Publication and Exchange of Data and Models

or: making the most out of your analysis

Sabine Kraml
LPSC Grenoble

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

PCD Data Science Basics — Oct 25, 2022

Outline

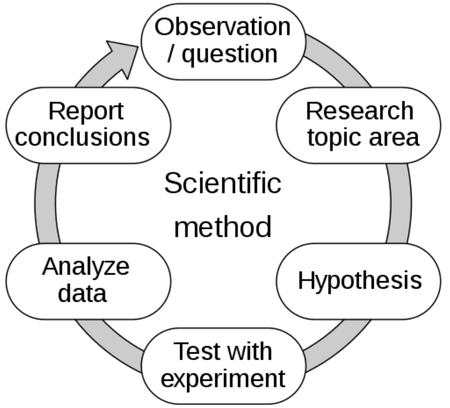
- Introduction and motivation
- Publication of statistical models
- Analysis preservation and reuse
- Specifics about ML models
- Conclusions

Motivation

https://en.wikipedia.org/wiki/Reproducibility

Reproducibility and reuse of scientific results

- Reproducibility is a major principle underpinning the scientific method. It means that results obtained by an experiment or an observational study or in a statistical analysis of a data set should be achieved again with a high degree of reliability when the study is replicated.
- Goes back to Robert Boyle, a pioneer of the experimental method, who maintained that the foundations of knowledge should be constituted by experimentally produced facts, which can be made believable to a scientific community by their reproducibility.



https://en.wikipedia.org/wiki/Scientific_method

Scientific results should be documented in such a way that their deduction is fully transparent.

- Requires a detailed description of the methods used to obtain the data and making the full dataset and the code to calculate the results easily accessible.
- In computational sciences: Any results should be documented by making all data and code available in such a way that the computations can be executed again with identical results.

Reproducibility can be distinguished from replication, as referring to reproducing the same results using the same dataset.

Motivation

Reproducibility and reuse of scientific results

Reproducibility also enables reuse

e.g., updating constraints, testing new hypotheses, performing combinations and/or fits, etc.

→ new research based on existing data and analyses

→ longer shelf life, more scientific impact

Snowmass 2021

US Community Study on the Future of Particle Physics

To achieve their full scientific impact, HEP experiments need to integrate **extensive data and analysis preservation efforts** into their publication processes, alongside the communication of results **in reusable form** and preservation of data products, and making event-level data publicly available.

Without this, the influence of the hundreds of published analyses from the LHC, HL-LHC, EIC, and other future experiments will be **limited mainly to the physics ideas in vogue** at the time the collaboration collected their data. The public investment in experimental programs underscores the importance of going beyond the original paper publication and ensuring that analyses continue providing scientific value in perpetuity.

Executive summary from “Data and Analysis Preservation, Recasting and Reinterpretation”
arXiv:2203.10057

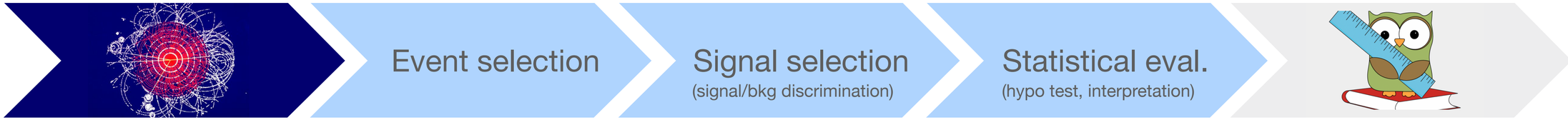
On the individual scale

To achieve its full scientific impact, your analysis needs to integrate extensive data and analysis preservation efforts, alongside the communication of results in reusable form and preservation of data products.

Without this, the influence of your work will be limited to the physics idea en vogue → the hypothesis pursued in the paper. The public investment → ***your own intellectual effort*** in the study underscores the importance of going beyond the original paper publication and ensuring that analyses continue providing scientific value in perpetuity.

Adaptation of Executive summary from “Data and Analysis Preservation, Recasting and Reinterpretation” arXiv:2203.10057

Typical particle physics analysis



Types of reuse (“reinterpretation”)

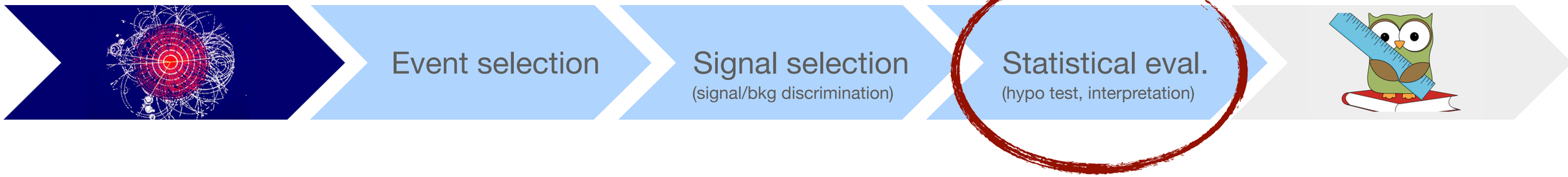
generally referred to as “recasting”

updates of existing analyses using e.g. more precise theoretical calculations, improved experimental calibrations, or a different probability model	**kinematic reinterpretation** considering a different physical process with a different phase space distribution, which might have different efficiencies and acceptances
parametric reinterpretation reparametrizing the likelihood through rescaling, without altering the efficiencies and acceptances that might modify the distributions	**combinations of analyses or datasets** in model surveys, global fits or global averages

approach taken e.g. when reusing simplified model results from the LHC, or in in the context of Higgs signal strengths

includes global Effective Field Theory (EFT) and Beyond-the-Standard Model (BSM) analyses, the reuse of datasets for the determination of parton distribution functions (PDFs), etc.

Typical particle physics analysis



Types of reuse ("reinterpretation")

generally referred to as "recasting"

updates of existing analyses using e.g. more precise theoretical calculations, improved experimental calibrations, or a different probability model	**kinematic reinterpretation** considering a different physical process with a different phase space distribution, which might have different efficiencies and acceptances
parametric reinterpretation reparametrizing the likelihood through rescaling, without altering the efficiencies and acceptances that might modify the distributions	**combinations of analyses or datasets** in model surveys, global fits or global averages

approach taken e.g. when reusing simplified model results from the LHC, or in in the context of Higgs signal strengths

includes global Effective Field Theory (EFT) and Beyond-the-Standard Model (BSM) analyses, the reuse of datasets for the determination of parton distribution functions (PDFs), etc.

Publication of statistical models

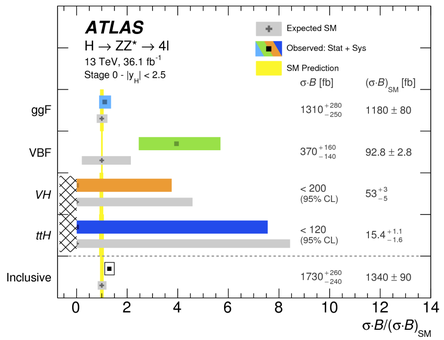
“Publishing statistical models: Getting the most out of particle physics experiments”,
K. Cranmer, SK, H. Prosper (eds) et al., arXiv:2109.04981

Statistical models

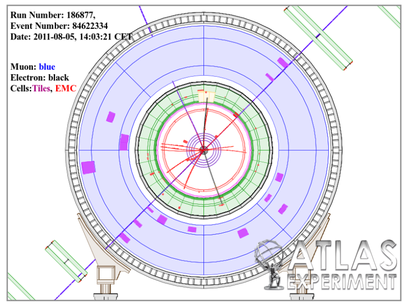
In particle physics experiments, the statistical nature of the data is typically quantified by ascribing to it a probability, as specified by a statistical model $p(\text{data}|\text{theory})$

$$p(\text{theory}|\text{data}) = \frac{p(\text{data}|\text{theory})}{p(\text{data})}\, p(\text{theory})$$

posterior, interpretation



experiment



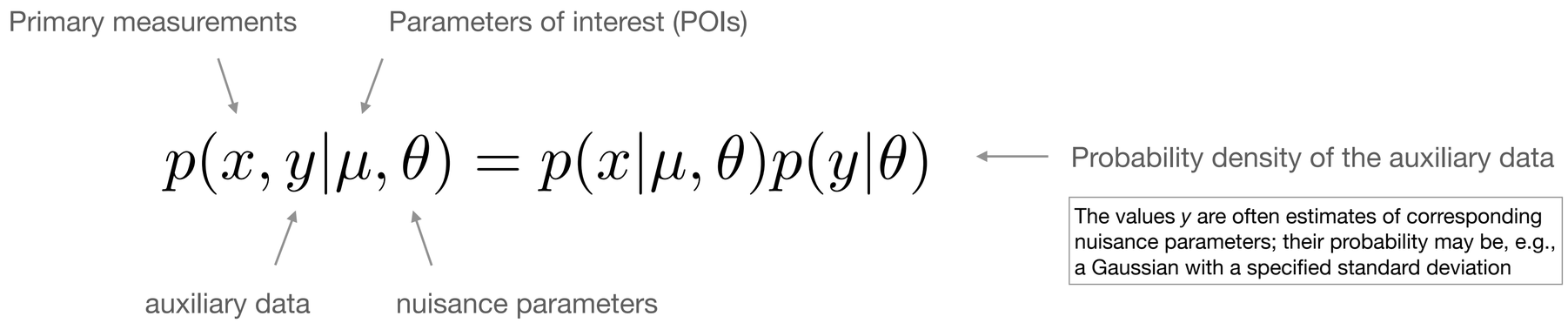
prior

$$\mathcal{L} = -\tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\bar{\Psi}\not{D}\Psi + \text{h.c.} + \bar{\Psi}_i y_{ij} \Psi_j \phi + \text{h.c.} + |D_\mu \phi|^2 - V(\phi)$$

Illustration courtesy Lukas Heinrich

Statistical models

In particle physics experiments, the statistical nature of the data is typically quantified by ascribing to it a probability, as specified by a statistical model *p*(data|theory)



- Describes the **probabilistic dependence of the observable data** on the **parameters of interest** and the **nuisance parameters**.
- When **observed data** are entered into the statistical model, this becomes the **likelihood function**.

Measurements reported as "*observed value ± 1sigma*" *(plus correlations)*

w/o proper statistical model, a Gaussian approximation is forced onto the reuse of experimental results

$$\chi^2(\mu) = \frac{1}{N_{\rm dat}} \sum_{ij=1}^{N_{\rm dat}} \left(\mathcal{O}_i^{(\rm th)}(\mu) - \mathcal{O}_i^{(\rm exp)}\right) \left(\rm cov^{-1}\right)_{ij} \left(\mathcal{O}_j^{(\rm th)}(\mu) - \mathcal{O}_j^{(\rm exp)}\right)$$

$$\rho_{xy} = Correlation\ (x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}.}$$

In principle OK if large enough statistics (→CLT). However:

- Symmetric Gaussian uncertainties are often a simplification; does not hold when systematics dominate
- Often lack of information on correlations
- Non-positive-definite covariance matrices on HEPData
- Lack of breakdown of correlated systematic sources; different naming for systematic sources complicates combining processes
- Correlations between processes often not available
- In case of asymmetric uncertainties, assumptions have to be made about the shape of the likelihood function

Severely affects many areas:
- Parton distribution functions
- Effective field theory fits
- Higgs physics
- BSM searches
- Heavy flavour physics
- Global averages
- BSM Global fits
- ….

Examples discussed in arXiv:2109.04981

Measurements reported as "*observed value ± 1sigma*" *(plus correlations)*

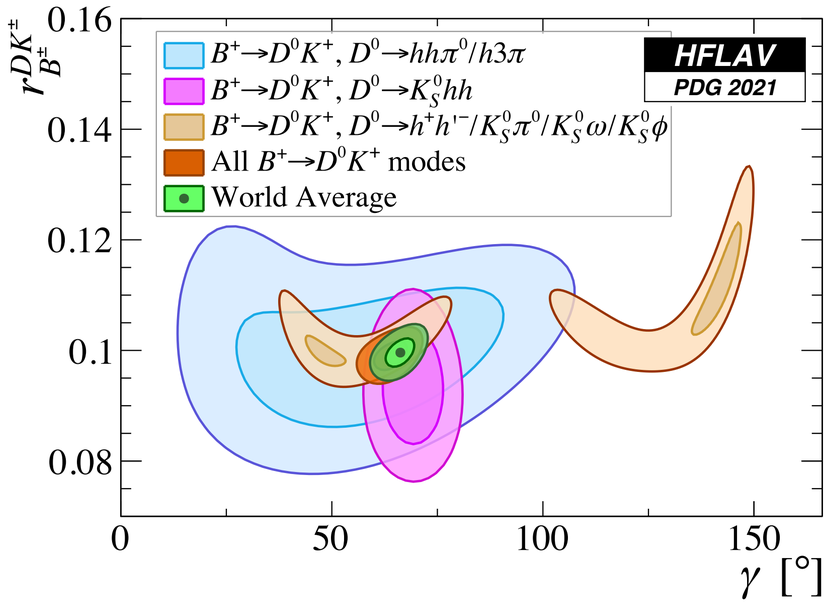
w/o proper statistical model, a Gaussian approximation is forced onto the reuse of experimental results

$$\chi^2(\mu) = \frac{1}{N_{\mathrm{dat}}} \sum_{ij=1}^{N_{\mathrm{dat}}} \left(\mathcal{O}_i^{(\mathrm{th})}(\mu) - \mathcal{O}_i^{(\mathrm{exp})}\right) \left(\mathrm{cov}^{-1}\right)_{ij} \left(\mathcal{O}_j^{(\mathrm{th})}(\mu) - \mathcal{O}_j^{(\mathrm{exp})}\right)$$

$$\rho_{xy} = Correlation\,(x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}.}$$

In principle OK if large enough statistics (→CLT). However:

- Symmetric Gaussian uncertainties are often a simplification; does not hold when systematics dominate
- Often lack of information on correlations
- Non-positive-definite covariance matrices on HEPData
- Lack of breakdown of correlated systematic sources; different naming for systematic sources complicates combining processes
- Correlations between processes often not available
- In case of asymmetric uncertainties, assumptions have to be made about the shape of the likelihood function



