Searching for HWW anomalous couplings with simulation-based inference

Marta Silva (IST/LIP), Inês Ochoa (LIP), Patricia Muíño (IST/LIP), Ricardo Barrué (IST/LIP)

Extended Higgs Sector subgroup meeting

November 19, 2024

TÉCNICO LISBOA
LIP
ATLAS EXPERIMENT
REPÚBLICA PORTUGUESA

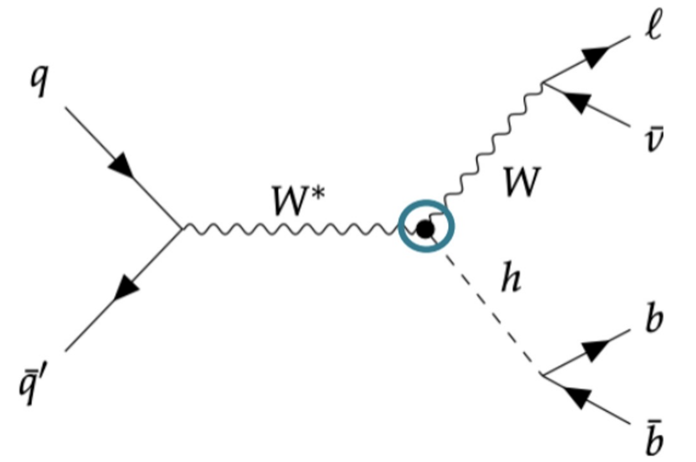
Motivation

One of the major questions left unaddressed by the SM is the **observed asymmetry between matter and antimatter in the Universe.**

Sources of **charge-parity (CP) violation beyond the SM (BSM)** are required to explain this puzzle

The presence of CP-odd components in the Higgs boson couplings is predicted by many BSM theories

Goal: search for CP violation in the **HWW interaction via leptonic WH production:**



Regarding the SM as a low-energy effective field theory (SMEFT):

$$\mathcal{L}_{\mathrm{SMEFT}} = \mathcal{L}_{\mathrm{SM}} + \frac{1}{\Lambda}\sum_{i=1}^{N_{d5}} c_i^{(5)}\mathcal{O}_i^{(5)} + \frac{1}{\Lambda^2}\sum_{i=1}^{N_{d6}} c_i^{(6)}\mathcal{O}_i^{(6)}$$

relevant for Higgs physics

c_i - **Wilson Coefficients;**

O_i - operators with the same SM symmetries

CP-odd operator:

$$\tilde{\mathcal{O}}_{WH} = \frac{c_{W\tilde{H}}}{\Lambda^2} H^{\dagger} H \tilde{W}^I_{\mu\nu} W^{I\mu\nu}$$

CP-even operator:

$$\mathcal{O}_{HW} = \frac{c_{HW}}{\Lambda^2} H^{\dagger} H W^I_{\mu\nu} W^{I\mu\nu}$$

Simulation-based inference

The ultimate goal of an EFT analysis is to establish exclusion limits on the parameters of interest θ.

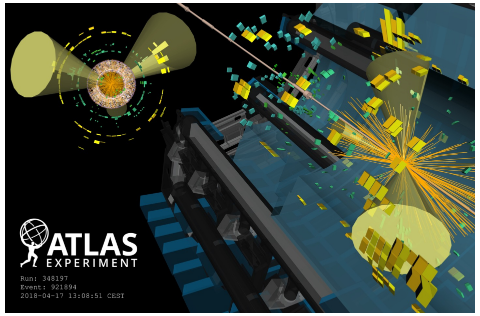
Need to construct the **Likelihood Function:** $p(x|\theta) = \boxed{\int dz_d \int dz_s \int dz_p} \; p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta)$

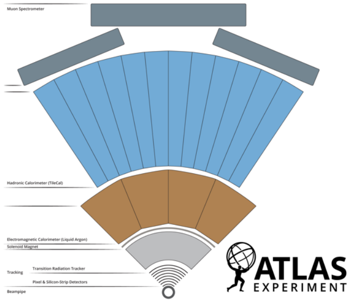
"How likely is an observation x described by the theory parameter θ "

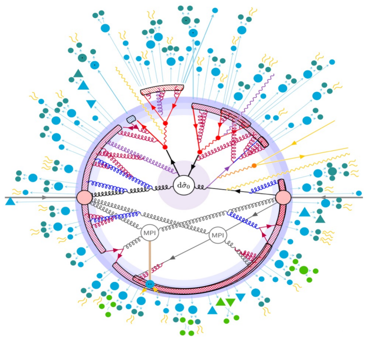
It's infeasible to calculate the integral over this enormous latent space

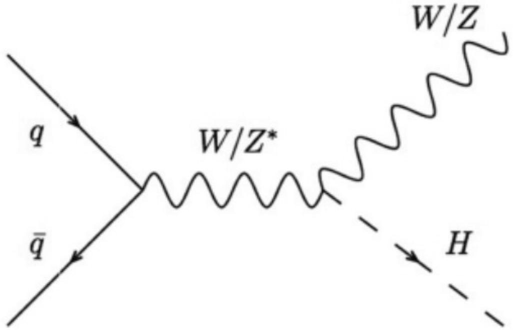
Leads to **likelihood-free** (LFI) or **simulation-based inference** (SBI) ← **Intractable likelihood**

Observables	Detector interactions	Shower splittings	Parton-level momenta	Theory parameters
x ←	z_d ←	z_s ←	z_p ←	θ









Monte Carlo (MC) simulation ←

Inference →

Addressing the likelihood intractability

arXiv:1805.12244

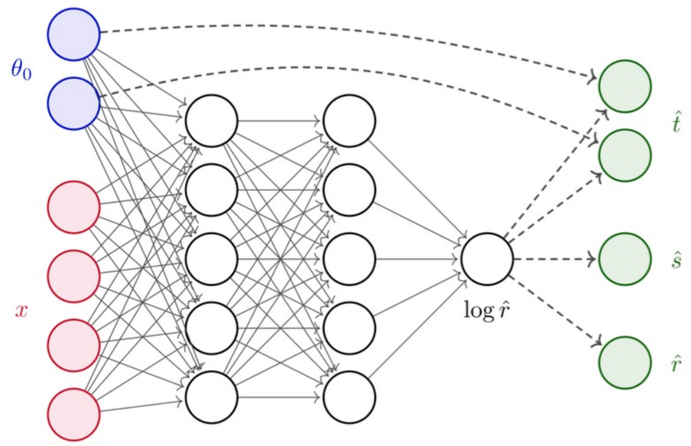
arXiv:1805.00020

Classical methods rely on using one or two observables as summary statistics or approximations of the shower and detector effects

We propose using **neural networks** to estimate:

- The **likelihood ratio**, $r(x)$
- A **locally optimal observable (score)**, $t(x)$

Using a trick known as "mining gold"



Data augmentation: additional information can be extracted from MC simulations and used to define loss functions that when minimized converge to the true likelihood ratio/score

Joint Likelihood ratio:

How much more likely is data to be described by θ_0 than θ_1

$$r(x,z|\theta_0,\theta_1) = \frac{p(x,z|\theta_0)}{p(x,z|\theta_1)} = \boxed{\frac{p(z_p|\theta_0)}{p(z_p|\theta_1)}}$$

Joint score:

Quantifies change of likelihood in parameter space

$$t(x,z|\theta) \equiv \nabla_\theta \log p(x,z|\theta) = \boxed{\frac{\nabla_\theta p(z_p|\theta)}{p(z_p|\theta)}}$$

Both quantities can be calculated by evaluating the matrix elements

SALLY vs ALICE(S)

arXiv:1805.12244

arXiv:1805.00020

SALLY

- **Goal:** learn a detector-level optimal observable **(score)** at the SM point

Requirements:

- Joint score
- Mean squared error loss function

A study of the SALLY sensitivity for $c_{H\tilde{W}}$ was published in JHEP04(2024)014 by R.Barrué (LIP)

Starting point for this study!

Problem: relies on the assumption that the parameter θ is close to the SM

VS

ALICE

- **Goal:** learn the likelihood ratio as a function of x and θ

Requirements:

- Joint likelihood ratio
- Improved cross-entropy loss function

ALICES

- **Goal:** learn the likelihood ratio as a function of x and θ

Requirements:

- Joint likelihood ratio
- Joint score
- Improved cross-entropy loss function

Analysis Overview

arXiv:1907.10621

1. Event Generation (MadGraph):

Signal samples: $\mathrm{WH}(lvbb)$; SMEFTsim3; $\Lambda = 1$ TeV

- LO reweighting + morphing technique to calculate event weights at any parameter point θ

Background samples: semileptonic $t\bar{t}$; single top s-channel; $W + (b)$-jets

2. Parton shower (Pythia8) and detector simulation (Delphes)

3. Chose observables/inputs for NN

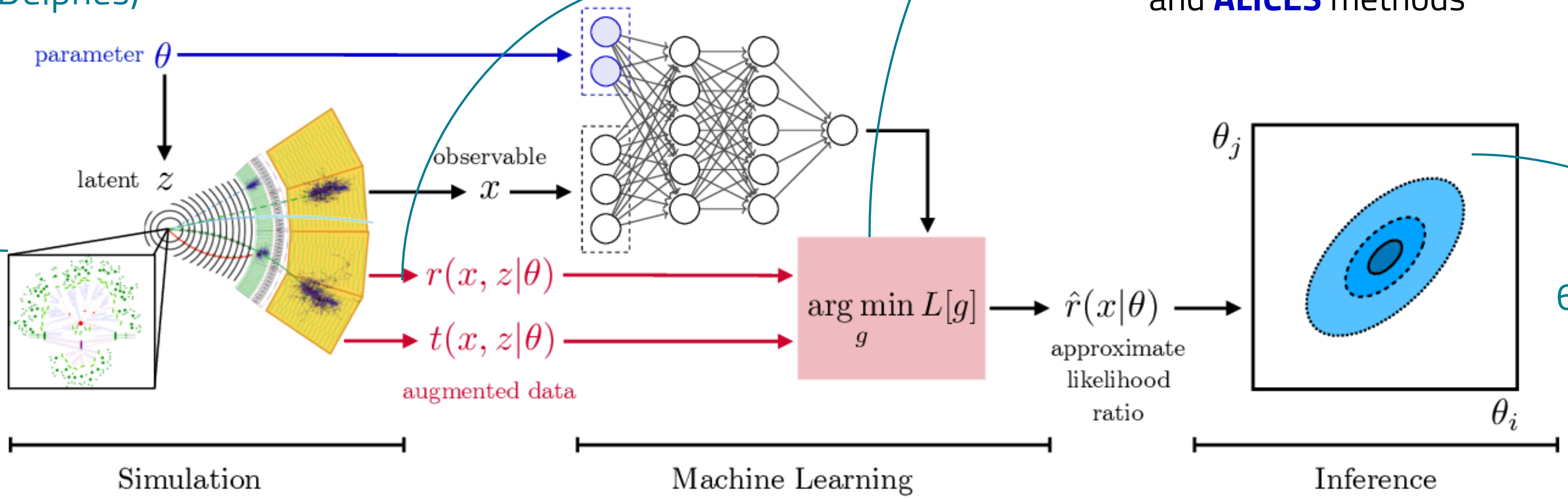
4. Data unweighting and augmentation

- Drawn from MC samples with probabilities proportional to the event weights

5. Train a Neural Network with a suitable loss function

- Ensembles of 5 NNs for the **SALLY**, **ALICE**, and **ALICES** methods

6. Set limits on θ



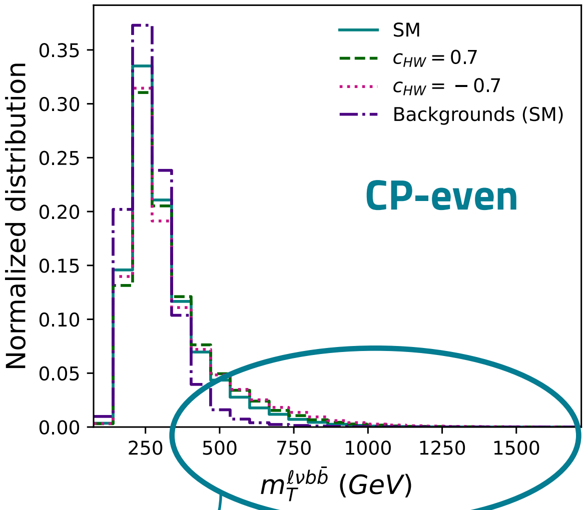
Energy-dependent and angular observables

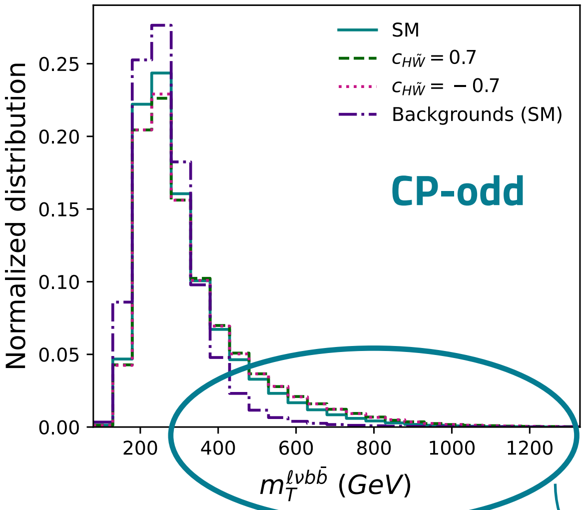
arXiv:1409.5449

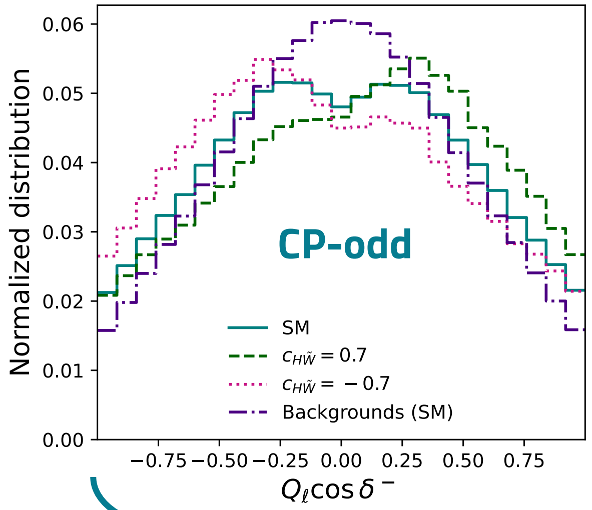
Not sensitive to sign of c_{HW} and $c_{H\widetilde{W}}$: changes in observables come mainly from **EFT quadratic terms**

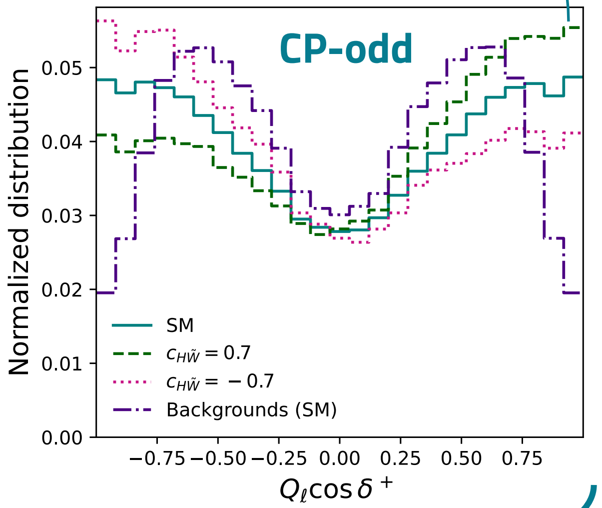
$$\cos\delta^+ = \frac{\vec{p}_\ell^{\,W} \cdot (\vec{p}_H \times \vec{p}_W)}{|\vec{p}_\ell^{\,W}||(\vec{p}_H \times \vec{p}_W)|}$$

$\vec{p}_\ell^{\,W}$: momentum of lepton in W boson rest frame









Sensitivity to non-zero c_{HW} and $c_{H\widetilde{W}}$: S/B increased in high $m_T^{\ell\nu b\bar{b}}$ (and p_T^W) regions w.r.t. SM

- **Symmetric for SM** signal and backgrounds, **asymmetric for $c_{H\widetilde{W}} \neq 0$**
- **Can extract sign** of $c_{H\widetilde{W}}$

EFT scenarios studied

Benchmarks for the ML methods:

- $Q_\ell \cos\delta^+$, $m_T^{\ell\nu b\bar{b}}$, $Q_\ell \cos\delta^+ \otimes m_T^{\ell\nu b\bar{b}}$

- **1D studies:** $c_{H\widetilde{W}}$ and c_{HW} independently constrained

- **2D studies:** both coefficients were used as inputs to the NNs

Training datasets

Signal Only (SO) | Signal + Backgrounds (S+B)

Input variables to the Neural Networks:

- E, p_x, p_y, p_z of final state particles;
- p_T, η, θ, ϕ of final state particles;
- x and y componentes + absolute value of E_T^{miss}
- $\Delta\phi$ and ΔR between relevant objects
- m_{bb}
- $Q_\ell \cos\delta^+$ and $Q_\ell \cos\delta^-$
- p_z^ν

Kinematic Only (KO)

