From likelihoods to measurements: analysis of the ttH production with the ATLAS detector

David Muñoz Pérez

Student Seminar

11th April 2025

Introduction

- The goal of this seminar is to clarify a bit some of the concepts that are constantly referred to when speaking about LHC physics measurements.

Main contents

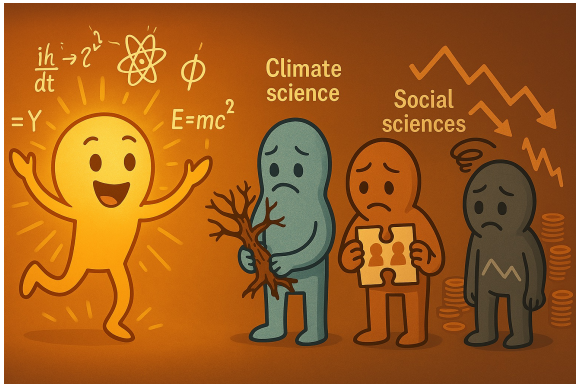
- Why do we have to use simulations in particle-physics experiments?
- The likelihood for measurements with binned data.
- How do we implement statistical uncertainties?
- Methods for statistical inference in LHC measurements.
- Review of some concepts: signal and control regions, the discovery significance.
- Application of all these concepts to one specific ATLAS analysis → the measurement of the Higgs boson production cross-section in association with top quarks (ttH) in multi-lepton final states using the ATLAS Run 2 dataset

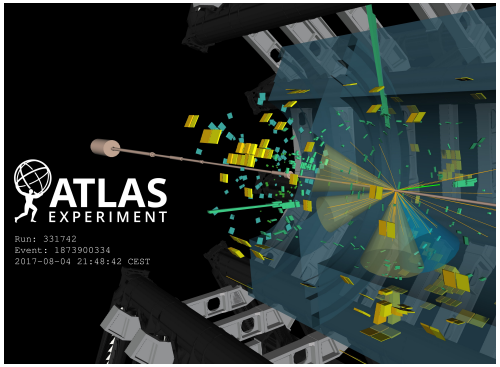
- The frequentist approach will be assumed in most of the statements made during the talk.

From pp collisions to the public results

Which is the goal of a particle-physics experiment?

- Among the sciences, **particle physics has a very well established theoretical basis**. Thanks to Quantum Field Theory, we can predict phenomena not only for the Standard Model (SM) but also for physics beyond that.
- This is a luxury that many other fields do not have e.g. climate science, social sciences, economics, etc, where the high complexity of the problems difficults predictive power.
- So which is the goal of a particle-physics experiment? In general, it is to make statements about SM (or BSM) parameters α given some data x.





$$\frac{m_f}{v} H \bar{f} f$$

Top Yukawa coupling

- The connection between α and x is the likelihood $L_x(\alpha) = p(x|\alpha)$ → probability of data x given specific values for the theory parameters α.
- Then, given the observed data x, we find the specific values of α that maximize the likelihood → **We need the functional form of the likelihood**.

The likelihood

Likelihood for n observed events:

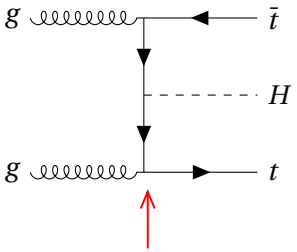
$$p_{\text{full}}(\mathcal{D}|\alpha) = \text{Pois}(n|\epsilon L\,\sigma(\alpha)) \prod_i p(x_i|\alpha)$$

- For a pp collider such as the LHC, the functional form of the likelihood for one event looks like this:

$$p(x|\alpha) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p\; p(x|z_d)\, p(z_d|z_s)\, p(z_s|z_p)\, p(z_p|\alpha)$$

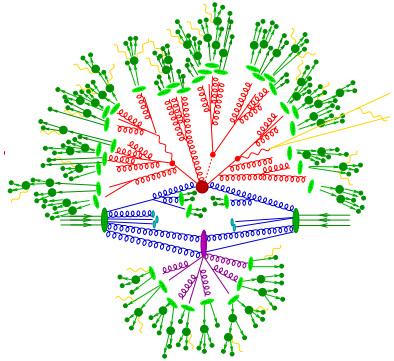
Latent variables → Variables that cannot be measured

parton level, $p(z_p|\alpha)$

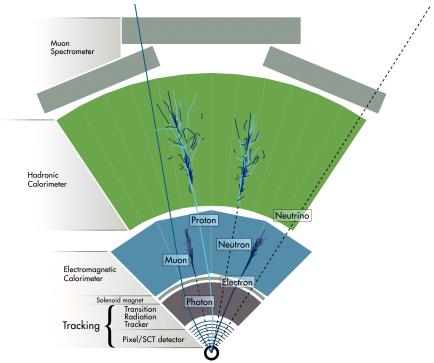


The dependence on theory parameters α is here.