Examples discussed in arXiv:2109.04981

Measurements reported as "*observed value ± 1sigma*" *(plus correlations)*

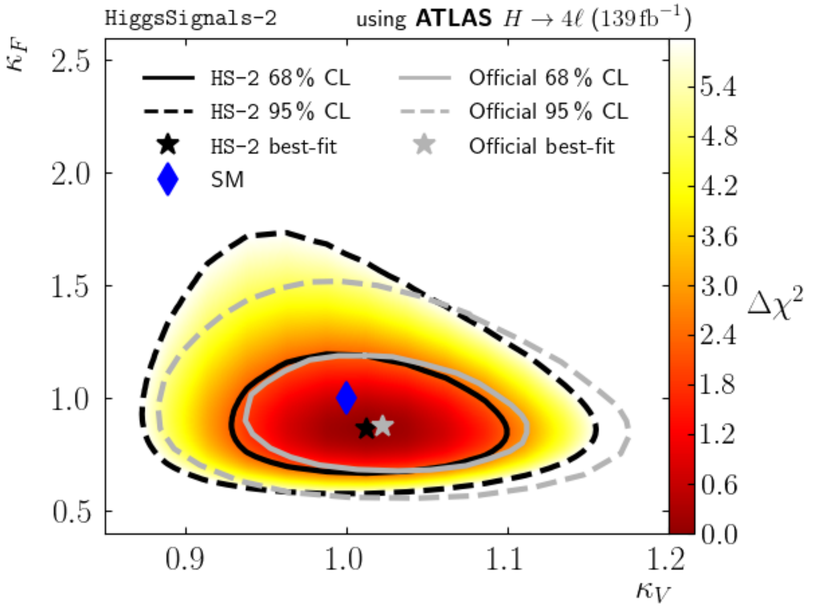
w/o proper statistical model, a Gaussian approximation is forced onto the reuse of experimental results

$$\chi^2(\mu) = \frac{1}{N_{\text{dat}}} \sum_{ij=1}^{N_{\text{dat}}} \left(\mathcal{O}_i^{(\text{th})}(\mu) - \mathcal{O}_i^{(\text{exp})}\right) \left(\text{cov}^{-1}\right)_{ij} \left(\mathcal{O}_j^{(\text{th})}(\mu) - \mathcal{O}_j^{(\text{exp})}\right)$$

$$\rho_{xy} = Correlation\,(x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}.}$$

In principle OK if large enough statistics (→CLT). However:

- Symmetric Gaussian uncertainties are often a simplification; does not hold when systematics dominate
- Often lack of information on correlations
- Non-positive-definite covariance matrices on HEPData
- Lack of breakdown of correlated systematic sources; different naming for systematic sources complicates combining processes
- Correlations between processes often not available
- In case of asymmetric uncertainties, assumptions have to be made about the shape of the likelihood function



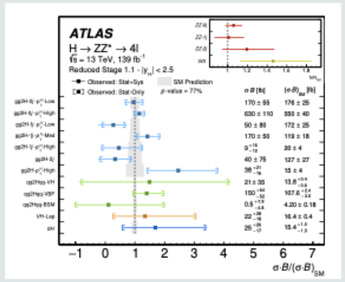
Examples discussed in arXiv:2109.04981

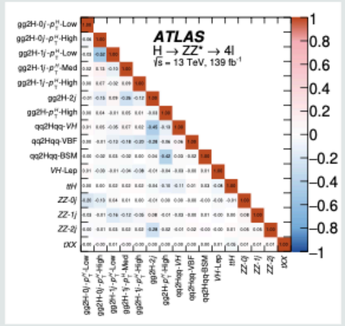
Higgs Measurements — Limitations of Reconstructed LLHs

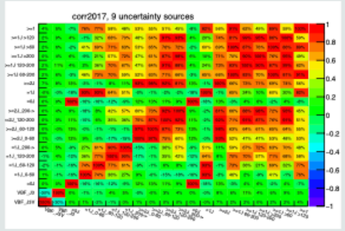
Example: ATLAS $H \to ZZ \to 4\ell$ [ATLAS 2004.03447]

HiggsSignals implementation

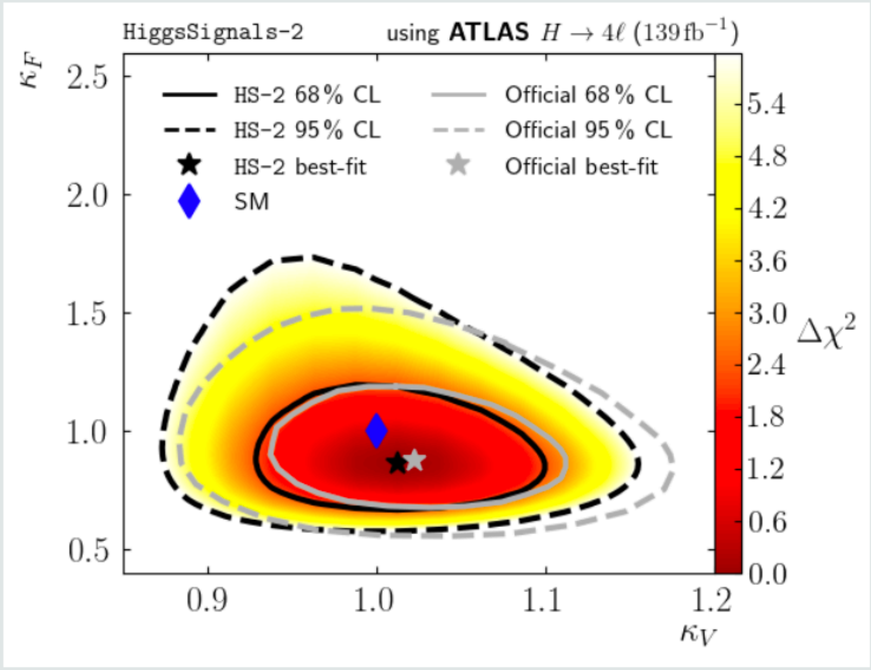
- measurements (12-bin STXS)
- experimental correlations
- theory correlations [2017 Scheme]







Performance of HiggsSignals compared to official κ-fit.

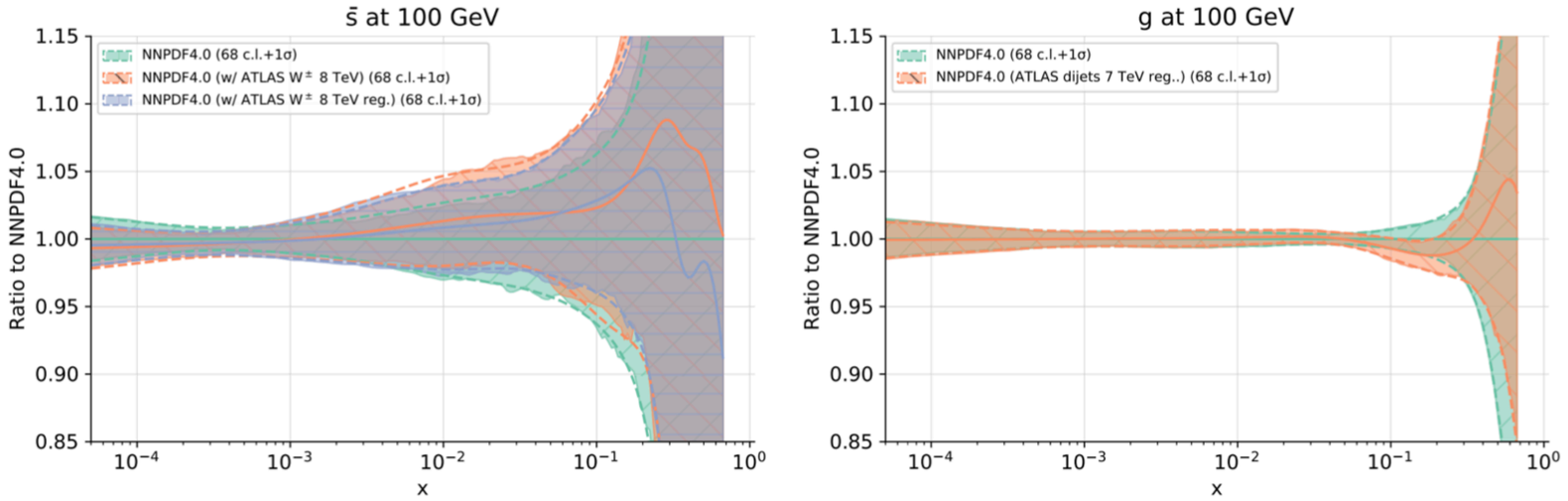


Jonas Wittbrodt

Publication of Statistical Models
Hands-on workshop 8-12 Nov 2021
https://indico.cern.ch/event/1088121/

ill-defined covariance matrices

One particularly worrisome consequence of the **lack of open likelihoods** is that results may become very sensitive to **small variations of the official correlation model**



Assess impact in fit by transforming the original covariance matrix into a matrix with the **same eigenvectors** but with **clipped eigenvalues** below some cut-off: **stable PDFs** with much lower χ^2

Dataset	$N_{\rm dat}$	$Z_{\rm orig}$	$\chi^2_{\rm orig}$	$\chi^2_{\rm reg}$
ATLAS W, Z 7 TeV CC ($\mathcal{L} = 4.6$ fb^{-1})	46	9.01	1.89	0.93
ATLAS W 8 TeV (*)	22	11.28	3.50	1.15
CMS dijets 7 TeV	54	4.70	1.81	1.73
ATLAS dijets 7 TeV	90	9.93	2.14	0.92
CMS 3D dijets 8 TeV (*)	122	4.47	1.50	0.92

minimal modification of correlation model, large impact in fit quality, PDFs stable

Juan Rojo on PDF fits

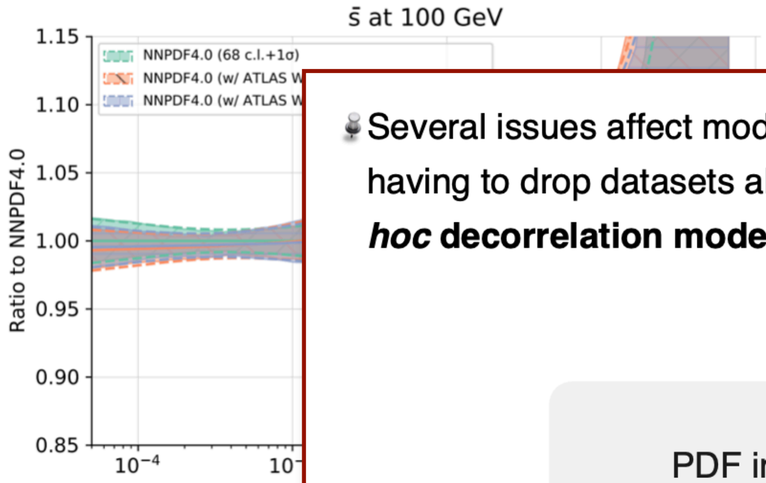
Key component of predictions for particle, nuclear, and astro-particle experiments.

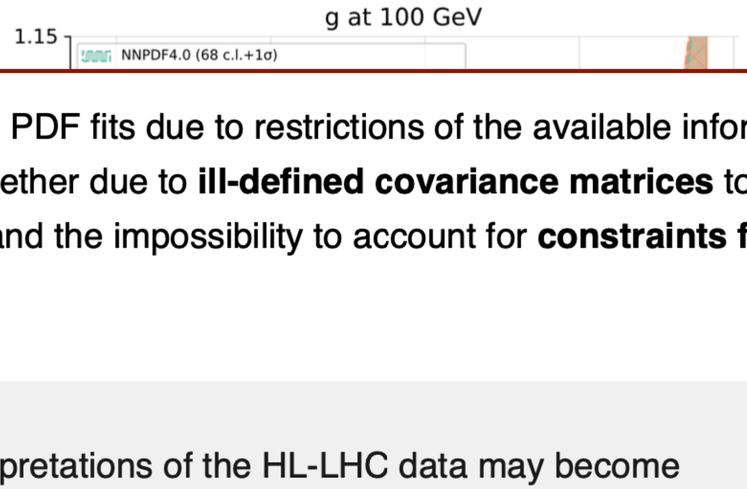
Address fundamental questions in QCD.

Publication of Statistical Models
Hands-on workshop 8-12 Nov 2021
https://indico.cern.ch/event/1088121/

ill-defined covariance matrices

One particularly worrisome consequence of the **lack of open likelihoods** is that results may become very sensitive to **small variations of the official correlation model**





- Several issues affect modern PDF fits due to restrictions of the available information, from having to drop datasets altogether due to **ill-defined covariance matrices** to implementing ***ad-hoc* decorrelation models** and the impossibility to account for **constraints from search data**

PDF interpretations of the HL-LHC data may become seriously hampered, or even impossible altogether, unless **experiments release their full statistical models**

Assess impa

eigenvectors

Dataset				
ATLAS W, Z 7 TeV CC ($\mathcal{L} = 4.6$ fb^{-1})	46	9.01	1.89	0.93
ATLAS W 8 TeV (*)	22	11.28	3.50	1.15
CMS dijets 7 TeV	54	4.70	1.81	1.73
ATLAS dijets 7 TeV	90	9.93	2.14	0.92
CMS 3D dijets 8 TeV (*)	122	4.47	1.50	0.92

minimal modification of correlation model, large impact in fit quality, PDFs stable

Juan Rojo on PDF fits

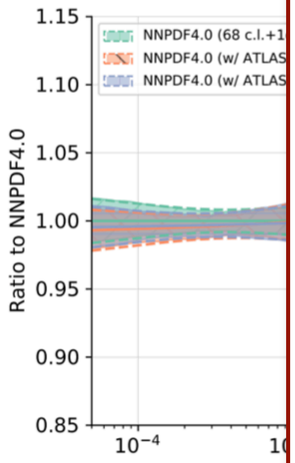
Key component of predictions for particle, nd astro-particle nts.

fundamental s in QCD.