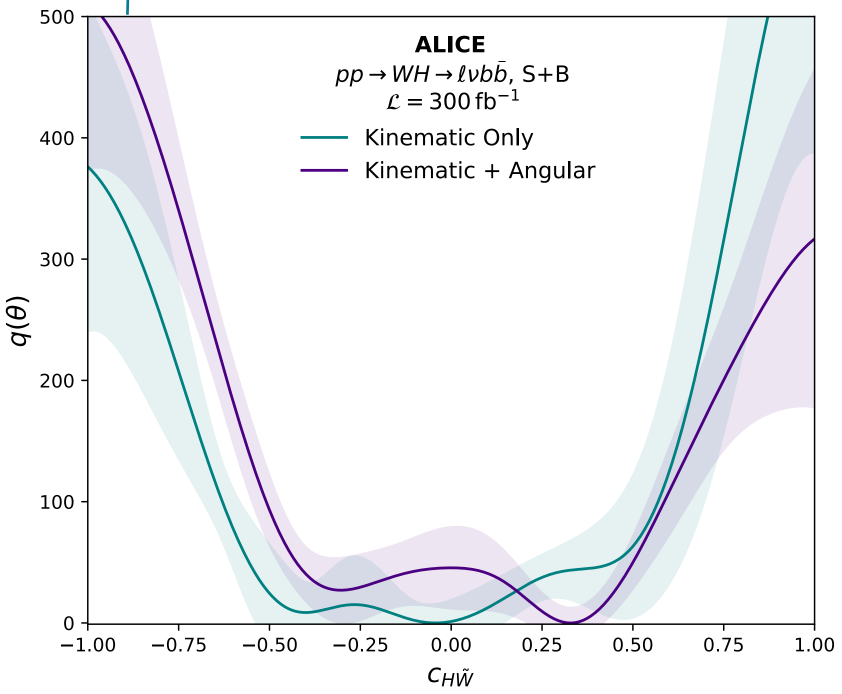
Kinematic + Angular (K+A)

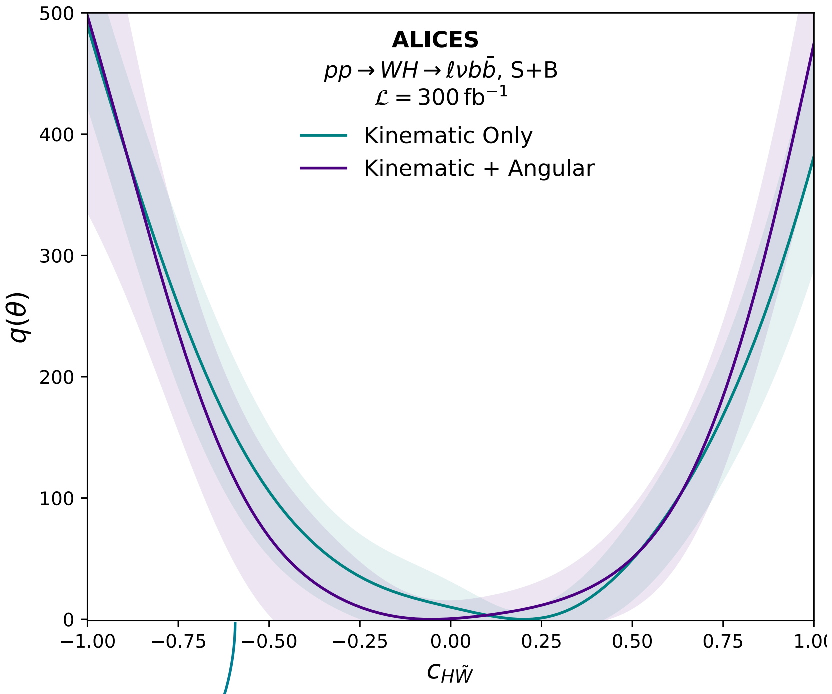
The effect of the angular observables

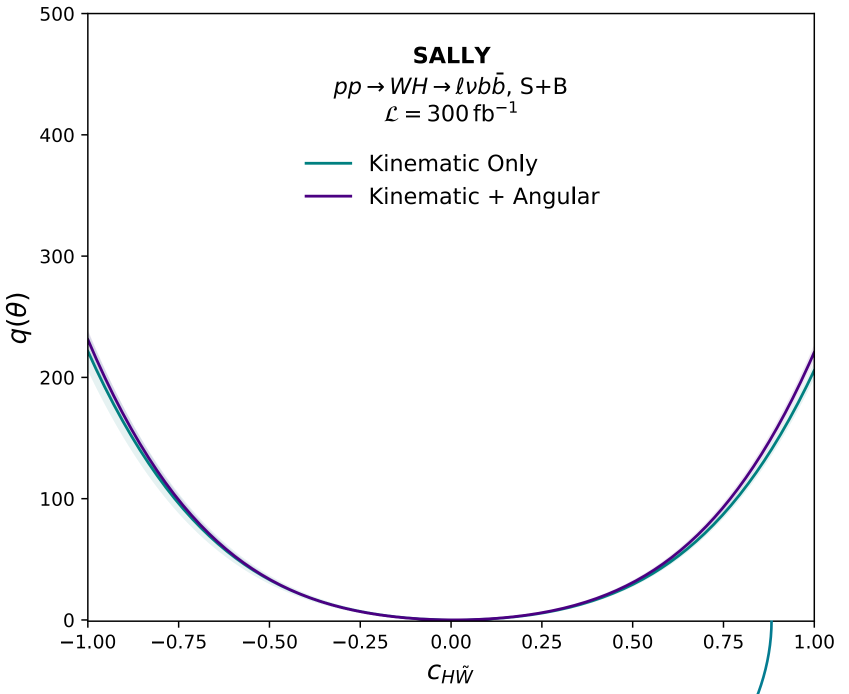
Overall, including the angular observable helped the estimation tasks

For both the SO and S+B datasets, the **ALICE** method **failed** to learn the likelihood function

Highlights the **importance** of having the additional **joint score information**







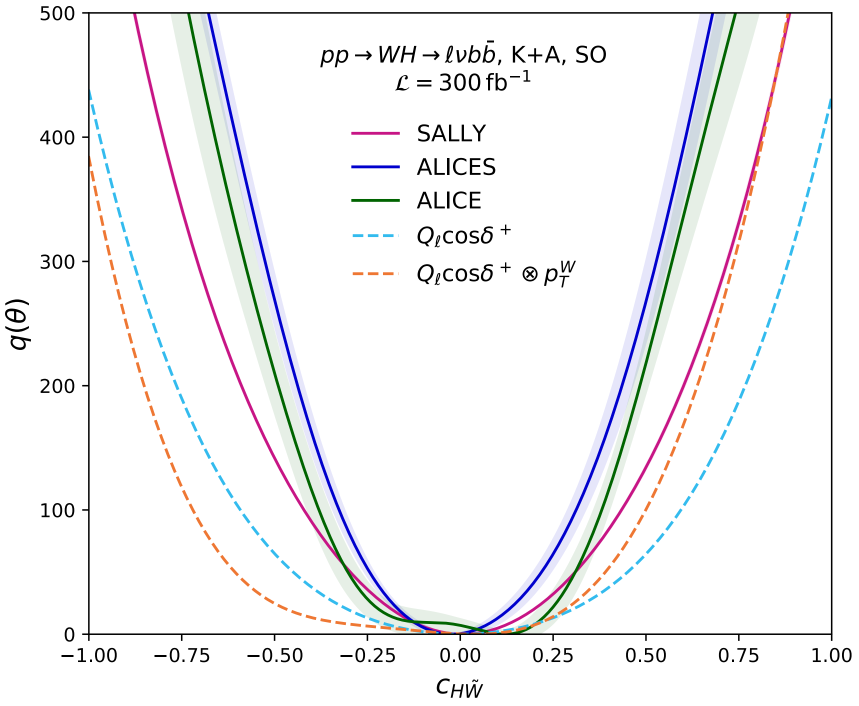
Better capturing the quadratic symmetry of the likelihood

MLE approached the true value + ↓ SD

1D studies: CP-odd results

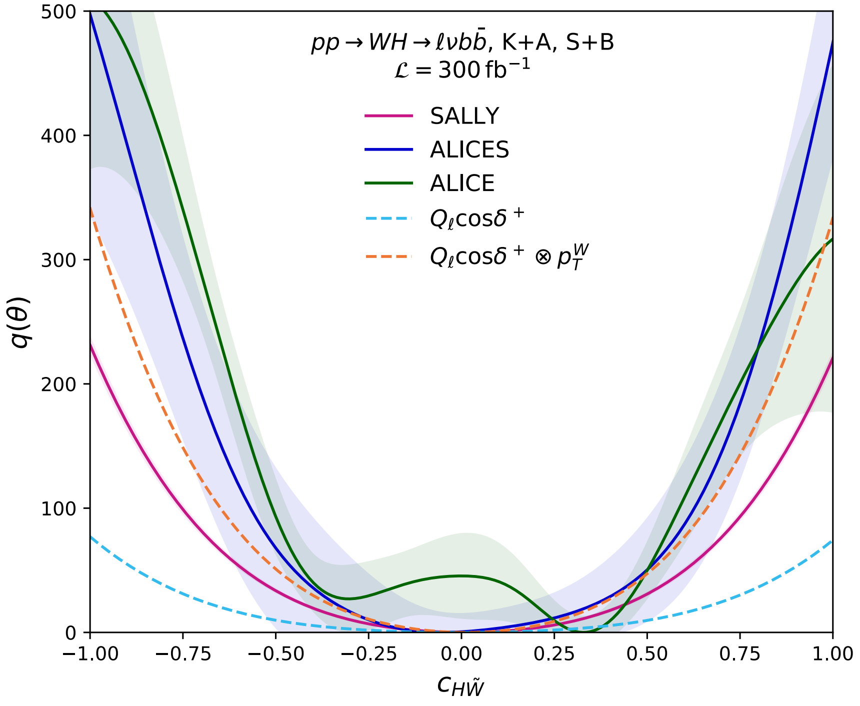
- **SALLY** and **ALICES** <u>outperform</u> the $Q_\ell \cos\delta^+$ **histogram**
- **ALICES** provides <u>tighter limits</u> than **SALLY** ⟹ **Trade-off:** ↑ variance + MLEs sometimes ≠ SM

Signal Only



ALICES > **SALLY** > $Q_\ell \cos\delta^+ \otimes m_T^{\ell\nu b\bar{b}}$

Signal + Backgrounds



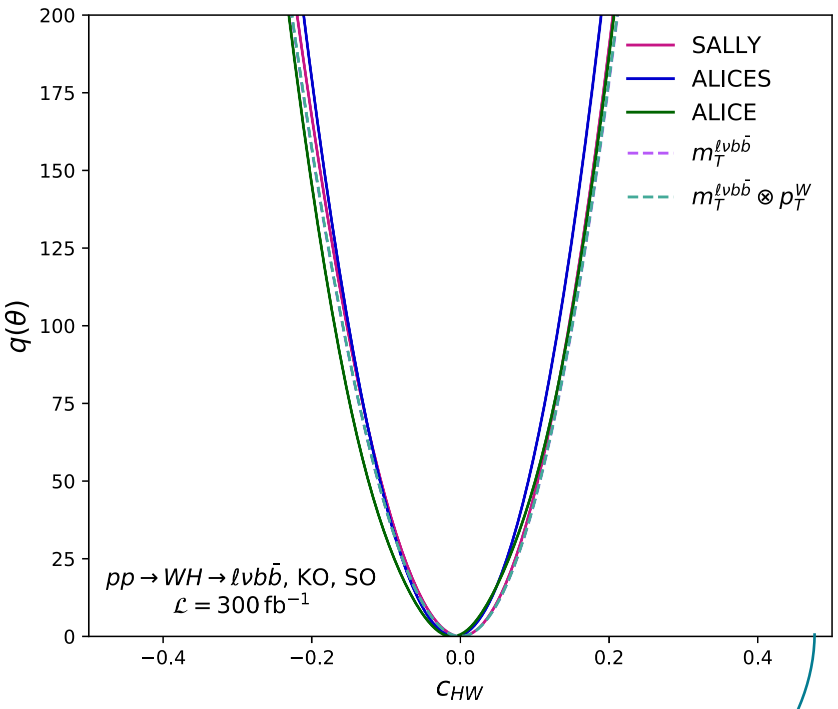
ALICES > $Q_\ell \cos\delta^+ \otimes m_T^{\ell\nu b\bar{b}}$ > **SALLY**

1D studies: CP-even results

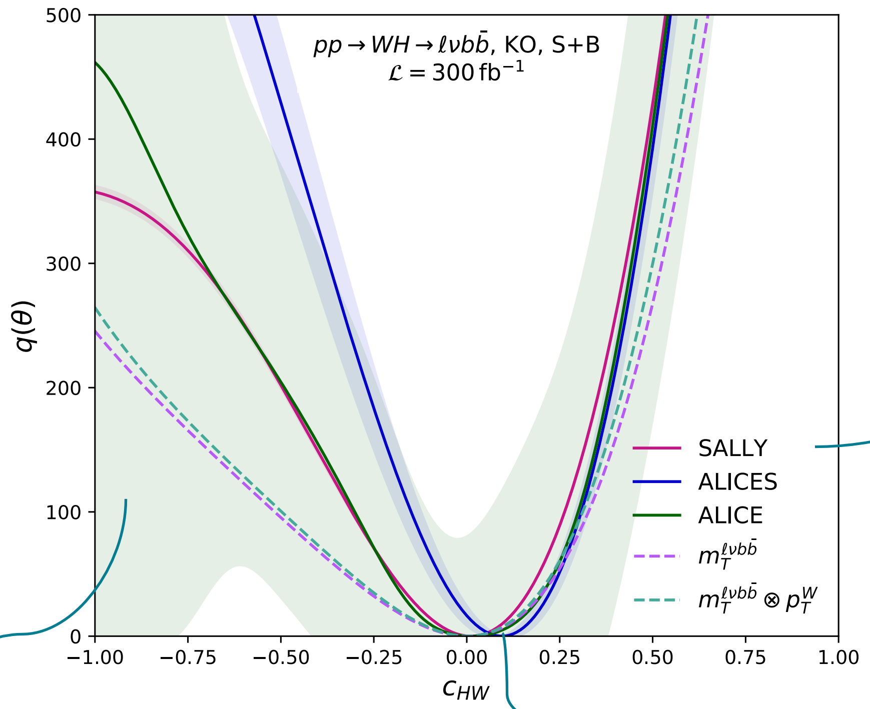
The SALLY method yields the most reliable outcomes

No significant advantage in using ML-methods over summary statistics

Signal Only



Signal + Backgrounds



SALLY > $m_T^{\ell\nu b\bar{b}}$

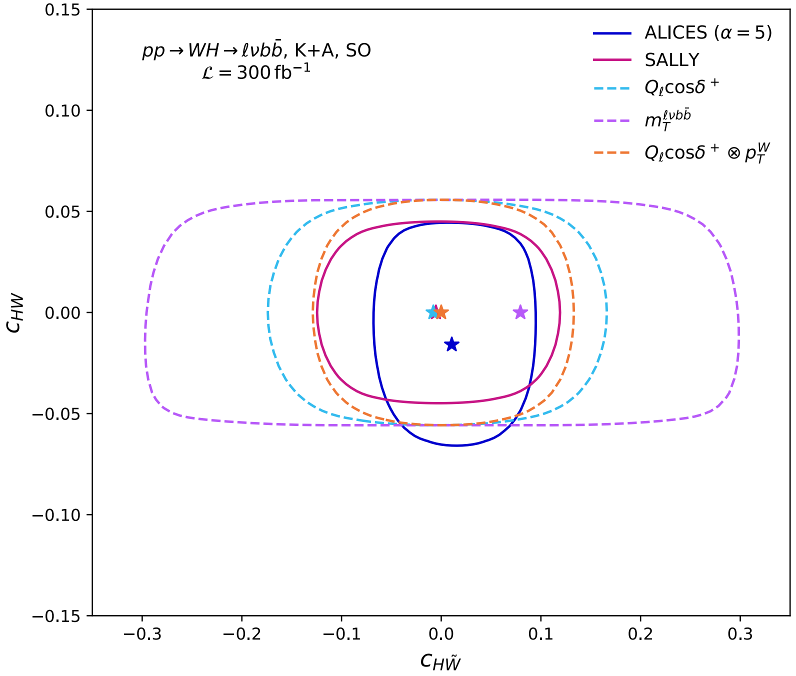
The rate term of the llr is the primary contributor to the results

