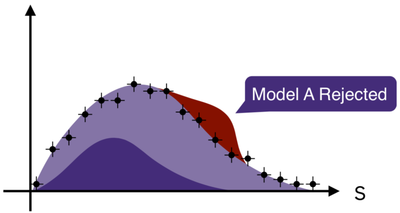
- Most material in this talk adapted from other talks:
  - K. Cranmer: https://indico.cern.ch/event/962997/
  - N. Wardle: https://indico.cern.ch/event/1012319/
- Interface between experimental physics and phenomenology
  - Not something I usually work on.
  - Imposes requirements on my work.
- Reinterpretation wants impose reproducibility needs
  - But not going to discuss this here.

Analysis in a Nutshell

- Filter your collision events
  - Subset relevant to test the hypotheses we are considering.
  - Example: standard model and "new physics Model A", with a free parameter θ_A.
- Design a summary statistic s
  - Distinguish between the different hypotheses.
- Build a statistical model p

$$p(s|\text{model A}, \theta_A)$$

- Test the hypothesis, report results, **publish the statistical model** p.



Note that I am describing a "search for new physics" analysis.

A measurement of a standard model parameter is different.

Interpreting the Results

If Model A has a free parameter θ_A, it is really a **family of models**.

- Example: standard model before Higgs discovery, where m_H was a free parameter.
- Example: a Z' model, mass of the Z' is a free parameter, spin and couplings fixed.

Dealing with a parameterised family of models:

- Consider different values of θ_A.
- Make multiple pairwise $\{H_0, H_1(\theta_A)\}$ hypothesis tests.
- Invert the hypothesis test to obtain a **confidence interval** on θ_A.

Easy for one parameter, complicated for multidimensional case.

Historical Example: D0 Search for RS Gravitons

RS1 Model

- mass of the RS1 graviton M_1
- coupling constant $\tilde{k} = k/\overline{M}_{\text{Pl}}$

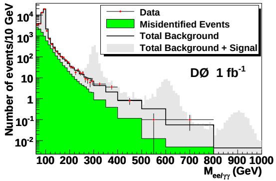
Search for $p\bar{p} \to G \to ee/\gamma\gamma$ process.

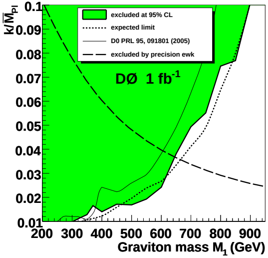
Selection: $\sim$ events with two opposite-sign electrons or two photons in acceptance.

- p_{T}, η, quality cuts

Statistical model:

- s: $\sim$ number of $M_{ee/\gamma\gamma}$ events in range.
- H_0 = SM, H_1 = SM + RS1(M_1,$\tilde{k}$)
- Exclude points in (M_1, $\tilde{k}$) space.





Phys. Rev. Lett. 100:091802, 2008

(Re-)interpreting the Results

The statistical model $p(s|\text{Model A}, \theta_A)$ is great for combinations and studies within Model A, but it isn't useful for answering questions about a different Model B!

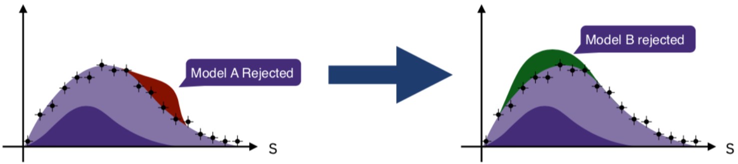
- The efficiency, acceptance, and distribution $p(s|\text{Model B}, \theta_B)$ for the new signal will all be different.

Sometimes Models A and B are similar, and the original analysis will be sensitive to Model B!

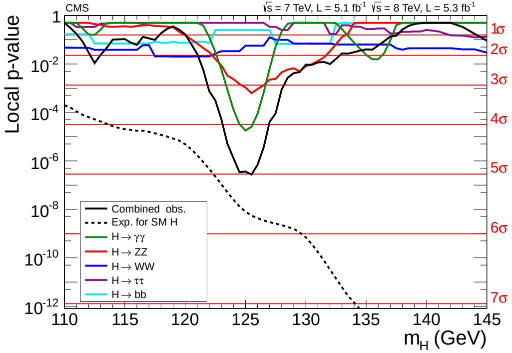
- For instance, $G \to ee$ and $Z' \to ee$ are not that different.