Publication of Statistical Models
Hands-on workshop 8-12 Nov 2021
https://indico.cern.ch/event/1088121/

ill-defined covariance matrices

One particularly worrisome consequence of the **lack of open likelihoods** is that results may becom



Assess imp
eigenvectors

Dataset				
ATLAS W, Z 7 T				
ATLAS W 8 TeV				
CMS dijets 7 Te				
ATLAS dijets 7 TeV	90	9.93	2.14	0.92
CMS 3D dijets 8 TeV (*)	122	4.47	1.50	0.92

- Availability of **open likelihoods** would make possible PDF determinations where statistical and systematic uncertainties are always accounted for by means of the appropriate model

$$\mu_i = \bar{\mu}_i \pm \sigma_{\text{poiss,i}} \pm \sigma_{\text{syst,gauss,i}} \pm \sum \sigma_{\text{syst,gauss,ij}} \pm \sum \sigma_{\text{theo,ij}} \pm \sum \sigma_{\text{model,ij}}$$

central value	*stat error (poisson)*	*syst error (uncorrelated, gaussian)*	*syst error (correlated, gaussian)*	*theory error (gaussian? uniform?)*	*model error (correlated, not gaussian)*

$$\max_{\boldsymbol{\theta}} \mathscr{L}\left(\mu_i, \boldsymbol{\theta}\right) = \mathscr{L}_{\text{gauss}} \times \mathscr{L}_{\text{poiss}} \times \mathscr{L}_{\text{uniform}} \times \mathscr{L}_{\text{theory}} \times \cdots$$

- This is not only **technically correct**, but also allows including much more information on PDF fits as compared to what is currently available e.g. by means of **search data**

Juan Rojo on PDF fits

ent of
or particle,
astro-particle

damental
QCD.

atistical Models
Hands-on workshop 8-12 Nov 2021
https://indico.cern.ch/event/1088121/

(Profile) likelihoods: very useful but not sufficient

- **In the likelihood, the data is baked in**
  - cannot evaluate likelihood on new data
  - cannot sample from the model (pseudo-data, toy MC)
- **In profile likelihoods, nuisance parameter are fixed**
  - cannot statistically combine profile likelihoods targeting parameters of interest if they share nuisances
  - cannot update constraint terms (auxiliary measurements)

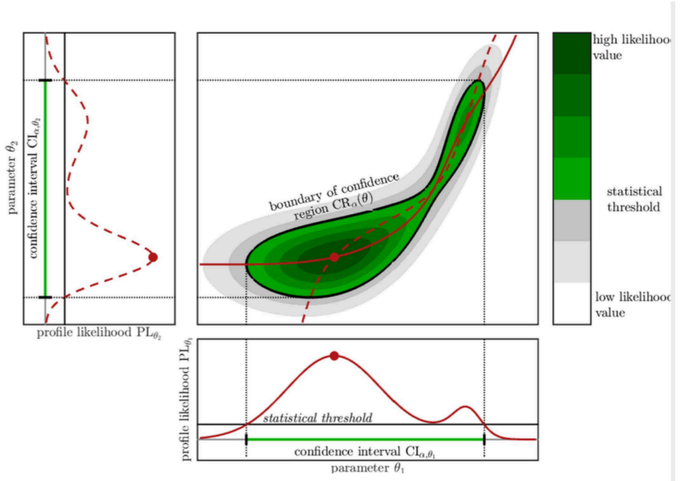


Illustration courtesy Lukas Heinrich

Simplified statistical models (with combined nuisances) also have such problems

- **Reparametrization in terms of different parameters of interest is not always possible**
  - parametrization in terms of quantities such as masses, cross sections, widths, branching fractions, etc., is often more useful than a parametrization in terms of theory-model (Lagrangian) parameters
  - risk of introducing dependencies which can result in a loss of information

Full statistical model

The **complete probability model for the analysis**; includes dependence on the data x,y, the parameters of interest μ and nuisance parameters θ, access to the individual terms and the ability to generate pseudo-data (“toy Monte Carlo”).

$$p(n,x,y|\mu,\theta) = \prod_{i=1}^{N_c} \left[\text{Pois}(n_i \mid \nu_i(\mu,\theta)) \prod_{j=1}^{n_i} p_i(x_{ij}|\mu,\theta) \right] p(y|\theta) \rightarrow L(\mu,\theta)$$

Likelihood: The value of the statistical model for a given fixed dataset as a function of the parameters

pdf of auxiliary data *y*

Access to the individual components ν_{ik} and p_{ik} is needed for certain types of reinterpretations, e.g.

- changing the distributions $p_{ik}(x_{ij}|\mu,\theta)$ because a different physical process with a different phase space distribution is being considered;
- updates of existing interpretations using more precise theoretical calculations or improved experimental calibrations, or both.

$$p_i(x_{ij}|\mu,\theta) = \sum_k \frac{\nu_{ik}(\mu,\theta)}{\nu_i(\mu,\theta)} p_{ik}(x_{ij}|\mu,\theta)\,,$$

The probability to measure x_{ij} in channel i event j

$$\nu_i(\mu,\theta) = \sum_k \nu_{ik}(\mu,\theta)$$

Essential information for analysis preservation and reuse

ATLAS took the leap

ATL-PHYS-PUB-2019-029

… and started to publish plain-text serialisation of HistFactory workspaces in JSON format

- Provides background estimates, changes under systematic variations, and observed data counts at the same fidelity as used in the experiment.

	Description	Modification	Constraint Term c_χ	Input
constrained	Uncorrelated Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Pois}\left(r_b = \sigma_b^{-2} \middle\vert \rho_b = \sigma_b^{-2}\gamma_b\right)$	σ_b
	Correlated Shape	$\Delta_{scb}(\alpha) = f_p\left(\alpha \middle\vert \Delta_{scb,\alpha=-1}, \Delta_{scb,\alpha=1}\right)$	$\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$	$\Delta_{scb,\alpha=\pm1}$
	Normalisation Unc.	$\kappa_{scb}(\alpha) = g_p\left(\alpha \middle\vert \kappa_{scb,\alpha=-1}, \kappa_{scb,\alpha=1}\right)$	$\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$	$\kappa_{scb,\alpha=\pm1}$
	MC Stat. Uncertainty	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Gaus}\left(a_{\gamma_b} = 1 \middle\vert \gamma_b, \delta_b\right)$	$\delta_b^2 = \sum_s \delta_{sb}^2$
	Luminosity	$\kappa_{scb}(\lambda) = \lambda$	$\text{Gaus}\,(l = \lambda_0 \vert\, \lambda, \sigma_\lambda)$	$\lambda_0, \sigma_\lambda$
free	Normalisation	$\kappa_{scb}(\mu_b) = \mu_b$		
	Data-driven Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$		

Rate modifications defined in HistFactory for bin *b*, sample *s*, channel *c*.

- Usage: RooFit, **pyhf**
- Target: long-term data/analysis preservation, reinterpretation purposes

ATLAS EXPERIMENT

ATLAS PUB Note

ATL-PHYS-PUB-2019-029

21st October 2019

CERN

Reproducing searches for new physics with the ATLAS experiment through publication of full statistical likelihoods

The ATLAS Collaboration

The ATLAS Collaboration is starting to publicly provide likelihoods associated with statistical fits used in searches for new physics on HEPData. These likelihoods adhere to a specification first defined by the `HistFactory` p.d.f. template. This note introduces a JSON schema that fully describes the `HistFactory` statistical model and is sufficient to reproduce key results from published ATLAS analyses. This is per-se independent of its implementation in `ROOT` and it can be used to run statistical analysis outside of the `ROOT` and `RooStats/RooFit` framework. The first of these likelihoods published on HEPData is from a search for bottom-squark pair production. Using two independent implementations of the model, one in `ROOT` and one in pure Python, the limits on the bottom-squark mass are reproduced, underscoring the implementation independence and long-term viability of the archived data.

pyhf

∂ifferentiable ℒikelihoods

ATLAS took the leap

ATL-PHYS-PUB-2019-029

… and started to publish plain-text serialisation of HistFactory workspaces in JSON format

- Provides background estimates, changes under systematic variations, and observed data counts at the same fidelity as used in the experiment.

	Description	Modification	Constraint Term c_χ	Input
constrained	Uncorrelated Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Pois}\left(r_b = \sigma_b^{-2} \middle\vert \rho_b = \sigma_b^{-2}\gamma_b\right)$	σ_b
	Correlated Shape	$\Delta_{scb}(\alpha) = f_p\left(\alpha \middle\vert \Delta_{scb,\alpha=-1}, \Delta_{scb,\alpha=1}\right)$	$\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$	$\Delta_{scb,\alpha=\pm1}$
	Normalisation Unc.	$\kappa_{scb}(\alpha) = g_p\left(\alpha \middle\vert \kappa_{scb,\alpha=-1}, \kappa_{scb,\alpha=1}\right)$	$\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$	$\kappa_{scb,\alpha=\pm1}$
	MC Stat. Uncertainty	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Gaus}\left(a_{\gamma_b} = 1 \middle\vert \gamma_b, \delta_b\right)$	$\delta_b^2 = \sum_s \delta_{sb}^2$
	Luminosity	$\kappa_{scb}(\lambda) = \lambda$	$\text{Gaus}\,(l = \lambda_0 \vert\, \lambda, \sigma_\lambda)$	$\lambda_0, \sigma_\lambda$
free	Normalisation	$\kappa_{scb}(\mu_b) = \mu_b$		
	Data-driven Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$		

Rate modifications defined in HistFactory for bin *b*, sample *s*, channel *c*.

- Usage: RooFit, **pyhf**
- Target: long-term data/analysis preservation, reinterpretation purposes

pyhf differentiable likelihoods

ATLAS took the leap

ATL-PHYS-PUB-2019-029

… and started to publish plain-text serialisation of HistFactory workspaces in JSON format

- Provides background estimates, changes under systematic variations, and observed data counts at the same fidelity as used in the experiment.

	Description	Modification	Constraint Term c_χ	Input
constrained	Uncorrelated Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Pois}\left(r_b = \sigma_b^{-2} \middle\vert \rho_b = \sigma_b^{-2}\gamma_b\right)$	σ_b
	Correlated Shape	$\Delta_{scb}(\alpha) = f_p\left(\alpha \middle\vert \Delta_{scb,\alpha=-1}, \Delta_{scb,\alpha=1}\right)$	$\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$	$\Delta_{scb,\alpha=\pm1}$
	Normalisation Unc.	$\kappa_{scb}(\alpha) = g_p\left(\alpha \middle\vert \kappa_{scb,\alpha=-1}, \kappa_{scb,\alpha=1}\right)$	$\text{Gaus}\,(a = 0 \vert\, \alpha, \sigma = 1)$	$\kappa_{scb,\alpha=\pm1}$
	MC Stat. Uncertainty	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Gaus}\left(a_{\gamma_b} = 1 \middle\vert \gamma_b, \delta_b\right)$	$\delta_b^2 = \sum_s \delta_{sb}^2$
	Luminosity	$\kappa_{scb}(\lambda) = \lambda$	$\text{Gaus}\,(l = \lambda_0 \vert\, \lambda, \sigma_\lambda)$	$\lambda_0, \sigma_\lambda$
free	Normalisation	$\kappa_{scb}(\mu_b) = \mu_b$		
	Data-driven Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$		

Rate modifications defined in HistFactory for bin *b*, sample *s*, channel *c*.

- Usage: RooFit, **pyhf**

- Target: long-term data/analysis preservation, reinterpretation purposes

HEPData

Resources

Likelihood available

Search for charginos and neutralinos in all-hadronic final states	SUSY	Accepted by PRD	17-AUG-21	13	139 fb^{-1}
4-top xsec measurement	TOPQ	Accepted by JHEP	22-JUN-21	13	139 fb^{-1}
Search for gluinos, stops and electroweakinos in RPV models in final states with 1L and many jets	SUSY	Accepted by EPJC	17-JUN-21	13	139 fb^{-1}
Search for charginos and neutralinos in final states with 3L and MET	SUSY	Accepted by EPJC	03-JUN-21	13	139 fb^{-1}
Measurement of ttZ cross sections in Run 2	TOPQ	Eur. Phys. J. C 81 (2021) 737	23-MAR-21	13	139 fb^{-1}
Search for third-generation scalar leptoquarks decaying to a top quark and a tau lepton	EXOT	JHEP 06 (2021) 179	27-JAN-21	13	139 fb^{-1}
Search for squarks and gluinos in final states 1L, jets and MET	SUSY	Eur. Phys. J. C 81 (2021) 600	05-JAN-21	13	139 fb^{-1}
Search for charginos and neutralinos in RPV models in final states with 3L (or more)	SUSY	Phys. Rev. D 103, (2021) 112003	20-NOV-20	13	139 fb^{-1}
Search for displaced leptons	SUSY	Phys. Rev. Lett. 127 (2021) 051802	13-NOV-20	13	139 fb^{-1}
Search for squarks and gluinos in final states with 0L, jets and MET	SUSY	JHEP 02 (2021) 143	27-OCT-20	13	139 fb^{-1}
Measurement of the ttbar production cross-section in the lepton+jets channel at 13 TeV	TOPQ	Phys. Lett. B 810 (2020) 135797	24-JUN-20	13	139 fb^{-1}
Stop pair, long-lived; displaced vertex and displaced muon	SUSY	Phys. Rev. D 102 (2020) 032006	26-MAR-20	13	136 fb^{-1}
Chargino-neutralino pair; 3 leptons, weak-scale mass splittings	SUSY	Phys. Rev. D 101 (2020) 072001	18-DEC-19	13	139 fb^{-1}
Chargino-neutralino pair, slepton pair; soft leptons	SUSY	Phys. Rev. D 101 (2020) 052005	28-NOV-19	13	139 fb^{-1}
Staus; taus	SUSY	Phys. Rev. D 101 (2020) 032009	15-NOV-19	13	139 fb^{-1}
Chargino-neutralino pair; Higgs boson in final state, 2 b-jets and 1 lepton	SUSY	Eur. Phys. J. C 80 (2020) 691	19-SEP-19	13	139 fb^{-1}
Stop pair, sbottom pair, gluino pair; two same-sign leptons or three leptons	SUSY	JHEP 06 (2020) 46	18-SEP-19	13	139 fb^{-1}
Sbottom; b-jets	SUSY	JHEP 12 (2019) 060	08-AUG-19	13	139 fb^{-1}