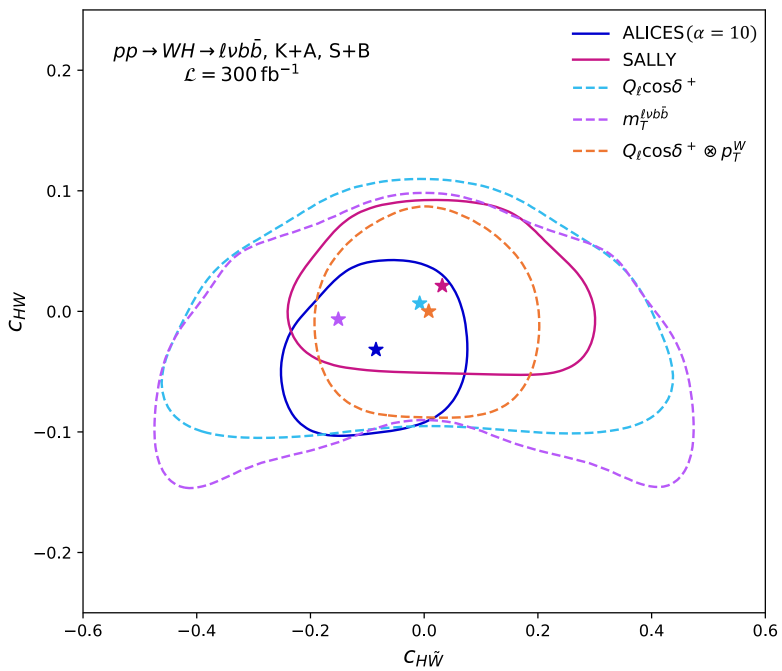
ALICE: ↑↑↑ variance

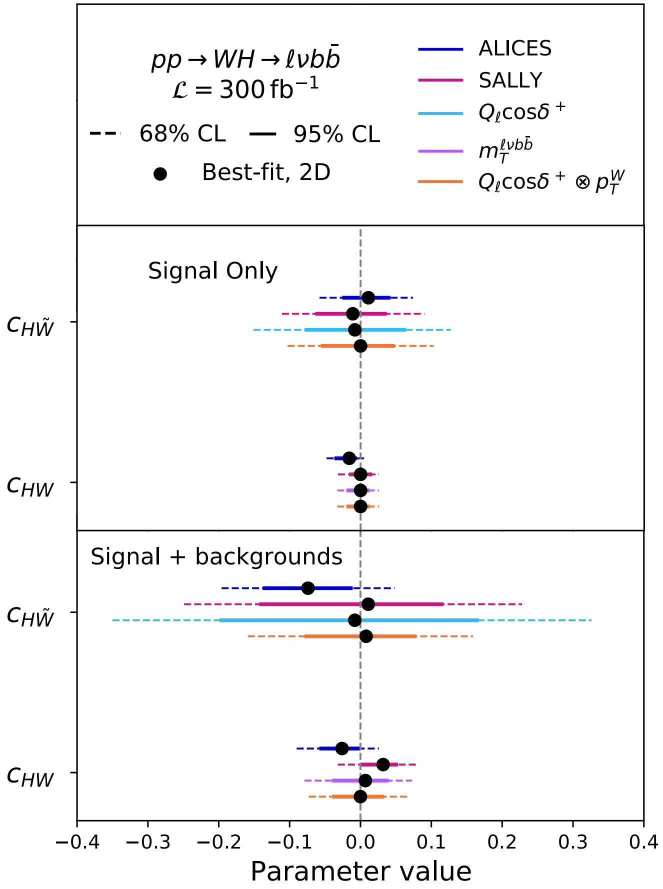
ALICES: MLE ≠ SM

2D studies: results

- Once again, for both datasets, the **ML-methods yielded tighter limits than** the best ones obtained with a **1D summary statistic.**
- The results from the **SALLY** method were **similar** to those from the **2D histogram**...
  - ... but SALLY can probe many couplings simultaneously!
- Contrary to what was expected from the literature, the **ALICES** method did not trivially outperform **SALLY** ☹
  - However, addressing **ALICES** difficulties could potentially take us beyond the sensitivity of both **SALLY** and **2D histograms**







Major Challenges

Is Physics An Obstacle?

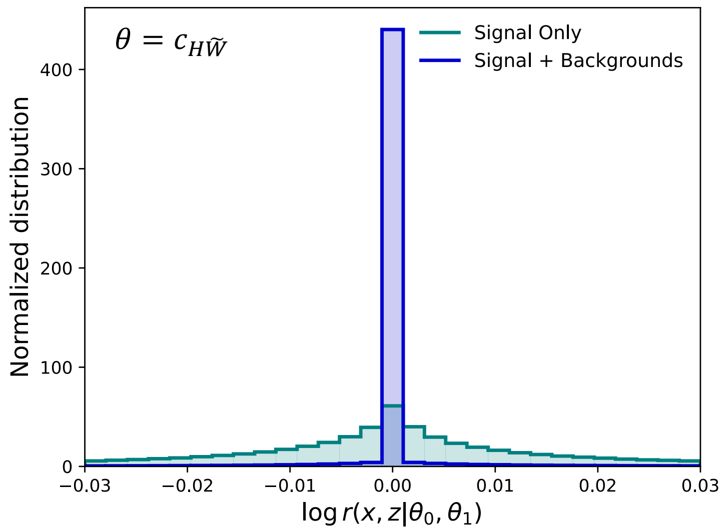
- **ALICES** learns the correct marginalization of $r(x, z|\theta_0, \theta_1)$

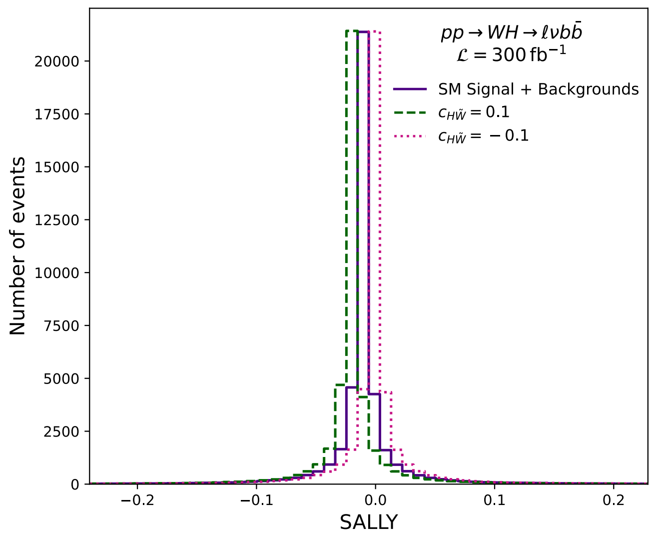
This quantity takes extremely small values

The SM and BSM points are very kinematically alike

Any bounds derived from the joint likelihood ratio are **highly sensitive to minor variations** in the NN output

SALLY learns an optimal observable near the SM and can capture the differences between SM and BSM points, **even for small values of $c_{H\widetilde{W}}$ and c_{HW}**





The Demand for More Powerful Computational Resources

The computational resources required to implement the ML–based inference techniques are a **significant limitation**

The **ALICES** sampling step was **one of the biggest obstacles** throughout this work

Training and evaluating these methods require substantial **resources** and **time**

Conclusion

- As Run 3 advances, the application of ML-based inference methods, such as ALICES and SALLY, **are very promising** in probing HWW anomalous couplings with higher sensitivity and precision.
  - These techniques offer the potential to **improve upon the traditional methods and current results from the ATLAS and CMS collaborations.**
- The advantages of these techniques come with the **trade-off of increased complexity** and **resource demands.**
  - Large amounts of training data are needed to effectively train Neural Networks.
  - Converging to the true likelihood ratio can be difficult when BSM signals are similar to SM ones.
- This work highlights the importance of **addressing the shortcomings of these techniques** (e.g. training stability and computational efficiency) to fully realize their potential.

Future Work

- Repeating this study in a higher p_T region

    - Increased sensitivity to BSM couplings
    - Increased signal-to-background ratio

- Training and evaluating these methods are very computationally demanding

    We are applying for special access to HPC+GPUs

Thanks!

Any questions?

Backup

Classical methods to constrain EFTs

Likelihood-free inference techniques

Traditional

Use one or two observables as **summary statistics** (e.g. invariant masses, angular observables, etc)

↓

Probability density function estimated by filling histograms.

↓

Problem: information loss

Matrix Element Based

Approximates the shower and detector effects by a transfer function (e.g. $p(x|z) = \delta(x - z)$)

↓

Examples: Matrix Element Method and Optimal Observables

↓

Problem: numerically expensive and not realistic

Machine Learning

Use Neural Networks to estimate the likelihood or likelihood ratio

↓

- By using ML we do not rely on simplifications
- Using the joint quantities also allows for training more precise estimators with less training data and computational resources.

Morphing Technique

Generating samples for each possible parameter θ is extremely time-consuming and impractical.

Solution: Morphing technique to calculate event weights at any parameter point

The morphing tecnhique relies on the fact that the matrix element squared is a polynomial function of the theory paramater (the Wilson coefficient):

$$|\mathcal{M}|^2 = \sum_{k=0}^{3} c_k A_k = \vec{c} \cdot \vec{A}, \quad \text{with} \quad \vec{c} = \left\{1, c, c^2\right\} \quad \text{and} \quad \vec{A} = (A_0, A_1, A_2)$$

Example: Measurement of a single BSM parameter

$$|\mathcal{M}|^2 (z_p \mid \theta) = \underbrace{1}_{w_0(\theta)} \underbrace{|\mathcal{M}_{SM}|^2 (z_p)}_{f_0(z_p)} + \underbrace{\theta}_{w_1(\theta)} \underbrace{2 \operatorname{Re} \mathcal{M}_{SM}^{\dagger} (z_p) \mathcal{M}_{BSM} (z_p)}_{f_1(z_p)} + \underbrace{\theta^2}_{w_2(\theta)} \underbrace{|\mathcal{M}_{BSM}|^2 (z_p)}_{f_2(z_p)}$$

By simulating samples from different values of c, one can write a vector of squared matrix elements $|M|^2_{simulated}$ which depend on the coupling values c via a morphing matrix C:

$$|\mathcal{M}|^2_{\text{simulated}} = C \times \vec{A}, \text{ with } C = \left[\vec{c}_{\text{Sample 1}}, \vec{c}_{\text{Sample 2}}, \ldots\right]$$

If the number of simulated samples is equivalent to the dimensionality of $\vec{A}$, the above relation can be inverted and one can calculate the matrix element for any coupling value as a linear combination of the previously simulated matrix elements:

$$|\mathcal{M}|^2 = \vec{c} \cdot \vec{A} = \vec{c} \cdot \left(\vec{C}^{-1} \cdot |\vec{\mathcal{M}}|^2{}_{\text{simulated}}\right) = \sum_j \underbrace{\left(\vec{C}_j^{-1} \cdot \vec{c}\right)}_{w_j(\vec{c})} |\mathcal{M}|_j^2$$

Morphing weights that can be used to interpolate to any parameter point

Additional formulas

Integrated luminosity

Full likelihood function: $p_{\text{full}}(x|\theta) = \text{Pois}(n|L\sigma(\theta)) \prod_i p(x_i|\theta)$, where $Pois(n|\lambda) = \lambda^n e^{-\lambda}/n!$

observed number of events Cross section

Joint Likelihood ratio:

$$r(x,z|\theta_0,\theta_1) \equiv \frac{p(x,z|\theta_0)}{p(x,z|\theta_1)} = \frac{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta_0)}{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta_1)}$$

$$= \frac{p(z_p|\theta_1)}{p(z_p|\theta_0)} = \frac{d\sigma(z_p|\theta_0)}{d\sigma(z_p|\theta_1)} \frac{\sigma(\theta_1)}{\sigma(\theta_0)}$$

Joint score:

$$t(x,z|\theta) \equiv \nabla_\theta \log p(x,z|\theta) = \frac{p(x|z_d)p(z_d|z_s)p(z_s|z_p)\nabla_\theta p(z_p|\theta)}{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta)}$$

$$= \frac{\nabla_\theta d\sigma(z_p|\theta)}{d\sigma(z_p|\theta)} - \frac{\nabla_\theta \sigma(\theta)}{\sigma(\theta)}$$

Parton-level event weights:

$$d\sigma(z_p|\theta) = \frac{(2\pi)^4 f_1(x_1,Q^2) f_2(x_2,Q^2)}{8x_1x_2s} |\mathcal{M}|^2(z_p|\theta) d\Phi(z_p).$$

LO reweighting:

$$w_{\text{new}} = \frac{|\mathcal{M}_{\text{new}}|^2}{|\mathcal{M}_{\text{orig}}|^2} w_{\text{orig}}$$

Slide borrowed from Alexander Held1

The likelihood ratio trick

- Consider **samples** x_i simulated under both hypotheses with **labels** y_i

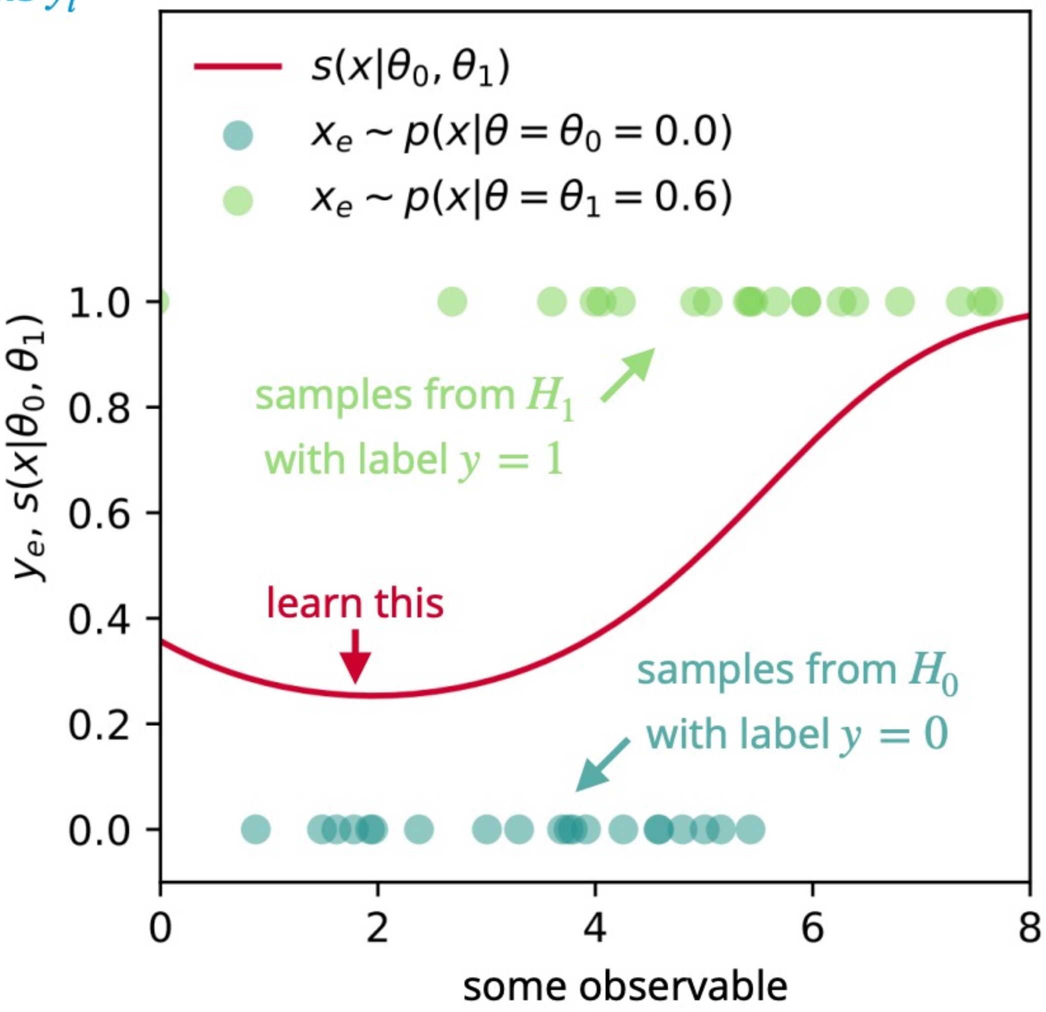
- **Find function** $s(x)$ that minimizes binary cross-entropy
  - $L[s] = -\frac{1}{N}\sum_i (y_i \log s(x_i) + (1-y)\log(1-s(x_i)))$

- A solution to that is $s(x \mid \theta_0, \theta_1) = \dfrac{p(x \mid \theta_1)}{p(x \mid \theta_0) + p(x \mid \theta_1)}$

- We can **find this function** with **standard ML** methods

- Then: $r(x \mid \theta_1) = \dfrac{p(x \mid \theta_1)}{p(x \mid \theta_0)} = \dfrac{s(x \mid \theta_0, \theta_1)}{1 - s(x \mid \theta_0, \theta_1)}$

- We can **learn the optimal observable** with ML
  - without ever knowing $p(x \mid \theta_i)$ directly (!)



arXiv:1805.00020

SALLY (Score Approximates Likelihood LocallY)

Score Estimator (SALLY):

- **Goal:** learn score as a function of x at θ_{SM}

- Uses joint score $t(x, z|\theta_{ref})$:

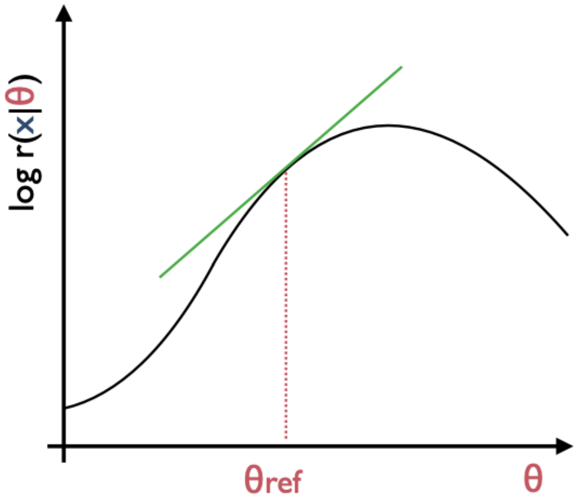
$$\text{NN} : x \to \hat{t}(x) \approx \nabla_\theta \log(x|\theta)|_{\theta_{SM}}$$

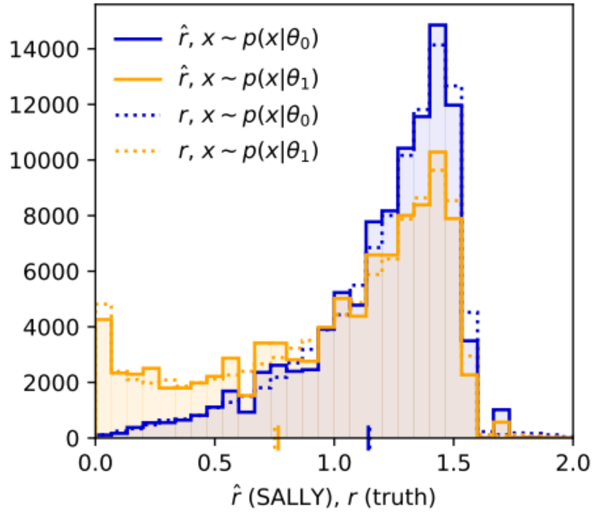
With the mean squared error (MSE) loss function

$$L_{\text{SALLY}}[\hat{s}(x|\theta_0, \theta_1)] = -\frac{1}{N} \sum_{(x_i,z_i)\sim p(x_i,z_i)} \left| t(x,z|\theta_0) - \hat{t}(x|\theta_0) \right|^2$$

Close to the Standard Model:

- The score is the sufficient statistics.
- Knowing $t(x)|_{\theta_{SM}}$ is as powerful as knowing $r(x|\theta)$.
- SALLY is a machine-learning version of an Optimal Observable.
- Can be used to fill histograms for different hypotheses and calculate likelihood ratios from them.





