Unbinned inclusive cross-section measurements with machine-learned systematic uncertainties

Lisa Benato, Cristina Giordano, Claudius Krause, Ang Li, Robert Schöfbeck, Dennis Schwarz, Maryam Shooshtari, and Daohan Wang

Institute for High Energy Physics, Austrian Academy of Sciences
Dominikanerbastei 16, 1010 Vienna, Austria

(Received 22 May 2025; accepted 22 July 2025; published 18 September 2025)

We introduce a novel methodology for addressing systematic uncertainties in unbinned inclusive cross-section measurements and related collider-based inference problems. Our approach incorporates known analytic dependencies on parameters of interest, including signal strengths and nuisance parameters. When these dependencies are unknown, as is frequently the case for systematic uncertainties, dedicated neural network parametrizations provide an approximation that is trained on simulated data. The resulting machine-learned surrogate captures the complete parameter dependence of the likelihood ratio, providing a near-optimal test statistic. As a case study, we perform a first-principles inclusive cross-section measurement of $H \to \tau\tau$ in the single-lepton channel, utilizing simulated data from the FAIR universe Higgs uncertainty challenge. Results in Asimov data, from large-scale toy studies, and using the Fisher information demonstrate significant improvements over traditional binned methods. Our computer code guaranteed optimal log-likelihood-based unbinned method (GOLLUM) for machine-learning and inference is publicly available.

DOI: 10.1103/zwzt-1rrw

I. INTRODUCTION

In recent years, the Large Hadron Collider (LHC) [1] has shifted from a discovery instrument to a precision measurement machine, allowing unprecedented accuracy in the measurements of parameters of the Standard Model (SM) and the processes and phenomena it predicts. A toolkit of statistical methods employs the likelihood function to extract parameter values, perform hypothesis tests, and set exclusion limits [2–5]. The exact likelihood for high-dimensional experimental data is, however, usually computationally intractable. Traditionally, LHC analyses eschew this issue by reducing high-dimensional measurements to low-dimensional summary statistics via binning and with the help of physically motivated observables. The likelihood is then tractable, but the dimensional reduction inevitably causes information loss, resulting in suboptimal performance.

To exploit the LHC's full potential, we must avoid information loss and work with a high-dimensional, unbinned likelihood function. While that function remains intractable in experimental settings, it can be inferred from detailed simulations with modern machine learning (ML) tools [6–10], a technique known as (neural) simulation-based inference (SBI) [11]. SBI eliminates the need for binning and summary statistics, while allowing unbinned approximations related to the effects of the parton shower, the hadronization, and the detector response. With this, it opens new avenues for precision analyses [12]. Additionally, once the initial training phase is complete, likelihood-based inference is computationally fast.

Although phenomenological studies have explored SBI [13–20], large-scale deployment in experimental analyses remains limited by the relative simplicity of tools for addressing systematic uncertainties. In Ref. [21], the various methods for estimating uncertainties in binned analyses are promoted to the unbinned case. The ML-based parametrizations of systematic uncertainties, either in groups of correlated effects or one by one, allow for a flexible and refinable approach that largely follows the workflow of the binned case. Analysis development begins with the most important processes and systematic effects and can later be improved step-by-step without invalidating earlier partial results. Building on this work, we demonstrate here how the methodology of Ref. [21] can be applied to unbinned cross-section measurements.

We showcase the developments in the FAIR universe Higgs uncertainty challenge [22] (FAIR-HUC), achieving highly competitive results and tying *ex aequo* for first place. Our code guaranteed optimal log-likelihood-based unbinned method (GOLLUM) is publicly available [23].

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

The ATLAS Collaboration [24] recently explored unbinned cross-section measurements with classifier-based modeling of systematic uncertainties [25,26]. In contrast, we learn, in principle, arbitrarily accurate parametric models of systematic effects and also include the cases where they do not factorize.

The structure of this paper is as follows. In Sec. II, we introduce the $H \to \tau\tau$ process and describe the FAIR-HUC simulated datasets, including the available event-level features, the event selections, and the systematic uncertainties. We develop the profiled likelihood ratio test statistic, both in a binned reference case and in the fully unbinned setting, in Sec. III. Using the FAIR-HUC dataset, we demonstrate how refinable modeling incorporates known analytic dependencies into an ansatz for the likelihood function. Unknown dependencies are captured via a neural network-based parametrization of several systematic effects whose details we describe in Sec. IV. Section V presents results obtained from applying these techniques to the FAIR-HUC datasets, including studies on Asimov data [2], large-scale toy studies, and the Fisher information matrix. We conclude in Sec. VI.

II. PROCESSES AND DATASETS

A. The $H \to \tau\tau$ process

The study of the $H \to \tau\tau$ process at the LHC has been systematically pursued by the ATLAS [24] and CMS [27] Collaborations across multiple datasets and center-of-mass energies. Before the discovery of the Higgs (H) boson, searches in the 7 TeV dataset obtained during Run I operations by CMS [28] and ATLAS [29] provided upper limits as a function of the H boson mass. After the Higgs boson discovery, the first evidence of the $H \to \tau\tau$ decay was reported using the combined 7 and 8 TeV datasets [30,31]. The observation of the process with a significance of 5σ was achieved by combining earlier results with Run II data at 13 TeV [32,33]. Subsequently, inclusive cross-section measurements were performed [34,35]. Both collaborations refined the differential measurements by targeting the boosted selections and the simplified template cross section (STXS) [36–38]. These results collectively establish the $H \to \tau\tau$ decay as a key probe of the Higgs boson's properties at the LHC.

The $H \to \tau\tau$ final state is classified by the leptonic (τ_{lep}) or hadronic (τ_{had}) decays of the τ leptons. Leptonic decays follow $\tau \to \ell\nu_\ell\nu_\tau$, where $\ell = (e, \mu)$, while hadronic decays produce charged and neutral hadrons along with a neutrino. The fully leptonic final state ($\tau_{\text{lep}}\tau_{\text{lep}}$) has a clean signature but suffers from reduced mass resolution due to multiple neutrinos. In contrast, hadronic decays allow for better momentum reconstruction with a single neutrino but face larger backgrounds from misidentified jets. The fully hadronic channel ($\tau_{\text{had}}\tau_{\text{had}}$), in particular, has the highest branching fraction but is significantly affected by jet misidentification. The semileptonic channel ($\tau_{\text{lep}}\tau_{\text{had}}$) balances signal yield and background suppression, benefiting from efficient identification of the τ_{lep} while retaining good mass resolution from the τ_{had}. Combining all channels optimizes sensitivity to $H \to \tau\tau$, with the semileptonic final state typically contributing most.

B. Features, event samples, and selections

As the signal process, the FAIR-HUC datasets contain the semileptonic $H \to \tau_{\text{lep}}\tau_{\text{had}}$ process generated with PYTHIA 8.2 [39] and processed with DELPHES 3.5.0 [40]. Each event contains a lepton with transverse momentum (p_{T}) above 20 GeV and with a pseudorapidity (η) satisfying $|\eta^{\ell}| \leq 2.5$. The τ_{had} satisfies[1] $p_{\text{T}}^{\tau_{\text{had}}} > 26$ GeV and $|\eta^{\tau_{\text{had}}}| < 2.69$.

The 28 available event features capture key kinematic properties of the τ_{had}, the lepton, the jets, and the neutrino system, along with derived quantities constructed from these primary observables.

The primary features include the transverse momentum ($p_{\text{T}}^{\tau_{\text{had}}}$, p_{T}^{ℓ}, $p_{\text{T}}^{j_1}$, $p_{\text{T}}^{j_2}$), pseudorapidity ($\eta^{\tau_{\text{had}}}$, η^{ℓ}, η^{j_1}, η^{j_2}), and azimuthal angle ($\phi^{\tau_{\text{had}}}$, ϕ^{ℓ}, ϕ^{j_1}, ϕ^{j_2}) of the τ_{had}, the lepton, and up to two leading jets. Additionally, the missing transverse momentum ($\vec{p}_{\text{T}}^{\,\text{miss}}$) and its azimuthal angle (ϕ^{miss}) are included.

Derived observables are computed from these primary features and incorporate information relevant for distinguishing $H \to \tau\tau$ signal events from background processes. These include kinematic quantities such as the transverse mass of the lepton and missing transverse energy [$m_{\text{T}}(\ell, \vec{p}_{\text{T}}^{\,\text{miss}})$], the visible invariant mass of the τ_{had} and the lepton (m_{vis}), the modulus of the vector sum of the τ_{had}, lepton, and missing transverse momentum (p_{T}^{H}), the invariant mass of the two leading jets ($m^{j_1 j_2}$), the number of reconstructed jets (N_j), and the sum of transverse jet momenta ($\sum_{\text{jets}} p_{\text{T}}$). Additional variables exploit correlations between objects, such as the pseudorapidity separation ($\Delta\eta^{j_1 j_2}$) and product ($\eta^{j_1} \cdot \eta^{j_2}$) between the two leading jets, the angular separation $\Delta R(\tau_{\text{had}}, \ell)$ between the τ_{had} and the lepton, and event-level features such as the modulus of the total transverse momentum sum ($p_{\text{T}}^{\text{tot}}$) and the sum of transverse momenta of the τ_{had}, lepton, and all jets ($\sum p_{\text{T}}$). In traditional analyses, the observables $\Delta\eta^{j_1 j_2}$, $\eta^{j_1} \cdot \eta^{j_2}$, and $m^{j_1 j_2}$ are used to identify signal events origination from vector-boson fusion (VBF) production. Furthermore, two centrality observables are included: the azimuthal centrality of the missing transverse energy with respect to the τ_{had} and the lepton (C_{ϕ}^{miss}) and the pseudorapidity centrality of the lepton with respect to the two jets (C_{η}^{ℓ}). The ratio of the transverse momenta of the lepton and the τ_{had} is also used ($p_{\text{T}}^{\ell}/p_{\text{T}}^{\tau_{\text{had}}}$).

[1]These thresholds are similar to the choices in Ref. [33].

TABLE I. List of event features used in the analysis. Primary features correspond to direct kinematic observables, while derived features incorporate correlations and high-level event properties.

Symbol	Description	Symbol	Description
$p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$	Transverse momentum of the τ_{had}	p_{T}^{ℓ}	Transverse momentum of the lepton
$\eta^{\tau_{\mathrm{had}}}$	Pseudorapidity of the τ_{had}	η^{ℓ}	Pseudorapidity of the lepton
$\phi^{\tau_{\mathrm{had}}}$	Azimuthal angle of the τ_{had}	ϕ^{ℓ}	Azimuthal angle of the lepton
$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$	Missing transverse momentum	ϕ^{miss}	Azimuthal angle of missing transverse momentum
$p_{\mathrm{T}}^{j_1}$	Transverse momentum of the leading jet	$p_{\mathrm{T}}^{j_2}$	Transverse momentum of the subleading jet
η^{j_1}	Pseudorapidity of the leading jet	η^{j_2}	Pseudorapidity of the subleading jet
ϕ^{j_1}	Azimuthal angle of the leading jet	ϕ^{j_2}	Azimuthal angle of the subleading jet
N_j	Number of reconstructed jets	$\sum_{\mathrm{jets}} p_{\mathrm{T}}$	Sum of transverse momenta of all jets
$m_{\mathrm{T}}(\ell, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$	Transverse mass of lepton and missing p_{T}	m_{vis}	Visible invariant mass of τ_{had} and ℓ
$p_{\mathrm{T}}^{\mathrm{H}}$	Modulus of vector sum of $\tau_{\mathrm{had}}, \ell, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$	$m^{j_1 j_2}$	Invariant mass of the two leading jets
$\Delta\eta^{j_1 j_2}$	Pseudorapidity separation of leading jets	$\eta^{j_1} \cdot \eta^{j_2}$	Product of leading jet pseudorapidities
$p_{\mathrm{T}}^{\mathrm{tot}}$	Modulus of vector sum of $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, $p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$, p_{T}^{ℓ}, $p_{\mathrm{T}}^{j_1}$, $p_{\mathrm{T}}^{j_2}$	$\sum p_{\mathrm{T}}$	Scalar sum of $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, $p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$, p_{T}^{ℓ}, and all jets
C_{ϕ}^{miss}	Azimuthal centrality of $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ w.r.t. $\tau_{\mathrm{had}}, \ell$	C_{η}^{ℓ}	Pseudorapidity centrality of lepton w.r.t. jets
$\Delta R(\tau_{\mathrm{had}}, \ell)$	Angular separation between τ_{had} and ℓ	$p_{\mathrm{T}}^{\ell}/p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$	Ratio of transverse momenta of lepton and τ_{had}

These engineered features are designed to enhance sensitivity to the Higgs boson decay topology and improve background rejection. Participants in the FAIR-HUC are free to utilize, modify, or omit these derived features based on their optimization strategy. All features are summarized in Table I with further implementation details provided in Ref. [22].

Besides the $\mathrm{H} \to \tau\tau$ signal, the FAIR-HUC dataset contains three background components: $\mathrm{Z} \to \tau\tau$, top quark pair production ($\mathrm{t\bar{t}}$), and diboson (VV) processes. The $\mathrm{Z} \to \tau\tau$ background is irreducible, as it shares the same final state but differs due to the absence of Higgs-mediated production. The $\mathrm{t\bar{t}}$ background arises from semileptonic top decays, where hadronic τ decays and missing transverse momentum resemble the signal topology. The VV processes, primarily WW, contribute when both bosons decay leptonically or when one produces the τ_{had}. These backgrounds contribute to distinct kinematic regions and must be separated from the signal and from each other for an accurate measurement of the signal strength. Backgrounds from instrumental effects, nonprompt leptons, or fake leptons are not included.