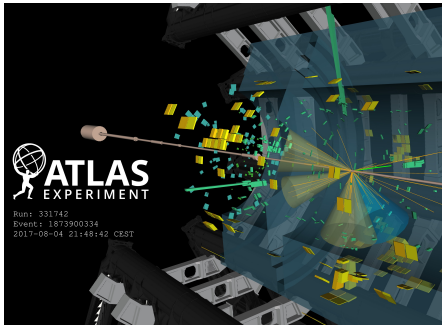
parton shower, $p(z_s|z_p)$



detector interactions, $p(z_d|z_s)$



data i.e. sensor read-outs, $p(x|z_d)$

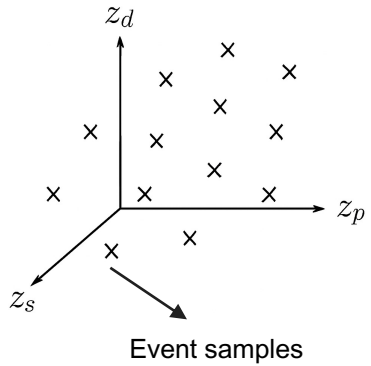


- We do know the functional form of the probability density functions $p(x|z_d)$, $p(z_d|z_s)$, $p(z_s|z_p)$ and $p(z_p|\alpha)$.
- However, to get $p(x|\alpha)$, we need to integrate over the z_d,z_s,z_p phase space compatible with the measurement x → **An integral over such an enormous phase space cannot be computed in practice**.
- We have an **intractable likelihood** i.e. we cannot make statistical inference with it → Solution: **a simulation-based approach**.

Simulation-based approach

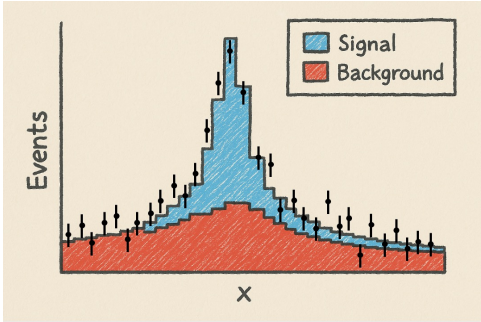
$$p(x|\alpha) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \; p(x|z_d)\, p(z_d|z_s)\, p(z_s|z_p)\, p(z_p|\alpha)$$

- We know $p(x|z_d)$, $p(z_d|z_s)$, $p(z_s|z_p)$ and $p(z_p|\alpha)$. We just can't compute the integral over the whole phase-space!
- Solution: we simulate points in the z_d,z_s,z_p phase space i.e. we approximate $p(x|\alpha)$ by **"sampling"** the probability density function $p(x|z_d)\, p(z_d|z_s)\, p(z_s|z_p)\, p(z_p|\alpha)$.
- The software that generates all the event samples is referred to as a Monte Carlo generator.
- By simulating many events, one can build histograms to estimate $p(x|\alpha)$.



Additional problem:

- The data x is not one variable, as shown in the illustration. It corresponds to millions of variables, corresponding to the millions of sensor read-outs installed in the detector.
- Histograms are hit by the curse of dimensionality → number of event samples needed scales exponentially with the dimension of the observation.
- Solution: find one variable or a small set of variables (summary statistics) that condenses the relevant information about the theory parameter to be measured e.g. if you want to measure the mass of the Z boson in the $Z \to \mu\mu$ channel, use the invariant mass of two muons as observable.
- An efficient and precise particle reconstruction is vital for finding a good summary statistics.
- Also, machine learning allows to separate signal from backgrounds → Output score as summary statistics v.

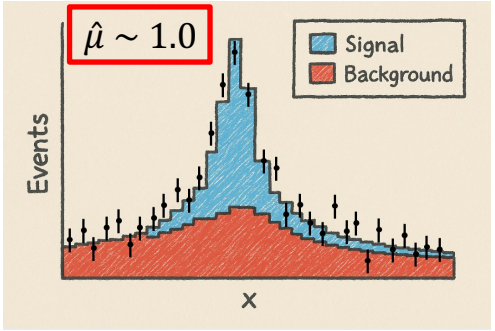


The binned likelihood

- Using histograms we can build the **binned likelihood**. But what is its exact functional form?

$$p(n_i|\alpha) = \prod_{i=1}^{N} \text{Poisson}\,(n_i|e_i(\alpha))$$

→ The Poisson distribution describes the probability of observing n_i events in bin i when expecting e_i.



- Notice the expected number of events in each bin, e_i, is what we are simulating with Monte Carlo → $\boldsymbol{e_i}$ **depends on the theory parameters** $\boldsymbol{\alpha}$.
- When one is searching for a certain process, the parameter of interest α is typically chosen to be the so-called signal strength μ, with e_i parametrised as

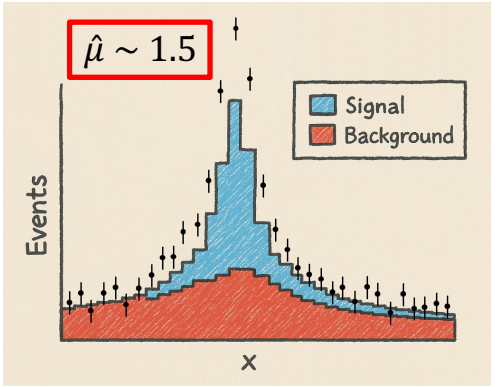
$$e_i(\mu) = \mu \cdot s_i + b_i$$

$\mu = 1 \rightarrow$ Signal is present, with the SM cross-section.

$\mu = 0 \rightarrow$ Signal is not present.

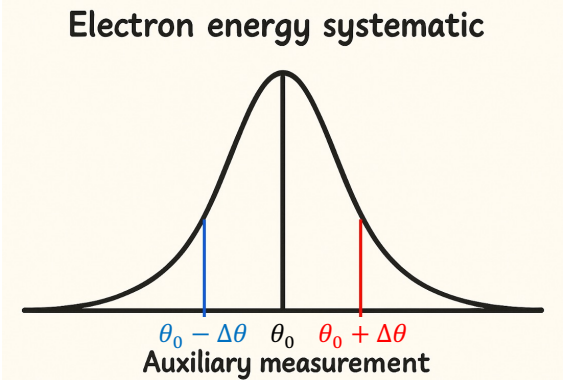
How do we measure the physical parameter μ?