Capture the definition of the summary $s(x)$ and the event selection, reuse/reinterpret the existing analysis!

- Prediction for the null and observation in the data.



Combining Results



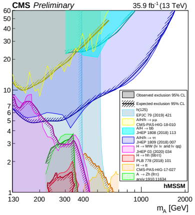
Combinations within the same model:

- Different channels:
  $H \to \gamma\gamma, ZZ, WW, \tau\bar{\tau}, b\bar{b}$
- Different datasets: $\sqrt{s} = 7$, 8, 13 TeV
- Different colliders: LEP, Tevatron, LHC
- Different experiments: colliders, telescopes, . . .

How do you combine results when you're not part of the experimental collaborations?

- Need access to the **experimental likelihoods**.

How do the experiments present results?



Simplified model limits

Inclusive cross-sections

Simplified likelihoods

REANA/RECAST workflows

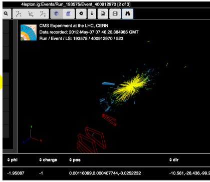
Exclusion limits/contours in UV-complete models

Signal strengths / Amplitudes

Unfolded / Fiducial cross-sections

Full likelihoods

Open data



Most assumptions (easy to use)

No assumptions (more expert level)

From N. Wardle and P. Owen at Reinterp2021

Extreme Case 1: Full UV-complete Model

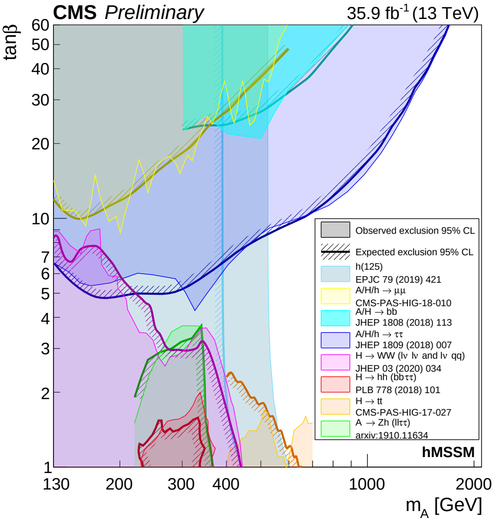
Most sensible for well-established model (SM).

- May prefer to present signal strengths μ and amplitudes instead (see later).

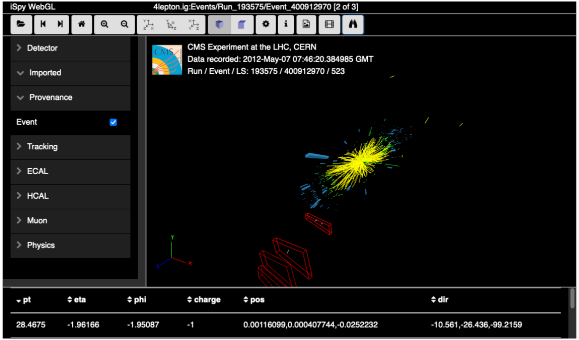
Usually also done for benchmark scenarios:

- hMSSM SUSY scenario
- M_h125 SUSY scenario (see arxiv:1808.07542)

But brings back the question:
what if I want to study a model that has a very similar signature?



Extreme Case 2: Open Data



http://opendata.cern.ch/

Release raw experimental data.

- Usually after embargo period.

Pro:

- Allows for maximum data preservation and reusability.

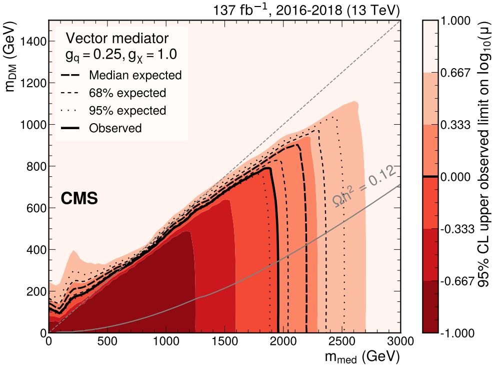
Cons:

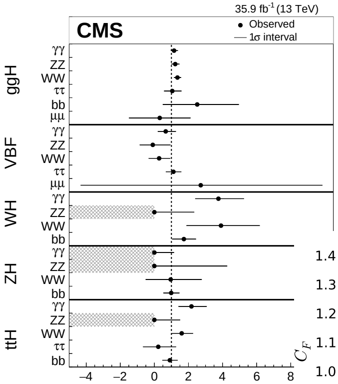
- Petabytes of data
- Large computing power needs (comparable with WLCG).
- High complexity of code for data reconstruction and analysis.