We employ the features to define two signal-enriched regions (SRs), targeting the $\mathrm{H} \to \tau\tau$ process, and three signal-depleted control regions (CRs), which help constrain uncertainties in the rates of background processes and the calibration of physics objects. In three of these regions, we perform unbinned cross-section measurements that fully exploit the high-dimensional feature correlation.

The SR `lowMT-VBFJet` aims to select the VBF Higgs boson production by requiring $p_{\mathrm{T}}^{j_1} > 50$ GeV and $p_{\mathrm{T}}^{j_2} > 30$ GeV, referred to as the `VBFJet` selection. No explicit requirement on $\Delta\eta^{j_1 j_2}$ is applied, as this variable is used in the training of ML surrogates. To suppress the dominant $\mathrm{t\bar{t}}$ background, we impose $m_{\mathrm{T}} \leq 70$ GeV. The second SR, `lowMT-noVBFJet-ptH100`, targets gluon fusion (ggH), which predominantly produces the H boson at high p_{T}. This region is defined by $p_{\mathrm{T}}^{\mathrm{H}} > 100$ GeV, $m_{\mathrm{T}} \leq 70$ GeV, and an overlap removal implemented by inverting the `VBFJet` selection. The CR `highMT-VBFJet` combines the `VBFJet` selection with $m_{\mathrm{T}} > 70$ GeV and is strongly enriched in $\mathrm{t\bar{t}}$ production due to the absence of a $\Delta\eta^{j_1 j_2}$ requirement. The $\mathrm{Z} \to \tau\tau$ production is enriched in the `highMT-noVBFJet` region by selecting events with $m_{\mathrm{T}} > 70$ GeV and inverting the `VBFJet` selection. This region can be further split to target $\mathrm{t\bar{t}}$ production and VV production individually using multivariate discriminants explained in Sec. VA. The `lowMT-noVBFJet-ptH0to100` region selects a small yield of low-p_{T} ggH events with $p_{\mathrm{T}}^{\mathrm{H}} \leq 100$ GeV and $m_{\mathrm{T}} < 70$ GeV. Because it is background-dominated and contains 95% of the total yield, its main purpose is to constrain the overall normalization. The event selections are summarized in Table II.

C. Systematic uncertainties

The systematic uncertainties of the FAIR-HUC can be categorized into "calibration"-type and "normalization"-type. The calibration-type uncertainties affect event reconstruction and selection, propagating from the primary features to the derived quantities. These include the jet energy scale (`jes`), associated with a nuisance parameter ν_{jes}, and the tau energy scale (`tes`), associated with ν_{tes}. When obtaining confidence intervals, these nuisance parameters contribute a penalty term to the likelihood from a hypothetical auxiliary measurement that is approximated with a standard normal distribution constrained to

TABLE II. Summary of the event selections. The "UB" label indicates whether the selection is included in the unbinned analysis. Signal regions (SR) and control regions (CR) are treated the same. The predicted Poisson yields correspond to $\mathcal{L}\sigma$ where $\mathcal{L}$ denotes the luminosity and are reported for each subprocess and region. The last column shows the signal-to-background ratio. The total yields in the "inclusive" row are slightly lower than the total yields in Ref. [22] because of selections on $p_{\mathrm{T}}^{\mathrm{H}}$ and p_{T}^{ℓ} mandated by FAIR-HUC. The regions marked with $*$ are subsets of the `highMT-noVBFJet` region and target the $\mathrm{t\bar{t}}$ and VV normalization via requirements on multivariate discriminants explained in Sec. VA.

Region	Requirements	Type	Poisson yield $\mathcal{L}\sigma$				S/B
			$\mathrm{H}\to\tau\tau$	$\mathrm{Z}\to\tau\tau$	$\mathrm{t\bar{t}}$	VV	
`lowMT-VBFJet`	$p_{\mathrm{T}}^{j_1} > 50$ GeV $p_{\mathrm{T}}^{j_2} > 30$ GeV $m_{\mathrm{T}} \leq 70$ GeV	UB, SR	225.8	41 280.4	15 425.9	313.4	3.96×10^{-3}
`highMT-VBFJet`	$p_{\mathrm{T}}^{j_1} > 50$ GeV $p_{\mathrm{T}}^{j_2} > 30$ GeV $m_{\mathrm{T}} > 70$ GeV	UB, CR	14.7	721.7	16 768.6	193.2	8.30×10^{-4}
`lowMT-noVBFJet-ptH100`	$p_{\mathrm{T}}^{\mathrm{H}} > 100$ GeV $m_{\mathrm{T}} \leq 70$ GeV veto on `VBFJet`	UB, SR	57.0	17 379.8	674.6	79.0	3.14×10^{-3}
`lowMT-noVBFJet-ptH0to100`	$p_{\mathrm{T}}^{\mathrm{H}} \leq 100$ GeV $m_{\mathrm{T}} \leq 70$ GeV veto on `VBFJet`	CR	642.5	837 928.7	3 360.1	1 438.5	7.62×10^{-4}
`highMT-noVBFJet`	$m_{\mathrm{T}} > 70$ GeV veto on `VBFJet`	CR	26.0	3 826.9	5 054.3	1 409.4	2.53×10^{-3}
`inclusive`			966.0	901 137.5	41 283.4	3 433.5	1.02×10^{-3}
`highMT-noVBFJet-tt*`	$m_{\mathrm{T}} > 70$ GeV veto on `VBFJet` $\hat{f}_{\mathrm{t\bar{t}}} > 0.4$	CR	3.2	203.7	3 821.8	258.5	7.36×10^{-4}
`highMT-noVBFJet-VV*`	$m_{\mathrm{T}} > 70$ GeV veto on `VBFJet` $\hat{f}_{\mathrm{t\bar{t}}} \leq 0.4$ $\hat{f}_{\mathrm{VV}} > 0.5$	CR	2.8	165.7	292.2	724.5	2.3×10^{-3}

±10 standard deviations. A variation of one standard deviation in `jes` and `tes` corresponds to 1% shifts in the jet and tau energies, respectively. The "soft $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$" energy scale, associated with ν_{met}, follows a log-normal distribution with mean zero and standard deviation $\sigma_{\mathrm{met}} = 1$ constrained to the interval [0, 5]. A computer code for event modification is provided [22] that recomputes primary and derived features as a function of ν_{jes}, ν_{tes}, and ν_{met}. The motivation for these choices and further details are provided in Ref. [22]. For toy studies, the nuisance parameter likelihood function is interpreted as a probability density function and sampled.

Log-normal uncertainties are used to vary the normalization of each process, scaling the yield by $(1+\alpha)^{\nu}$, where α is a constant and ν the corresponding nuisance parameter, associated with a standard normal distribution constrained to ±10 standard deviations. Three normalization uncertainties are considered. The parameter ν_{bkg} modifies the overall background normalization with a small impact of $\alpha_{\mathrm{bkg}} = 0.001$. The $\mathrm{t\bar{t}}$ background normalization uncertainty has $\alpha_{\mathrm{t\bar{t}}} = 0.02$, while the VV normalization uncertainty is estimated at $\alpha_{\mathrm{VV}} = 0.25$ [22]. These normalization uncertainties scale the background yields without altering their individual kinematic distributions but have indirect shape effects on the cumulative background.

In total, the model includes seven parameters, abbreviated by $\boldsymbol{\nu}$: the signal strength μ, three nuisance parameters related to calibration uncertainties (ν_{jes}, ν_{tes}, ν_{met}), and three nuisance parameters (ν_{bkg}, $\nu_{\mathrm{t\bar{t}}}$, ν_{VV}) controlling the background normalization.

Figure 1 shows examples of distributions of kinematic features in the `lowMT-VBFJet` region as provided in the training data. Consistent with results from ATLAS and CMS, this region will dominate the sensitivity to the signal strength. The effects of systematic uncertainties at the 1σ level exceed the small signal contribution almost everywhere. With signal-to-background ratios at most at the percent level, sophisticated methods are required to reliably isolate the small $\mathrm{H}\to\tau\tau$ contribution and to reduce the uncertainties.

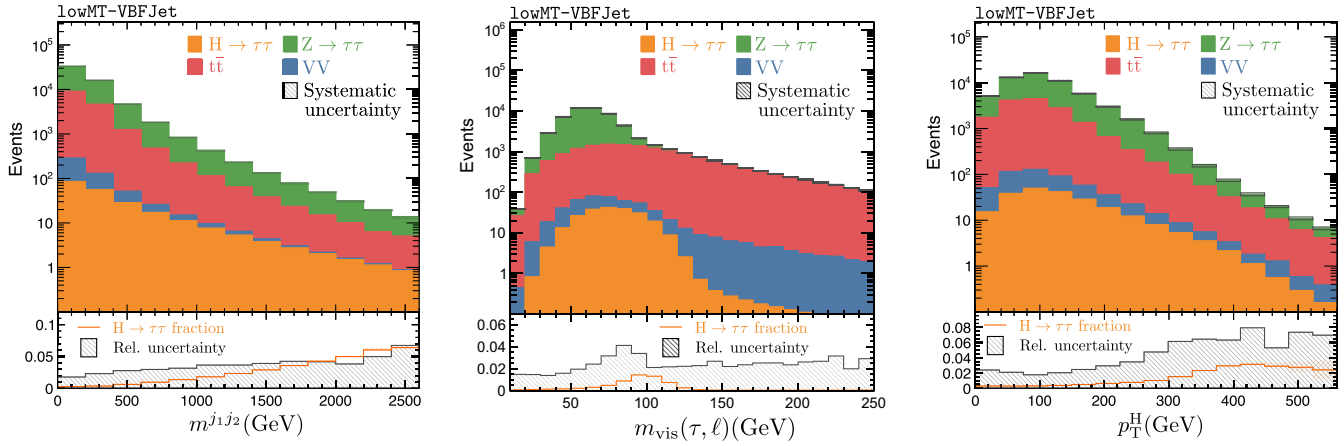

FIG. 1. Training data distributions of $m^{j_1 j_2}$ (left), $m_{\mathrm{vis}}(\tau,\ell)$ (middle), and $p_{\mathrm{T}}^{\mathrm{H}}$ (right) in the main signal region `lowMT-VBFJet`. The total systematic uncertainty, obtained from uncorrelated $\pm 1\sigma$ variations, is shown as a hashed band. In the bottom pad, the relative total uncertainty is compared to the fraction of the $\mathrm{H} \to \tau\tau$ signal.

Generating event samples for specific points in the model's parameter space is straightforward and allows, in principle, the evaluation of the likelihood. However, for limit setting, we must evaluate the likelihood continuously and efficiently, and we must do so for unseen data. The following two sections describe how we address this challenge in the SBI context. The FAIR-HUC systematics represent the leading uncertainties for a $\mathrm{H} \to \tau\tau$ cross-section measurement, but real-world analyses can include hundreds of nuisance parameters. Refinable modeling—the ability to improve the description of individual backgrounds and/or systematic effects without affecting other parts of the analysis—is, therefore, practically important.

The statistical treatment of additional uncertainty sources for unbinned analyses, such as renormalization and factorization scale variations or b-tagging efficiencies, is discussed in Ref. [21].

III. MODELING

A. The extended likelihood ratio test statistic

In this section, we discuss the general statistical setup for unbinned modeling. It is suitable for FAIR-HUC, but more widely applicable to general unbinned LHC data analyses.

We aim to set confidence intervals on the signal strength parameter μ of the $\mathrm{H} \to \tau\tau$ process. To this end, we compute the profiled likelihood ratio test statistic

$$q_\mu(\mathcal{D}) = -2\log\frac{\max_{\boldsymbol{\nu}} L(\mathcal{D}|\mu,\boldsymbol{\nu})}{\max_{\mu,\boldsymbol{\nu}} L(\mathcal{D}|\mu,\boldsymbol{\nu})} \tag{1}$$

of the observed unbinned dataset $\mathcal{D}$, comprising a variable-length list of high-dimensional feature vectors $\boldsymbol{x}_i$, $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{N_{\mathrm{obs}}}$. In our case, $\boldsymbol{x}$ has 28 components, as listed in Table I. In the absence of $\boldsymbol{\nu}$, the Neyman-Pearson lemma [41] states that $q_\mu(\mathcal{D})$ is the optimal test statistic for all values of μ. No such guarantee is known in the presence of nuisance parameters, but the profiled likelihood ratio is nevertheless widely adopted because of its asymptotic behavior: If μ is the true value, q_μ is asymptotically distributed as a χ^2 distribution with one degree of freedom for any value of $\boldsymbol{\nu}$ according to Wilks' theorem [42]. The approximate confidence interval (μ_{16}, μ_{84}) is given by the solutions to $q_\mu(\mathcal{D}) = 1$ which are obtained numerically by scanning μ and minimizing $\boldsymbol{\nu}$ for each value of μ.