- We simply find the value of μ that maximizes the binned likelihood $p(n_i|\mu)$.
- Such value is called the maximum-likelihood estimator (MLE) of $\mu \rightarrow \hat{\mu}$



What about systematic uncertainties?

- Systematic uncertainties include
  - Theory uncertainties e.g. on the calculation of the theoretical cross-section of the simulated process.
  - Experimental uncertainties e.g. uncertainties on the measured energy of the different objects → These are obtained from dedicated **auxiliary measurements**.



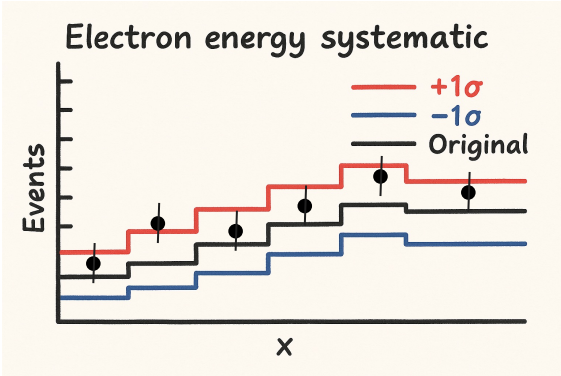
- How are systematics uncertainties implemented into the binned likelihood? They are implemented as **nuisance parameters** θ_j $(j = 1, \dots, M)$, representing the M auxiliary measurements that define the different systematic uncertainties

$$p(n_i|\mu,\vec{\theta}) = \prod_{i=1}^{N} \text{Poisson}\left(n_i|e_i(\mu,\vec{\theta})\right) \prod_{j=1}^{M} \mathcal{N}(\theta_0, \Delta\theta|\theta_j)$$

$$e_i(\mu,\vec{\theta}) = \mu \cdot s_i(\vec{\theta}) + b_i(\vec{\theta})$$

Poisson counting term of the likelihood

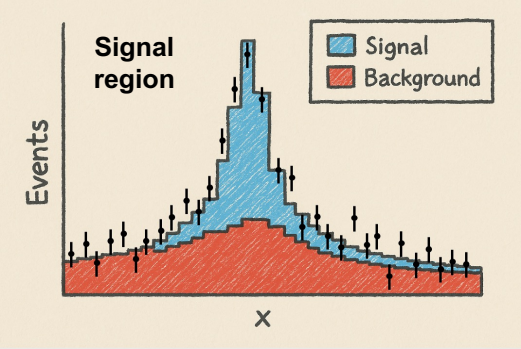
Gaussian constrain term of the likelihood

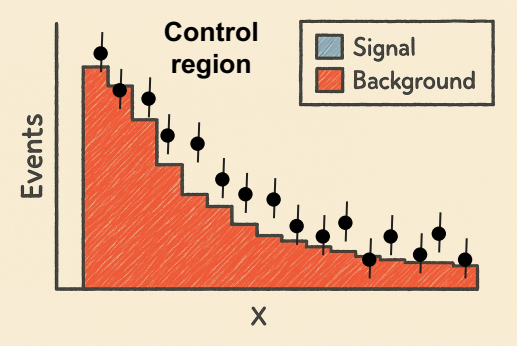


- Nuisance parameters θ_j are present in both the Poisson counting term and in the Gaussian auxiliary measurement term → This may generate a competition between both terms when maximizing the likelihood

Signal regions and control regions

- **Signal region** → Region of the phase space defined by certain selection cuts that try to maximize the $S/\sqrt{B}$ ratio e.g. if you are looking for the Higgs boson in the $H \to \gamma\gamma$ channel, one of the selection cuts for your SR will be to have $m_{\gamma\gamma}$ around the Higgs-boson mass peak.





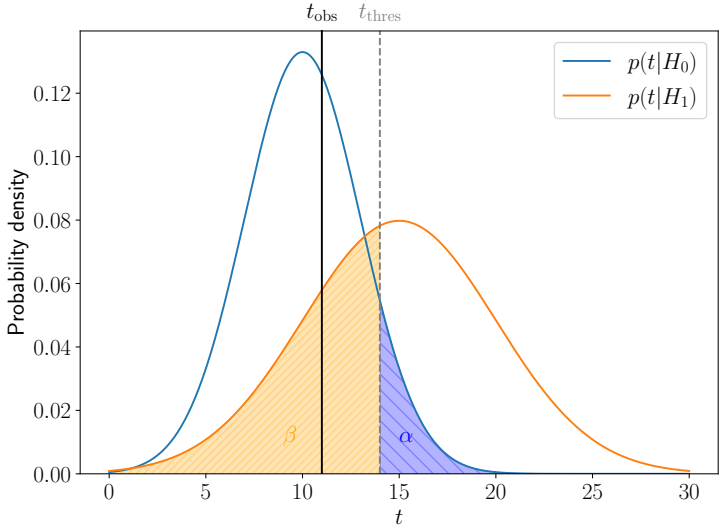
- Even applying the most efficient selection cuts for your SR, in general you will still have contributions from other processes that you are not targetting in your measurement (backgrounds).
- Instead of assuming that the background behaves exactly as predicted by the SM, one can define a control region for such background.
- **Control region** → Region of the phase-space enriched with events of a certain background process. It is used to fit such background process to data i.e. providing a more realistic estimate of the number of background events.