ALICES (Approximate Likelihood with Improved Cross-entropy Estimator and Score)

arXiv:1808.00973

Likelihood Ratio Estimator (ALICES):

- **Goal:** learn likelihood ratio as a function of x and θ

- Uses joint likelihood ratio $r(x, z|\theta)$ and joint score $t(x, z|\theta)$:

$$\text{NN} : (x, \theta) \rightarrow \hat{r}(x|\theta) \approx p(\text{x}|\theta)/p(x|\theta_{SM})$$

With the improved cross-entropy loss function

ALICE

$$L_{\text{ALICES}}[\hat{s}(x|\theta_0,\theta_1)] = \boxed{-\frac{1}{N}\sum_{(x_i,z_i)\sim p(x_i,z_i)} \Big[s(x_i,z_i|\theta_0,\theta_1)\log(\hat{s}(x_i)) + (1-\underline{s(x_i,z_i|\theta_0,\theta_1)})\log(1-\hat{s}(x_i))}$$

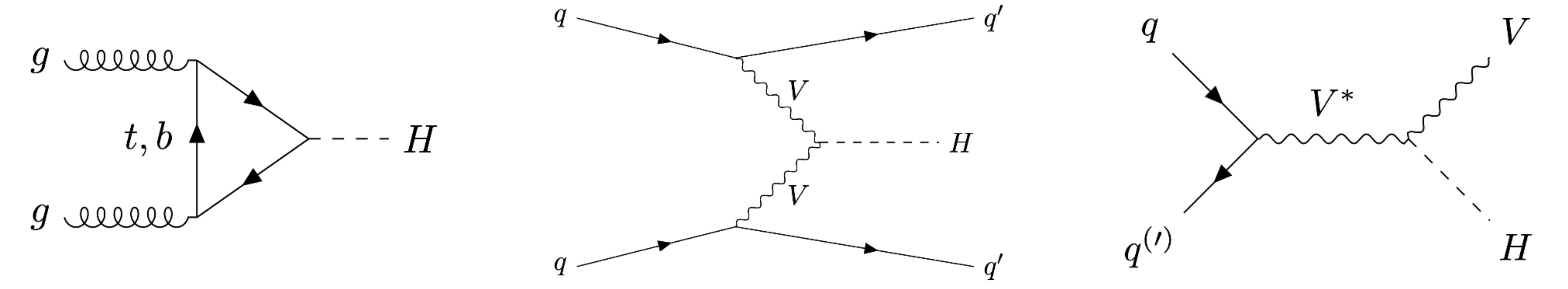
$$+\alpha(1-y_i)\left|t(x_i,z_i|\theta_0,\theta_1) - \nabla_\theta \log\left(\frac{1-\hat{s}(x_i|\theta,\theta_1)}{\hat{s}(x_i|\theta,\theta_1)}\right)\Big|_{\theta_0}\right|^2\Bigg]$$

$$s(x,z|\theta_0,\theta_1) = \frac{p(x,z|\theta_1)}{p(x,z|\theta_0) + p(x,z|\theta_1)}$$

The ALICE/ALICES methods are expected to exhibit superior performance:

- They use the complete event information for reconstructing the likelihood ratio
- Do not rely on the assumption that the parameter θ is close to the SM
- According to the literature, cross-entropy losses are expected to have lower variance and increased robustness to outliers compared to the standard cross-entropy loss or the Mean Squared Error loss.

HWW interaction vertex



(a) Gluon-gluon fusion (ggF) (b) Vector boson fusion (VBF) (c) Associated production (VH)

- VBF does not allow access to the HWW vertex independently of the HZZ vertex.

- Regarding the H → WW decay:

  - Exhibits lower sensitivity due to the Higgs boson being always on-shell.

  - The invariant mass of the WW system must always match the mass of the Higgs, constraining the energy transfer to particles in the final state.

  - Involves two neutrinos in the final state, posing a significant challenge.

Parton-level validation study

To validate these methods, we need to choose a setup in which we **can calculate the true likelihood ratio/score**

$$\boxed{\begin{aligned} r(x,z) &\approx r(x) \\ t(x,z) &\approx t(x) \end{aligned}}$$

Simplified process where all initial and final state flavors are specified:

Need access to the matrix element information

$$\boxed{u\bar{d} \to W^+h \to \mu^+\nu_\mu b\bar{b}}$$

10^6 events generated at the SM point + <u>200k events generated at the other 2 benchmarks (BSM points)</u>

Reweighting to far-away points in parameter space can lead to large event weights and thus large statistical fluctuations

Monte Carlo samples

- $\mathrm{WH}(l\nu bb)$ signal samples; MadGraph and SMEFTsim3; $\Lambda = 1$ TeV
  - Signal events generated at $(c_{H\widetilde{W}}, c_{HW}) = (0{,}0)$ and reweighted to obtain event weights for benchmark points
  - Maximum range used in the morphing basis optimization: $|c_{H\widetilde{W}}| \leq 1.2$; $|c_{HW}| \leq 1.0$

- $\frac{1}{5}$ of the signal samples were directly generated at the benchmark points to mitigate large statistical fluctuations that can arise from reweighting events to distance points in parameter space

- No reweighting or morphing was applied to the background samples

Optimized morphing basis points

	Validation	1D (CP-odd)	1D (CP-even)	2D	
Coefficient	$c_{H\tilde{W}}$	$c_{H\tilde{W}}$	c_{HW}	$c_{H\tilde{W}}$	c_{HW}
Benchmark 1	1.150	1.150	0.940	-0.902	0.420
Benchmark 2	-1.035	-1.035	-0.972	-0.234	0.970
Benchmark 3	-	-	-	-1.120	-0.764
Benchmark 4	-	-	-	0.720	-0.873
Benchmark 5	-	-	-	1.150	0.630

Number of generated events

	Validation	1D (CP-odd)	1D (CP-even)	2D
SM Signal	1.0×10^6	4.0×10^6	4.0×10^6	8.0×10^6
Backgrounds	-	6.0×10^6	6.0×10^6	12.0×10^6
BSM Signal	0.2×10^6	1.6×10^6	1.6×10^6	8.0×10^6
Total	1.2×10^6	11.6×10^6	11.6×10^6	28.0×10^6

Event selection

Observable	Cut
Transverse momentum of lepton/light quarks (charm or lighter)	$p_{T,\ell}, p_{T,j} > 10$ GeV
Missing transverse energy	$E_T^{\text{miss}} > 25$ GeV
Transverse momentum of b-quarks	$p_{T,b} > 35$ GeV
Pseudorapidity of charged lepton, b-quarks and light quarks	$\lvert\eta_{\ell,b,j}\rvert < 2.5$
Angular distance between decay particles	$\Delta R_{bb,b\ell,bj,\ell j,jj} > 0.4$
Invariant mass of b-quark pair	$80 \text{ GeV} < m_{bb} < 160$ GeV
Transverse momenta of light quarks	$p_{T,j} < 30$ GeV

Generator-level cuts

Cut	Signal (SM)	$t\bar{t}$	$W+$ jets	Single top
$p_{T,\ell}, p_{T,j} > 10$ GeV	96.77	87.12	93.83	93.83
$E_T^{\text{miss}} > 25$ GeV	76.17	70.03	56.17	74.41
$p_{T,b} > 35$ GeV	50.05	52.08	1.91	50.6
$\lvert\eta_{\ell,b,j}\rvert < 2.5$	35.42	39.14	1.25	35.01
$\Delta R_{bb,b\ell,bj,\ell j,jj} > 0.4$	34.18	36.46	0.99	33.9
$80 \text{ GeV} < m_{bb} < 160$ GeV	34.31	13.2	0.46	11.39
$p_{T,j} < 30$ GeV	34.25	0.28	0.46	11.38

Cumulative efficiencies (in %)

ALICES sampling

SAMPLING SETUP

n_{samples} ; $\theta_0 \sim$ Gaussian or flat prior ; n_{θ_0} ; $\theta_1 =$ SM

Sampling from θ_0: $x \sim p(x|\theta_0)$

- $\frac{n_{\text{samples}}}{2}$ sampled from θ_0 with probabilities proportional to the event weights
- n_{θ_0} drawn from the prior
- Sampling $\frac{n_{\text{samples}}/2}{n_{\theta_0}}$ for each θ_0 drawn from the chosen prior
- The same values of θ_0 are used when calculating the joint likelihood ratio

Sampling from θ_1: $x \sim p(x|\theta_1)$

- $\frac{n_{\text{samples}}}{2}$ sampled from θ_1 with probabilities proportional to the event weights
- n_{θ_0} drawn from the prior solely to calculate the joint quantities

Calculate the joint quantities for every sampled event:

$$r(x,z|\theta_0,\theta_1) = \frac{p(x,z|\theta_0)}{p(x,z|\theta_1)} \quad t(x,z|\theta) = \nabla_\theta \log p(x,z|\theta)$$

Sampling Setup for each EFT scenario

	Validation	1D (CP-odd)		1D (CP-even)		2D	
		Signal Only	Signal + Backgrounds	Signal Only	Signal + Backgrounds	Signal Only	Signal + Backgrounds
n_{samples}	10^6	5×10^6	10^7	5×10^6	10^7	5.5×10^6	11.5×10^6
Prior	Gaussian ($\mu = 0$, $\sigma = 0.4$) or Uniform ($[-1.2, 1.2]$)	Gaussian ($\mu = 0$, $\sigma = 0.4$)		Gaussian ($\mu = 0$, $\sigma = 0.3$)		Gaussian ($\mu = 0$, $\sigma = 0.4$) and Gaussian ($\mu = 0$, $\sigma = 0.3$)	
n_{θ_0}	1000 and 10000	10000				10000 + 10000	

Training settings

- Training dataset defined as 80% of the total generated samples and further split into 75% for training and 25% for validation.

Architecture for SALLY, ALICE and ALICES (validation) and SALLY in the 1D/2D studies

	Validation	1D + 2D	
		Signal Only	Signal + Backgrounds
Hidden layers	[50]	[50]	[100 , 100]
Epochs	50	100	50
Activation	ReLU	ReLU	Tanh
Optimizer		AMSGrad	
Batch size		128	
Learning rate		$10^{-3} \to 10^{-4}$	

For ALICE and ALICES

2D study (S+B): $\alpha = 10$ (optimized)
Else: $\alpha = 5$

- Standardize inputs (zero mean + unit variance)
- Early stopping
- Ensemble of 5 NNs → makes the predictions more robust to different random seeds
- Different unweighted dataset for training each NN → ensemble variance reflects the uncertainty in the NN outputs due to finite training sample sizes

Setting Limits

We want to decide between two hypothesis: $H_0\colon \theta = \theta_{SM}$ and $H_1\colon \theta \neq \theta_{SM}$

The best test statistic to discriminate between two hypotesis is:

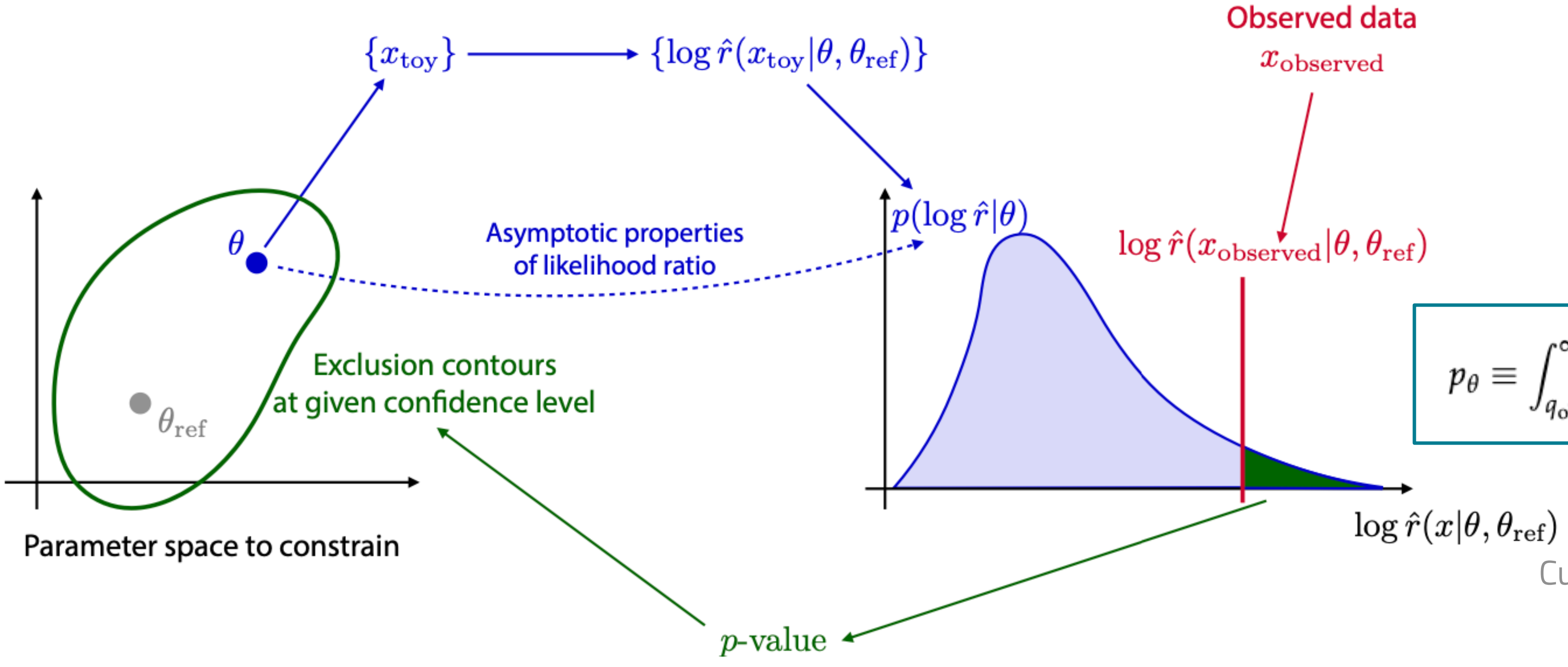
$$\boxed{q(\theta) = -2\sum_e \log r\left(x_e|\theta,\hat{\theta}\right)}$$

Maximum-likelihood estimator

This can be converted into a p-value:

$$\boxed{p_\theta \equiv \int_{q_{\mathrm{obs}(\theta)}}^{\infty} dq\, p(q|\theta) = 1 - F_{\chi^2}(q_{\mathrm{obs}(\theta)}|k),}$$

Cumulative distribution function of the chi-squared distribution

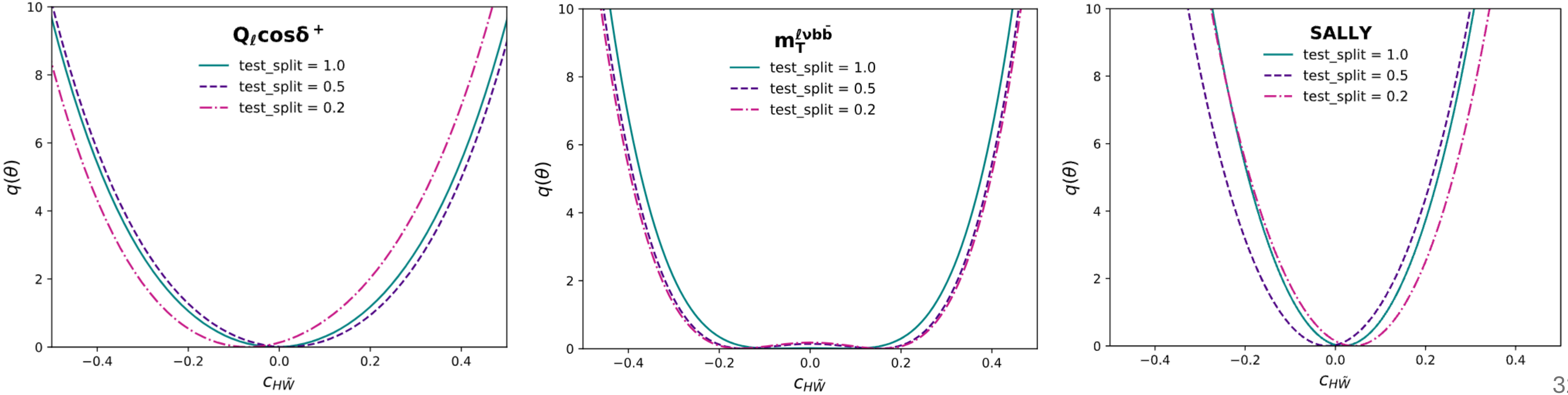


Represents the probability, assuming H_0, of observing data at least as extreme as predicted by H_0