Next, we derive the profiled likelihood ratio, which is the main test statistic used in this work. We assume the likelihood is "extended" by a Poisson factor with observation N_{obs} [43]. The remaining factor, the probability to observe a number of N_{obs} independently and identically distributed events $\boldsymbol{x}_i \sim p(\boldsymbol{x}|\mu,\boldsymbol{\nu})$, then leads to

$$L(\mathcal{D}|\mu,\boldsymbol{\nu}) = \mathrm{P}(N_{\mathrm{obs}}|\mathcal{L}\sigma(\mu,\boldsymbol{\nu}))\prod_{i=1}^{N_{\mathrm{obs}}} p(\boldsymbol{x}_i|\mu,\boldsymbol{\nu}). \tag{2}$$

where $\mathcal{L}$ denotes the luminosity, $\sigma(\mu,\boldsymbol{\nu})$ the total cross section, and P the probability density function of the Poisson process. We convert the normalized probability density function $p(\boldsymbol{x}|\mu,\boldsymbol{\nu})$ to the differential cross sections via

$$\mathrm{d}\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu}) = \sigma(\mu,\boldsymbol{\nu})p(\boldsymbol{x}|\mu,\boldsymbol{\nu})\mathrm{d}\boldsymbol{x} \tag{3}$$

so we have $\sigma(\mu,\boldsymbol{\nu}) = \int \mathrm{d}\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu})$ for the inclusive cross section. To simplify the profiling, we divide both the numerator and denominator likelihoods in Eq. (1) by a reference likelihood at $(\mu,\boldsymbol{\nu}) = (1,\mathbf{0})$ [21], corresponding to a signal strength of $\mu = 1$ and the nominal value for the nuisance parameters. This yields

$$q_\mu(\mathcal{D}) = \min_{\boldsymbol{\nu}} u(\mathcal{D}|\mu,\boldsymbol{\nu}) - \min_{\mu,\boldsymbol{\nu}} u(\mathcal{D}|\mu,\boldsymbol{\nu}), \tag{4}$$

where

$$-\frac{1}{2}u(\mathcal{D}|\mu,\boldsymbol{\nu}) = -\mathcal{L}(\sigma(\mu,\boldsymbol{\nu}) - \sigma(1,\mathbf{0})) + \sum_{i=1}^{N_{\text{obs}}} \log\left(\frac{\mathrm{d}\sigma(\boldsymbol{x}_i|\mu,\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}_i|1,\mathbf{0})}\right) \tag{5}$$

is the log-likelihood ratio to the reference. The first term in Eq. (4) is the best-fit value of u for a given μ over the nuisance parameters, while the second term corresponds to the best-fit value, also including μ in the fit.

The term in the logarithm in Eq. (5) is the differential cross-section ratio (DCR). With the inclusive cross section and the DCR, we can evaluate Eq. (5). These quantities, therefore, are the goal of unbinned SBI-based strategy. We compute the quantities with ML surrogates in Sec. IV.

In a binned approach, where $\mathcal{D}$ is represented by a number of observed counts ($N_{\text{obs},i}$) associated with Poisson bins labeled by i, Eq. (5) simplifies to

$$-\frac{1}{2}u_{\text{binned}}(\mathcal{D}|\mu,\boldsymbol{\nu}) = -\mathcal{L}(\sigma(\mu,\boldsymbol{\nu}) - \sigma(1,\mathbf{0})) + \sum_{i=1}^{N_{\text{bins}}} N_{\text{obs},i} \log\left(\frac{\sigma_i(\mu,\boldsymbol{\nu})}{\sigma_i(1,\mathbf{0})}\right). \tag{6}$$

The per-bin cross-section values $\sigma_i(1,\mathbf{0})$ and $\sigma_i(\mu,\boldsymbol{\nu})$ can be approximately evaluated from simulated (training) data. The continuous interpolation as a function of $\boldsymbol{\nu}$ is achieved by (log-)polynomial approximations [4,5], and we will obtain it for FAIR-HUC in the following section. The total cross section is given as the sum over bins, $\sigma(\mu,\boldsymbol{\nu}) = \sum_{i=1}^{N_{\text{bins}}} \sigma_i(\mu,\boldsymbol{\nu})$.

B. Binned likelihoods

In this section, we compute the necessary interpolations for the binned likelihood function defined in Eq. (6), covering the case of an arbitrary number of Poisson-distributed observations indexed by i as commonly applied in traditional analyses. We extend the nuisance parameter dependence to higher-order polynomials in $\boldsymbol{\nu}$ and allow systematic effects on binned yields to be nonfactorizable. Although the discussion is tailored to the seven parameters in FAIR-HUC, the presented methodology is general and provides a guardrail for the unbinned case.

The binned case also illustrates conceptual similarities between the two approaches. Crucially, both binned and unbinned methods allow for iterative refinements without invalidating previous intermediate results. In the binned case, this is straightforward: additional subleading (background) contributions can be added to the Poisson yield as needed, and further systematic effects can be incorporated via nuisance parameters that rescale the affected predictions. It is of practical importance that this flexibility extends to the unbinned case.

We construct the model of binned Poisson yields by separating into contributions from the signal and the background processes as

$$\sigma_i(\mu,\boldsymbol{\nu}) = \mu\sigma_{i,\text{H}}(\boldsymbol{\nu}_{\text{calib}}) + \sum_{p=\text{Z},\text{t}\bar{\text{t}},\text{VV}} \sigma_{i,p}(\boldsymbol{\nu}). \tag{7}$$

The normalization uncertainties enter the per-bin per-process cross sections as

$$\sigma_{i,\text{Z}}(\boldsymbol{\nu}) = (1+\alpha_{\text{bkg}})^{\nu_{\text{bkg}}}\sigma_{i,\text{Z}}(\boldsymbol{\nu}_{\text{calib}}), \tag{8a}$$

$$\sigma_{i,\text{t}\bar{\text{t}}}(\boldsymbol{\nu}) = (1+\alpha_{\text{bkg}})^{\nu_{\text{bkg}}}(1+\alpha_{\text{t}\bar{\text{t}}})^{\nu_{\text{t}\bar{\text{t}}}}\sigma_{i,\text{t}\bar{\text{t}}}(\boldsymbol{\nu}_{\text{calib}}), \tag{8b}$$

$$\sigma_{i,\text{VV}}(\boldsymbol{\nu}) = (1+\alpha_{\text{bkg}})^{\nu_{\text{bkg}}}(1+\alpha_{\text{VV}})^{\nu_{\text{VV}}}\sigma_{i,\text{VV}}(\boldsymbol{\nu}_{\text{calib}}), \tag{8c}$$

where we have introduced the abbreviation $\boldsymbol{\nu}_{\text{calib}} = \{\nu_{\text{tes}}, \nu_{\text{jes}}, \nu_{\text{met}}\}$ which we use to parametrize the calibration-type systematic uncertainties dependence.

We next parametrize $\sigma_{i,p}(\boldsymbol{\nu}_{\text{calib}})$ for $p = \{\text{H}, \text{Z}, \text{t}\bar{\text{t}}, \text{VV}\}$. In standard binned analyses as implemented with, e.g., the COMBINE package [5], systematic up- and down variations corresponding to $\nu = \pm 1$ are used to obtain a continuous interpolation [4] of the dependence of the simulated yields on the nuisance parameters. Typically, a log-polynomial expansion is used, and the nuisance parameter dependence is assumed to factorize. Here, we do not use this assumption and include cross-terms in a log-polynomial expansion, including the mixed terms, concretely

$$\begin{aligned}\log\frac{\sigma_{i,p}(\boldsymbol{\nu}_{\text{calib}})}{\sigma_{i,p}(\mathbf{0})} &= \nu_{\text{tes}}\hat{\Delta}_{i,p,\text{tes}} + \nu_{\text{jes}}\hat{\Delta}_{i,p,\text{jes}} + \nu_{\text{met}}\hat{\Delta}_{i,p,\text{met}} \\ &\quad + \nu_{\text{tes}}^2\hat{\Delta}_{i,p,\text{tes,tes}} + \nu_{\text{jes}}^2\hat{\Delta}_{i,p,\text{jes,jes}} \\ &\quad + \nu_{\text{met}}^2\hat{\Delta}_{i,p,\text{met,met}} + \nu_{\text{tes}}\nu_{\text{jes}}\hat{\Delta}_{i,p,\text{tes,jes}} \\ &\quad + \nu_{\text{jes}}\nu_{\text{met}}\hat{\Delta}_{i,p,\text{jes,met}} + \nu_{\text{tes}}\nu_{\text{met}}\hat{\Delta}_{i,p,\text{tes,met}} \\ &= \nu_A\hat{\Delta}_{i,p,A}.\end{aligned} \tag{9}$$

The first three coefficients correspond to the log-linear dependence, the three terms in the second line to the log-quadratic dependence and the remainder constitute the mixed terms. For each bin i and process p, we thus need to determine 9 constants $\hat{\Delta}_{i,p,A}$ that approximate the left-hand side To simplify the notation, we label the 9 linear, quadratic, and mixed coefficients as

$$\begin{aligned}A = \{&\{\text{tes}\},\{\text{jes}\},\{\text{met}\},\{\text{tes,tes}\},\{\text{jes,jes}\},\{\text{met,met}\},\\ &\{\text{tes,jes}\},\{\text{jes,met}\},\{\text{tes,met}\}\},\end{aligned} \tag{10}$$

which defines the last equality in Eq. (9) where we use the Einstein sum convention for the new index A. Note that ν_A

also contains terms linear and quadratic in $\boldsymbol{\nu}_{\text{calib}}$, explicitly

$$\nu_A = \{\nu_{\text{tes}}, \nu_{\text{jes}}, \nu_{\text{met}}, \nu_{\text{tes}}^2, \nu_{\text{jes}}^2, \nu_{\text{met}}^2, \nu_{\text{tes}}\nu_{\text{jes}}, \nu_{\text{jes}}\nu_{\text{met}}, \nu_{\text{tes}}\nu_{\text{met}}\}. \tag{11}$$

Note also that we are free to drop individual terms from Eq. (10) if it simplifies the analysis. Finally, Eq. (10) extends to an arbitrary number of nuisance parameters and polynomial order straightforwardly so that the following discussion and the ML strategy are generally applicable.

We determine the constant coefficients $\hat{\Delta}_{i,p,A}$ from simulated yields $\sigma_{i,p}(\boldsymbol{\nu}_{\text{calib}})$, where $\boldsymbol{\nu}_{\text{calib}} = \mathbf{0}$, as usual, corresponds to the nominal calibration of the training data. The three-dimensional parameter points $\boldsymbol{\nu}_{\text{calib}}$ come from a training set $\mathcal{V}$, which we must choose such that all coefficients $\hat{\Delta}_{i,p,A}$ can be determined. Given $\mathcal{V}$, we obtain the constants by minimizing

$$\chi^2 = \sum_{\boldsymbol{\nu}\in\mathcal{V}} \left(\log \frac{\sigma_{i,p}(\boldsymbol{\nu})}{\sigma_{i,p}(\mathbf{0})} - \nu_A \hat{\Delta}_{i,p,A} \right)^2 \tag{12}$$

independently for each bin i and process p. Here, we omit the subscript "calib" for readability. The solution is obtained analytically by differentiation [21] as

$$\hat{\Delta}_{i,p,A} = \left[\sum_{\boldsymbol{\nu}\in\mathcal{V}} \boldsymbol{\nu}\boldsymbol{\nu}^{\mathrm{T}} \right]^{-1}_{AB} \left[\sum_{\boldsymbol{\nu}\in\mathcal{V}} \boldsymbol{\nu} \log \frac{\sigma_{i,p}(\boldsymbol{\nu})}{\sigma_{i,p}(\mathbf{0})} \right]_B . \tag{13}$$

From the first factor in Eq. (13), it follows that the condition on $\mathcal{V}$ is that the 9×9 matrix $[\sum_{\boldsymbol{\nu}\in\mathcal{V}} \boldsymbol{\nu}\boldsymbol{\nu}^{\mathrm{T}}]_{AB}$ is invertible. As before, A and B index the nine basis terms used in the polynomial expansion. It is always possible to find a suitable $\mathcal{V}$ [21]; specifically, we select $\nu = -3, -2, -1, 0, 1, 2, 3$ for `tes` and `jes`, and $\nu_{\text{met}} = 0$, 1, 2, 3. From all combinations of these integer values, we retain the 44 parameter vectors also satisfying $|\nu_{\text{tes}}| + |\nu_{\text{jes}}| + |\nu_{\text{met}}| \leq 3$. The resulting $\mathcal{V}$ is sufficient to evaluate Eq. (13).

There is considerable flexibility in this procedure. Note, for example, that we can estimate $\hat{\Delta}_{i,p,A}$ separately for groups of nuisance parameters, provided that we do not include mixed terms that spoil the factorization of the effects of these groups in Eq. (10). In fully realistic settings, this allows the estimation of systematic effects independently of each other.

This completes the construction of the binned likelihood function, summarized as follows. The background normalization uncertainties are implemented using Eqs. (8a)–(8c), where $\sigma_{i,p}(\boldsymbol{\nu}_{\text{calib}})$ is computed as

$$\sigma_{i,p}(\boldsymbol{\nu}_{\text{calib}}) = \exp(\nu_A \hat{\Delta}_{i,p,A}) \sigma_{i,p}(1, \mathbf{0}). \tag{14}$$

The three-dimensional $\boldsymbol{\nu}_{\text{calib}}$ leads to nine coefficients $\hat{\Delta}_{i,p,A}$ up to quadratic order per bin and process, each labeled by A as given in Eq. (10). The $\hat{\Delta}_{i,p,A}$ are determined from Eq. (13), using systematically varied yields evaluated at an adequate set of parameter points $\mathcal{V}$. With these quantities, the test statistic in Eq. (4) and the binned likelihood ratio in Eq. (6) can be evaluated.

Table III summarizes the parametrization of systematic dependencies in the cross sections of the five regions defined in Table II up to log-quadratic order and each region treated as a single bin. It shows that ν_{tes} and ν_{jes} have impacts of $10^{-3} - 10^{-2}$ dominated by the log-linear term, while ν_{met} is smaller overall but with a non-negligible log-quadratic term. Mixed contributions are generally small.