$$p(n_i|\mu,\vec{\theta}) = \prod_{i=1}^{N} \text{Poisson}\left(n_i|e_i(\mu,\vec{\theta})\right) \prod_{j=1}^{M} \mathcal{N}(\theta_0, \Delta\theta|\theta_j)$$

$$e_i(\mu,\vec{\theta}) = \mu \cdot s_i(\vec{\theta}) + N_b \cdot b_i(\vec{\theta})$$

Hypothesis testing and the sigmas thing

- When targeting the discovery of a process, two different hypothesis are tested:
  - Null hypothesis H_0 → Corresponds to $\mu = 0$ i.e. background-only hypothesis.
  - Alternative hypothesis H_1 → Corresponds to $\mu = 1$ i.e. signal+background hypothesis.
- Test statistic $\mathrm{t}(x)$ → Scalar function of the data x. A simple choice of $\mathrm{t}(x)$ could be the total number of observed events $\mathrm{t}(x) = n$.

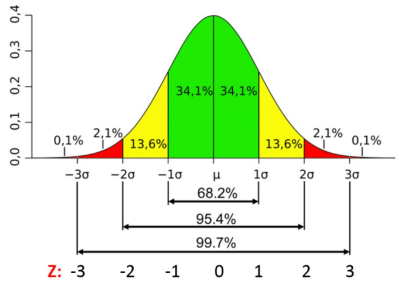


Example (see figure)

- We expect $\sim 9(15)$ events under the assumption of $H_0(H_1)$.
- **Before looking at data**, we define the critical region α, which defines the region for which we will reject the H_0 hypothesis → Normally α is a very low number to be quite sure when rejecting H_0.
- Then, **we look at data** → We observe 11 events i.e. $t_{obs} = 11$ → **We cannot reject the background-only hypothesis**.

Observed discovery significance

- The t_{obs} value determines the significance of the background-only hypothesis rejection.
- In our example, if we had observed 20 events i.e. $t_{obs} = 20$, we would have rejected H_0 with a good significance.

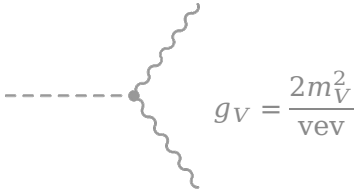


The Standard-Model Higgs boson

The SM Higgs boson

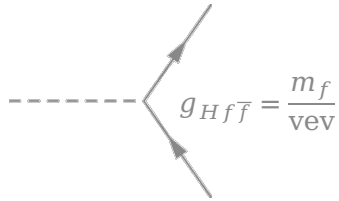
- After the discovery of the Higgs boson by ATLAS and CMS in 2012, the measurement of its properties has been a priority.
- The coupling between the Higgs boson and other particles is defined by the particle's mass and type. Three types of couplings to massive particles:

Gauge couplings to vector bosons

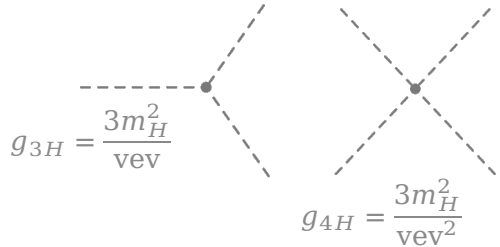


$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\not{F}\not{D}\not{\chi} + \chi_i y_{ij}\chi_j \phi + h.c + |D_\mu \phi|^2 - V(\phi)$$

Yukawa couplings to fermions e.g. $Ht\bar{t}$ coupling



Self-couplings of the Higgs field

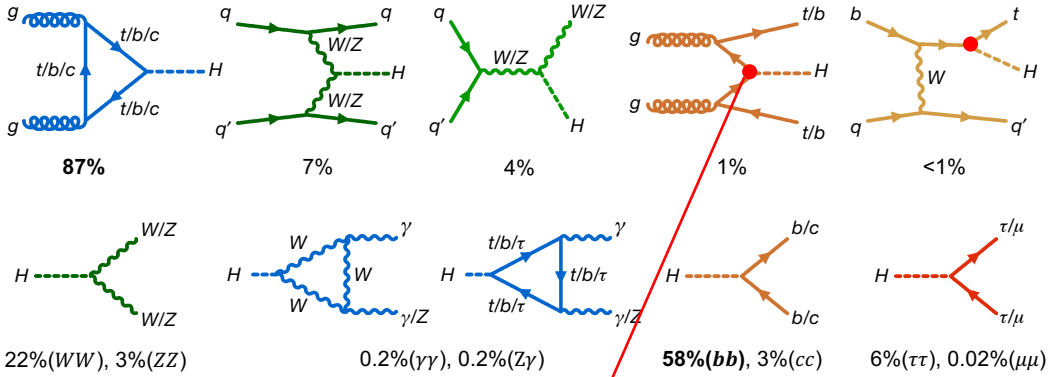


- Precision measurements of couplings are crucial:
  - Test of spontaneous symmetry-breaking mechanism.
  - Test for Standard Model predictions in the Higgs sector (modified couplings, CPV, etc).