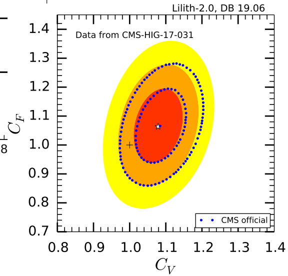
Simplified Models

Assume only a few particles are relevant to the signature in question.

- Simplified dark matter: mediator Z', dark matter χ.
- Simplified SUSY: pair production of
  - Gluinos $\tilde{g}$, with $\tilde{g} \to q\bar{q} + \tilde{\chi}_1^0$.
  - Squarks $\tilde{q}$, with $\tilde{q} \to q + \tilde{\chi}_1^0$
  - Chargino-neutralino ($\tilde{\chi}_1^\pm$ and $\tilde{\chi}_2^0$) or sleptons $\tilde{\ell}$, with $\tilde{\chi}_1^\pm/\tilde{\chi}_2^0 \to W/Z + \tilde{\chi}_1^0$.







arxiv:1908.03952

Very common on Higgs physics results:

$$\mu_i = \frac{\sigma_i}{(\sigma_i)_{\rm SM}} \quad \text{and} \quad \mu^f = \frac{{\rm BR}^f}{({\rm BR}^f)_{\rm SM}}$$

Assume only total rate of $ii \to H \to ff$ modified by new physics.

Profile likelihood:

$$-2\log L(\boldsymbol{\mu}) = (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^T C^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})$$

$$C = \frac{1}{4}\left[\boldsymbol{\sigma}^+ + \boldsymbol{\sigma}^-\right] \cdot \rho \cdot \left[\boldsymbol{\sigma}^+ + \boldsymbol{\sigma}^-\right]$$

Reparameterise in terms of coupling modifiers:

$$\mu_i, \mu^f \to \mu_i\left(C_V, C_F\right), \mu^f\left(C_V, C_F\right)$$

Caveat: "variable Gaussian" approximation.

Unfolding (1)

Detector simulation and reconstruction

- Collaborations compare reconstructed (real) data with reconstructed (simulated) data.
- Reconstruction is, in a sense, a partial way to get back to the original particles.
- **Unfolding:** correcting for smearing effects using a nonparametric estimator.

A way to "undo" the detector effects

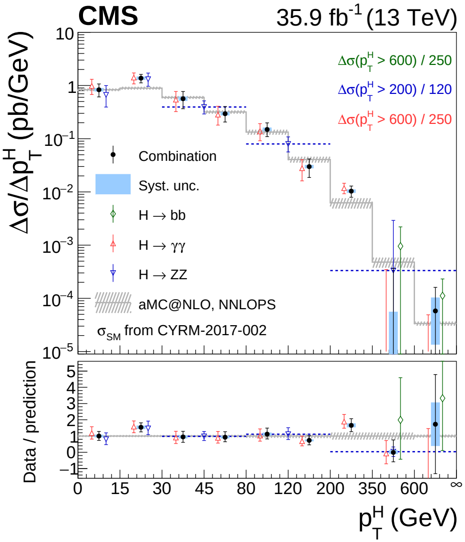
- Let $\boldsymbol{y}$ be a histogram of smeared observations and $\boldsymbol{\lambda}$ the mean particle-level histogram.
- Unfolding then consists of solving the Poisson regression problem $\boldsymbol{y} \sim \mathrm{Pois}(\boldsymbol{K}\boldsymbol{\lambda})$. where the elements of the smearing matrix $\boldsymbol{K}$ are given by

$$K_{ij} = P(\text{smeared event in bin } i | \text{true event in bin } j)$$

- Efficiency to observe an event originating from the true bin j is given by $\epsilon_j = \sum_i K_{ij}$.

Note: in mathematics, signal processing, image processing this is usually called deconvolution.