Asymptotic limits in Madminer

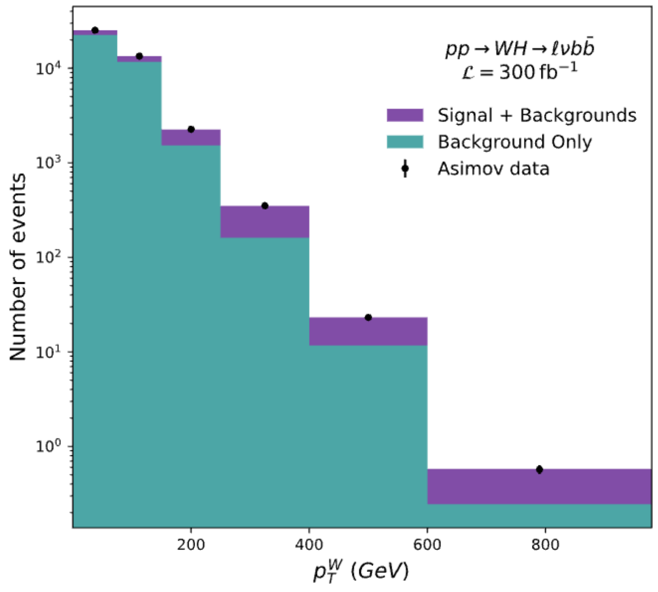
- p-values calculated using an Asimov dataset build from the test partition.
  - For ALICES, the NN is evaluated for multiple values of θ and the rate information is added
  - For SALLY, inference is performed similarly to histograms of summary statistics

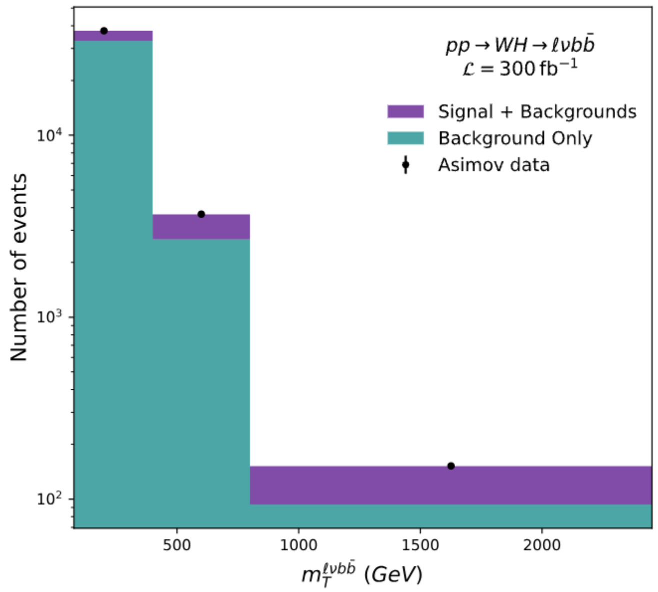
In Madminer these histograms are constructed from the training partition → Leads to fluctuations in the SM template compared to the Asimov histogram → We used the entire dataset to build the histograms (not ideal for SALLY)

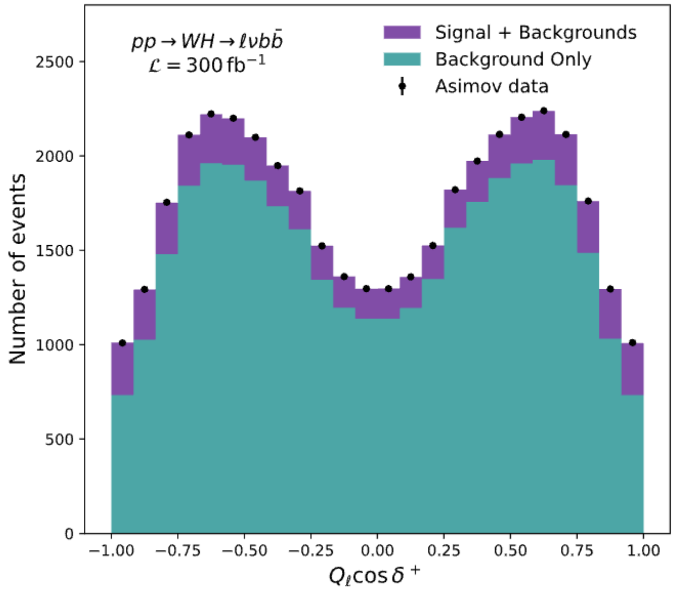


Binning & likelihood scans

- SALLY and $Q_\ell \cos\delta^+$ – 25 bins
- p_T^W (bins) $= (0-75, 75-150, 150-250, 250-400, 400-600, 600\ -\infty)$ GeV
- $m_T^{\ell\nu b\bar{b}}$ (bins) $= (0-400,\ 400-800,\ 800\ -\infty)$ GeV
- $c_{H\widetilde{W}}$ scanned over [-1.2,1.2] and c_{HW} over [-1.0,1.0] using 303 points across these ranges. For the 2D studies, 35 points were considered in each direction.
- Likelihood fits interpolated using spline functions

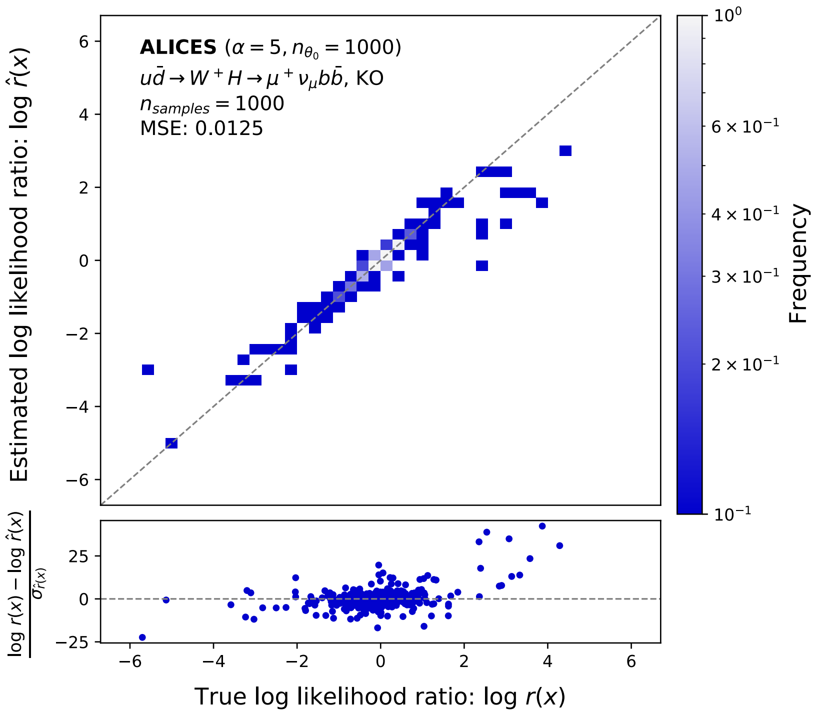


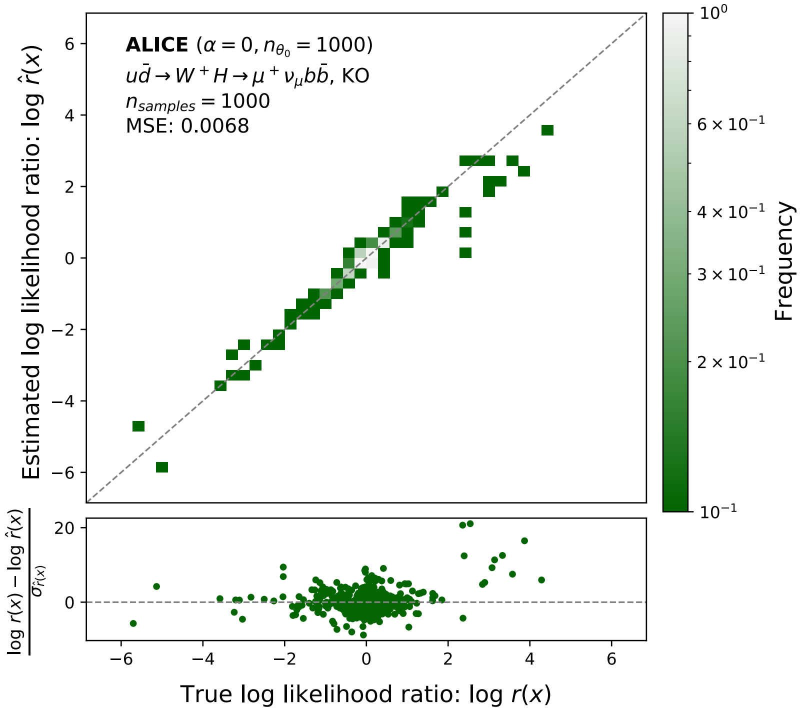


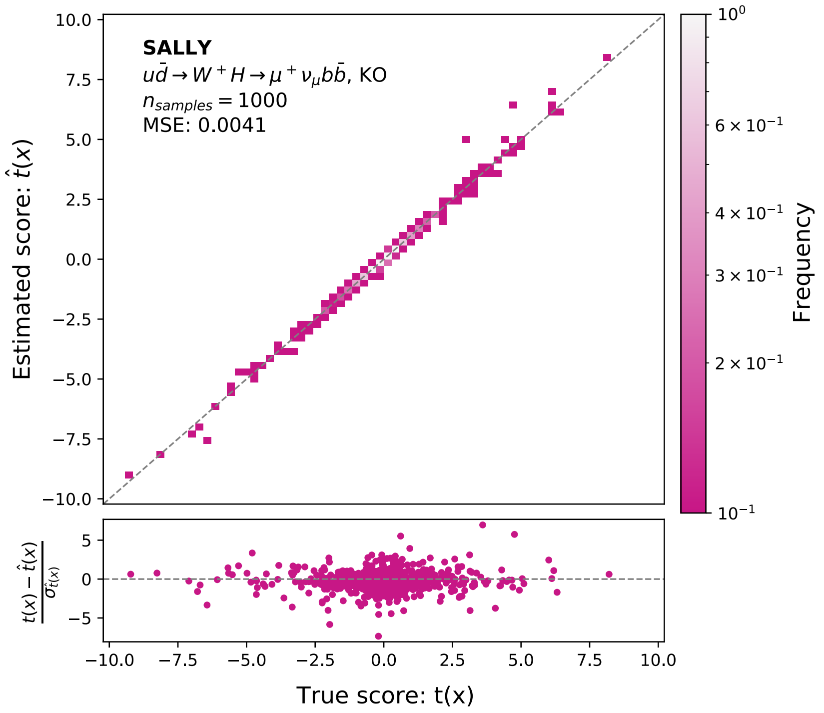


Validation at parton-level (I)

- An ensemble of 5 NN was trained using the **SALLY**, **ALICE**, and **ALICES** methods, and the Mean Squared Error (MSE) was used to compare each method's sensitivity.
- The estimated quantities closely align with the true values, confirming that these inference techniques yield reliable results (at least in the truth-level scenario).



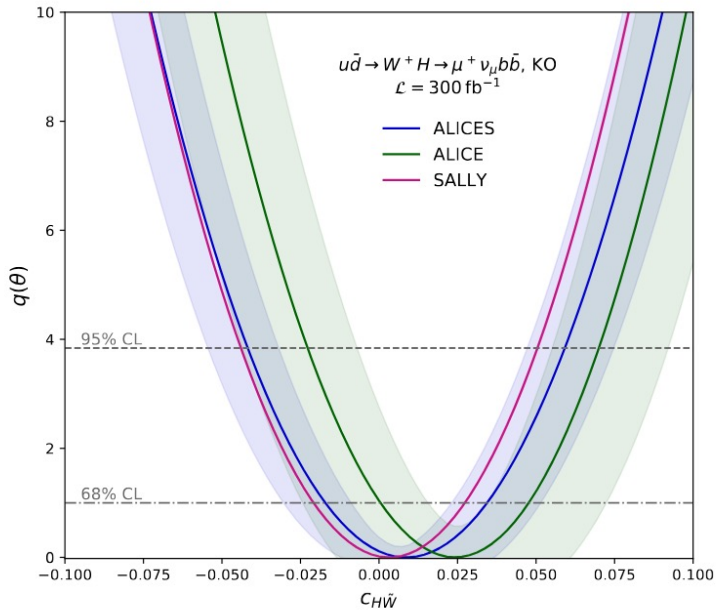




- The **ALICES** and **ALICE** MSE were consistently higher when using a Uniform Prior in the sampling.
  - A Gaussian Prior was chosen for the subsequent studies

Validation at parton-level (II)

Method	MSE
SALLY	0.0041
ALICES ($\alpha = 5, n_{\theta_0} = 1000$) w/ Gaussian ($\mu = 0$, $\sigma = 0.4$)	0.0125
ALICE ($\alpha = 0, n_{\theta_0} = 1000$) w/ Gaussian ($\mu = 0$, $\sigma = 0.4$)	0.0068
ALICES ($\alpha = 5, n_{\theta_0} = 1000$) w/ Uniform([-1.2,1.2])	0.0523
ALICE ($\alpha = 0, n_{\theta_0} = 1000$) w/ Uniform([-1.2,1.2])	0.0167



Method	Central value	SD	68% CL	95% CL
SALLY	0.002	0.001	[-0.022, 0.026]	[-0.046, 0.050]
ALICES ($\alpha = 5, n_{\theta_0} = 1000$) w/ Gaussian ($\mu = 0$, $\sigma = 0.4$)	0.010	0.229	[-0.019, 0.034]	[-0.043, 0.058]
ALICE ($\alpha = 0, n_{\theta_0} = 1000$) w/ Gaussian ($\mu = 0$, $\sigma = 0.4$)	0.024	0.572	[0.000, 0.046]	[-0.024, 0.070]

Validation at parton-level (III)

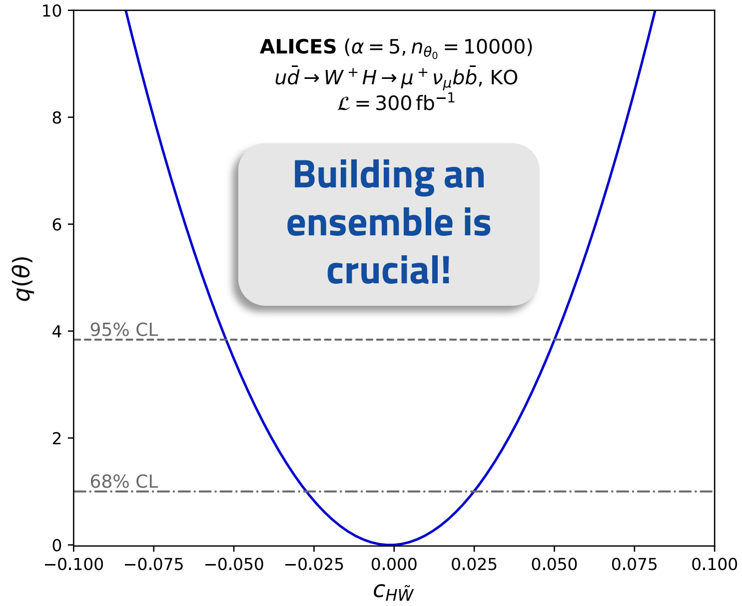
ALICE and ALICES exhibited a high variance between the 5 estimators contrary to SALLY

A different dataset was used for each estimator

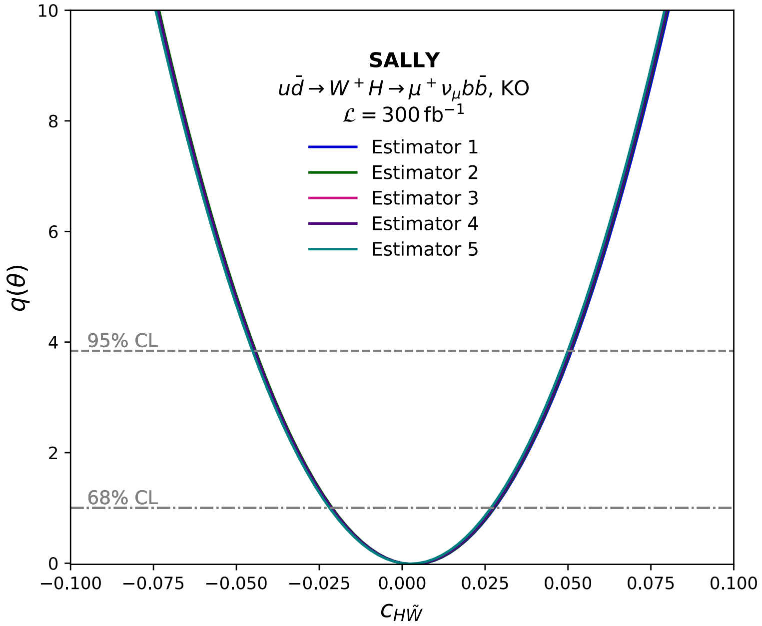
Part of the variance associated with the different θ populations