The SM Higgs boson: production and decay at the LHC

- The SM Higgs couplings, the collided partices (pp in the case of the LHC) and the collision energy determine the predicted Higgs production and decay rates:

Main Higgs production and decay channels for pp collisions at $\sqrt{s} = 13$ TeV ($m_H = 125$ GeV)

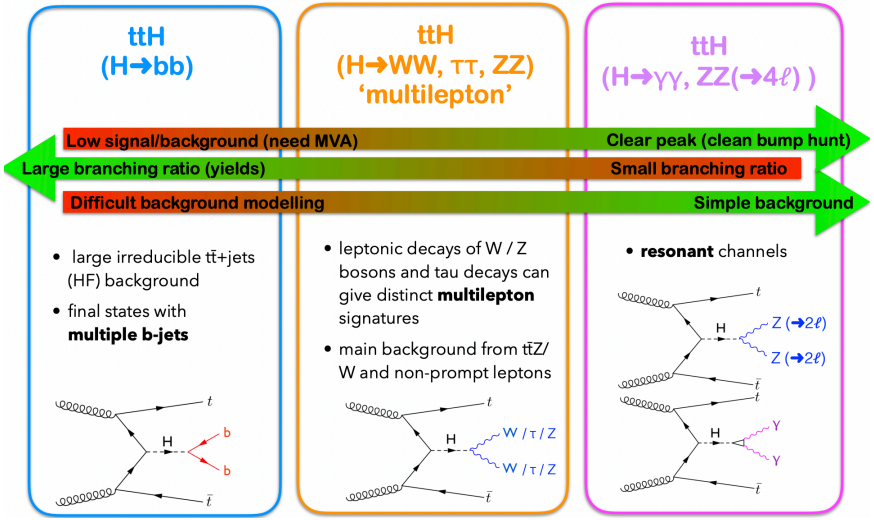


$t\bar{t}H$ provides a window to the **direct measurement of the top-Higgs coupling** (gluon-gluon fusion provides larger cross-section but the loop could include BSM particles and bias the coupling measurement)

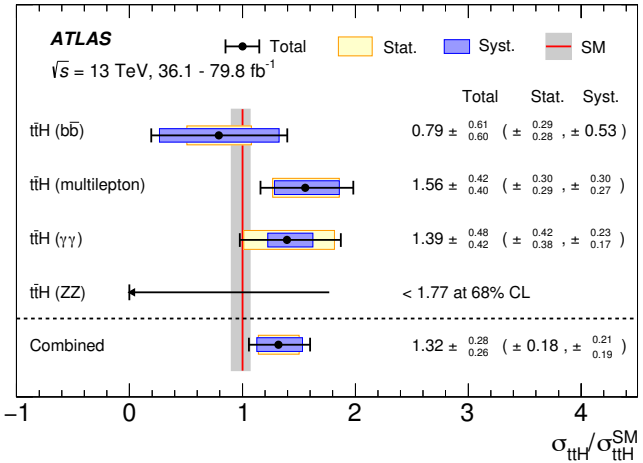
$t\bar{t}H$ production in multi-lepton final states

Why $t\bar{t}H$? Which $t\bar{t}H$?

- Direct measurement of the top-Higgs Yukawa coupling.
- Test of the electroweak symmetry breaking.
- Helps with sensitivity to Higgs self-coupling.
- Allows to measure the CP properties of the top-Higgs interaction → Potential source of CP violation contributing to the observed baryon asymmetry.
- $t\bar{t}H$ production measurement is typically splitted in different analyses depending on the Higgs decay mode:



$t\bar{t}H$ observation (Phys. Lett. B 784 (2018) 173)



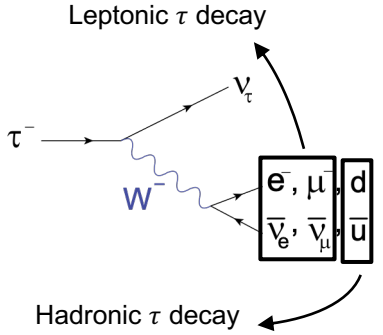
- $t\bar{t}H$ cross-section measurements using the full Run 2 dataset (140 fb^{-1} at $\sqrt{s} = 13$ TeV) have been provided in the $H \to ZZ^* \to 4\ell$ (Eur. Phys. J. C 80 (2020) 957), $H \to \gamma\gamma$ (JHEP 07 (2023) 088) and $H \to b\bar{b}$ (arXiv:2407.10904) channels. **Not yet in the $H \to$ multi-lepton (ML) channel !!**

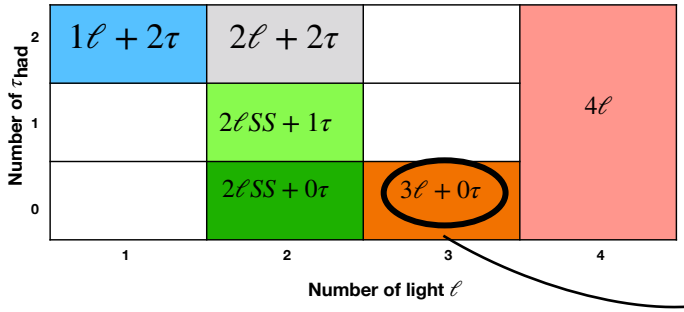
$t\bar{t}H$ ML analysis

- The $t\bar{t}H$ ML channel targets Higgs decays that yield multiple leptons in the final state i.e. $H \to WW$, $H \to ZZ^*$ (not to 4ℓ) and $H \to \tau\tau$.