Unfolding (2)



Pros:

- ☐ Removes the need to model the detector to compare to theory.
- ☐ Systematic uncertainties included in the measurements.

Cons:

- ☐ Typically $\boldsymbol{K}$ is an ill-conditioned matrix!
  - Classical estimators of $\boldsymbol{\lambda}$ are very sensitive to the Poisson fluctuations in $\boldsymbol{y}$.
  - Regularization technique

  $\Rightarrow$ can introduce bias.
- ☐ Often involves Gaussian approximations.
- ☐ How to deal with ML-based quantities?

Likelihood Refresher (1)

Given a **probability model** $p(X|\theta)$ and data x_0:

- The **likelihood function** is a function of the parameter θ only, and its value is given by $L(\theta) = p(X = x_0|\theta)$
- Notice that that $L(\theta)$ doesn't describe the distribution in X.
- Technically the likelihood function doesn't have enough enough information to generate synthetic data (a.k.a. "toy Monte Carlo"), which is needed for most frequentist statistical procedures.

Notice: HEP practitioners often use the term "likelihood" to mean the full probability model.

Likelihood Refresher (2)

General form for experimental likelihood

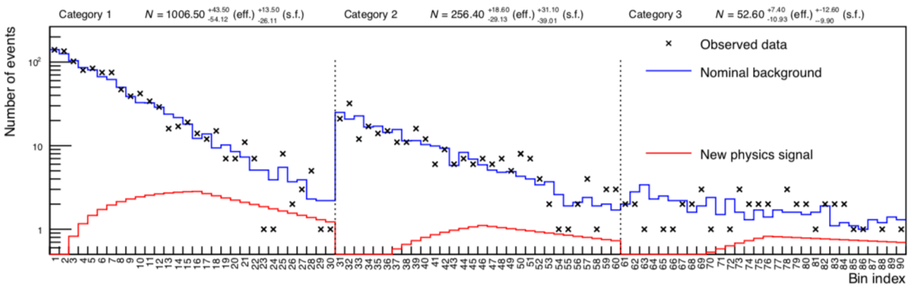
$$L(\boldsymbol{\alpha}, \boldsymbol{\delta})\pi(\boldsymbol{\delta}) = \prod_{I=1}^{P} \mathrm{Pr}\left(n_I^{\mathrm{obs}} \mid n_I(\boldsymbol{\alpha}, \boldsymbol{\delta})\right) \pi(\boldsymbol{\delta})$$

- $\boldsymbol{\alpha}$: parameters of interest
  - Mass of new hypothetical particle
  - Cross-section for new process
- $\boldsymbol{\delta}$: nuisance parameters
- Profiled LH ratio: one param. of interest: $\boldsymbol{\alpha} = \mu$, common multiplier for signal yield.
- Sum over signal / background contributions

$$n_I(\mu, \boldsymbol{\delta}) \to \mu \cdot \sum_{\text{sigs}} n_{s_k,I} + \sum_{\mathrm{bkgs}} n_{b_k,I}(\boldsymbol{\delta}) \to \mu \cdot n_{s,I} + n_{b,I}(\boldsymbol{\delta})$$

- Binned likelihood: $\mathrm{Pr}(n|\lambda) = \dfrac{\lambda^n}{n!} e^{-\lambda}$
- Nuisance parameter “in-situ” measurements: $\pi(\boldsymbol{\delta})$

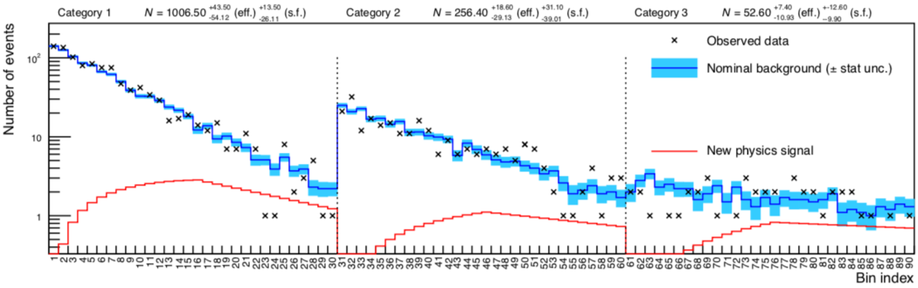
Example Search for New Physics (1)



Imagine a (rather simplified) model inspired by a typical search for a new physics signature.