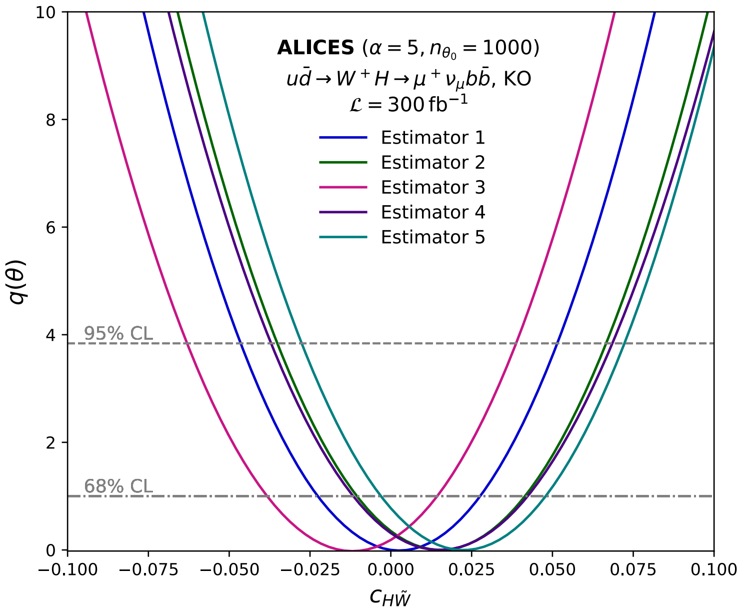
The remaining variance is attributed to the increased complexity of the ALICE(S) loss function compared to SALLY

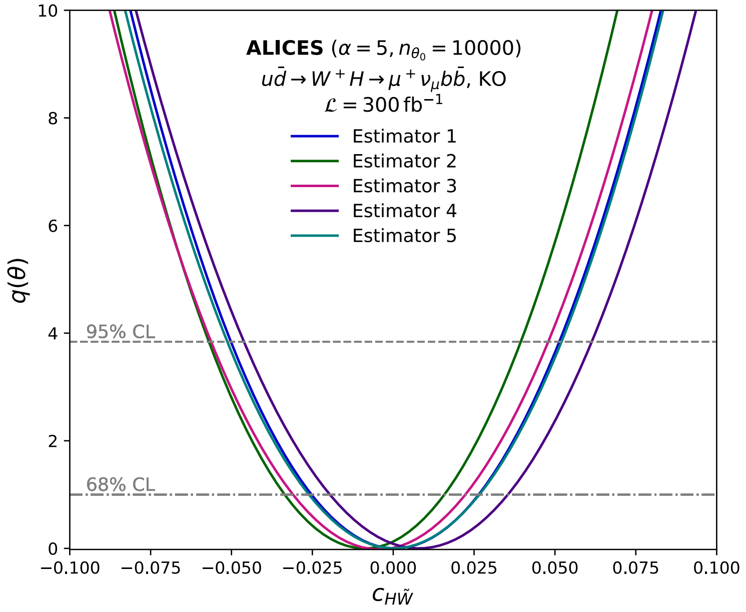
More susceptible to outliers by learning the θ dependency



Building an ensemble is crucial!

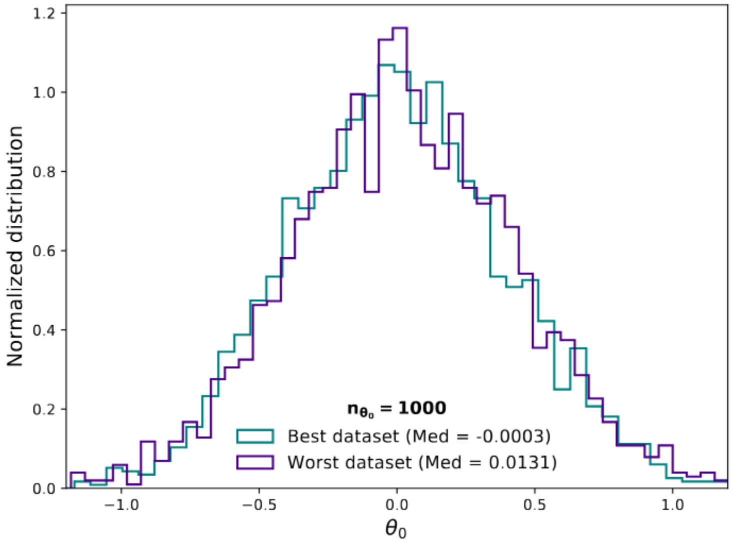


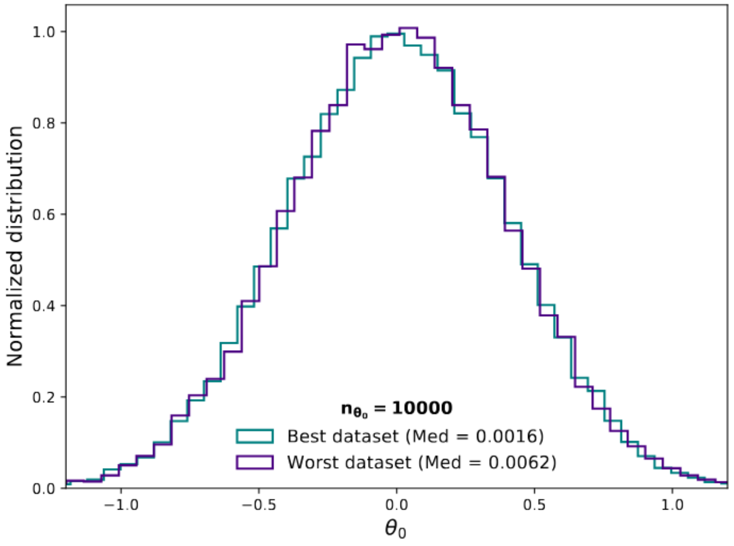




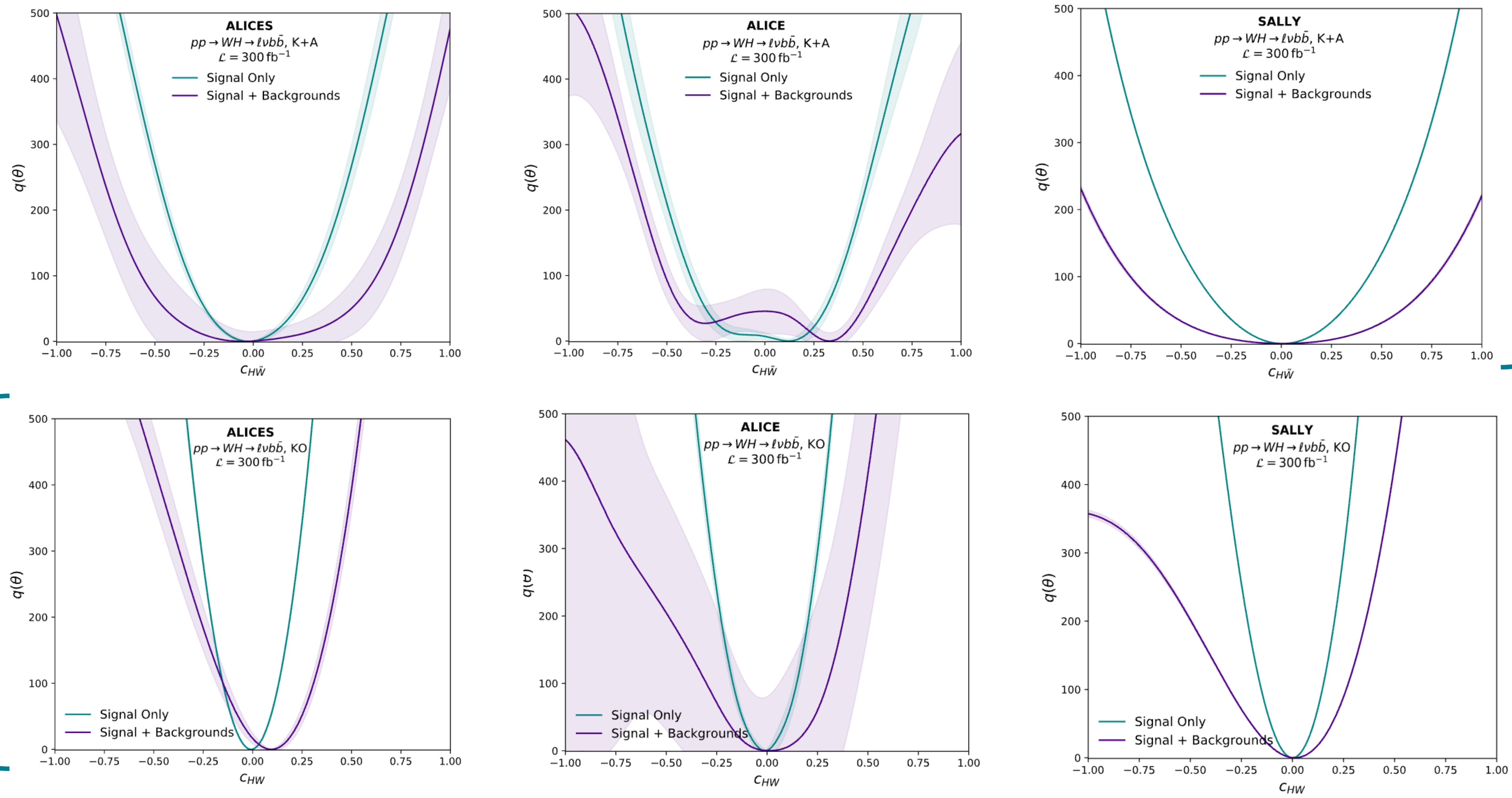
Validation at parton-level (IV)

ALICES w/ Gaussian(0,0.4)	Central value	SD	68% CL	95% CL
$n_{\theta_0} = 1000$, 5 datasets	0.010	0.229	[-0.019, 0.034]	[-0.043, 0.058]
$n_{\theta_0} = 1000$, *best dataset* (med = -0.0003)	-0.002	0.093	[-0.029, 0.024]	[-0.053, 0.048]
$n_{\theta_0} = 1000$, *worst dataset* (med = 0.0131)	0.007	0.034	[-0.019, 0.029]	[-0.043, 0.053]
$n_{\theta_0} = 10000$, 5 datasets	-0.002	0.056	[-0.029, 0.024]	[-0.053, 0.048]





1D results – signal only vs signal + backgrounds



CP-odd

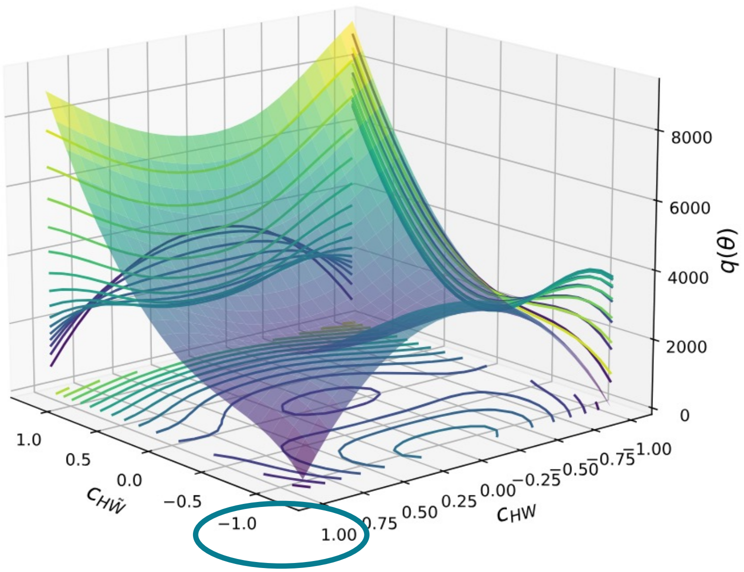
CP-even

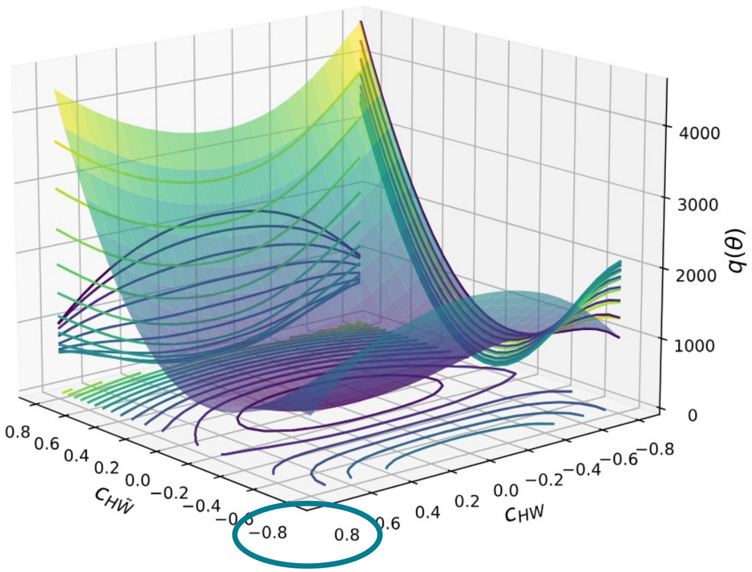
1D results – extra results (II)

Wilson Coefficient: $c_{H\tilde{W}}$								
Method	Signal Only				Signal + Backgrounds			
	Central value	SD	68% CL	95% CL	Central value	SD	68% CL	95% CL
$Q_\ell \cos\delta^+$	-0.002	-	[-0.074,0.070]	[-0.142, 0.137]	-0.005	-	[-0.194,0.182]	[-0.348, 0.343]
$Q_\ell \cos\delta^+ \otimes p_T^W$	0.038	-	[-0.046, 0.103]	[-0.156, 0.158]	0.012	-	[-0.072, 0.094]	[-0.151, 0.170]
Kinematic Only								
ALICES	0.019	0.120	[-0.014, 0.050]	[-0.046, 0.082]	0.202	6.700	[0.149, 0.247]	[0.091, 0.288]
ALICE	0.132	14.213	[0.089, 0.166]	[0.024, 0.197]	-0.043	15.711	[-0.084, -0.005]	[-0.122, 0.034]
SALLY	0.007	0.000	[-0.036, 0.050]	[-0.077, 0.091]	0.012	0.011	[-0.094, 0.113]	[-0.187, 0.209]
Kinematic + Angular Observables								
ALICES	-0.024	0.368	[-0.058, 0.010]	[-0.089, 0.043]	-0.053	15.824	[-0.120, 0.022]	[-0.180, 0.110]
ALICE	0.120	3.179	[0.086, 0.149]	[0.046, 0.178]	0.329	12.647	[0.302, 0.350]	[0.278, 0.372]
SALLY	0.007	0.000	[-0.038, 0.048]	[-0.079, 0.089]	0.007	0.004	[-0.096, 0.110]	[-0.190, 0.204]

Wilson Coefficient: c_{HW}								
Method	Signal Only				Signal + Backgrounds			
	Central value	SD	68% CL	95% CL	Central value	SD	68% CL	95% CL
$m_T^{\ell\nu b\bar{b}}$	0.000	-	[-0.017, 0.014]	[-0.031, 0.029]	0.000	-	[-0.041, 0.036]	[-0.079, 0.072]
$m_T^{\ell\nu b\bar{b}} \otimes p_T^W$	0.000	-	[-0.017, 0.014]	[-0.031, 0.029]	0.000	-	[-0.038, 0.036]	[-0.074, 0.070]
Kinematic Only								
ALICES	-0.007	0.098	[-0.024, 0.005]	[-0.036, 0.019]	0.094	0.724	[0.070, 0.115]	[0.048, 0.137]
ALICE	-0.012	2.817	[-0.026, 0.002]	[-0.041, 0.017]	0.012	82.855	[-0.029, 0.050]	[-0.065, 0.086]
SALLY	0.000	0.000	[-0.017, 0.014]	[-0.031, 0.029]	0.012	0.023	[-0.017, 0.038]	[-0.046, 0.062]

3D likelihood surfaces (ALICES)