The unbinned estimates will follow the same procedure except that the constants $\hat{\Delta}_{i,p,A}$ become $\boldsymbol{x}$-dependent functions which we implement with neural networks and train by minimizing a cross-entropy (CE) loss function instead of a χ^2.

C. Unbinned likelihoods

Comparing Eqs. (5) to (6), the unbinned likelihood can be viewed as the continuum limit of the binned Poisson likelihood[2] where the granularity of the cross-section ratio in the logarithm becomes arbitrarily fine [16]. In analogy to the binned case, we therefore define an additive fully differential model

$$\mathrm{d}\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu}) = \mu\, \mathrm{d}\sigma_{\mathrm{H}}(\boldsymbol{x}|\boldsymbol{\nu}_{\text{calib}}) + \sum_{p=\mathrm{Z,t\bar{t},VV}} \mathrm{d}\sigma_p(\boldsymbol{x}|\boldsymbol{\nu}). \tag{15}$$

The relation to the binned Poisson yields is given by

$$\sigma_{i,p}(\boldsymbol{\nu}) = \int_{\Delta \boldsymbol{x}_i} \frac{\mathrm{d}\sigma_p(\boldsymbol{x}|\boldsymbol{\nu})}{\mathrm{d}\boldsymbol{x}} \mathrm{d}\boldsymbol{x} \tag{16}$$

with $\Delta\boldsymbol{x}_i$ the phase-space region associated with bin i. It is important that Eq. (15) is available via event generators only in the "forward" mode; we may specify $\boldsymbol{\nu}$ and then obtain event samples, but we cannot use those to evaluate, e.g., the DCR in Eq. (5) for newly observed data. This was, in fact, true already in the binned case: To obtain a parametrization that is continuous in $\boldsymbol{\nu}$, we used the log-polynomial ansatz in Eq. (9) and had to solve a system of equations for the per-bin constants $\hat{\Delta}_{i,p,A}$ based on training data at certain values of $\boldsymbol{\nu}$. A similar interpolation is implemented in COMBINE [5].

Our next goal is to define a fully unbinned model. We start with specifying the analytic dependence of the normalization-type systematic uncertainties in complete analogy with the binned case as

[2]A first-principles derivation is provided in, e.g., Ref. [21].

TABLE III. Systematic dependence of the four processes ($H \to \tau\tau$, $Z \to \tau\tau$, $t\bar{t}$, and VV) in the five regions. Each row represents the coefficients of a quadratic polynomial describing the impact of the nuisance parameters ν_{tes}, ν_{jes}, and ν_{met}.

	$\nu_{tes} \times 10^{-3}$	$\nu_{jes} \times 10^{-3}$	$\nu_{met} \times 10^{-5}$	$\nu_{tes}^2 \times 10^{-5}$	$\nu_{jes}^2 \times 10^{-5}$	$\nu_{met}^2 \times 10^{-5}$	$\nu_{tes}\nu_{jes} \times 10^{-5}$	$\nu_{jes}\nu_{met} \times 10^{-5}$	$\nu_{tes}\nu_{met} \times 10^{-5}$
`lowMT-VBFJet`									
$H \to \tau\tau$	6.03	13.0	−3.40	−9.91	−16.5	−9.15	4.63	5.28	−0.451
$Z \to \tau\tau$	9.89	22.0	2.57	−12.8	−16.0	−6.87	7.25	2.03	−1.78
$t\bar{t}$	8.78	7.15	−6.36	−8.89	−18.1	−13.7	2.38	0.466	−3.84
VV	11.0	20.4	11.3	−6.43	−4.85	−6.97	2.60	−59.9	14.5
`highMT-VBFJet`									
$H \to \tau\tau$	−1.95	8.85	29.3	23.3	67.7	148	−59.3	−10.3	2.60
$Z \to \tau\tau$	8.32	3.60	22.5	34.8	149	333	−119	−11.7	29.5
$t\bar{t}$	6.25	6.50	−0.402	−6.40	−8.06	14.6	−2.05	−1.05	−0.172
VV	6.57	16.5	−24.6	−4.91	−18.7	13.4	4.71	3.45	−10.9
`highMT-noVBFJet`									
$H \to \tau\tau$	−10.2	−3.83	−18.4	8.18	27.9	351	−18.6	2.41	−6.94
$Z \to \tau\tau$	−7.65	−2.56	46.7	2.35	46.4	930	−37.5	5.82	20.2
$t\bar{t}$	5.32	−24.0	−23.9	−5.14	9.06	7.79	−3.07	−9.88	3.33
VV	6.99	−3.16	7.56	−11.7	−2.86	5.61	−6.29	8.78	5.92
`lowMT-noVBFJet-ptH100`									
$H \to \tau\tau$	5.38	2.49	5.27	−7.84	−11.9	24.4	9.03	−4.20	1.61
$Z \to \tau\tau$	7.06	12.9	−9.71	−8.88	−19.9	51.6	9.48	−5.45	−1.04
$t\bar{t}$	7.50	−2.96	82.6	−4.46	−1.11	4.63	12.1	26.5	5.17
VV	10.2	9.49	341	27.6	−19.0	−72.5	9.88	66.8	33.3
`lowMT-noVBFJet-ptH0to100`									
$H \to \tau\tau$	6.42	−4.83	−0.726	−11.1	−0.15	−16.2	−0.20	0.205	−0.44
$Z \to \tau\tau$	12.5	−1.34	−2.28	−22.4	−0.644	−4.67	0.137	−0.055	−1.57
$t\bar{t}$	8.40	−28.4	15.0	−10.9	2.93	−11.3	−3.67	6.35	−8.47
VV	14.2	−4.07	−4.42	−10.8	−0.172	−8.79	5.44	0.067	1.59

$$d\sigma_Z(\boldsymbol{x}|\boldsymbol{\nu}) = (1+\alpha_{bkg})^{\nu_{bkg}} d\sigma_Z(\boldsymbol{x}|\boldsymbol{\nu}_{calib}), \tag{17a}$$

$$d\sigma_{t\bar{t}}(\boldsymbol{x}|\boldsymbol{\nu}) = (1+\alpha_{bkg})^{\nu_{bkg}} (1+\alpha_{t\bar{t}})^{\nu_{t\bar{t}}} d\sigma_{t\bar{t}}(\boldsymbol{x}|\boldsymbol{\nu}_{calib}), \tag{17b}$$

$$d\sigma_{VV}(\boldsymbol{x}|\boldsymbol{\nu}) = (1+\alpha_{bkg})^{\nu_{bkg}} (1+\alpha_{VV})^{\nu_{VV}} d\sigma_{VV}(\boldsymbol{x}|\boldsymbol{\nu}_{calib}). \tag{17c}$$

We note that Eq. (5) only asks for the DCR, which we can write as

$$\frac{d\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu})}{d\sigma(\boldsymbol{x}|1,\boldsymbol{0})} = \mu \frac{d\sigma_H(\boldsymbol{x}|\boldsymbol{\nu}_{calib})}{d\sigma(\boldsymbol{x}|1,\boldsymbol{0})} + (1+\alpha_{bkg})^{\nu_{bkg}} \left(\frac{d\sigma_Z(\boldsymbol{x}|\boldsymbol{\nu}_{calib})}{d\sigma(\boldsymbol{x}|1,\boldsymbol{0})} + (1+\alpha_{t\bar{t}})^{\nu_{t\bar{t}}} \frac{d\sigma_{t\bar{t}}(\boldsymbol{x}|\boldsymbol{\nu}_{calib})}{d\sigma(\boldsymbol{x}|1,\boldsymbol{0})} + (1+\alpha_{VV})^{\nu_{VV}} \frac{d\sigma_{VV}(\boldsymbol{x}|\boldsymbol{\nu}_{calib})}{d\sigma(\boldsymbol{x}|1,\boldsymbol{0})} \right). \tag{18}$$

We expand the DCRs for $p = H, Z, t\bar{t}, VV$ on the right-hand side (rhs) as

$$\frac{d\sigma_p(\boldsymbol{x}|\boldsymbol{\nu}_{calib})}{d\sigma(\boldsymbol{x}|1,\boldsymbol{0})} = \frac{d\sigma_p(\boldsymbol{x}|\boldsymbol{\nu}_{calib})}{d\sigma_p(\boldsymbol{x}|\boldsymbol{0})} \frac{d\sigma_p(\boldsymbol{x}|\boldsymbol{0})}{d\sigma(\boldsymbol{x}|1,\boldsymbol{0})} \simeq \hat{S}_p(\boldsymbol{x}|\boldsymbol{\nu}_{calib}) \hat{g}_p(\boldsymbol{x}) \tag{19}$$

where the second line uses the ML approximations

$$\hat{g}_p(\boldsymbol{x}) \simeq \frac{d\sigma_p(\boldsymbol{x}|\boldsymbol{0})}{d\sigma(\boldsymbol{x}|1,\boldsymbol{0})} = \frac{d\sigma_p(\boldsymbol{x}|\boldsymbol{0})}{\sum_q d\sigma_q(\boldsymbol{x}|\boldsymbol{0})} \quad \text{and} \tag{20}$$

$$\hat{S}_p(\boldsymbol{x}|\boldsymbol{\nu}_{calib}) \simeq \frac{d\sigma_p(\boldsymbol{x}|\boldsymbol{\nu}_{calib})}{d\sigma_p(\boldsymbol{x}|\boldsymbol{0})}. \tag{21}$$

The sum over q extends over H, Z, $t\bar{t}$, and VV. Note that none of these quantities depend on μ or the normalization-type nuisance parameters. Because the denominator in Eq. (20) is just the total nominal differential cross section, the quantity $\hat{g}_p(\boldsymbol{x})$ predicts the $\boldsymbol{x}$-dependent DCR of process p to the total differential cross section. Similarly, the parametric estimator $\hat{S}_p(\boldsymbol{x}|\boldsymbol{\nu}_{calib})$ predicts the relative

dependence of the differential cross section of a process p on the nuisance parameters $\boldsymbol{\nu}_{\text{calib}}$ and $\boldsymbol{x}$.

In summary, our ML surrogate model of the DCR is

$$\begin{aligned}\frac{\mathrm{d}\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0})} \simeq \hat{R}(\boldsymbol{x}|\mu,\boldsymbol{\nu}) = \mu\hat{g}_{\text{H}}(\boldsymbol{x})\hat{S}_{\text{H}}(\boldsymbol{x}|\boldsymbol{\nu}_{\text{calib}}) \\ + (1+\alpha_{\text{bkg}})^{\nu_{\text{bkg}}}\big(\hat{g}_{\text{Z}}(\boldsymbol{x})\hat{S}_{\text{Z}}(\boldsymbol{x}|\boldsymbol{\nu}_{\text{calib}}) \\ + (1+\alpha_{\text{t}\bar{\text{t}}})^{\nu_{\text{t}\bar{\text{t}}}}\hat{g}_{\text{t}\bar{\text{t}}}(\boldsymbol{x})\hat{S}_{\text{t}\bar{\text{t}}}(\boldsymbol{x}|\boldsymbol{\nu}_{\text{calib}}) \\ + (1+\alpha_{\text{VV}})^{\nu_{\text{VV}}}\hat{g}_{\text{VV}}(\boldsymbol{x})\hat{S}_{\text{VV}}(\boldsymbol{x}|\boldsymbol{\nu}_{\text{calib}})\big). \qquad (22)\end{aligned}$$

We devise training strategies for $\hat{g}_p(\boldsymbol{x})$ and $\hat{S}_p(\boldsymbol{x}|\boldsymbol{\nu}_{\text{calib}})$ in Sec. IV.

While our DCR surrogate captures the full dependence on all seven model parameters, practical applications typically require incremental refinement. Binned analysis development starts with the dominant backgrounds and leading uncertainties and then adds subleading components and systematic effects to the Poisson yields. Our approach extends this staged workflow to unbinned modeling, allowing the gradual inclusion of new effects, in analogy to the binned case.

For instance, unmodeled systematics can be incorporated into Eq. (22) by multiplying with additional factors of $\hat{S}$, using per-process approximations as in Eq. (21). If mixed terms between the new and nominal systematics are negligible, this single factor suffices, and the existing ML surrogates remain valid.

Similarly, unaccounted background contributions can be included. For example, we can refine the existing differential cross section by adding a new component $\mathrm{d}\sigma_{\text{new}}$,

$$\mathrm{d}\sigma'(\boldsymbol{x}|\mu,\boldsymbol{\nu}) = \mathrm{d}\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu}) + \mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\boldsymbol{\nu}). \qquad (23)$$

We can then rewrite the updated DCR in terms of the original model as

$$\begin{aligned}\frac{\mathrm{d}\sigma'(\boldsymbol{x}|\mu,\boldsymbol{\nu})}{\mathrm{d}\sigma'(\boldsymbol{x}|1,\mathbf{0})} &= \frac{\mathrm{d}\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu}) + \mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0}) + \mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\mathbf{0})} \\ &= \frac{\frac{\mathrm{d}\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0})} + \frac{\mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\boldsymbol{\nu})}{\mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\mathbf{0})}\frac{\mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\mathbf{0})}{\mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0})}}{1 + \frac{\mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\mathbf{0})}{\mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0})}} \\ &\simeq \frac{\hat{R}(\boldsymbol{x}|\mu,\boldsymbol{\nu}) + \hat{S}_{\text{new}}(\boldsymbol{x}|\boldsymbol{\nu})\hat{g}_{\text{new}}(\boldsymbol{x})}{1+\hat{g}_{\text{new}}(\boldsymbol{x})} \qquad (24)\end{aligned}$$

where the only additional surrogates needed are

$$\begin{aligned}\hat{g}_{\text{new}}(\boldsymbol{x}) &\simeq \frac{\mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\mathbf{0})}{\mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0})}, \quad \text{and} \\ \hat{S}_{\text{new}}(\boldsymbol{x}|\boldsymbol{\nu}) &\simeq \frac{\mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\boldsymbol{\nu})}{\mathrm{d}\sigma_{\text{new}}(\boldsymbol{x}|\mathbf{0})}. \qquad (25)\end{aligned}$$

Expressing the refined DCR in Eq. (24) entirely in terms of the original DCR and only two surrogates for the new component mirrors the flexibility of the binned approach and facilitates the refinable modeling for unbinned analyses.