- Single source of background (can also think of this as the sum of all backgrounds)
- The data (observations) are divided into regions
  - 3 categories for the data $\rightarrow$ each category has 30 bins
  - Increasing S/B with bin-number, within each category

Example Search for New Physics (2)



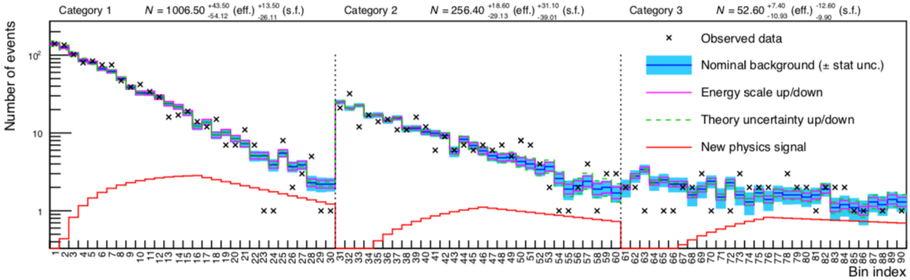
Two uncertainties on the background yields (N)

- "Efficiency" and "scale factor" (data/MC correction)

Each bin has an uncertainty which is uncorrelated between bins

- For instance, from limited simulated sample size that is used to estimate n_I

Example Search for New Physics (3)



Another two uncertainties correlated between bins

- "Energy scale" and "theory" uncertainty

Total: 94 nuisance parameters

Example Search for New Physics (4)

Expected number of background events in a given bin I is the fraction of events in that bin (f_I) multiplied by the total number of events (N): $n_I \equiv f_I(\boldsymbol{\delta})N(\boldsymbol{\delta})$.

$\boldsymbol{\delta}$ are nuisance parameters representing independent sources of uncertainty.

- Here we have 94 of them.

Uncertainties in the normalisation (N) usually follow lognormals: $N(\boldsymbol{\delta}) = N^0 \cdot \prod_j (1 + K_j)^{\delta_j}$.

Similarly for uncorrelated bin-by-bin uncertainties: $\dfrac{n_I(\boldsymbol{\delta})}{n_I^0} = \prod_j (1 + \epsilon_{Ij})^{\delta_j}$.

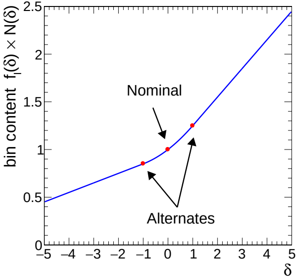
K_j and ϵ_{Ij} represent the relative size and direction of the uncertainty.

Example Search for New Physics (5)

Effects of correlated systematic uncertainties on n_I are more complicated.

Model them using quadratic (linear) interpolation (extrapolation) functions:

$$f_I(\boldsymbol{\delta}) = f_I^0 \cdot \frac{1}{F(\boldsymbol{\delta})} \prod_j p_{Ij}(\delta_j), \text{ w/ } F(\boldsymbol{\delta}) = \sum_I f_I(\boldsymbol{\delta}).$$



$$p_{Ij}(\delta_j) = \begin{cases} \frac{1}{2}\delta_j(\delta_j - 1)\kappa_{Ij}^- - (\delta_j - 1)(\delta_j + 1) + \frac{1}{2}\delta_j(\delta_j + 1)\kappa_{Ij}^+ & \text{for } |\delta_j| < 1 \\ \left[\frac{1}{2}\left(3\kappa_{Ij}^+ + \kappa_{Ij}^-\right) - 2\right]\delta_j - \frac{1}{2}\left(\kappa_{Ij}^+ + \kappa_{Ij}^-\right) + 2 & \text{for } \delta_j > 1 \\ \left[2 - \frac{1}{2}\left(3\kappa_{Ij}^- + \kappa_{Ij}^+\right)\right]\delta_j - \frac{1}{2}\left(\kappa_{Ij}^+ + \kappa_{Ij}^-\right) + 2 & \text{for } \delta_j < -1 \end{cases}$$