ATLAS full statistical models HistFactory JSON format

→ statistical evaluation through JSON patching

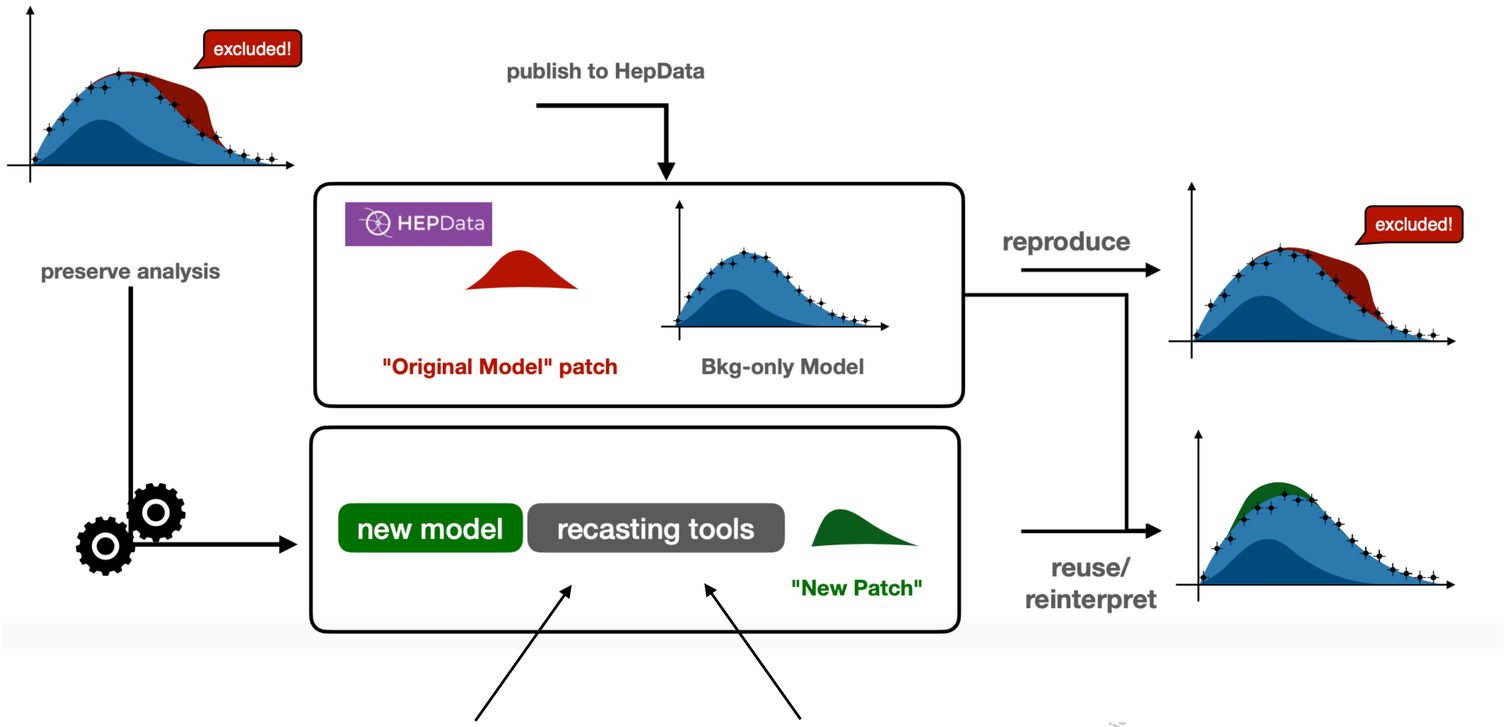


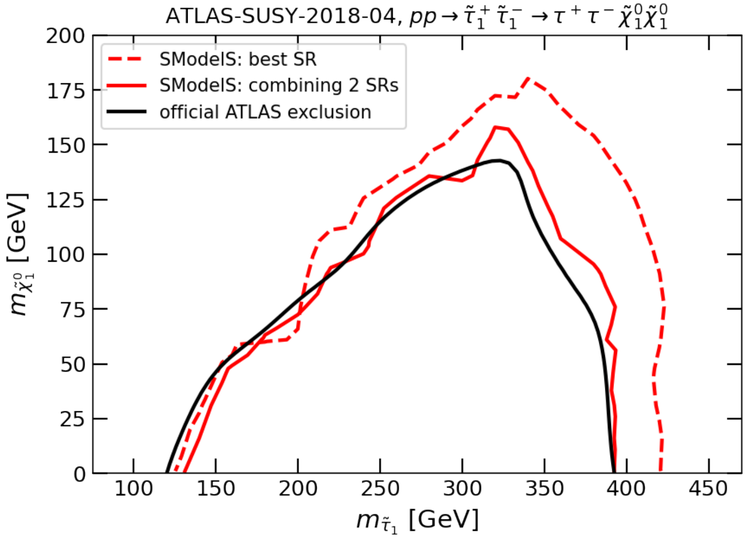
Illustration by **Lukas Heinrich**
Hands-on workshop 8 Nov 2021

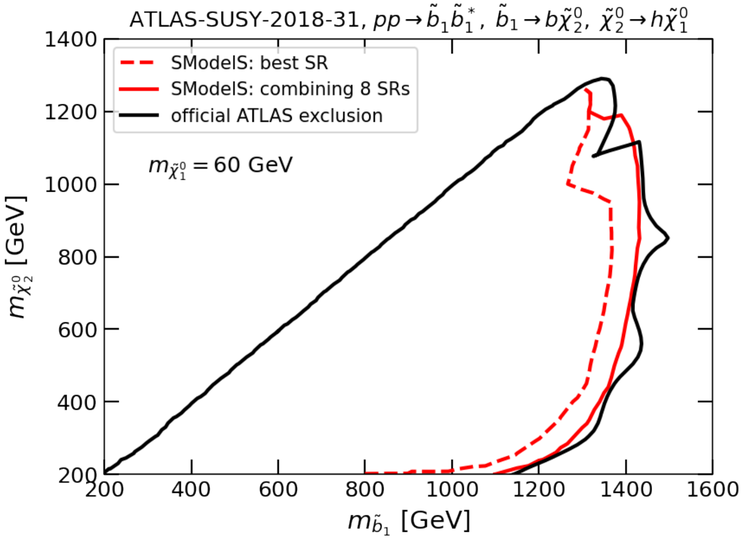
G. Alguero, J. Araz, B. Fuks, SK
arXiv:2206.14870

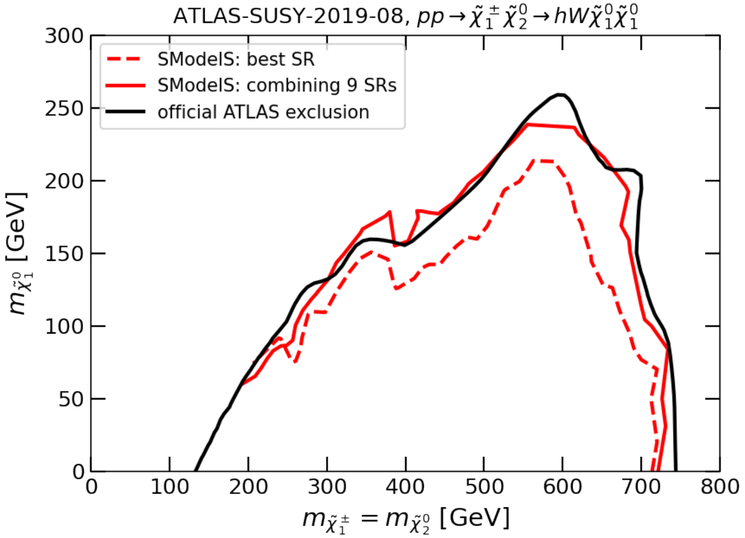
Interfaced to pyhf since SModelS v1.2.4 (now v2.2)
G. Alguero, SK, W. Waltenberger, arXiv:2009.01809

The `Simplify` python tool can be used to create simplified statistical models from full ones by merging all background contributions and combining all nuisance parameters into a single one; may yield equivalent results at much lower CPU cost — needs testing case-by-case!

Improvements w.r.t. “best signal region” approach







G. Alguero, SK, W. Waltenberger, arXiv:2009.01809

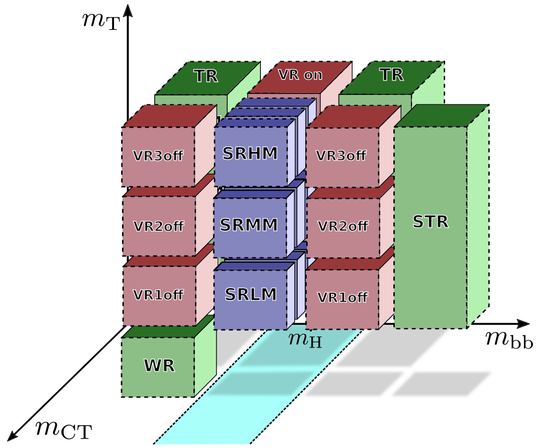
(

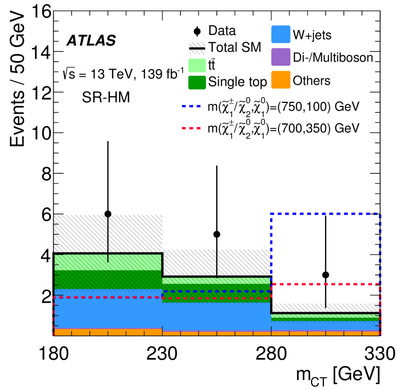
Beyond the Standard Model (BSM) searches at the LHC

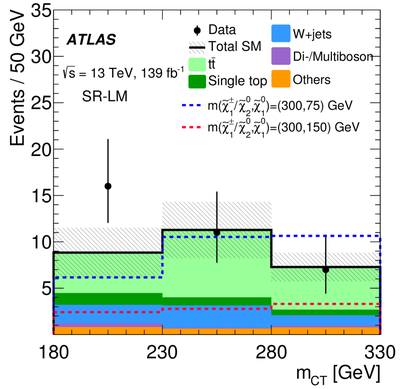
BSM searches are typically performed in **specific bins of kinematic distributions, so-called signal regions (SRs)**

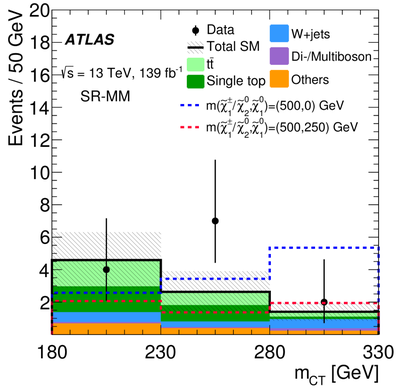
- designed to maximise the number of events from the hypothesised signal with respect to the number of “background” events originating from Standard Model processes.
- control and validation regions are defined in the phase space where no or very little signal from new physics is expected.
- A **statistical analysis** is then performed **to evaluate the confidence level of the hypothesised BSM scenario**, and claim evidence for or set a limit on the new particles of this scenario.

Example from ATLAS search for electroweakinos
in final states with 1 lepton + H(→bb) + MET
ATLAS-SUSY-2019-08







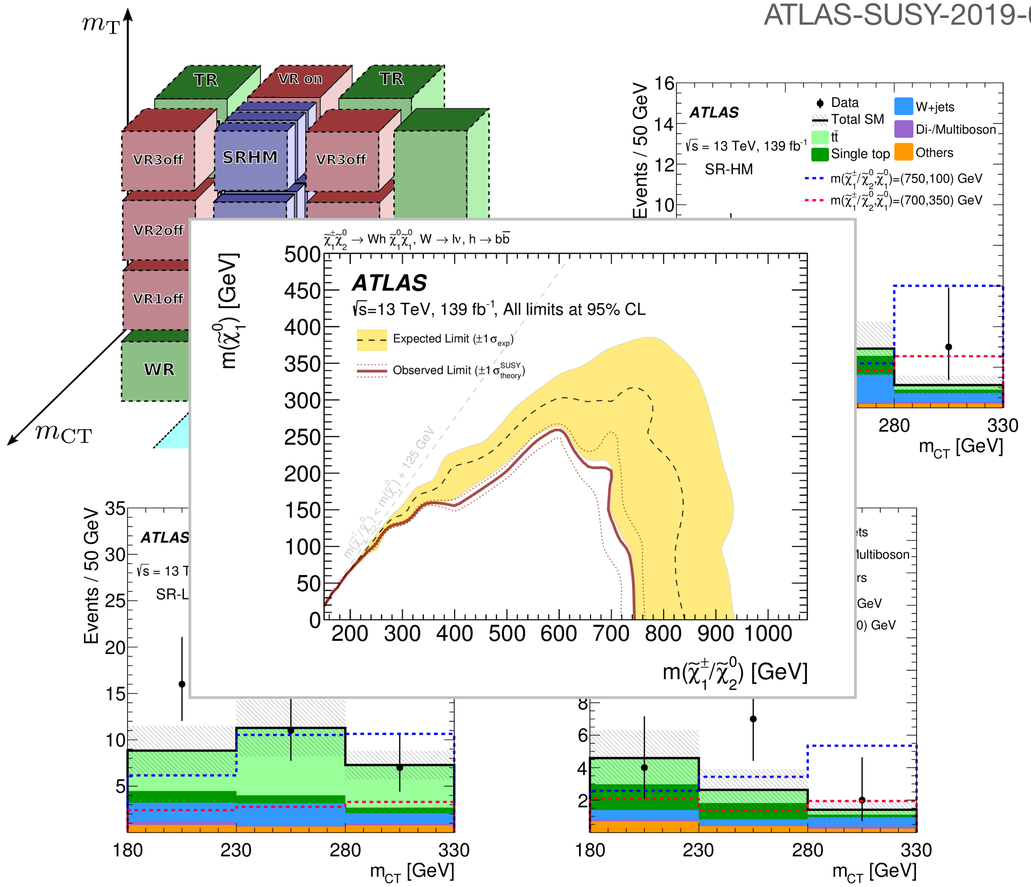


Beyond the Standard Model (BSM) searches at the LHC

BSM searches are typically performed in **specific bins of kinematic distributions, so-called signal regions (SRs)**

- designed to maximise the number of events from the hypothesised signal with respect to the number of “background” events originating from Standard Model processes.
- control and validation regions are defined in the phase space where no or very little signal from new physics is expected.
- A **statistical analysis** is then performed **to evaluate the confidence level of the hypothesised BSM scenario**, and claim evidence for or set a limit on the new particles of this scenario.

Example from ATLAS search for electroweakinos
in final states with 1 lepton + H(→bb) + MET
ATLAS-SUSY-2019-08



Beyond the Standard Model (BSM) searches at the LHC

BSM searches are typically performed in **specific bins of kinematic distributions, so-called signal regions (SRs)**

- designed to maximise the number of events from the hypothesised signal with respect to the number of "background" events originating from Standard Model processes.
- control and validation regions are defined in the phase space where no or very little signal from new physics is expected.
- A **statistical analysis** is then performed **to evaluate the confidence level of the hypothesised BSM scenario**, and claim evidence for or set a limit on the new particles of this scenario.