D. Inclusive cross-section parametrization

If Eqs. (20) and (21) are available, we can efficiently evaluate the inclusive cross-section difference, which is the final ingredient required in Eq. (5). Using a nominal simulation at $\mu = 1$ and $\boldsymbol{\nu} = 0$, i.e., an event sample $\mathcal{D}_{1,\mathbf{0}} \sim \mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0})$, we obtain

$$\begin{aligned}\mathcal{L}\big(\sigma(\mu,\boldsymbol{\nu}) - \sigma(1,\mathbf{0})\big) &= \mathcal{L}\int\left(\frac{\mathrm{d}\sigma(\boldsymbol{x}|\mu,\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0})} - 1\right)\frac{\mathrm{d}\sigma(\boldsymbol{x}|1,\mathbf{0})}{\mathrm{d}\boldsymbol{x}}\mathrm{d}\boldsymbol{x} \\ &\approx \sum_{\{w_i,\boldsymbol{x}_i\}\in\mathcal{D}_{1,\mathbf{0}}} w_i(\hat{R}(\boldsymbol{x}_i|\mu,\boldsymbol{\nu}) - 1). \qquad (26)\end{aligned}$$

The event weights w_i implement the cross-section normalization such that

$$\sum_{D^{(p)}_{1,\mathbf{0}}} w_i = \mathcal{L}\sigma_p(\mathbf{0}), \qquad (27)$$

where $D^{(p)}_{1,\mathbf{0}}$ is the sample with nominal simulation of process p in Table 1 of Ref. [22]. Because $\sum_p \hat{g}_p(\boldsymbol{x}) = 1$ by construction, it follows from Eq. (22) that $\hat{R}(\boldsymbol{x}|1,\mathbf{0}) = 1$, which expresses the trivial fact that the DCR of $(1,\mathbf{0})$ to the reference hypothesis, which is characterized by the same parameters, is one.

This completes the construction of the unbinned surrogate model, which is, in principle, ready for inference once $\hat{g}_p(\boldsymbol{x})$ and $\hat{\Delta}_{p,A}(\boldsymbol{x})$ are available.

In practice, Eq. (26) must be evaluated using the training dataset. During the profiling of nuisance parameters when computing q_μ, many evaluations are required. Although the surrogate functions $\hat{g}_p(\boldsymbol{x})$ and $\hat{\Delta}_{p,A}(\boldsymbol{x})$ are independent of the model parameters and can thus be precomputed and stored even for large samples, the computational cost of repeatedly evaluating Eq. (26) can still become prohibitive for large samples. Additionally, the expression involves products of several terms that are close to unity near the nominal parameter point, $(\mu,\boldsymbol{\nu}) = (1,\mathbf{0})$, potentially causing numerical instabilities. To mitigate the computational cost and to improve numerical stability, we separate the evaluation into a numerically stable, analytically calculable part and interpolate the residual dependence on $\boldsymbol{\nu}_{\text{calib}}$. Specifically, we rearrange Eq. (26) as

$$\sum_{\{w_i,\boldsymbol{x}_i\}\in\mathcal{D}_{1,\boldsymbol{\nu}}}\big(\hat{R}(\boldsymbol{x}_i|\mu,\boldsymbol{\nu})-1\big)w_i=\mu\mathrm{CSI}_{\mathrm{H}}(\boldsymbol{\nu}_{\mathrm{calib}})+(1+\alpha_{\mathrm{bkg}})^{\nu_{\mathrm{bkg}}}\mathrm{CSI}_{\mathrm{Z}}(\boldsymbol{\nu}_{\mathrm{calib}})+(1+\alpha_{\mathrm{bkg}})^{\nu_{\mathrm{bkg}}}(1+\alpha_{\mathrm{t\bar{t}}})^{\nu_{\mathrm{t\bar{t}}}}\mathrm{CSI}_{\mathrm{t\bar{t}}}(\boldsymbol{\nu}_{\mathrm{calib}})$$
$$+(1+\alpha_{\mathrm{bkg}})^{\nu_{\mathrm{bkg}}}(1+\alpha_{\mathrm{VV}})^{\nu_{\mathrm{VV}}}\mathrm{CSI}_{\mathrm{VV}}(\boldsymbol{\nu}_{\mathrm{calib}})+(\mu-1)\mathrm{CSIC}_{\mathrm{H}}+\big((1+\alpha_{\mathrm{bkg}})^{\nu_{\mathrm{bkg}}}-1\big)\mathrm{CSIC}_{\mathrm{Z}}$$
$$+\big((1+\alpha_{\mathrm{bkg}})^{\nu_{\mathrm{bkg}}}(1+\alpha_{\mathrm{t\bar{t}}})^{\nu_{\mathrm{t\bar{t}}}}-1\big)\mathrm{CSIC}_{\mathrm{t\bar{t}}}+\big((1+\alpha_{\mathrm{bkg}})^{\nu_{\mathrm{bkg}}}(1+\alpha_{\mathrm{VV}})^{\nu_{\mathrm{VV}}}-1\big)\mathrm{CSIC}_{\mathrm{VV}}\quad(28)$$

with

$$\mathrm{CSI}_p(\boldsymbol{\nu}_{\mathrm{calib}})=\sum_{\{w_i,\boldsymbol{x}_i\}\in\mathcal{D}_{1,\mathbf{0}}}w_i\hat{g}_p(\boldsymbol{x})\big(\hat{S}_p(\boldsymbol{x}|\boldsymbol{\nu}_{\mathrm{calib}})-1\big),\quad(29a)$$

$$\mathrm{CSIC}_p=\sum_{\{w_i,\boldsymbol{x}_i\}\in\mathcal{D}_{1,\mathbf{0}}}w_i\hat{g}_p(\boldsymbol{x}).\quad(29b)$$

Equation (28) is a cross-section interpolation (CSI) that reflects the known analytic dependence on μ and the normalization-type uncertainties. It is written such that training-data sums over many small terms remain numerically stable: the parenthesis in the last four terms in Eq. (28) and on the rhs of Eq. (29a) can be accurately determined by the `expm1` method in NUMPY [44]. The dependence on $\boldsymbol{\nu}_{\mathrm{calib}}$ is encoded in Eq. (29a), which we evaluate on a fine grid within three standard deviations for the three nuisance parameters in $\boldsymbol{\nu}_{\mathrm{calib}}$ and a coarser grid that fills the entire allowed parameter space. In total 6561 grid points are used and a three-dimensional quintic spline, implemented in the SCIPY-package [45], provides a fast interpolation separately for each process. The constant terms CSIC_p need only be evaluated once. We emphasize that this interpolation removes the main bottleneck in CPU consumption during the profiling and is numerically stable even for large simulated datasets.

If we sum the unbinned log-likelihood over the $N_r=3$ regions marked with UB in Table II, we arrive at

$$-\frac{1}{2}u_{\mathrm{UB}}(\mathcal{D}|\mu,\boldsymbol{\nu})=\sum_{r=1}^{N_r}\left[-\mathcal{L}\big(\sigma_r(\mu,\boldsymbol{\nu})-\sigma_r(1,\mathbf{0})\big)+\sum_{\boldsymbol{x}_i\in\mathcal{D}}\log\left(\frac{\mathrm{d}\sigma_r(\boldsymbol{x}_i|\mu,\boldsymbol{\nu})}{\mathrm{d}\sigma_r(\boldsymbol{x}_i|1,\mathbf{0})}\right)\right]$$
$$\approx\sum_{r=1}^{N_r}\left[-\sum_{\substack{\{w_i,\boldsymbol{x}_i\}\in\\ \mathcal{D}_{1,\mathbf{0}}\cap r}}w_i(\hat{R}_r(\boldsymbol{x}_i|\mu,\boldsymbol{\nu})-1)+\sum_{\boldsymbol{x}_i\in\mathcal{D}}\log\hat{R}_r(\boldsymbol{x}_i|\mu,\boldsymbol{\nu})\right]\quad(30)$$

where we denote by $\mathcal{D}_{1,\mathbf{0}}\cap r$ a set of *simulated* events at $(\mu,\boldsymbol{\nu})=(1,\mathbf{0})$ that satisfy the selection requirements of region r. The ratio in the logarithm in the first term pertains to region r, i.e., it is the same as in Eq. (18). In the second line, the DCR surrogate $\hat{R}_r(\boldsymbol{x}|\mu,\boldsymbol{\nu})$ from Eq. (22) is the same in both terms with multiclassifiers and neural network parametrizations learned separately for each region r. The observed dataset $\mathcal{D}$ only enters in the second term in both lines.

IV. MACHINE LEARNING

A. Calibrated multiclassifier

For estimating the DCR with $\hat{g}_p(\boldsymbol{x})$ in Eq. (20), we could use the CE loss function and train a multiclassifier with simulated samples $\mathcal{D}_p\sim\mathrm{d}\sigma_p(\boldsymbol{x}|\mathbf{0})$ at nominal values for the nuisance parameters. In Table II, however, we observe relative normalizations spanning three orders of magnitude. Therefore, a cross-section-weighted training is highly imbalanced.

To stabilize the training, we equalize the normalization by dividing each process by its total cross section. The resulting CE multiclassifier loss function is

$$L_{\mathrm{CE}}[\hat{f}_p]=-\sum_{p=1}^{N_p}\int\frac{\mathrm{d}\sigma_p(\boldsymbol{x})}{\sigma_p}\log\big(\hat{f}_p(\boldsymbol{x})\big)\quad(31)$$

where $\hat{f}_p(\boldsymbol{x})$ is a neural network with 28 input nodes and $N_p=4$ outputs. A Softmax output layer guarantees

$$\sum_{p=1}^{N_p}\hat{f}_p(\boldsymbol{x})=1.\quad(32)$$

The empirical version, suitable for implementation in computer code, is

$$L_{\mathrm{CE}}[\hat{f}_p]\approx-\sum_{p=1}^{N_p}\sum_{\{\boldsymbol{x}_i,w_i\}\in\mathcal{D}_p}\frac{w_i}{\mathcal{L}\sigma_p}\log\big(\hat{f}_p(\boldsymbol{x}_i)\big).\quad(33)$$

Functional minimization of Eq. (31) and treating Eq. (32) with a Lagrange multiplier shows that the minimum is attained for

$$\hat{f}_p^*(\boldsymbol{x})\simeq\frac{\mathrm{d}\sigma_p(\boldsymbol{x})/\sigma_p}{\sum_{q=1}^{N_p}\mathrm{d}\sigma_q(\boldsymbol{x})/\sigma_q},\quad(34)$$

which is nothing but the likelihood ratio, learned by a conventional multiclassifier.

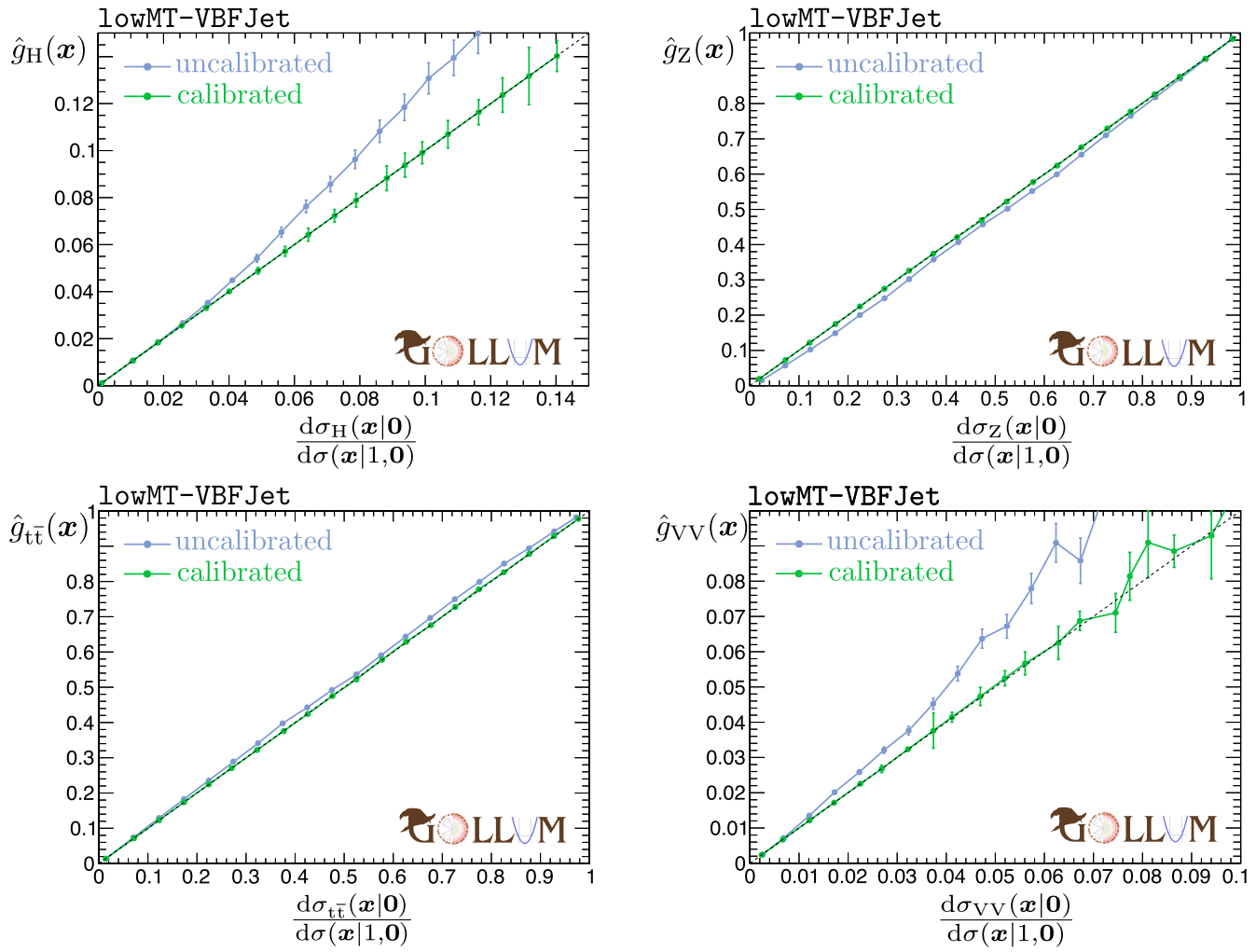

FIG. 2. Calibration of the DCR in the `lowMT-VBFJet` region of the four processes in FAIR-HUC: H → ττ (top left), Z → ττ (top right), tt̄ (bottom left), and VV (bottom right).