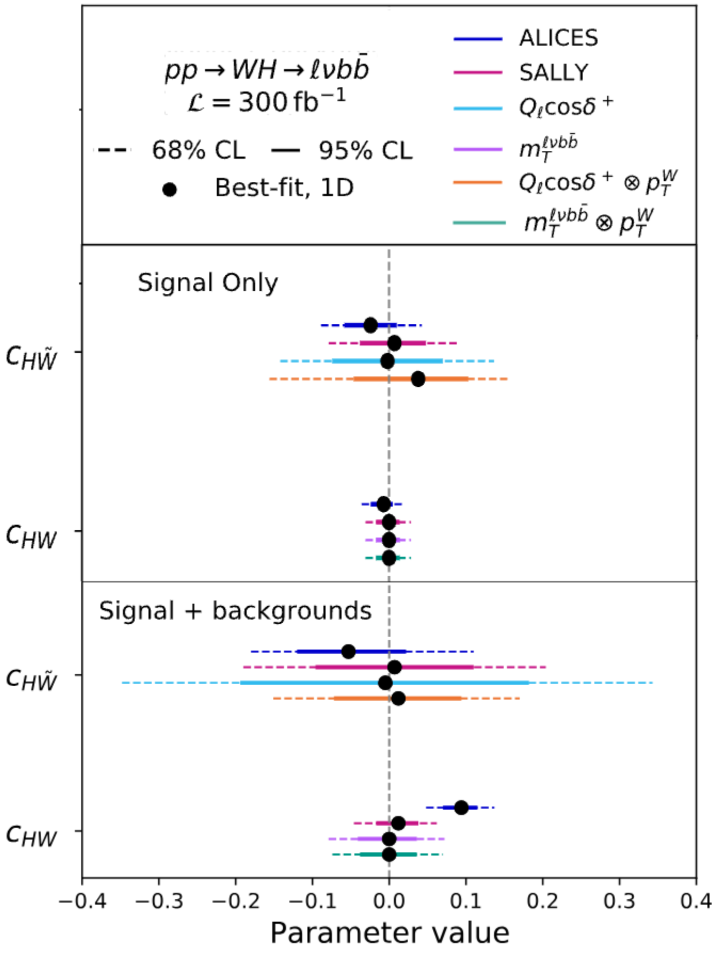


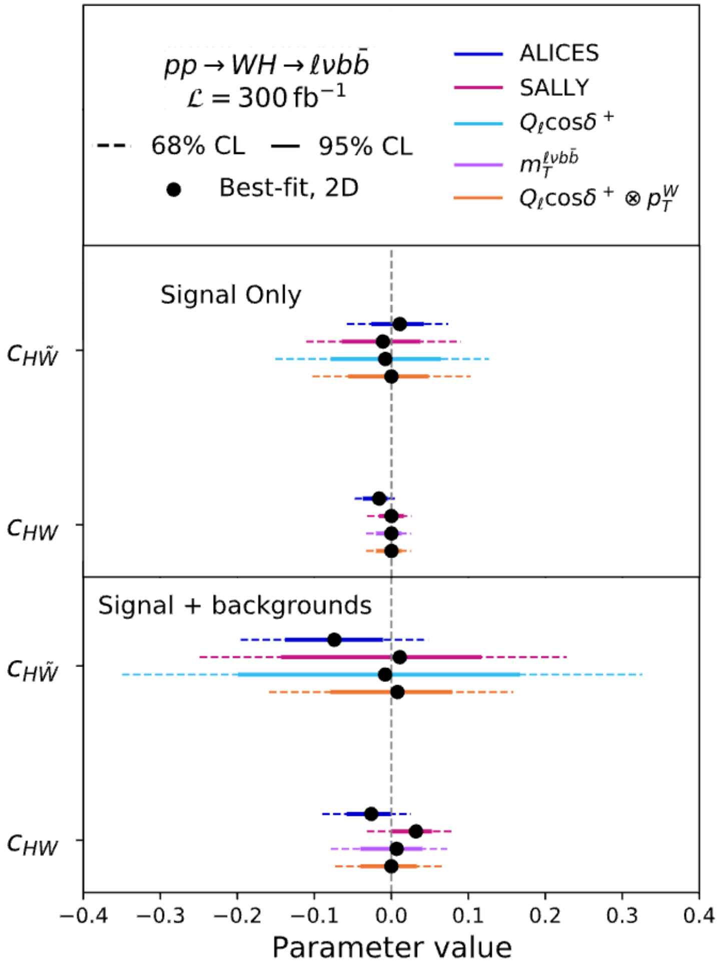
2D results – extra results (I)

Method	Signal Only				Signal + Backgrounds			
Wilson Coefficient: $c_{H\tilde{W}}$								
	Central value	SD	68% CL	95% CL	Central value	SD	68% CL	95% CL
$Q_\ell \cos\delta^+$	-0.008	-	[-0.079, 0.064]	[-0.151, 0.127]	-0.008	-	[-0.199, 0.167]	[-0.350, 0.326]
$Q_\ell \cos\delta^+ \otimes p_T^W$	0.000	-	[-0.056, 0.048]	[-0.103, 0.103]	0.008	-	[-0.079, 0.079]	[-0.159, 0.159]
Kinematic + Angular Observables								
ALICES	0.011	1.595	[-0.026, 0.042]	[-0.058, 0.074]	-0.074	9.670	[-0.138, -0.011]	[-0.196, 0.048]
SALLY	-0.011	0.041	[-0.064, 0.037]	[-0.111, 0.090]	-0.011	1.384	[-0.143, 0.117]	[-0.249, 0.228]

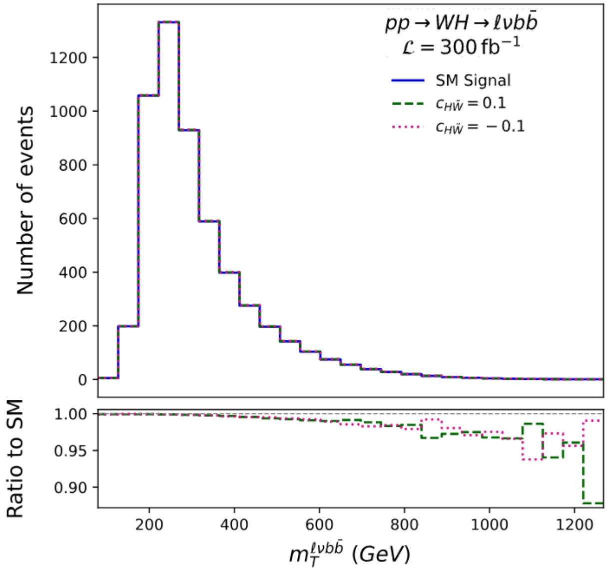
Method	Signal Only				Signal + Backgrounds			
Wilson Coefficient: c_{HW}								
	Central value	SD	68% CL	95% CL	Central value	SD	68% CL	95% CL
$m_T^{\ell\nu b\bar{b}}$	0.000	-	[-0.020, 0.013]	[-0.033, 0.026]	0.007	-	[-0.040, 0.040]	[-0.079, 0.073]
$Q_\ell \cos\delta^+ \otimes p_T^W$	0.000	-	[-0.020, 0.013]	[-0.033, 0.026]	0.000	-	[-0.040, 0.033]	[-0.073, 0.066]
Kinematic + Angular Observables								
ALICES	-0.016	1.644	[-0.037, -0.005]	[-0.048, 0.005]	-0.026	9.562	[-0.058, 0.000]	[-0.090, 0.026]
SALLY	0.000	0.001	[-0.016, 0.016]	[-0.032, 0.026]	0.032	1.070	[0.000, 0.053]	[-0.032, 0.079]

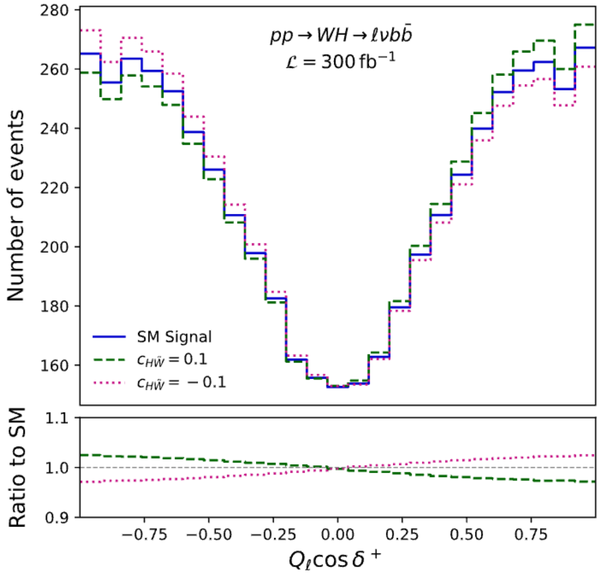
1D vs 2D results

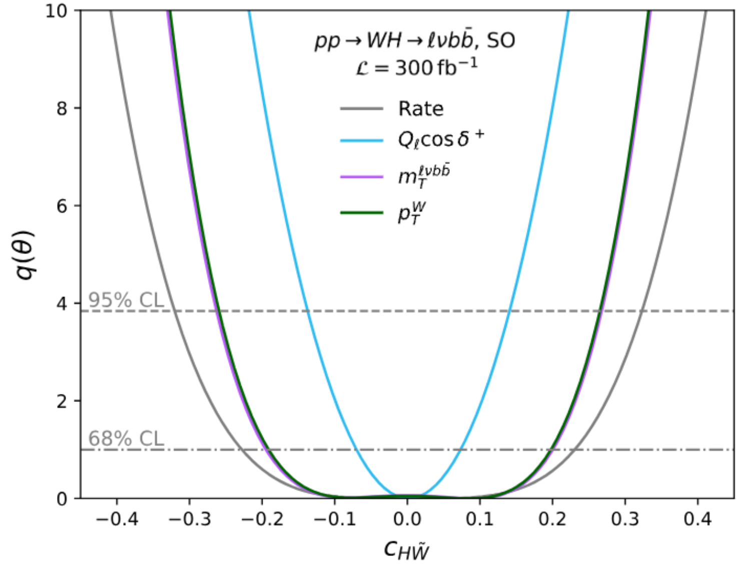


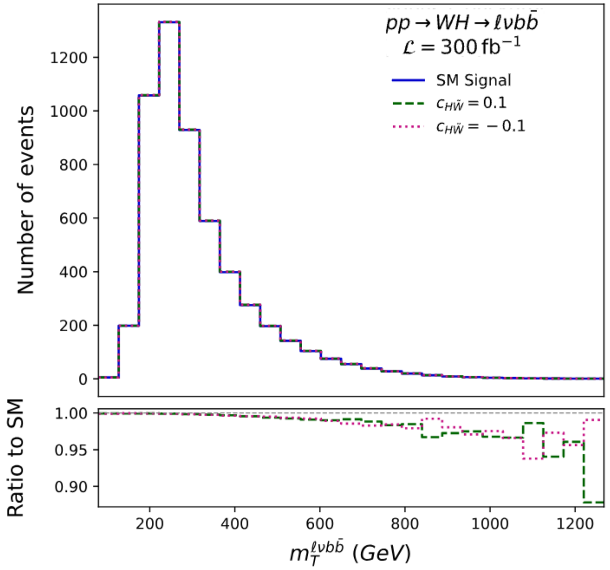


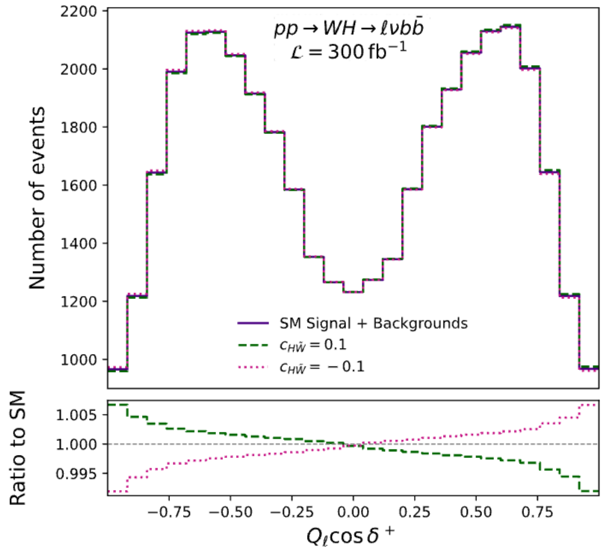
Is Physics na Obstacle? – extra plots CP-odd

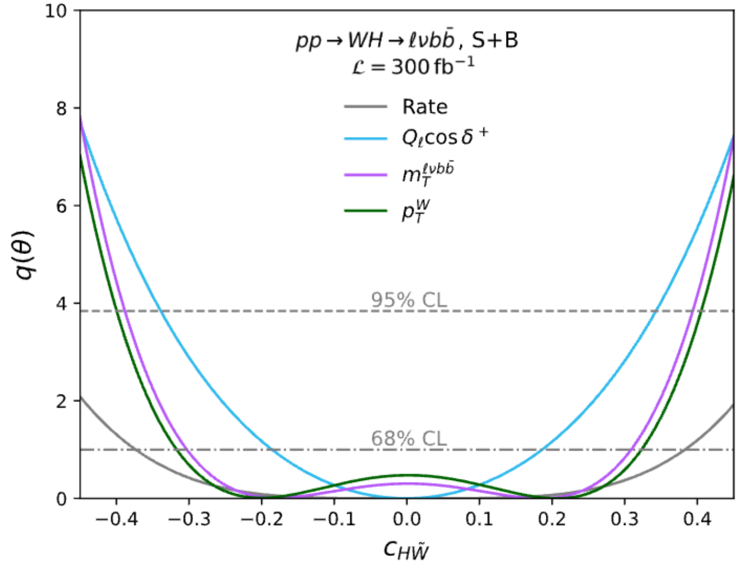




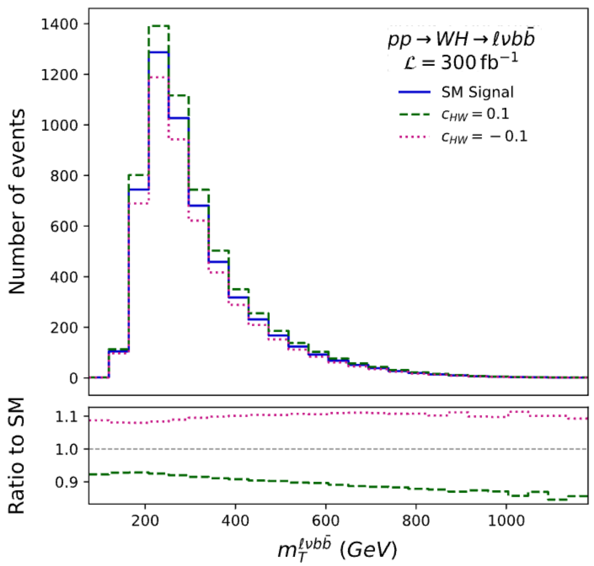


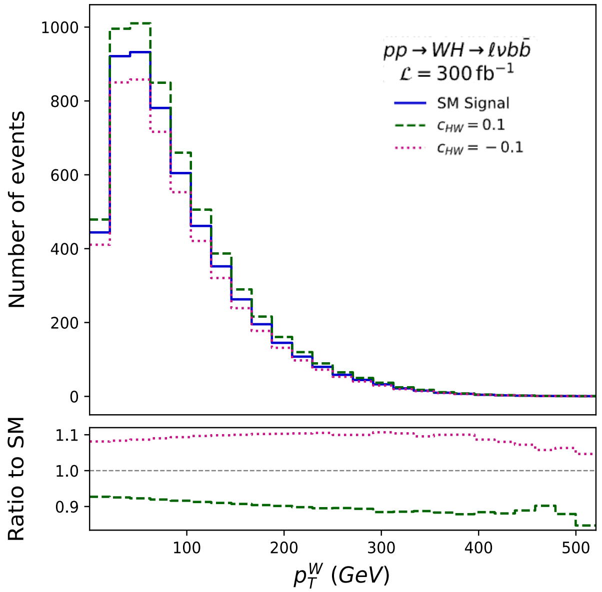


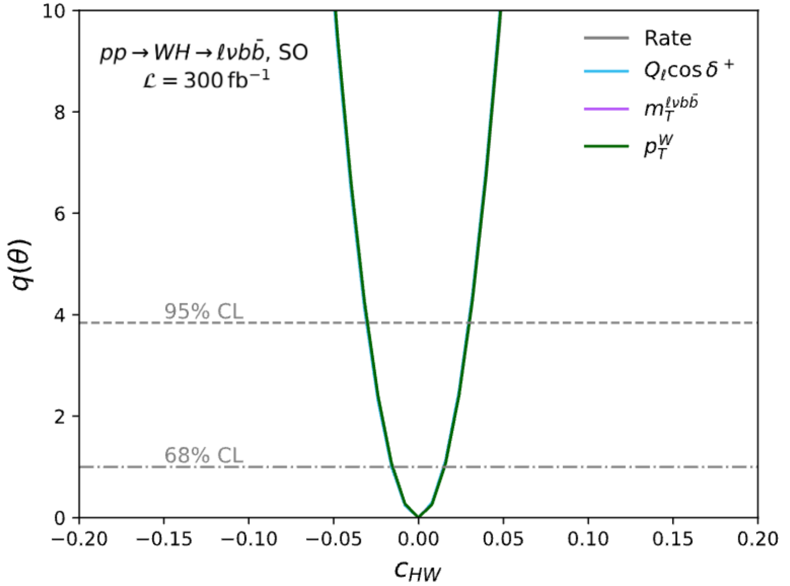


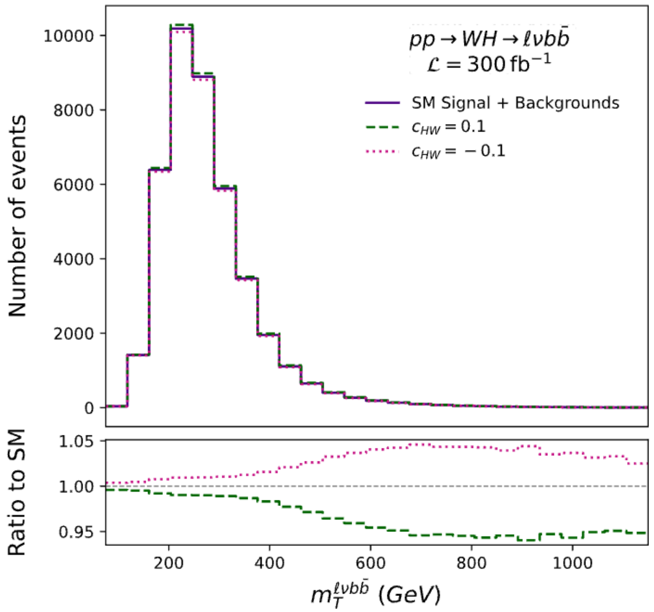


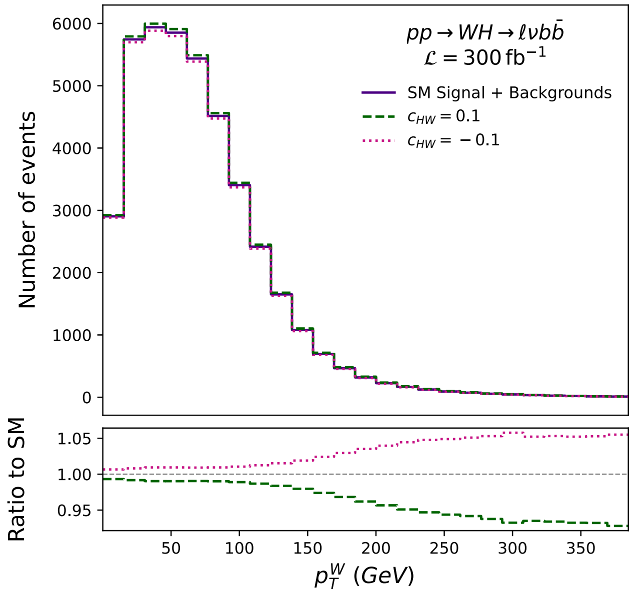
Is Physics an Obstacle? – extra plots CP-even