Reinterpretation studies outside the experimental collaborations **aim at reproducing this process** for BSM scenarios different from those considered in the original experimental publication.

- In the **absence of appropriate correlation information**, it is common practice to **use only the most sensitive SR (a.k.a. "best SR")** for the statistical evaluation, e.g. limit setting.

$$\mathcal{L}(\mu,\theta) = \frac{(\mu s + b + \theta)^{n_{\text{obs}}}\, e^{-(\mu s+b+\theta)}}{n_{\text{obs}}!}\, exp\left(-\frac{\theta^2}{2\delta^2}\right)$$

s, μ … signal, signal strength; *b* … background; θ … nuisances; $\delta^2 = \delta_s^2 + \delta_b^2$

- Problems: **1. only a part of the available data is used**, which can lead to false conclusions (typically loss of sensitivity); **2.** if the best SR changes from point to point in a scan, this can lead to **numerical instabilities in global fits**.

The statistical combination of disjoint SRs in reinterpretation studies uses more of the data of an analysis and gives more robust results than the single (best) SR approach.

CMS: covariance matrices

- Best-SR approach: assuming a Poisson distribution for the data and a Gaussian with variance of δ^2 for the nuisances, p(θ)

$$\mathcal{L}(\mu,\theta) = \frac{(\mu s + b + \theta)^{n_{\text{obs}}}\, e^{-(\mu s + b + \theta)}}{n_{\text{obs}}!}\, exp\left(-\frac{\theta^2}{2\delta^2}\right)$$

$\delta^2 = \delta_s^2 + \delta_b^2$
signal+background uncertainties

- **CMS** analyses sometimes provide a covariance matrix, which allow for the combination of disjoint SRs in a **simplified likelihood approach**

$$\mathcal{L}_S(\mu,\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu\cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu\cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{V}^{-1}\boldsymbol{\theta}\right)$$

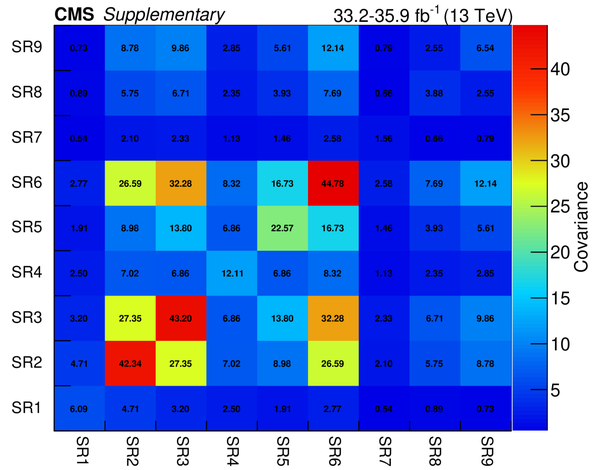
covariance matrix

[CMS NOTE-2017/001]

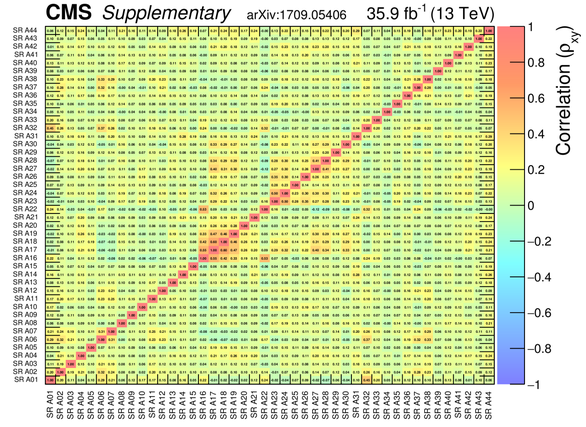
  - Implemented in SModelS and GAMBIT since a while; recently also in MadAnalysis 5

Much(!) better than best-SR, but caveat are non-Gaussian effects e.g. when systematic uncertainties dominate

CMS-SUS-16-048



CMS-SUS-16-039



$$\rho_{xy} = Correlation\,(x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}.}$$

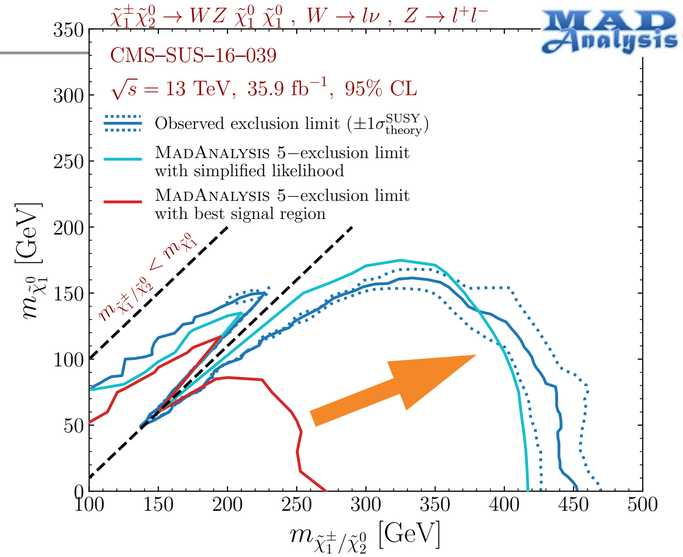
G. Alguero, J. Araz, B. Fuks, SK,
arXiv:2206.14870

CMS: covariance matrices

- Best-SR approach: assuming a Poisson distribution for the data and a Gaussian with variance of δ^2 for the nuisances, p(θ)

$$\mathcal{L}(\mu,\theta) = \frac{(\mu s + b + \theta)^{n_{\rm obs}}\, e^{-(\mu s+b+\theta)}}{n_{\rm obs}!}\, exp\left(-\frac{\theta^2}{2\delta^2}\right)$$

$\delta^2 = \delta_s^2 + \delta_b^2$
signal+background uncertainties

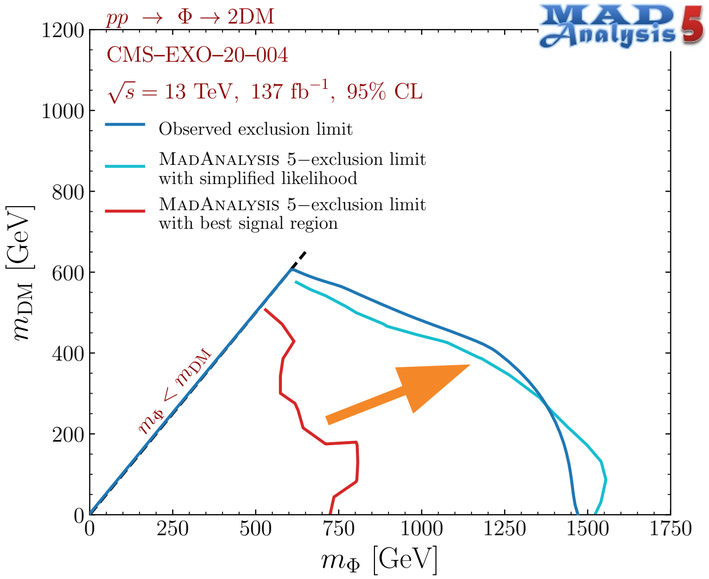


- **CMS** analyses sometimes provide a covariance matrix, which allows for the combination of disjoint SRs in a **simplified likelihood approach**

$$\mathcal{L}_S(\mu,\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu\cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu\cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{V}^{-1}\boldsymbol{\theta}\right)$$

[CMS NOTE-2017/001]

covariance matrix



- Implemented in SModelS and GAMBIT since a while; recently also in MadAnalysis 5

Much(!) better than best-SR, but caveat are non-Gaussian effects e.g. when systematic unc. dominate

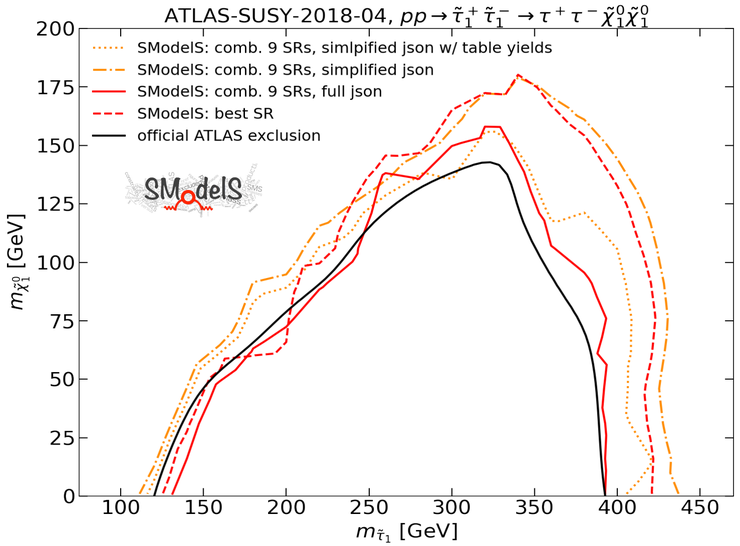
back to full statistical models

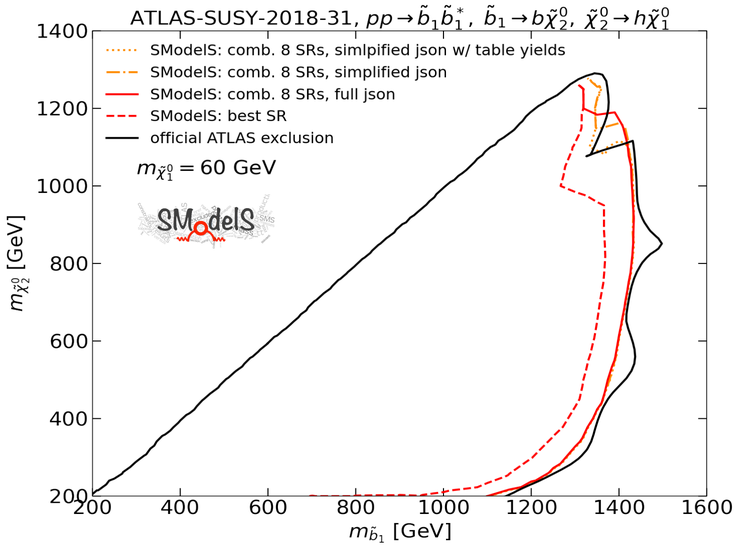
Simplify: from full to simplified likelihoods

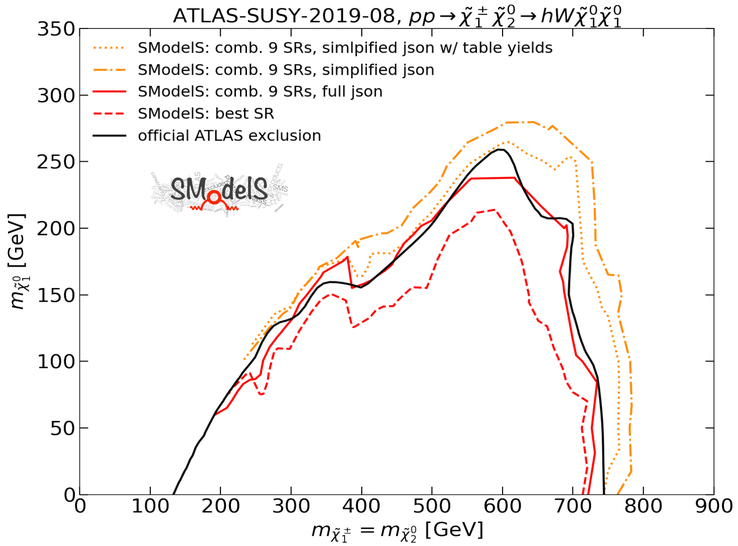
ATL-PHYS-PUB-2021-038

Computing e.g. a CLs value from the full statistical model with tens to hundreds of nuisance parameters is accurate but too CPU intensive for some use cases → can we simplify this?

The `Simplify` python tool creates simplified statistical models from full ones by **merging all background contributions and combining all nuisance parameters into a single one**; same HistFactory JSON format







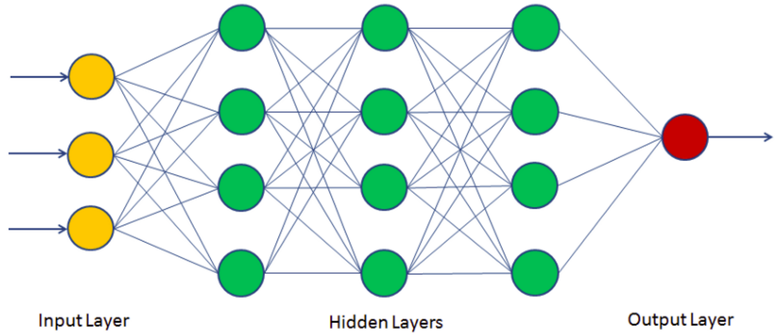
much lower CPU cost but not always a good approximation

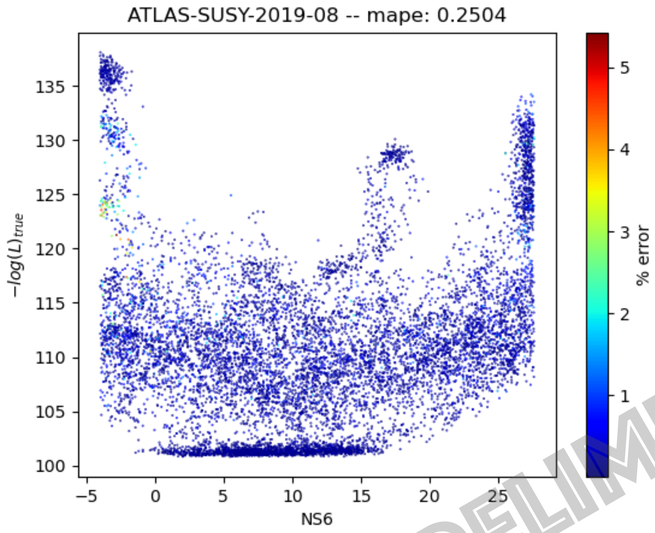
Machine-learn likelihoods

Work in progress!