To convert to the DCR, we invert the normalization that was included in the loss and arrive at

$$\hat{g}_p^*(\boldsymbol{x}) = \frac{\hat{f}_p^*(\boldsymbol{x})\sigma_p}{\sum_q \hat{f}_q^*(\boldsymbol{x})\sigma_q} \simeq \frac{\mathrm{d}\sigma_p(\boldsymbol{x}|\mathbf{0})}{\sum_q \mathrm{d}\sigma_q(\boldsymbol{x}|\mathbf{0})} \tag{35}$$

which now exactly corresponds to Eq. (20).

For FAIR-HUC, we train one multiclassifier for each of the three regions marked with UB in Table II using TENSORFLOW 2.11 [46]. All neural networks have a straightforward architecture with 28 input nodes, corresponding to the features listed in Table I, and four Softmax output nodes. We selected the three-layer, 64-unit architecture and its regularization parameters by first running a coarse grid search over hidden-layer sizes (32, 64, 128) and depths (2–4 layers) on a held-out validation split, then refining dropout (0.1–0.5) and penalties (`L1` = 0.1, `L2` = 0.05) through random search. All models were trained for up to 1000 epochs with early stopping (patience 10 epochs) monitored on validation loss.

An important step is the calibration of the multiclassifier $\hat{f}_p(\boldsymbol{x})$ and the corresponding DCR estimate $\hat{g}_p(\boldsymbol{x})$. A classifier is calibrated if its output $\hat{f}_p(\boldsymbol{x})$ accurately reflects the true fraction of events in the vicinity of $\boldsymbol{x}$ belonging to class p. Consequently, a value $\hat{g}_p(\boldsymbol{x}) \approx 0.1$ implies that approximately 10% of cross-section-weighted events near $\boldsymbol{x}$ originate from process p. Modern classifiers are typically not well calibrated [47,48], as illustrated in Fig. 2. For instance, the predicted signal DCR tends to exceed the true expected DCR.

Binary classifiers can be recalibrated after training by applying a monotonic function—such as isotonic regression (IREG)—to adjust the classifier outputs without altering decision boundaries [48]. Among several extensions of the isotonic regression to the multiclass case, we chose the following: first, we independently calibrate each of the four $\hat{g}_p(\boldsymbol{x})$ outputs using one-vs-rest binary isotonic

regression, implemented via `sklearn` [49]. Then, we retain the calibrated output for the signal-vs-background class ($p = \mathrm{H}$) and rescale the calibrated outputs of the three background classes so that the four DCR outputs sum to unity

$$\hat{g}_p(\boldsymbol{x}) = \begin{cases} \mathrm{IREG}(g^*_{\mathrm{H}}(\boldsymbol{x})), & \text{if } p = \mathrm{H} \\ \frac{\mathrm{IREG}(g^*_p(\boldsymbol{x}))(1-\mathrm{IREG}(g^*_{\mathrm{H}}(\boldsymbol{x})))}{\sum_{q=\mathrm{Z},\mathrm{t\bar{t}},\mathrm{VV}} \mathrm{IREG}(g^*_q(\boldsymbol{x}))}, & \text{otherwise.} \end{cases} \tag{36}$$

Prioritizing the calibration of the signal class ensures accurate estimation of the signal DCR, reducing biases from the dominant backgrounds and thereby improving precision in the extraction of μ.

Figure 2 illustrates the effect of calibration on the DCR distributions. The corrections for the $\mathrm{H} \to \tau\tau$ signal and the VV background are substantial. The corrections to the $\hat{g}_{\mathrm{Z}}$ are at most 2%. However, the large cross section of this background entails a significant degradation of the measurement of μ if the miscalibration is not corrected.

Figures 3–5 demonstrate the closure of the calibrated surrogate for several distributions in the `lowMT-VBFJet` region. The simulated (true) distributions of $p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$, $m_{\mathrm{vis}}(\tau, \ell)$, and $\Delta\eta^{j_1 j_2}$ (dashed lines) for the four processes $\mathrm{H} \to \tau\tau$, $\mathrm{Z} \to \tau\tau$, $\mathrm{t\bar{t}}$, and VV are compared to the predictions obtained from the surrogate $\hat{g}_p(\boldsymbol{x})$ (solid lines). Specifically, the surrogate predictions shown correspond to binned yields computed as $\sum_{\boldsymbol{x}_i \in \mathrm{bin}} w_i \hat{g}_p(\boldsymbol{x}_i)$. The left column shows spectra normalized to the expected event counts, while the right column compares their shapes. After calibration, the agreement between surrogate predictions and simulated spectra is excellent, spanning several orders of magnitude.

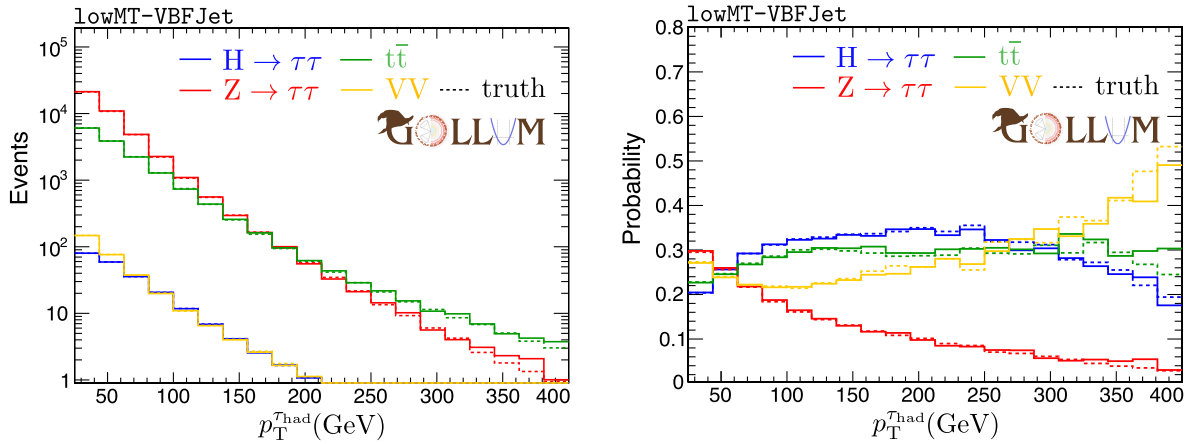

FIG. 3. Comparison of the true spectra of $p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$ (dashed lines) for the four processes in the `lowMT-VBFJet` region: $\mathrm{H} \to \tau\tau$ (blue), $\mathrm{Z} \to \tau\tau$ (red), $\mathrm{t\bar{t}}$ (green), and VV (yellow) with the prediction from the ML surrogates (solid lines). The left column shows the spectra normalized to the expected number of events, while the right column shows a comparison of the per-bin fractions of the area-normalized shapes of each process.

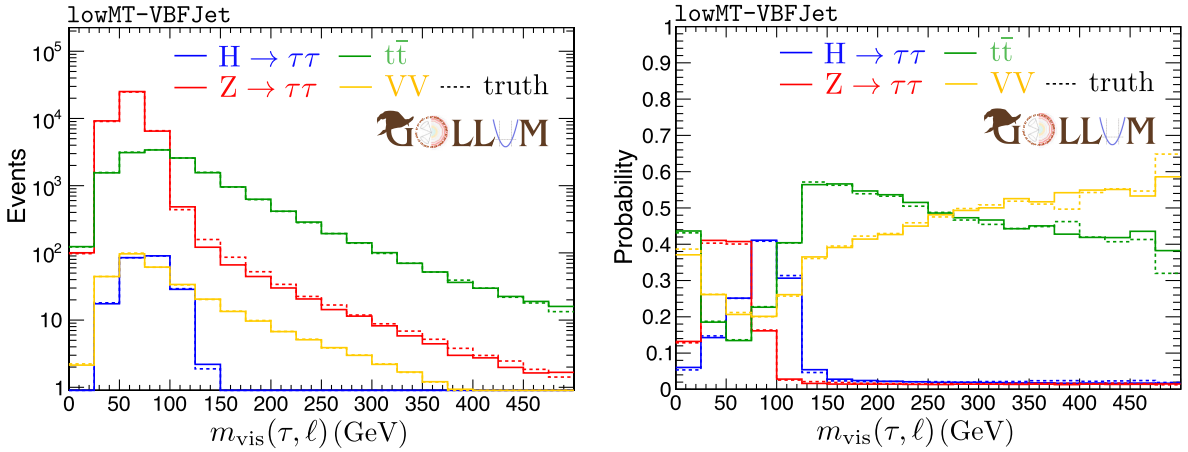

FIG. 4. Same as Fig. 3, but for $m_{\mathrm{vis}}(\tau, \ell)$.

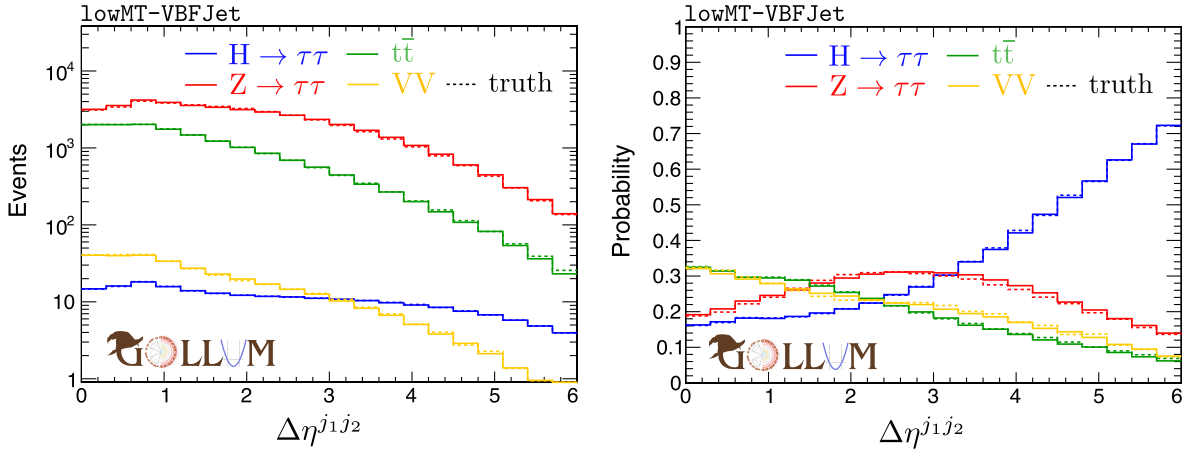

FIG. 5. Same as Fig. 3, but for $\Delta\eta^{j_1 j_2}$.

B. Neural network parametrization

Next, we train a parametric neural network that estimates the nuisance parameter dependence of the DCR in Eq. (21), following in broad terms Ref. [21] but using neural networks instead of a boosted tree algorithm. In this section, we remove notational clutter and solve the generic training task

$$\hat{S}(\boldsymbol{x}|\boldsymbol{\nu}) \simeq \frac{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{0})}. \tag{37}$$

The first step is to choose a sufficiently large set of nuisance parameter vectors $\boldsymbol{\nu} \in \mathcal{V}$, where the same considerations apply as in Sec. III B, and thus we keep the choice discussed there. For every $\boldsymbol{\nu} \in \mathcal{V}$, we obtain a systematically varied training sample $\mathcal{D}_{\boldsymbol{\nu}}$. The nominal sample $\mathcal{D}_{\boldsymbol{0}}$ is removed from $\mathcal{V}$ and plays a special role.