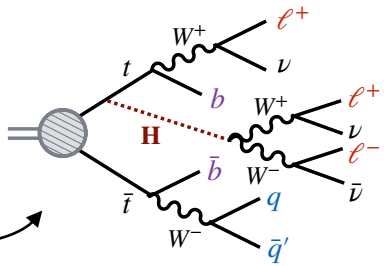
- Note: In ATLAS, we distinguish light leptons ℓ (e,μ), which traverse the full detector and interact with it, from τ leptons, which do not reach the detector material since they decay next to the interaction point. Hadronically-decaying τ leptons can be identified by looking at the resultant hadrons. However, leptonically-decaying τ leptons cannot be identified since light leptons coming from τ cannot be distinguised from light leptons coming from the hard-scattering.

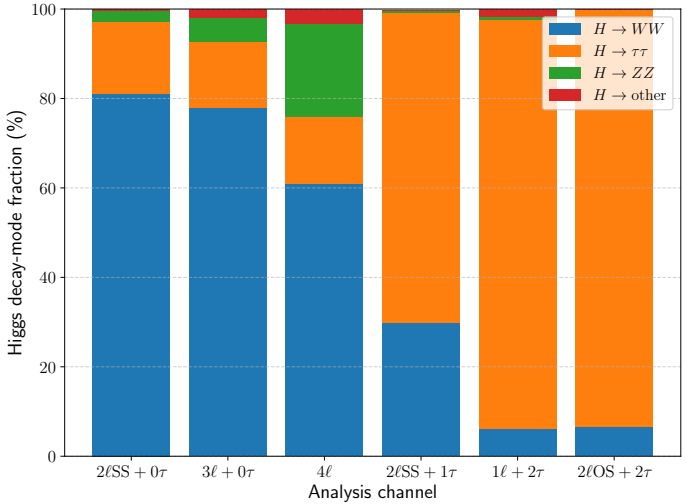
- We define several analysis channels depending on the number of selected light leptons ℓ (e,μ) and hadronically-decaying τ leptons (τ_{had}).





Example with 3 final-state leptons

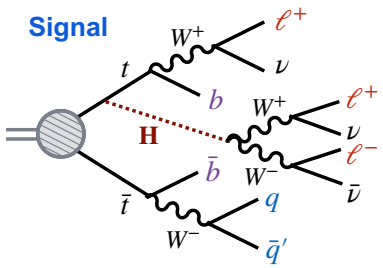




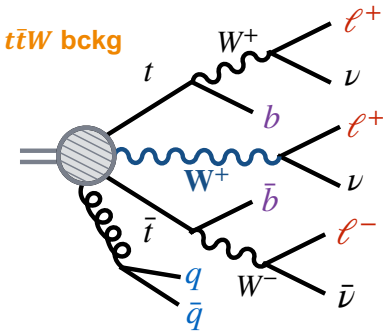
Event selection

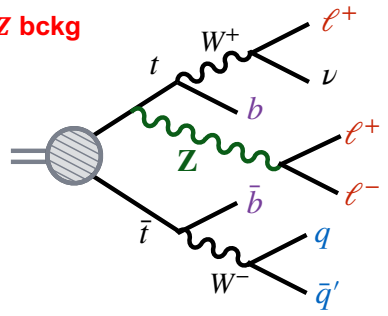
- My work mainly covered the analysis of the most sensitive final state i.e. the $\mathbf{3\ell + 0\tau_{had}}$ **final state**.

Event selection in the $3\ell + 0\tau_{had}$ channel

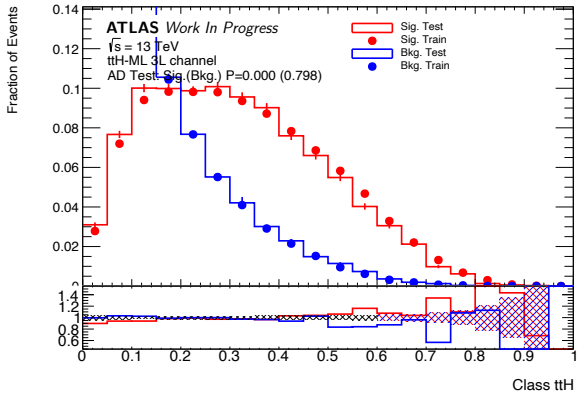


- 3 light leptons ℓ (e,μ)
- 0 hadronic τ leptons
- At least two jets
- At least one b-jet
- $|m_{\ell^+\ell^-} - m_Z| > 10$ GeV





- Use machine learning to build discriminants that separate signal from main backgrounds (summary statistics, remember that!?).
- A BDT is trained with 5 clases (signal and main backgrounds): $t\bar{t}H$ $t\bar{t}W$, $t\bar{t}Z$, WZ and $t\bar{t}$. This allows to define five BDT-output discriminants.



Regions definition

- The five BDT-output discriminants are used to split the available phase space into several regions.