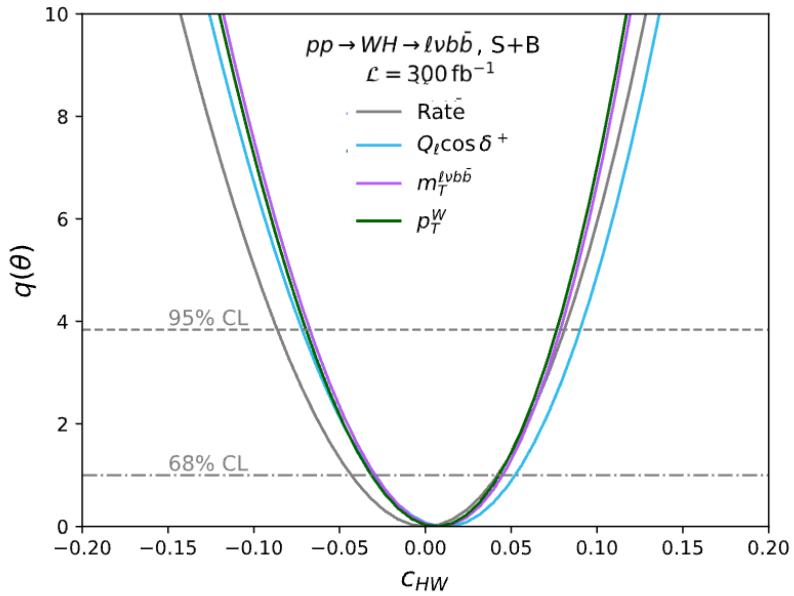




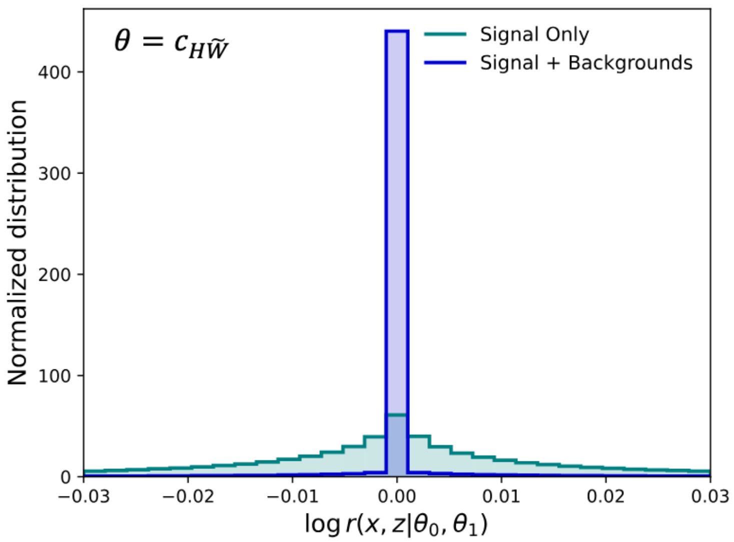


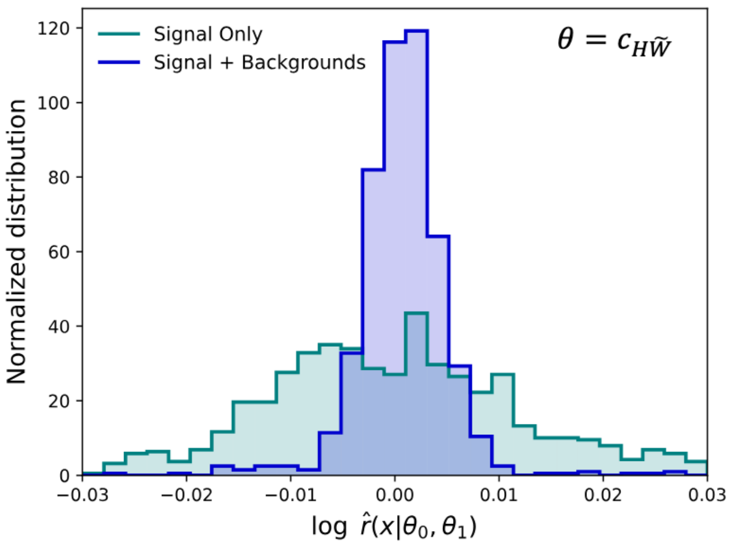


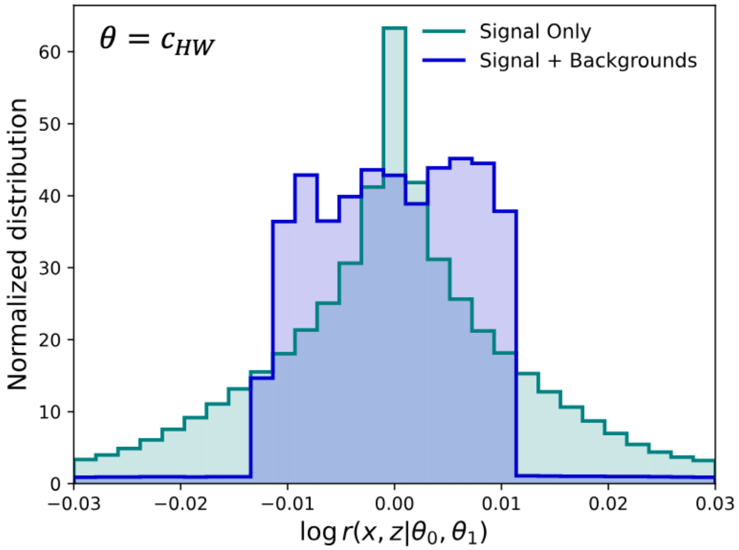


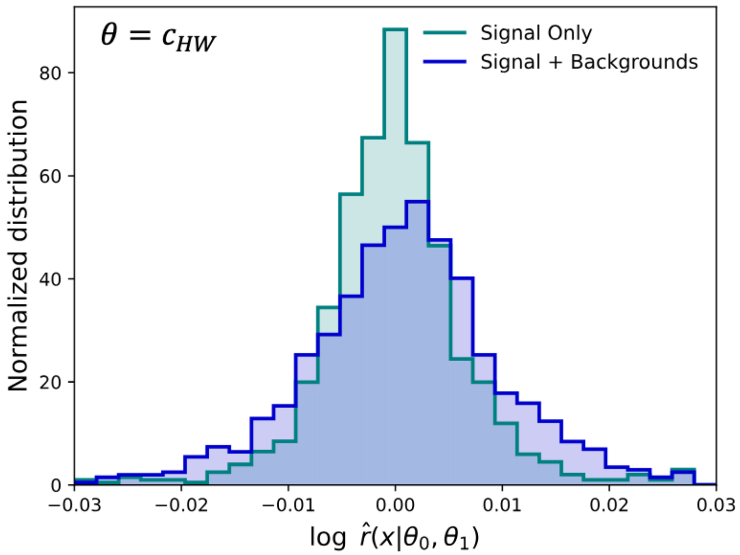


Is Physics an Obstacle? – Joint and estimated llr

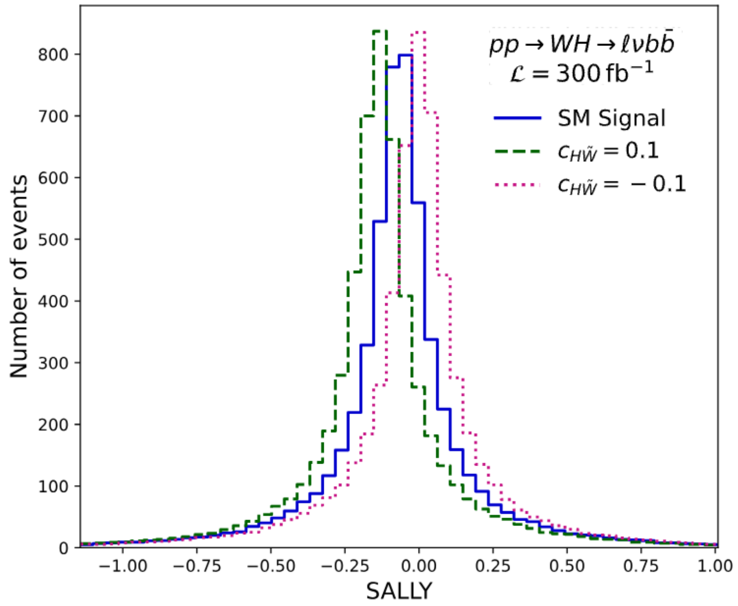


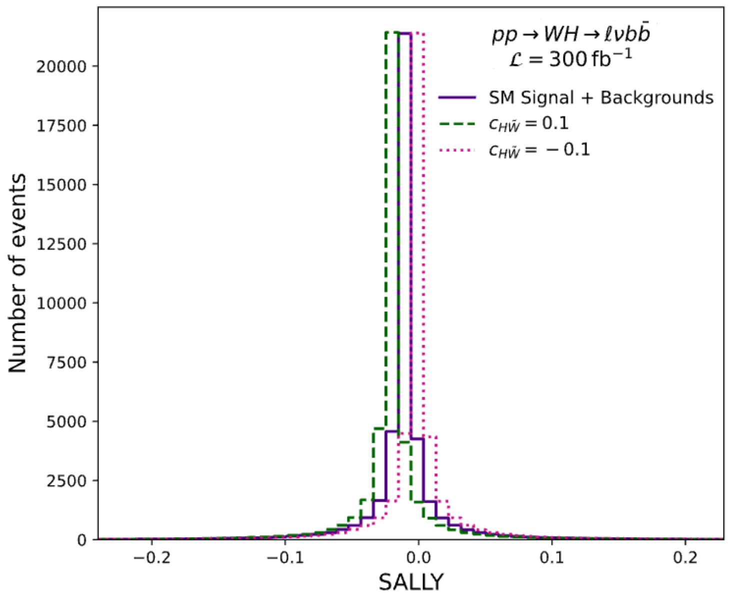


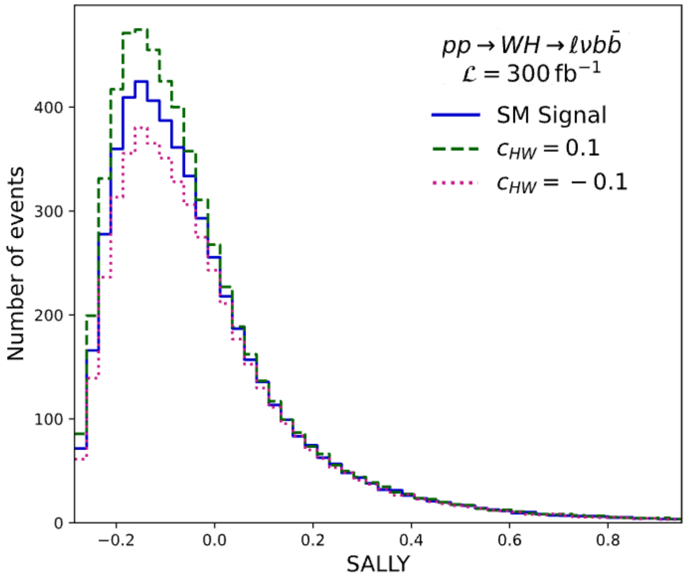


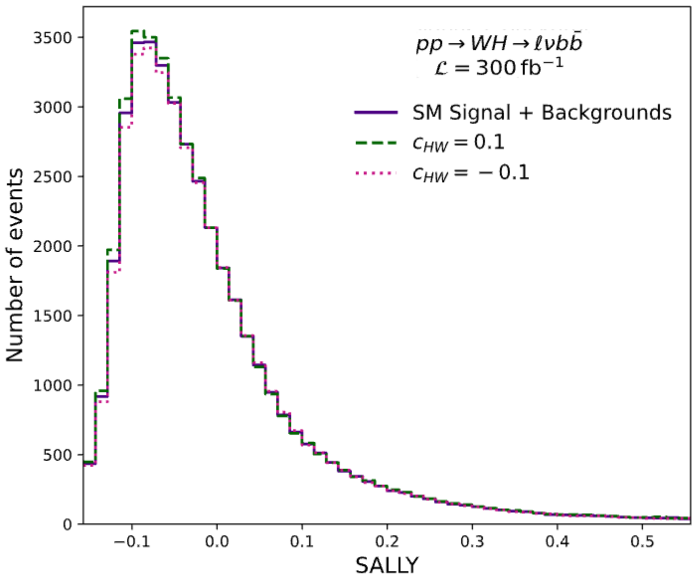


Is Physics na Obstacle? – SALLY









The Demand for More Powerful Computing Resources

Sampling time

Method	Time	
	Signal Only	Signal + Backgrounds
SALLY	4m	17m
ALICES	7h20m	16h03m

Training and Evaluation Times for the 5 NNs

Method	Training Time		Evaluation Time	
	Signal Only	Signal + Backgrounds	Signal Only	Signal + Backgrounds
SALLY	11h00m	1d 1h 14m	17m	36m
ALICES	23h05m	1d 9h 19m	15h29m	1d 20h 19m

Future Work / Ideas to overcome the challenges

- Repeating this study in a higher p_T region
  - Increased sensitivity to BSM couplings
  - Increased signal-to-background ratio

- The SO results were much more reliable and consistent among individual estimators

  Pass the training samples first through a classifier (prior to training) to reduce the number of background events

- One of the main challenges is that the sampling in the $\theta -$ space induces instabilities

  Factorize from the likelihood parametrization the θ dependency

- The ALICES sampling is based on inverse transform sampling and the most computationally intensive aspects arise from calculating the cumulative sum and the index search

  Explore other sampling techniques or calculating a binned cumulative distribution function

- Training and evaluating these methods are very computationally demanding

  We are applying for special access to HPC+GPUs

Calibration and diagnostics

The expectation value of the likelihood ratio assuming θ_1 to be true is given by:

$$\mathbb{E}[r(x|\theta_0,\theta_1)|\theta_1] = \int dx\, p(x|\theta_1)\, \frac{p(x|\theta_0)}{p(x|\theta_1)} = 1$$

A good estimator for the likelihood ratio should reproduce this property. We can numerically approximate this expectation value with:

$$\hat{R}(\theta) = \frac{1}{N} \sum_{x_e \sim \theta_1} \hat{r}(x_e|\theta,\theta_1) \approx 1 .$$

If a likelihood ratio estimator $\hat{r}_{raw}(x|\theta,\theta_1)$does not satisfy this condition, we can **calibrate** it by rescaling it as:

$$\hat{r}_{\text{cal}}(x|\theta,\theta_1) = \frac{\hat{r}_{\text{raw}}(x|\theta,\theta_1)}{\hat{R}_{\text{raw}}(\theta)} .$$

For a perfect estimator, we can even calculate the variance of the numeric calculation of the expectation value:

$$\text{var}[\hat{R}(\theta)] = \frac{1}{N} \left[\mathbb{E}\left[\hat{r}(x|\theta,\theta_1)|\theta\right] - 1\right]$$

Ensemble variance:

Train an ensemble of estimators with different training data and random seeds

Ensemble variance as a measure of uncertainty of the prediction

Reference hypothesis variation:

Any estimated likelihood ratio between two hypotheses $\theta_A$and θ_B should be independent of the choice of the reference hypothesis θ_1 used in the estimator $\hat{r}$.

$$\hat{r}(x|\theta_A,\theta_B) = \frac{\hat{r}(x|\theta_A,\theta_1)}{\hat{r}(x|\theta_B,\theta_1)}$$

To check the stability of the results we can train several independent estimators with different values of θ_1

Reweighting distributions:

A good estimator should satisfy: $p(x|\theta_0) \approx \hat{r}(x|\theta_0,\theta_1)\, p(x|\theta_1)$

We can draw samples from the 2 distributions and reweight one of them with $\hat{r}(x|\theta_0,\theta_1)$. If a classifier can distinguish between the sample from $\theta_1$and the reweighted one, $\hat{r}(x|\theta_0,\theta_1)$ is not a good approximation of $r(x|\theta_0,\theta_1)$