Humberto Reyes Gonzalez

- Machine-learn likelihoods from full statistical models as functions of signal counts in each SRs
- Neural networks: sequential multilayer perceptrons using `TensorFlow2`.
  - loss function: mean squared error
  - activation function: Leaky Relu with alpha=0.2.
- Example for ATLAS-SUSY-2019-08 (9 SRs):
  - 3 layers, 512 neurons, MAPE = 0.2504, Max PE: 5.429
- The NN models are
  - saved in ONNX (Open Neural Network Exchange) format using the `tf2onnx` converter,
  - then interfaced from SModelS using ONNX Runtime





PRELIMINARY

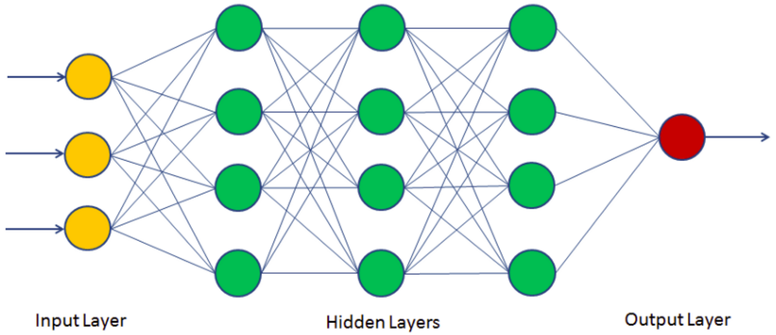
TensorFlow 2.0 ONNX ONNX RUNTIME

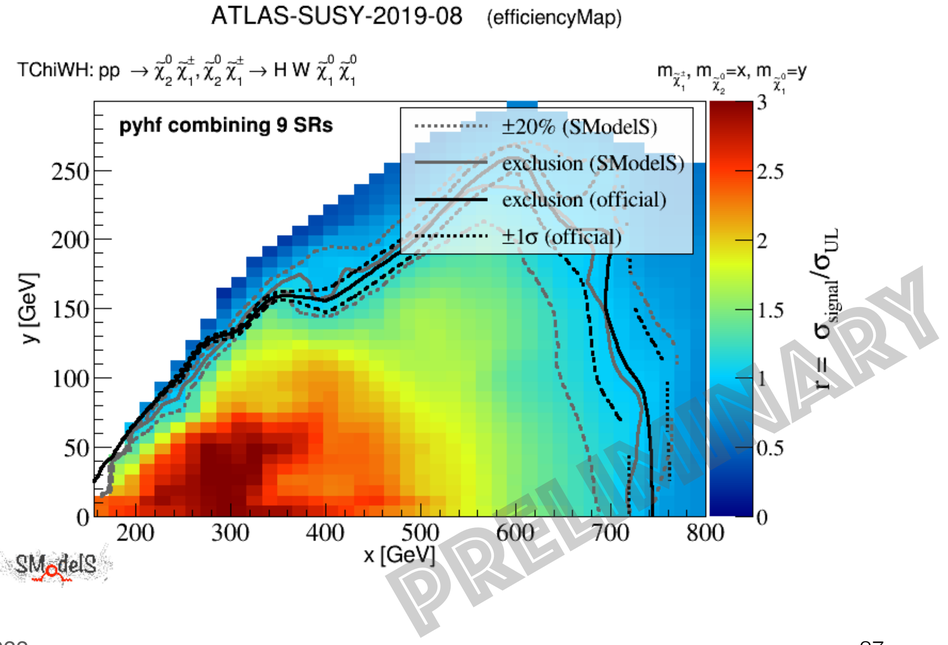
Machine-learn likelihoods

Work in progress!

Humberto Reyes Gonzalez

- Machine-learn likelihoods from full statistical models as functions of signal counts in each SRs
- Neural networks: sequential multilayer perceptrons using `TensorFlow2`.
  - loss function: mean squared error
  - activation function: Leaky Relu with alpha=0.2.
- Example for ATLAS-SUSY-2019-08 (9 SRs):
  - 3 layers, 512 neurons, MAPE = 0.2504, Max PE: 5.429
- The NN models are
  - saved in ONNX (Open Neural Network Exchange) format using the `tf2onnx` converter,
  - then interfaced from SModelS using ONNX Runtime





Analysis preservation and reuse

“Data and Analysis Preservation, Recasting and Reinterpretation”
S. Bailey et al., arXiv:2203.10057

Analysis preservation

Preservation of **analysis logic and workflows** enabling the reuse of the original analysis process and associated data products.

Full preservation

Exact software chain used to perform the analysis in the experiment.

- Full post-generation software stack from detector simulation and reconstruction to the physics analysis code.
- Complicated by the diversity of analysis software frameworks, even within a given experiment → container images
- High computing power requirements on any large-scale reuse

recast

Lightweight preservation

Fast simulation frameworks, independent of the original experiment software chain

- More modest CPU demands, but approximations involved reduce the precision.
- Several **public** frameworks enabling reinterpretation studies for the whole HEP community
- Two categories:
  - Particle-level (detector-unfolded) measurements
  - Reconstruction-level analyses (require emulation of detector effects)

RIVET

CHECKMATE

MADAnalysis 5

GAMBIT ColliderBit

ADL

LLP Recasting

....

Snowmass white paper on data and analysis preservation and reinterpretation

S. Bailey et al., arXiv:2203.10057

Analysis Preservation Recommendations

3.1: Ensure use of interoperable systems to maximise the preservability and reusability of experiment simulation and analysis software chains. This includes the use of version control, archival systems, containerisation, common software interfaces and data formats, and commitments from experimental collaborations and their host laboratories to maintain documentation and provide long-term support.

3.2: Ensure that all operational and in-preparation experiments have a planned and resourced programme for capture and long-term reproduction of their complete computational processing chain, including validation regression-tests.

3.3: Ensure
commu
process
umenta
for community consumption.

Ensure that **release of analysis preservation logic via public frameworks for the community to use** is **integrated with experiment publication and data-release processes**, to maximise analysis impact.

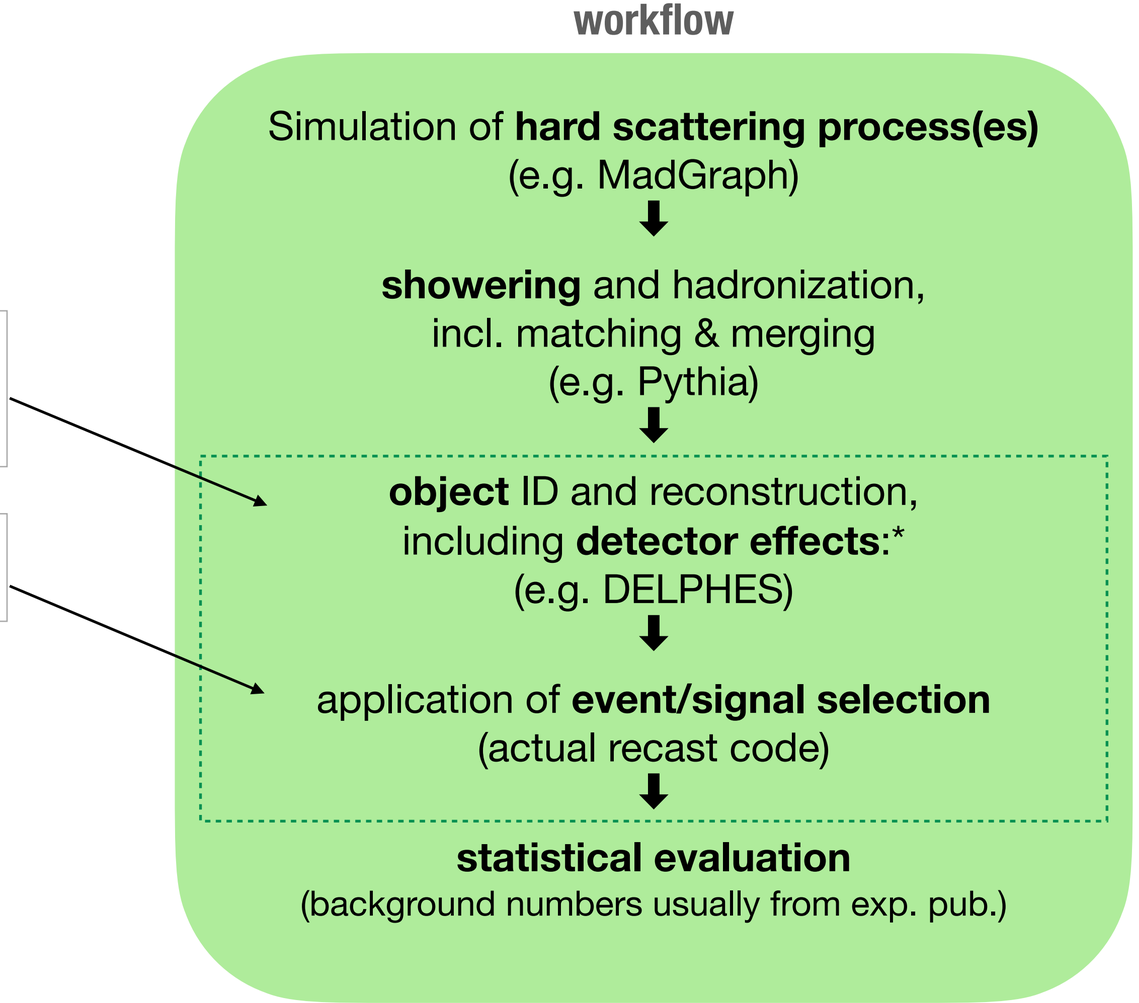
3.4: Support continuing development and uptake of new technologies for increasingly framework-independent analysis specifications, such as via declarative domain-specific analysis description languages.

Simulation-based reinterpretation (“recasting”)

- Aims at reproducing experimental analyses in Monte Carlo simulation
- Nowadays well established for traditional cut-based analyses. Information needed: cf. arXiv:2003.07868

object definitions; identification, tagging, reconstruction efficiencies

detailed preselection and signal (+control) region cuts



* except for detector-unfolded results (Rivet/Contur)

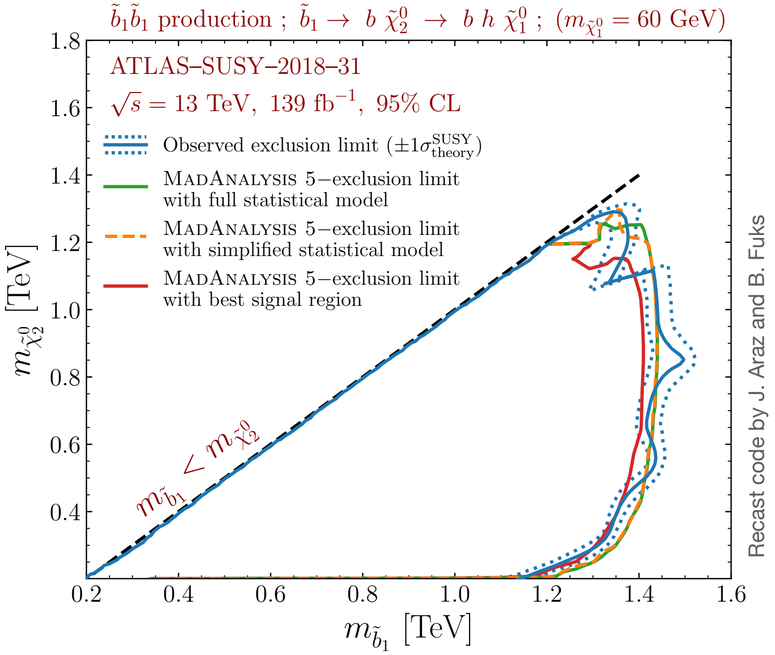
- However, more and more analyses exploit ML techniques to gain in sensitivity

e.g. ML-based taggers,
signal/bkg discrimination with ML classifiers

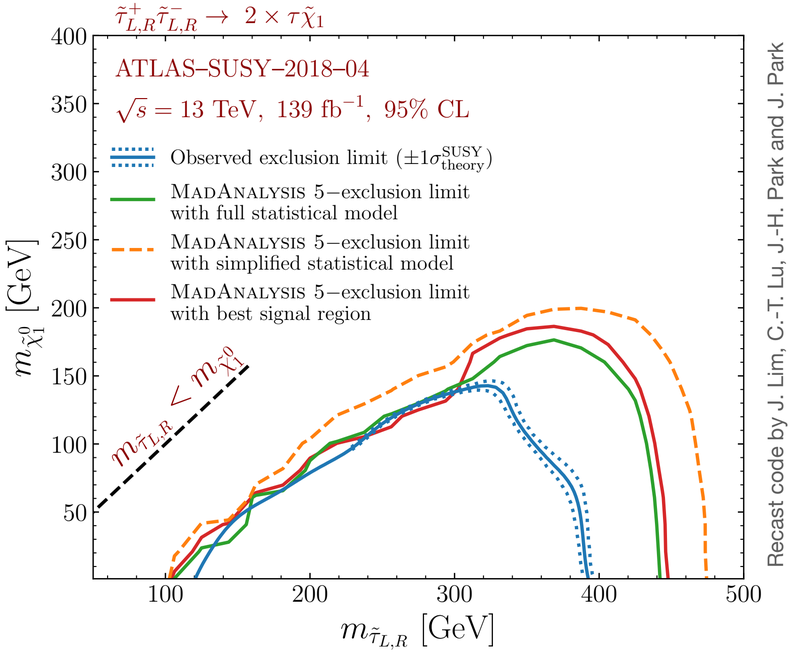
Pb: how can we reuse those?

ML as a bottleneck for reinterpretation?

Illustrative example from G. Alguero, J. Araz, B. Fuks, SK, arXiv:2206.14870



Conventional cut-based analysis.
All needed information is provided,
recasting works very well



Analysis employs a BDT for tau tagging;
resulting efficiencies given only approximately
→ serious differences in the recasting

(final weights ~30% too high)

ML as a bottleneck for reinterpretation?

- More and more analyses exploit ML techniques to gain in sensitivity.
- Serious difficulty for analysis preservation and reuse unless
  - resulting id/reco efficiencies can be (and are!) parametrised in terms of quantities accessible in a simulation, e.g., p_t, η, …
  - the actual ML model is published in appropriate form.