If we train a binary classifier between a specific $\mathcal{D}_{\boldsymbol{\nu}}$ and the nominal with the CE loss[3]

$$L_{\mathrm{CE}}[\hat{f}] = -\int \mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{0}) \log \hat{f}(\boldsymbol{x}) - \int \mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{\nu}) \log\bigl(1-\hat{f}(\boldsymbol{x})\bigr),$$

the network $\hat{f}(\boldsymbol{x})$ attains its minimum at

$$f^*(\boldsymbol{x}) = \left(1 + \frac{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{0})}\right)^{-1}. \tag{38}$$

[3]The FAIR-HUC datasets provide event weights that are normalized to the event yield $\mathcal{L}\sigma$, technically including an unspecified luminosity factor. Therefore, in principle, we only have access to $L_{\mathrm{CE}}[\hat{f}] = \mathcal{L}\int \mathrm{d}\sigma \cdots$. However, we can still drop this distracting constant in the notation because it does not affect the minimum.

While $f^*(\boldsymbol{x})$ is a monotonous function of the required DCR, it is not parametric in $\boldsymbol{\nu}$. To arrive at a fully parametric estimate, we combine the analytic structure in Eq. (38) with an ansatz for the $\boldsymbol{\nu}$-dependence chosen as $\exp(\nu_A \hat{\Delta}_A(\boldsymbol{x}))$ in analogy to Eq. (14). Taken together, the ansatz is

$$\hat{f}_{\boldsymbol{\nu}}(\boldsymbol{x}) = \frac{1}{1+\exp(\nu_A \hat{\Delta}_A(\boldsymbol{x}))}, \tag{39}$$

where the $\hat{\Delta}_A(\boldsymbol{x})$ are polynomial coefficient functions. These provide the unbinned generalization, replacing the per-bin constants in Eq. (13). They are implemented as neural networks and designed to learn the systematic effects continuously as a function of $\boldsymbol{x}$. The ensuing parametrization is also continuous in $\boldsymbol{\nu}$ and given by the combination $\nu_A \hat{\Delta}_A(\boldsymbol{x})$. The ν_A are polynomial in $\boldsymbol{\nu}$ and given by Eq. (10) as in the binned case.

Similar to Eq. (9), this expansion can encode arbitrary polynomial orders. For the FAIR-HUC, the nine output nodes of the neural network correspond to the nine components of $\hat{\Delta}_A(\boldsymbol{x})$ in the expansion

$$\begin{aligned}\nu_A \hat{\Delta}_A(\boldsymbol{x}) = {} & \nu_{\mathrm{tes}} \hat{\Delta}_{\mathrm{tes}}(\boldsymbol{x}) + \nu_{\mathrm{jes}} \hat{\Delta}_{\mathrm{jes}}(\boldsymbol{x}) + \nu_{\mathrm{met}} \hat{\Delta}_{\mathrm{met}}(\boldsymbol{x}) \\ & + \nu_{\mathrm{tes}}^2 \hat{\Delta}_{\mathrm{tes,tes}}(\boldsymbol{x}) + \nu_{\mathrm{jes}}^2 \hat{\Delta}_{\mathrm{jes,jes}}(\boldsymbol{x}) \\ & + \nu_{\mathrm{met}}^2 \hat{\Delta}_{\mathrm{met,met}}(\boldsymbol{x}) + \nu_{\mathrm{tes}} \nu_{\mathrm{jes}} \hat{\Delta}_{\mathrm{tes,jes}}(\boldsymbol{x}) \\ & + \nu_{\mathrm{jes}} \nu_{\mathrm{met}} \hat{\Delta}_{\mathrm{jes,met}}(\boldsymbol{x}) + \nu_{\mathrm{tes}} \nu_{\mathrm{met}} \hat{\Delta}_{\mathrm{tes,met}}(\boldsymbol{x}).\end{aligned} \tag{40}$$

Each network has 28 inputs, in accordance with the multiclassifier case. To train all coefficient functions simultaneously, we sum the CE loss over $\boldsymbol{\nu} \in \mathcal{V}$. A short calculation reveals

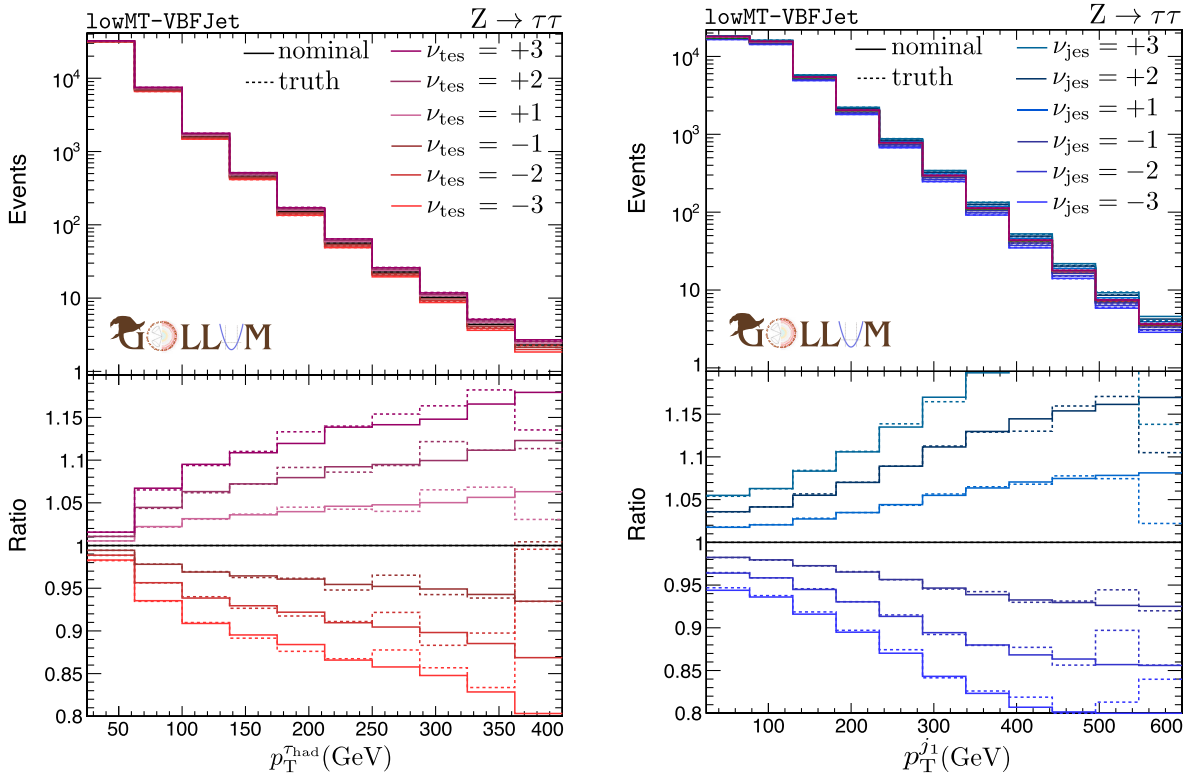

FIG. 6. Comparison of the $Z \to \tau\tau$ training data spectra in the `lowMT-VBFJet` region with the prediction from the parametric neural network. We show $p_T^{\tau_{had}}$ with `tes` variations (left) and $p_T^{j_1}$ with `jes` variations (right).

$$L[\hat{\Delta}_A] = \sum_{\nu\in\mathcal{V}}\left[\int \mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{0})\,\mathrm{Soft}^+\left(\nu_A\hat{\Delta}_A(\boldsymbol{x})\right) + \int \mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{\nu})\,\mathrm{Soft}^+\left(-\nu_A\hat{\Delta}_A(\boldsymbol{x})\right)\right] \tag{41a}$$

$$\approx \sum_{\nu\in\mathcal{V}}\left[\sum_{\{\boldsymbol{x}_i,w_i\}\in\mathcal{D}_0} w_i\,\mathrm{Soft}^+\left(\nu_A\hat{\Delta}_A(\boldsymbol{x}_i)\right) + \sum_{\{\boldsymbol{x}_i,w_i\}\in\mathcal{D}_\nu} w_i\,\mathrm{Soft}^+\left(-\nu_A\hat{\Delta}_A(\boldsymbol{x}_i)\right)\right], \tag{41b}$$

where $\mathrm{Soft}^+(x) = \log(1+\exp(x))$ and the last approximation defines the empirical version of the loss function. We emphasize that for each term in the sum over $\mathcal{V}$, the sample $\mathcal{D}_0$ in the first term is the same. The functional minimum of Eq. (41a) is given by

$$\hat{S}^*(\boldsymbol{x}|\boldsymbol{\nu}) = \exp\left(\nu_A\hat{\Delta}_A(\boldsymbol{x})\right) \simeq \frac{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{\nu})}{\mathrm{d}\sigma(\boldsymbol{x}|\boldsymbol{0})}, \tag{42}$$

which is the DCR we need for Eq. (21).

For FAIR-HUC, we train neural network parametrizations separately for each region and process, which amounts to a total of 12 networks. Each consists of two hidden layers with 128 nodes each and nine output nodes. Otherwise, the training closely follows the multiclassifier setup, using the Adam optimizer with a learning rate of 10^{-3}. We stress that the nuisance parameters do not enter the neural networks which solely predict $\hat{\Delta}_A(\boldsymbol{x})$. Instead, $\boldsymbol{\nu}$ enters Eq. (42).

Figure 6 shows examples of the parametric dependence learned for the $Z \to \tau\tau$ process in the `lowMT-VBFJet` selection. The approximations for `tes` and `jes` as functions of $p_T^{\tau_{had}}$ and $p_T^{j_1}$ are stable across the full phase space. At high tails, where stochastic fluctuations in the training data variations ("truth") become large, the model interpolates smoothly. We further assess the quality of the model through extensive toy-data studies in Sec. V D.

V. LIMIT SETTING

A. Binned control regions and test statistics

To exploit the constraining power of the control regions, we include the $N_{CR} = 3$ regions marked as CR in Table II. The `lowMT-noVBFJet0pt0to100` region is by far the largest. Tests showed that treating this region unbinned does not improve the measurement of μ, but as an inclusive control region, it mildly helps constrain because $(\mathcal{L}\sigma(1,\boldsymbol{0}))^{-1/2} \approx 10^{-3} \approx \alpha_{bkg}$.

The $\mathrm{t\bar{t}}$ and VV processes have larger uncertainties, motivating the inclusion of CRs based on `highMT-noVBFJet`. In this selection, we adopt the learned likelihood ratio $\hat{f}_{\mathrm{t\bar{t}}}$ and $\hat{f}_{\mathrm{VV}}$ from Eq. (34) to define CRs enriched in $\mathrm{t\bar{t}}$ and VV. An additional requirement of $\hat{f}_{\mathrm{t\bar{t}}} > 0.4$ defines `highMT-noVBFJet-tt` with a purity in $\mathrm{t\bar{t}}$ of approx. 90%. Similarly, the requirements $\hat{f}_{\mathrm{t\bar{t}}} \leq 0.4$ and $\hat{f}_{\mathrm{VV}} > 0.5$ select the VV contribution with a purity of 60%. We denote this region by `highMT-noVBFJet-VV`. Given $(\mathcal{L}\sigma_{\mathrm{t\bar{t}}}(1,\mathbf{0}))^{-1/2} \approx 0.015 < \alpha_{\mathrm{t\bar{t}}} = 0.02$ and $(\mathcal{L}\sigma_{\mathrm{VV}}(1,\mathbf{0}))^{-1/2} \approx 0.037 < \alpha_{\mathrm{VV}} = 0.25$, we expect significant constraints on $\nu_{\mathrm{t\bar{t}}}$ and ν_{VV}. The two regions are shown in Table II.

In summary, the log-likelihood ratio from the three CRs defined in this section is

$$-\frac{1}{2}u_{\mathrm{CR}}(\mathcal{D}|\mu,\boldsymbol{\nu}) = \sum_{i=1}^{N_{\mathrm{CR}}}\left[-\sum_{p}\mathcal{L}\big(\sigma_{i,p}(\mu,\boldsymbol{\nu}) - \sigma_{i,p}(1,\mathbf{0})\big) + N_{i,\mathrm{obs}}\log\left(\frac{\sum_p \sigma_{i,p}(\mu,\boldsymbol{\nu})}{\sum_p \sigma_{i,p}(1,\mathbf{0})}\right)\right], \tag{43}$$

with $\sigma_{i,p}(\mu,\boldsymbol{\nu})$ again defined by Eq. (7), using components computed from Eqs. (8), (13), and (14) and the sums over the processes p run over H, Z, $\mathrm{t\bar{t}}$, and VV.

Along the same lines, we also define a binned likelihood in regions that are otherwise treated as unbinned and marked with UB in Table II. Doing so provides a fully binned reference likelihood for performance comparisons. For each UB region, we construct the distribution of the signal node of the multiclassifier ($\hat{f}_{\mathrm{H}}$) using nominal simulation. Bins in this observable are defined as follows.