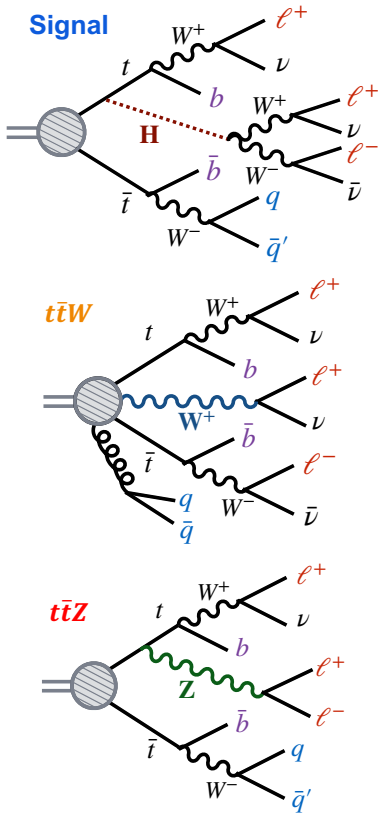
Regions	Selections
$t\bar{t}H$ SR	$t\bar{t}H > 0.2$.
tH SR	$tHjb > 0.25, t\bar{t}H < 0.2$.
$t\bar{t}W$ CR	$t\bar{t}W > 0.3, t\bar{t}H < 0.2, tHjb < 0.25$.
$t\bar{t}Z$ CR	$t\bar{t}Z > 0.45, t\bar{t}H < 0.2, tHjb < 0.25, t\bar{t}W < 0.3$.
VV CR	$VV > 0.65, t\bar{t}H < 0.2, tHjb < 0.25, t\bar{t}W < 0.3, t\bar{t}Z < 0.45$.
$t\bar{t}$ Region	$t\bar{t} > 0.25, t\bar{t}H < 0.2, tHjb < 0.25, t\bar{t}W < 0.3, t\bar{t}Z < 0.45, VV < 0.65$.
Other Region	$t\bar{t} < 0.25, t\bar{t}H < 0.2, tHjb < 0.25, t\bar{t}W < 0.3, t\bar{t}Z < 0.45, VV < 0.65$.

- The BDT does an amazing job separating signal from background events → it allows to define a much more sensitive phase space to measure $t\bar{t}H$ events: the $t\bar{t}H$ signal region (SR).

- Apart from the six regions defined using the BDT output scores, we define additional control regions to constrain the main backgrounds.

Expected number of events (SM)

	Pre-selection	$t\bar{t}H$ SR
$t\bar{t}H$	83 ± 9	56 ± 6
$t\bar{t}W$	200 ± 23	59 ± 6
$t\bar{t}Z/\gamma$	179 ± 6	76 ± 3
WZ	119 ± 9	8.7 ± 1.0
WW/ZZ	30 ± 7	1.8 ± 0.6
Fake-lepton (int. γ-conv.)	9 ± 5	2.4 ± 1.3
Fake-lepton (mat. γ-conv.)	15 ± 4	2.4 ± 0.4
Fake-lepton (HF μ)	25 ± 7	4.4 ± 3.0
Fake-lepton (HF e)	30 ± 11	5.3 ± 1.9
tZ	32.9 ± 3.3	5.4 ± 0.6
WtZ	16 ± 8	3.8 ± 1.9
$t\bar{t}t\bar{t}$	12 ± 5	5.0 ± 2.1
$t\bar{t}t$	1.5 ± 0.5	0.66 ± 0.23
$t\bar{t}WW$	10 ± 5	4.7 ± 2.3
VVV	3.9 ± 1.2	0.57 ± 0.19
VH	7.1 ± 2.3	2.1 ± 1.2
tHq	1.70 ± 0.27	0.40 ± 0.07
tWH	2.95 ± 0.32	1.50 ± 0.17
Total	**780 ± 40**	**241 ± 13**
$S/\sqrt{B}$	3.14	4.12



Example Feynman diagrams in the 3ℓ channel for the signal (ttH) and most-relevant background (ttW and ttZ) processes.

Input bins

unconstrained parameters, e.g. POI

prediction (summed over samples)

constraint term (e.g. Gaussian)