    Caveat: input variables need to be physics quantities that can be matched in a simulation

Two analyses where the latter has been attempted:

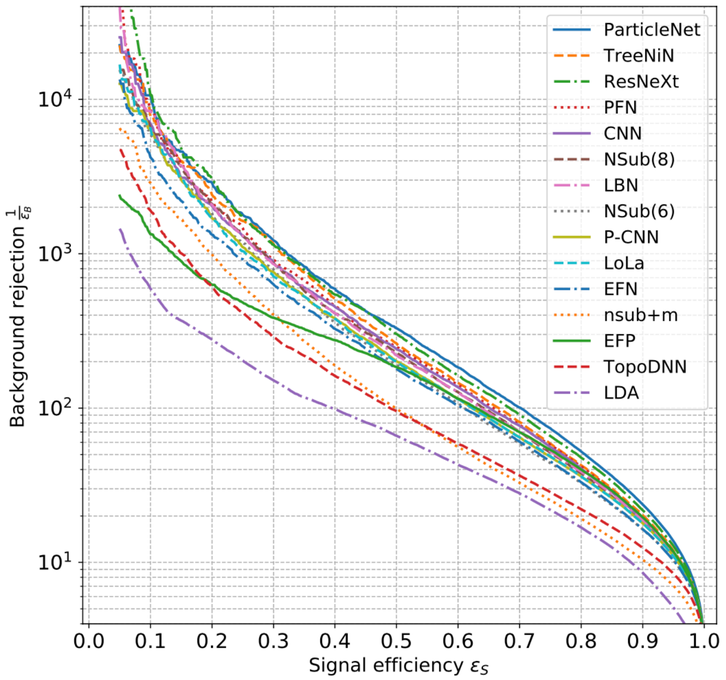
- **ATLAS-SUSY-2018-22** (0-lepton gluino/squark search) published **BDT weights** as XML file

  → RAMP seminar by Kenta Uno
- **ATLAS-SUSY-2019-04** (1-2 leptons + jets RPV search) published neural network as **ONNX** file

  → RAMP seminar by Javier Montejo Berlingen

RAMP: Reinterpretation Auxiliary Material Presentation

Example: top taggers



13 different algorithms: image-based (2), 4-vector-based (5), theory-inspired (6) taggers

"The Machine Learning Landscape of Top Taggers"
G. Kasieczka, T. Plehn et al., arXiv:1902.09914

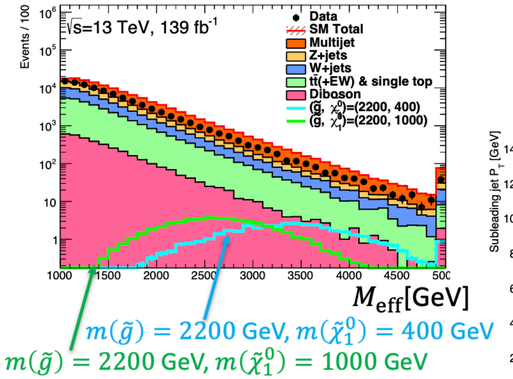
Kenta Uno on ATLAS-SUSY-2018-22

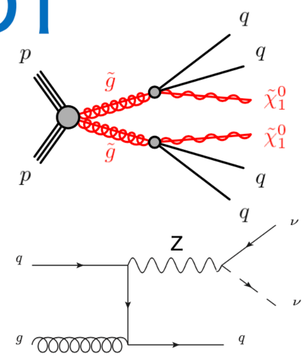
Analysis approach: BDT

Key variable in the previous analysis: M_{eff}

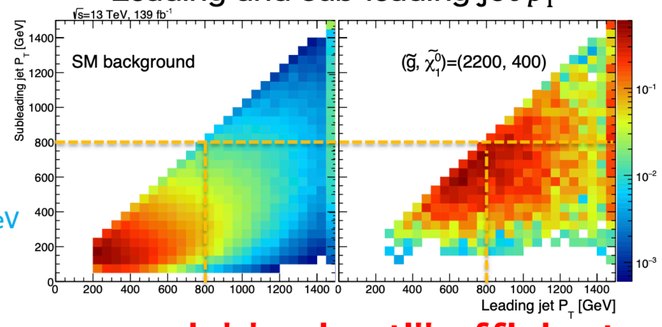
$$M_{\mathrm{eff}} = \sum_i |p_{\mathrm{T}}(i)| + E_{\mathrm{T}}^{\mathrm{miss}}$$

- However, it is not efficient as $\tilde{\chi}_1^0$ mass is large





Leading and sub-leading jet p_{T}



Using correlation between variables is still efficient.

→ BDT analysis is useful for SUSY search

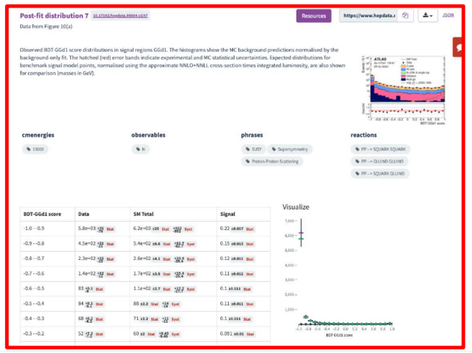
Discussion point 1

- The auxiliary material provided for the BDT reinterpretation
  - We published HepData materials [Link]
    - Provided xml files of BDT classifiers
      → ZeroLepton2018-SRBDT-weight.tar.gz
    - Also provided analysis snipped code
      → ZeroLeptonBDT2018.cxx
      → By using this code, you can use same BDT classifiers.
    - Full likelihood is provided in the hepdata material

※ Signal events: fast sim (reco)
※ Background events: full reco

We provide data, SM total and signal yields in each BDT score bin and the acceptance and efficiencies of BDT SRs for simplified GG models.

→ I think it might help everyone..



10 -12 input variables: jet $p_{\mathrm{T}}, \eta, E_{\mathrm{T}}^{\mathrm{miss}}, m_{\mathrm{eff}}$, Aplanarity

Javier Montejo Berlingen on ATLAS-SUSY-2019-04

Reinterpretation material

Additional resources

C++ code snippet with the implementation of the analysis selection at truth-level

- Can be used with SimpleAnalysis framework

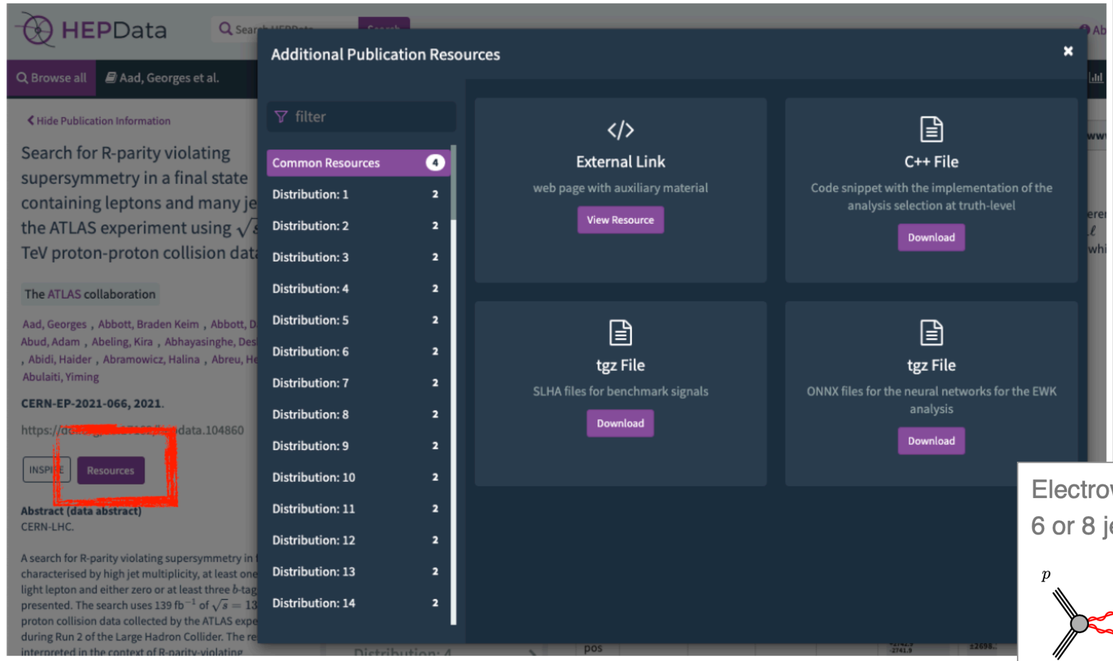
SLHA files for benchmark signals

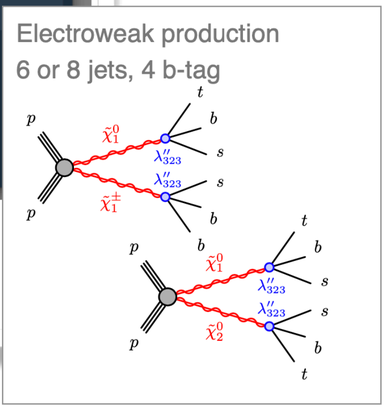
ONNX files for the neural networks for the EWK analysis

- Not possible to use lwtnn because the architecture of one of the layers is not supported

Upcoming: **ROOT workspaces**, containing data and the fitted background model in all SRs

- not using pyhf as it doesn't support our parameterised background model





Snowmass white paper on data and analysis preservation and reinterpretation

S. Bailey et al., arXiv:2203.10057

Reinterpretation and Recasting Recommendations

5.1: Encourage that reinterpretability and reuse be kept in mind early on in the ar
M
te
nc
th

Encourage that **reinterpretability and reuse be kept in mind early on in the analysis design**. This concerns, for instance, the **choice of input parameters in ML models**, the full specification of the fiducial phase space of a measurement in terms of the final state, including any vetos applied, and generally the **choice of non-overlapping regions and standard naming of shared nuisances to facilitate the combination of analyses**.

5.2: De
ucts, such as statistical models, with reinterpretation use-cases in mind.

5.3: Improve the coordination among the different public reinterpretation frameworks with the goal of a centralised database of recast codes, common input/output formats, and a unified statistical treatment.

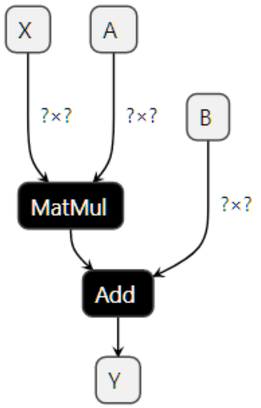
5.4: Encourage the FAIR-ification of codes and data products from (theory) reinterpretation studies outside the experimental collaborations at the same level of sophistication as asked for experimental analyses and results. Suitable repositories are, e.g., GitHub and Zenodo; appropriate versioning is essential.

Some more comments on reproducibility and reuse of ML models

Open Neural Network Exchange

https://onnx.ai/onnx/tutorial_python/

"defines all the necessary operations a machine learning model needs to implement its inference function"



- ONNX is an open format built to represent ML models.
  - aims at providing a common language any ML framework can use to describe its models.
  - makes it possible to deploy a model **independent from the learning framework** used to build it.

  The deployment of a ML model usually requires replicating the entire ecosystem used to train the model, most of the time with a *docker*. Once a model is converted into ONNX, the production environment only needs a runtime (C, java, python, javascript, ….) to execute the graph defined with ONNX operators.

- Converters exist for scikit-learn, tensorflow, pytorch, and others

  NB must be updated every time ONNX or the library they support have a new released version.

- Beware of custom layers, experimental features, etc.!

  may be troublesome for converter and/or runtime (interpreter)

For pertinent reuse, input variables must be clearly documented

Runtime (interpreter) must match ONXX version → possible issue for preservation?

Ltwnn

Lightweight Trained Neural Network

David Hohn > lwtnn

build passing | coverity passed | DOI 10.5281/zenodo.597221

What is this?

The code comes in two parts:

1. A set of scripts to convert saved neural networks to a standard JSON format
2. A set of classes which reconstruct the neural network for application in a C++ production environment

The main design principles are:

- **Minimal dependencies:** The C++ code depends on C++11, Eigen, and boost PropertyTree. The converters have additional requirements (Python3 and h5py) but these can be run outside the C++ production environment.
- **Easy to extend:** Should cover 95% of deep network architectures we would realistically consider.
- **Hard to break:** The NN constructor checks the input NN for consistency and fails loudly if anything goes wrong.

We also include converters from several popular formats to the `lwtnn` JSON format. Currently the following formats are supported:

- Scikit Learn
- Keras (most popular, see below)

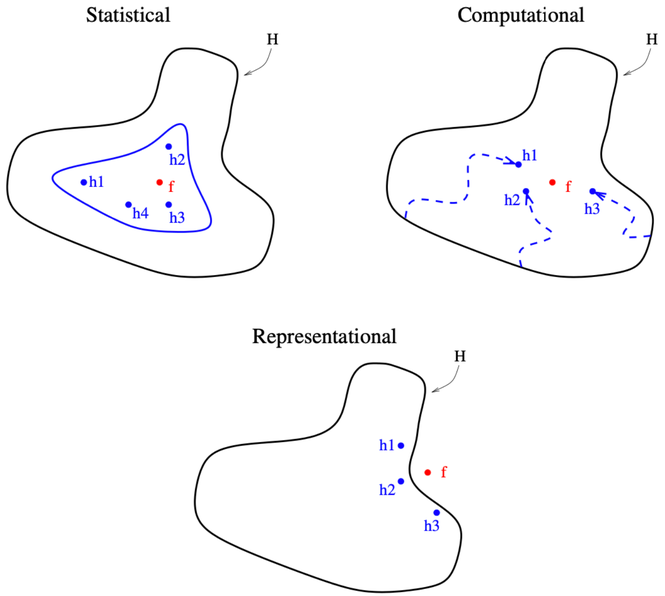
> "Our underlying assumption is that *training* and *inference* happen in very different environments: we assume that the **training environment** is flexible enough to support modern and **frequently-changing libraries**, and that the **inference environment is much less flexible**."