The Full Experimental Likelihood

Now we can finally write the likelihood for this search:

$$L(\mu, \boldsymbol{\delta})\pi(\boldsymbol{\delta}) = \prod_{I=1}^{90} P\left(n_I^{\text{obs}} \mid \mu \cdot n_{s,I} + n_{b,I}(\boldsymbol{\delta})\right) \cdot \prod_{j=1}^{94} e^{-\delta_j^2}$$

with

$$n_{b,I}(\boldsymbol{\delta}) = N_c^0 \cdot \prod_{k=1}^{2} (1 + K_k)^{\delta_k} \cdot f_I^0 \cdot \frac{1}{F(\boldsymbol{\delta})} \prod_{j=3}^{4} p_{I,j}\left(\delta_j\right) \cdot (1 + \epsilon_I)^{\delta_I}$$

Writing it in this general form means that we can publish the full likelihood in standard, plain, human-readable format!

But wait, how many terms to write the likelihood again? Answer: we need **729 inputs**.

$$L(\mu, \boldsymbol{\delta})\pi(\boldsymbol{\delta}) = \prod_{I=1}^{90} P\left(n_I^{\text{obs}} \mid \mu \cdot n_{s,I} + n_{b,I}(\boldsymbol{\delta}) \right) \cdot \prod_{j=1}^{94} e^{-\delta_j^2}$$

with

$$n_{b,I}(\boldsymbol{\delta}) = N_c^0 \cdot \prod_{k=1}^{2} (1 + K_k)^{\delta_k} \cdot f_I^0 \cdot \frac{1}{F(\boldsymbol{\delta})} \prod_{j=3}^{4} p_{I,j}(\delta_j) \cdot (1 + \epsilon_I)^{\delta_I}$$

- **90** observations;
- **90** expected signal yields;
- **9** normalisation terms (one term per category, 3×2+3=9);
- **450** terms for corr. uncertainties (90 functions, each needs 1+4 quantities to specify);
- **90** terms for uncorrelated uncertainties.

Simplified Likelihoods

The elementary components of systematic uncertainty sources are generally **independent** of each other. With that and the CLT we can write the simplified likelihood:

$$L_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{I=1}^{P} \Pr\left(n_I^{\mathrm{obs}} \mid n_{s,I}(\boldsymbol{\alpha}) + a_I + b_I\theta_I + c_I\theta_I^2\right) \cdot \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\rho}^{-1}\boldsymbol{\theta}}}{\sqrt{(2\pi)^P}}$$

The parameters of the L_{S} $(a_I, b_I, c_I, \rho_{IJ})$ have analytical expressions as a function of the variance and the skew of each elementary nuisance parameter, but can be deduced from the three first moments of the event yields n_I distributions!

It can be shown that if P is the number of bins and Q is the number of nuisance parameters:

- Number of inputs needed for full likelihood $\sim 2Q$ (at large Q).
  - But is constant for simplified likelihood!
- Number of inputs needed for simplified likelihood $(P^2)/2 + P$ (at large P).

Paper on the simplified likelihood framework: JHEP 04 2019 064

Patching the Likelihoods

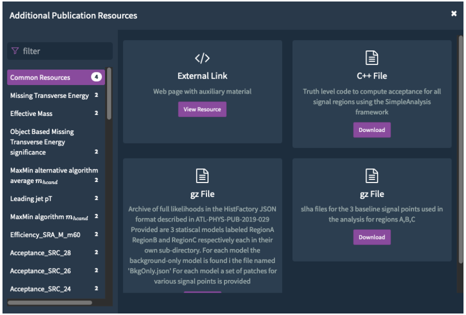
With the full or simplified likelihoods, you can describe an experimental analysis by:

- ☐ Prediction of the null hypothesis
- ☐ Observation of the data
- ☐ A set of **likelihood patches** describing the alternate hypotheses.

$$L(\mu, \boldsymbol{\delta})\pi(\boldsymbol{\delta}) = \prod_{I=1}^{90} P\left(n_I^{\text{obs}} \mid \mu \cdot n_{s,I} + n_{b,I}(\boldsymbol{\delta})\right) \cdot \prod_{j=1}^{94} e^{-\delta_j^2}$$

Simply write a likelihood patch for your model!

- ☐ Signal samples simulation still a minor issue.



JHEP 12 (2019) 060.

Additional resources available in HEPData.