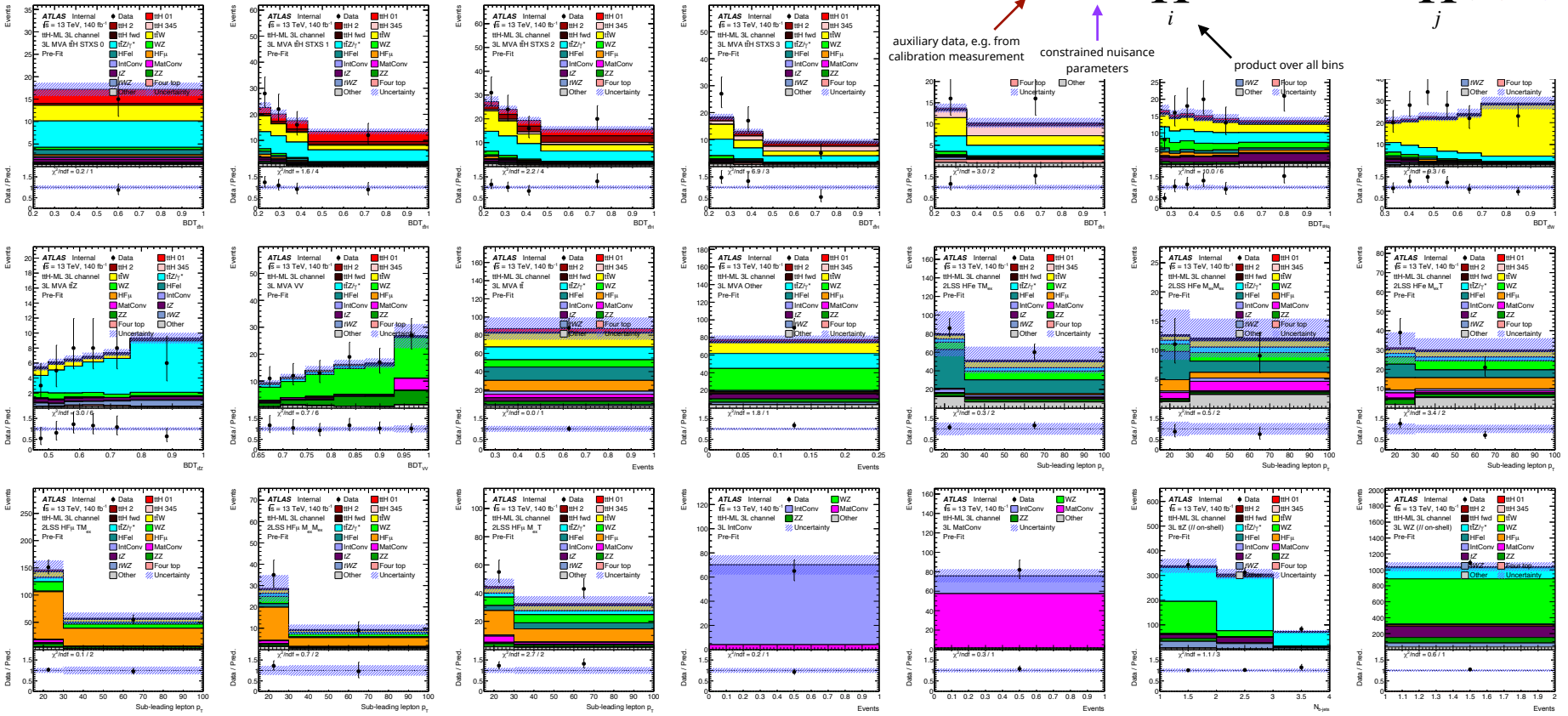
observed data

$$p(\vec{n}, \vec{a} \mid \vec{k}, \vec{\theta}) = \prod_i \text{Pois}(n_i \mid \nu_i(\vec{k}, \vec{\theta})) \cdot \prod_j c_j(a_j \mid \theta_j)$$

auxiliary data, e.g. from calibration measurement

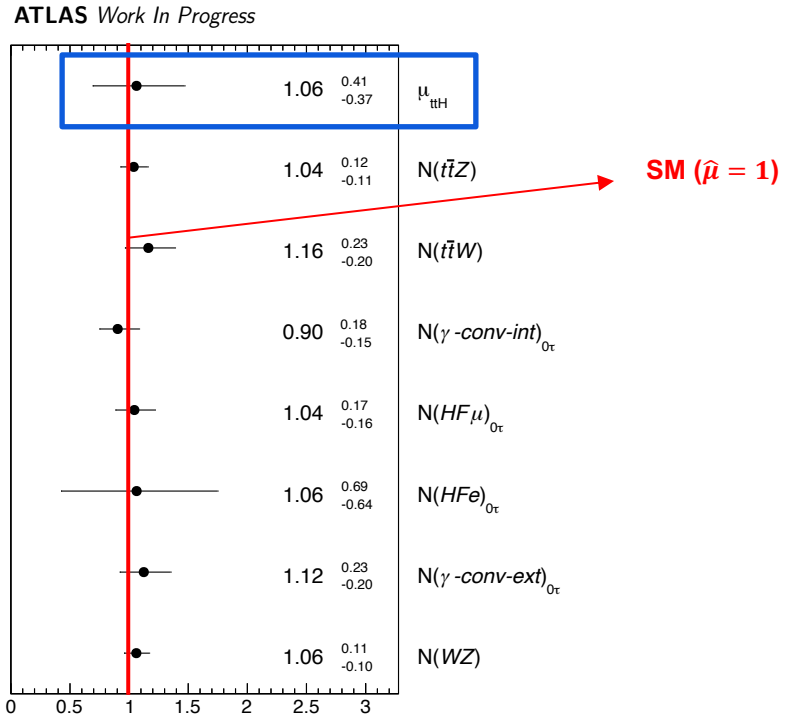
constrained nuisance parameters

product over all bins

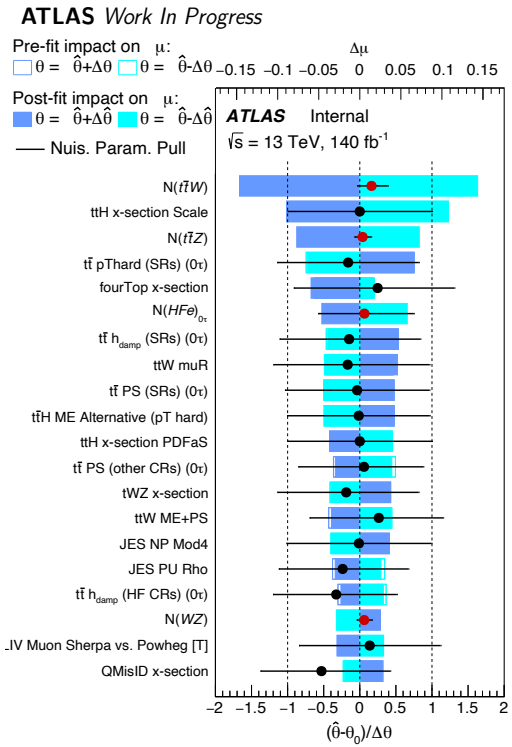


Results

- We found that the **observed (expected) significance** for a $t\bar{t}H$ excess in the $3\ell + 0\tau_{had}$ channel is $\mathbf{2.94\sigma}$ $\mathbf{(3.07\sigma)}$.



Most relevant nuisance parameters



Thanks for your attention :)

Some useful references

- Simulation-based inference methods for particle physics → https://arxiv.org/abs/2010.06439
- Particle Data Group (PDG) review on Statistics → https://pdg.lbl.gov/2024/reviews/rpp2024-rev-statistics.pdf
- HistFactory: A tool for creating statistical models for use with RooFit and RooStats → https://cds.cern.ch/record/1456844