Reproducibility of ML models

To reproduce a ML model, one needs

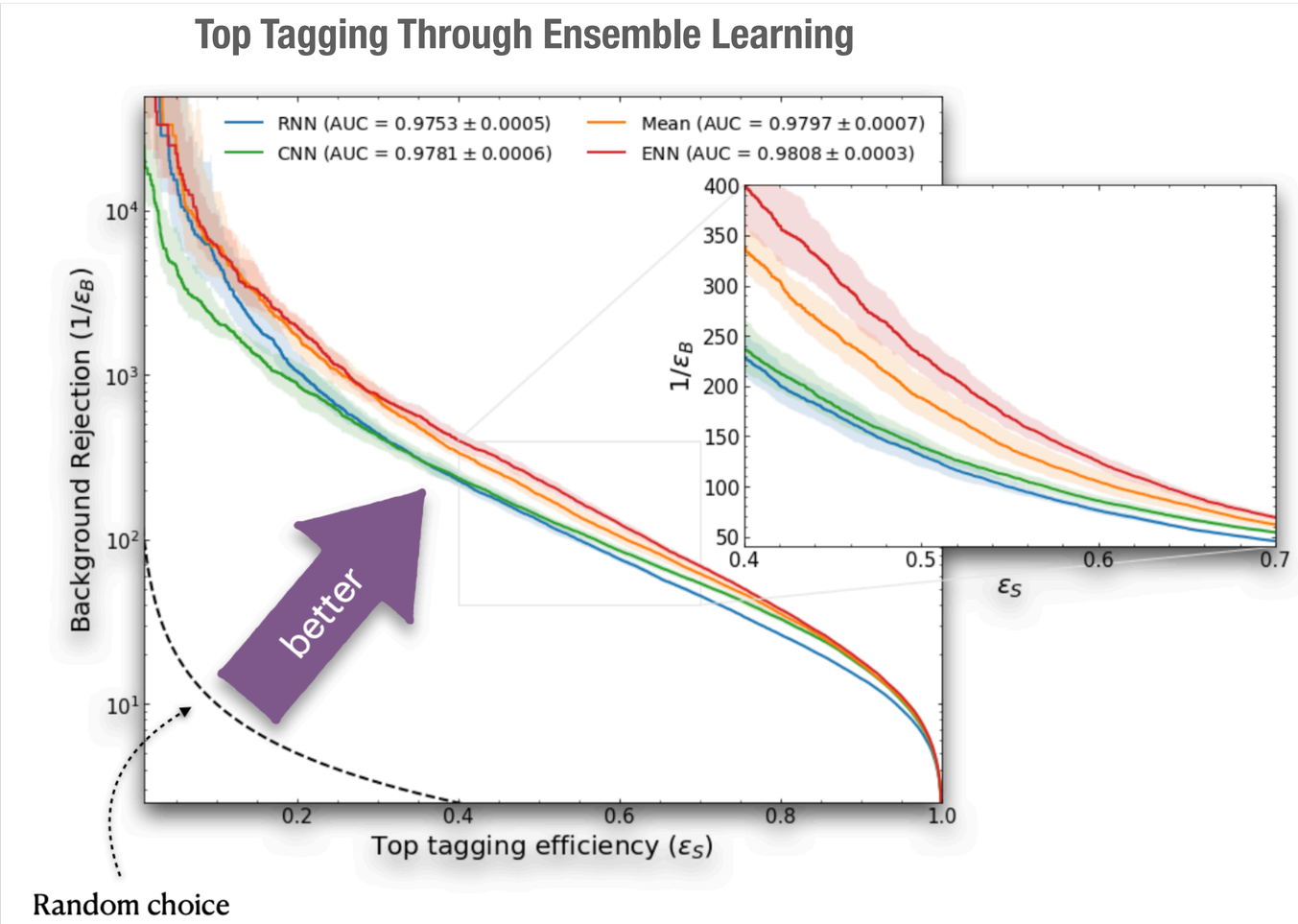
- The **precise ecosystem** used to train the model
  - Tools and their versions
  - Exact architecture (layers, neurons, …)
  - Activation and loss functions
  - Anything else specifying the algorithm
- The **training and validation data**
- The **initial conditions** (random seeds) for the weights

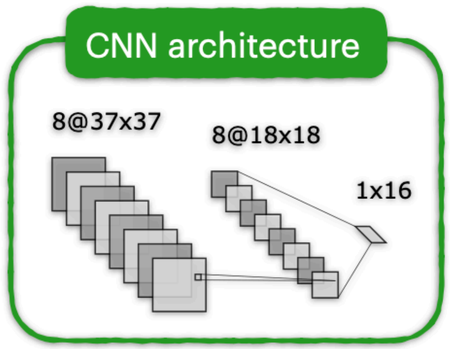
"A learning algorithm can be viewed as searching a space H of hypotheses to identify the best hypothesis in the space."

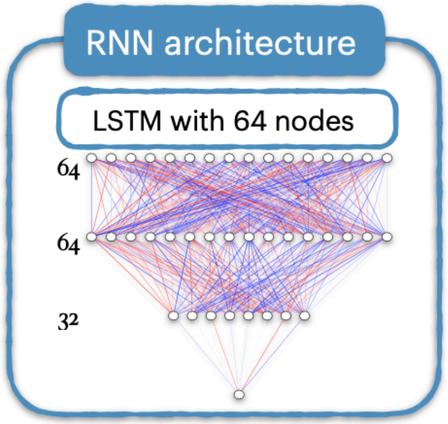


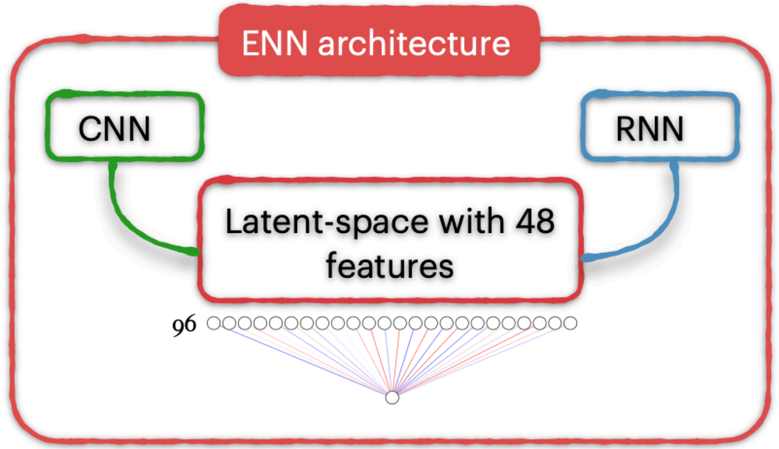
Thomas G. Dietterich
Ensemble Methods in Machine Learning

Effect of different initialisation → ensemble methods









Durham University

12

Jack Y. Araz - *Ensemble Neural Networks*

Conclusions

- The data and analyses from particle physics experiments are unique and of immense scientific value.
- Impact can go much beyond original paper publication.
- Proper preservation, enabling long-term future reuse, maximises the scientific return.
- It is worthwhile to keep reinterpretability and reuse in mind early on in analysis design.

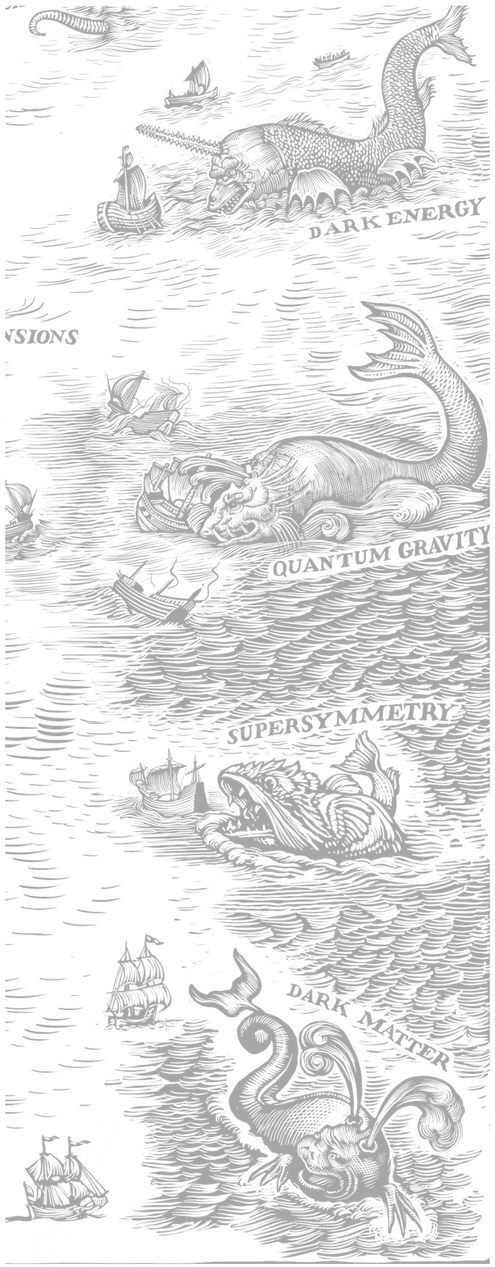


Illustration (c) C.Wormell from "A Map of the Invisible" by J.Butterworth

Next Reinterpretation-Forum workshop

(Re)interpretation of the LHC results for new physics

12–15 Dec 2022
CERN
Europe/Zurich timezone

Enter your search term

- Overview
- Timetable
- Registration
- Call for Abstracts
- Participant List

This is the 7th general workshop of the "Forum on the interpretation of the LHC results for BSM studies", or LHC Reinterpretation Forum (RIF) for short. Its aim is to review new developments on the tools, pheno, and the experimental sides, and to prepare for the Run 3 results of the LHC. In this context, major topics of this workshop will be:
i) the publication and reuse of statistical models,
ii) the reinterpretation of analyses that employ machine learning, and
iii) global analyses and global fits.

Continuing the conversation from the last workshop session, we would like to include general best practices for reinterpretation/reuse of experimental results beyond the LHC, and particularly welcome contributions regarding results from precision or astrophysical experiments.

https://indico.cern.ch/event/1197680/

Backup

CMS simplified likelihood

[CMS NOTE-2017/001]

$$\mathcal{L}(\mu,\boldsymbol{\theta}) = \mathcal{P}(\text{data}|\mu\cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta}))\cdot p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) \quad \Rightarrow \quad \mathcal{L}_S(\mu,\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu\cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu\cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{V}^{-1}\boldsymbol{\theta}\right)$$

Some simplifying assumptions must be made to reduce the complexity of the full probability density function $p(\tilde{\boldsymbol{q}}|\boldsymbol{q})$.

- The **constraints on the background contributions are Gaussian** such that the distribution of the number of background events is symmetric about the expectation, b_i, and its variance is independent of $\boldsymbol{q}$. Often, the background contributions are estimated from control regions in data with large sample sizes, which makes this assumption valid. The covariance, and therefore only the linear correlation, between the background contribution in each region is sufficient to approximate $p(\tilde{\boldsymbol{q}}|\boldsymbol{q})$ at least for values of $\boldsymbol{q}$ which are close to $\tilde{\boldsymbol{q}}$.
- The **numbers of events, n_i, are statistically independent from one another**. This is true when there are no events which are included in more than one search region and the estimates of the background contributions, b_i, and covariance matrix **V** have not been obtained from data which are statistically dependent on the data from any search region.
- The **systematic uncertainties in the signal model can be neglected**. The validity of this assumption will strongly depend on the specific BSM physics model being considered. Systematic uncertainties on the signal can be accounted for by adding appropriate nuisance parameters with Gaussian constraints as for the background contributions.

Reinterpretation: two approaches

Reproduce experimental analysis
in a Monte Carlo simulation

“recasting”

Reuse simplified model results
(σ_{95}, signal A×ε)

Assumes that A×ε doesn’t change
too much w.r.t. original model

CHECKMATE

MADAnalysis 5

RIVET
Contur

ADL

ColliderBit

**Test of
BSM hypothesis**

SModelS

Workflow

simulation-based recasting

Simulation of **hard scattering process(es)**
(e.g. MadGraph)

⬇

showering and hadronization,
incl. matching & merging
(e.g. Pythia)

⬇

emulation of **detector effects**:*
object reconstruction, efficiencies, …
(e.g. DELPHES)

⬇

application of **signal selection cuts**
(actual recast code)

⬇

statistical evaluation
(background numbers usually from exp. pub.)

simplified model approach (SModelS)

Create **SLHA file**
with mass spectrum and decay tables

⬇

add **production cross sections**
(SModelS xseccomputer or micrOMEGAs)

⬇

compute all **signal weights** :
σ×BR×BR…

⬇

sum **matching elements**

⬇

find and apply A×ε (or σ_{95}) values
from **experimental results**

⬇

statistical evaluation‡
(background numbers usually from exp. pub.)

* except for detector-unfolded results (Rivet/Contur)

‡ in case exp. result is σ_{95}: only allowed/excluded

Pro's and con's

simulation-based recasting

Simulation of **hard scattering process(es)**
(e.g. MadGraph)

⬇

Showering and hadronization,
incl. matching & merging
(e.g. Pythia)

⬇

emulation of **detector effects**:
object reconstruction, efficiencies, …
(e.g. DELPHES)

⬇

application of **signal selection cuts**
(actual recast code)

⬇

statistical evaluation
(background numbers usually from exp. pub.)

- **More generic** and often **more precise** than simplified model results; in principle applicable to any new signal
  caveat: control regions typically not included in react codes!
- Need to take care to **simulate all relevant processes**
  (not always obvious e.g. in scans of complex parameters spaces where dominant processes can change)
- **Very CPU expensive**
- So far **only cut-and-count** analyses are recasted
- **ATLAS / CMS** as well as Run1 (8 Tev) / Run 2 (13 TeV) analyses need to be **run separately**
- So far, **prompt and long-lived signatures need to be treated separately**
  → careful separation needed in models featuring both
  → response of prompt analyses to LLPs unclear / wrong
- Implementation and validation of new analyses is time-consuming and sometimes quite difficult

  → **Detailed information needed from experiment**
  analysis logic, object definitions, cuts, efficiencies, cut-flows, etc.

Pro's and con's

- Assumes that **signal acceptances** are to good approximation **the same** as in original experimental result.

  Valid for **simple rescaling** of production and decay rates (σ×BR); other cases need to be **verified**, e.g. spin or production mode dependence.

- Applicable beyond cut & count analyses
  (ML techniques)
- Advantages are simplicity and **speed!**
  → **very fast** b/c no MC simulation needed
  → well suited for large scans and model surveys
- **Large database** of experimental results
- ATLAS and CMS, Run1 and Run2, **prompt and long-lived results** all **treated simultaneously**
- Easy **classification of** unconstrained cross section, **missing topologies**
- **Often conservative:** coverage depends on variety of available simplified-model results

simplified model approach (SModelS)

Create **SLHA file**
with mass spectrum and decay tables

⬇

add **production cross sections**
(SModelS xseccomputer or micrOMEGAs)

⬇

compute all **signal weights** :
σ×BR×BR…

⬇

sum **matching elements**

⬇

find and apply A×ε (or σ_{95}) values
from **experimental results**

⬇

statistical evaluation‡
(background numbers usually from exp. pub.)

‡ in case exp. result is σ_{95}: only allowed/excluded