Tools of the Trade

Package	Refs.	Experimental inputs	Event input	Detector simulation	Inference/Output
GAMBIT (ColliderBit)	12, 99–101	Cut-flows, analysis logic, object-level efficiency functions, observed event numbers in signal regions, background covariance matrices	particle	BuckFast (smearing & efficiencies)	Detector-level distributions, signal region efficiencies, simplified likelihood for calculating exclusion limits/contours
CheckMATE	95, 96	Cut-flows, analysis logic, object-level efficiency functions, observed event numbers in signal regions	particle, parton	Delphes	Detector-level distributions, signal region efficiencies, ratio of predicted to excluded cross-section
MadAnalysis 5	17–19, 97, 98	Cut-flows, analysis logic, object-level efficiency functions, observed event numbers in signal regions, background covariance matrices, JSON likelihoods	particle	Delphes; customisable smearing	Detector-level distributions, signal region efficiencies, $1 - \mathrm{CL}_s$ values
Rivet	48, 49	Cut-flows, analysis logic, detector smearing & efficiency functions	particle	Customisable smearing	Truth/detector-level distributions
Contur	61	Unfolded (particle-level) differential cross-sections via Rivet	particle	N/A	Exclusion contours in BSM model space
ADL interpreters: adl2tnm, CutLang	20, 53, 54	analysis logic, external functions of complex variables, object or event level efficiencies	particle	External (Delphes, CMS and ATLAS simulations)	cutflows, event-by-event weights per region, histograms
Recast	8	Experiment-specific formats	parton	Experiment-owned (fast) simulation	p-values, upper limits, likelihood values

Table I. Summary of public frameworks for the reinterpretation of searches and measurements. The columns summarise the major inputs from the experiments used for the reinterpretation, how detector effects are modelled (if necessary) and the principle outputs in terms of performing statistical inference. Particle-level inputs specifically refer to files in HepMC format, whereas parton-level inputs specifically refer to LHE files (except in the case of Recast, which can also accept other internal ATLAS parton-level formats).

Package	Refs.	Experimental inputs	Model input	Inference/Output
SModelS	33, 35, 36	Simplified-model cross-section upper limits and efficiency maps from SUSY searches, background covariance matrices	SLHA or LHE (any BSM model with Z_2-like symmetry)	Ratio of predicted to excluded cross-section, exclusion CL (if efficiency maps are available)
HiggsBounds	90, 91	Model independent (exp. and obs.) 95% CL upper limits and exclusion likelihoods from BSM Higgs searches	masses, widths, cross-sections and BRs (or effective couplings) of all Higgs bosons	Ratio of predicted to excluded cross-section, allowed/excluded at 95% CL, χ^2 for specific searches
ZPEED	92	Observed event numbers in signal regions, background predictions, detector resolution and efficiencies	Model parameters	Likelihood values
DarkCast	93	Simplified-model production mechanism, cross-section upper limits or ratio map of observed to expected cross-sections for dark photon searches	couplings of new gauge bosons to the SM fermions	95% CL exclusion limits on couplings
DarkEFT	104	95% CL exclusion limits on dark sector searches and rare meson decay BRs	effective couplings for 4-fermion operators	95% CL exclusion limits on the effective coupling

Table II. Summary of public frameworks for the reinterpretation of searches and measurements (continued). The columns summarise the major inputs from the experiments used for the reinterpretation, the model inputs, and the principle outputs in terms of performing statistical inference.

See the full paper in SciPost Phys. 9 (2020) 2, 022

Conclusions

High-energy physics is the field that studies the smallest building blocks of matter.

It is equally powered by contributions from theorists, experimentalists, computer scientists, engineers... The harmonious cooperation of those different groups is vital to the success of the field.

From the theoretical side, the field has had continued, resounding success with the standard model of particles and fields. Extensions to the standard model continue be proposed, exploring new ideas and addressing additional data produced by other fields, like astronomy and cosmology.

From the experimental side, the field has moved to global collaborations that design, build and operate extremely large and complex detectors. The data taken with those detectors dwarfs all other scientific datasets to date, and allows to measure the properties of the particles and fields to unprecedented precision.

High-energy physics is a long term endeavour, with experiment time scales measured in decades. The field is already preparing for the challenges aheads, with new experiments being proposed all around the world. Finally, the LHC is scheduled to run at least until 2035.

Thanks

And. . .

We need YOU!!! Please join us at `https://sprace.org.br`