The weighted distribution of $\hat{f}_{\mathrm{H}}$ is integrated from a lower threshold up to its maximum value of one. We first set the lower integration bound to one, resulting in zero integral, and then gradually lower it, computing the $\mathrm{H} \to \tau\tau$ signal yield (S) and cumulative background yield (B) at each step. These yields define a simple proxy for the signal significance in the interval, $s = S/\sqrt{S+B}$. We continue lowering the threshold until $s = 2$ or until s reaches a maximum. At that point, a bin is formed from the integration interval, the corresponding phase space is removed, and the procedure is repeated until all events are binned. The resulting distribution in the `lowMT-VBFJet` region is shown in Fig. 7. This binning procedure roughly approximates a traditional analysis, with the classifier output used in a binned template fit carried out simultaneously in all UB regions. We tested finer variants of the binning procedure that led to similar results, as reported below. The corresponding binned reference log-likelihood is given by

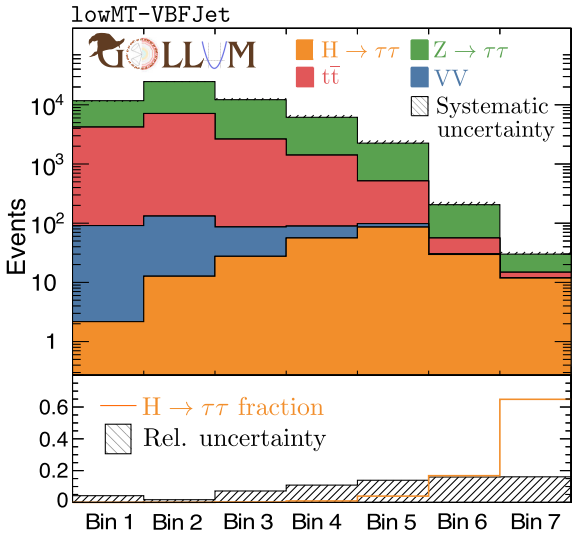

FIG. 7. The binned distribution of $\hat{f}_{\mathrm{H}}(\boldsymbol{x})$.

$$-\frac{1}{2}u_{\mathrm{B}}(\mathcal{D}|\mu,\boldsymbol{\nu}) = \sum_{r=1}^{N_{\mathrm{UB}}}\sum_{i=1}^{N_{\mathrm{bins,r}}}\left[-\sum_{p}\mathcal{L}\big(\sigma_{r,i,p}(\mu,\boldsymbol{\nu}) - \sigma_{r,i,p}(1,\mathbf{0})\big) + N_{r,i,\mathrm{obs}}\log\left(\frac{\sum_p \sigma_{r,i,p}(\mu,\boldsymbol{\nu})}{\sum_p \sigma_{r,i,p}(1,\mathbf{0})}\right)\right]. \tag{44}$$

Finally, the likelihood of the hypothetical auxiliary measurements that provide the FAIR-HUC nuisance parameter constraints is

$$u_{\mathrm{P}} = \nu_{\mathrm{bkg}}^2 + \nu_{\mathrm{t\bar{t}}}^2 + \nu_{\mathrm{VV}}^2 + \nu_{\mathrm{tes}}^2 + \nu_{\mathrm{jes}}^2 + \nu_{\mathrm{met}}^2 \tag{45}$$

and acts as a penalty during the profiling. We do not allow nuisance parameter values outside the boundaries discussed in Sec. II C.

With these ingredients, we define the unbinned and binned log-likelihoods as

$$u_{\mathrm{unbinned}} = u_{\mathrm{UB}} + u_{\mathrm{CR}} + u_{\mathrm{P}}, \tag{46a}$$

$$u_{\mathrm{binned}} = u_{\mathrm{B}} + u_{\mathrm{CR}} + u_{\mathrm{P}}, \tag{46b}$$

and use those in the profiling in Eq. (4). We emphasize that both unbinned and binned likelihoods include a binned component from the CRs (u_{CR}) and the penalty (u_{P}).

B. Asimov results

We use the Asimov dataset [2] to compute expected confidence intervals for all test statistics. The unbinned Asimov dataset is introduced in Ref. [16] and is adapted to our notation in Appendix A. The profiling is performed with the MINUIT package [50]. Figure 8 shows the Asimov

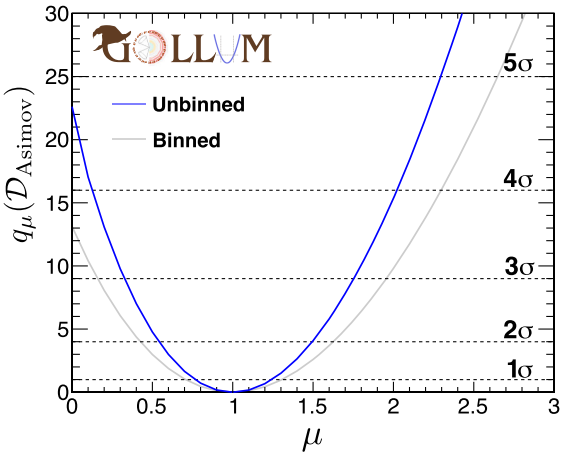

FIG. 8. The profiled likelihood test statistic $q_\mu(\mathcal{D}_{\text{Asimov}})$ for the binned and unbinned case with Asimov data.

expected likelihood test statistic for the (assumed) true values $\mu^{\text{true}} = 1$ and $\boldsymbol{\nu}^{\text{true}} = \mathbf{0}$. We observe no noticeable bias in the location of the minimum: the maximum likelihood estimate (MLE) is found at $(\hat{\mu}, \hat{\boldsymbol{\nu}}) = (1, \mathbf{0})$ with an accuracy of 10^{-3}. Repeating the Asimov fit for different values of μ^{true} in the range $0.1 \leq \mu \leq 3$ confirms this conclusion. Any residual imperfection in the ML surrogates thus has minimal impact on the result. However, their overall quality remains critical: For example, removing the isotonic regression calibration from Sec. IV A induces a bias on $\hat{\mu}$ in the range 0.3–0.4, with strong dependence on $\boldsymbol{\nu}^{\text{true}}$. This highlights the necessity of calibrating the classifier output using this or similar SBI methods [48].

The binned profiled likelihood ratio in Fig. 8 exceeds the 1σ threshold outside the interval $\mu = 1.00 \pm 0.21$. For the unbinned case, the interval narrows to 0.17, yielding a 20% improvement in the FAIR-HUC setup. The hypothesis $\mu = 0$ is excluded at 3.6σ for the binned case and 4.8σ for the unbinned case, marking a significant gain.

The impacts of the nuisance parameters on the extracted signal strength, shown in Fig. 9, highlight differences between the binned and unbinned approaches. We show prefit and postfit nuisance parameter values for both analyses based on an Asimov fit to the nominal simulation. Starting from the MLE, each nuisance parameter is varied until the likelihood ratio test statistic increases by one. This defines the postfit uncertainty. Refitting the signal strength at this configuration yields the corresponding impact.

In both the binned and unbinned fits, we observe significant postfit constraints on the normalization uncertainties. This is expected for FAIR-HUC, where the auxiliary likelihood is less constraining than in realistic applications with high-quality calibrations that reduce these impacts already prior to the fit.

Notably, the unbinned model provides slightly stronger constraints on the $\mathrm{t\bar{t}}$ background normalization, while other normalization uncertainties remain similar between the two approaches—indicating that the common binned control

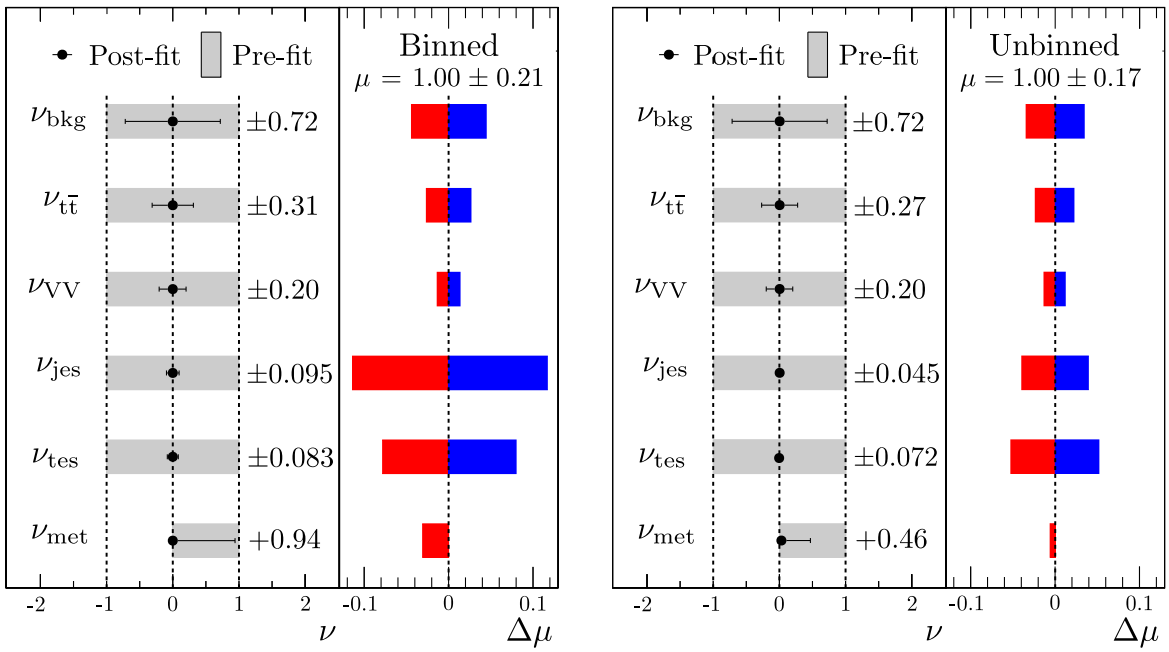

FIG. 9. Prefit and postfit nuisance parameter values for the binned analysis (left) and the unbinned analysis (right) in the Asimov fit to the nominal simulation. In each figure, the impacts on the measured signal strength are also shown. The red color corresponds to shifts of μ associated with upward variations of the nuisance parameters, and the blue color corresponds to downward variations. The `met` uncertainty is one-sided.

regions predominantly determine these parameters. The jes and tes nuisance parameters are constrained to 9.5% and 8.3% in the binned case, respectively, while the unbinned fit yields much tighter constraints of 4.5% and 5.2%. This suggests that the high-dimensional unbinned likelihood captures subtle shape information more effectively. Moreover, the largest impact in the binned fit arises from the jes nuisance (11.5%), which is reduced to 4% and becomes subleading in the unbinned model. The impact of the tes nuisance also improves, decreasing from 8.3% to 5.2%, a 37% reduction. Additionally, only the unbinned model exhibits a meaningful constraint on the met nuisance, which remains largely unconstrained in the binned case. The strong reduction in the impact of met in the unbinned analysis indicates a dominantly nonlinear influence, which the high-dimensional likelihood captures efficiently. Overall, the results demonstrate that the unbinned likelihood efficiently exploits the information distributed across the feature space. We conjecture that realistic measurements with typically more than 100 nuisance parameters will benefit even more.

Finally, we note that all positive variations of the nuisance parameters around the MLE (shown in red) reduce the signal strength, which is consistent with the training data where positive values of the nuisance parameters increase the background yield.

We show the correlation matrix of the model parameters in Fig. 10 for the unbinned models. Notably, the parameters ν_{tes} and ν_{bkg} exhibit a correlation coefficient of −0.53. This anticorrelation arises from a normalization change induced by the tes variation across all processes, which results from a shift that moves low-energetic τ_{had} candidates above the selection threshold. All other correlation coefficients have a magnitude of less than 0.2.

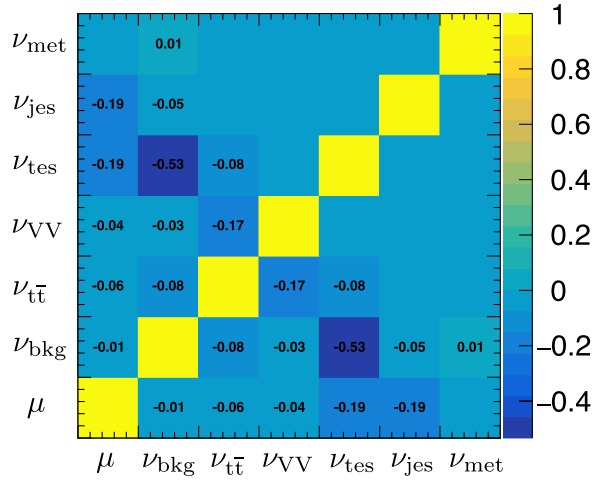

FIG. 10. The correlation of the seven model parameters at the best-fit point in the unbinned Asimov dataset. Correlations below an absolute value of 1% and along the diagonal are not shown.

C. Fisher information matrix

The learned likelihood ratio can be used to assess differences in the approaches for parameter points close to the true values. We can use the Fisher information matrix to express the likelihood in the vicinity of the fixed but unknown true parameter value $\boldsymbol{\theta}^{\text{true}}$ compactly as

$$L(\mathcal{D}|\boldsymbol{\theta}) \approx L(\mathcal{D}|\boldsymbol{\theta}^{\text{true}}) \exp\left[-\frac{1}{2}\Delta\boldsymbol{\theta}^T I(\boldsymbol{\theta}^{\text{true}})\Delta\boldsymbol{\theta}\right], \quad (47)$$

where we arrange the seven model parameters as $\boldsymbol{\theta} = (\mu, \boldsymbol{\nu})$ and $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}^{\text{true}}$. The Fisher information matrix $I(\boldsymbol{\theta})$, for general $\boldsymbol{\theta}$, is given by

$$I_{ab}(\boldsymbol{\theta}) = \left\langle \frac{\partial}{\partial\theta_a}\log p(\mathcal{D}|\boldsymbol{\theta}) \frac{\partial}{\partial\theta_b}\log p(\mathcal{D}|\boldsymbol{\theta}) \right\rangle_{\boldsymbol{\theta}}, \quad (48)$$

where a, b label the model parameters. The probability density function $p(\mathcal{D}|\boldsymbol{\theta})$ here is the same as the extended likelihood in Eq. (2) but is interpreted as a function of the dataset. The Fisher information matrix has many useful applications in estimation problems. For example, the Cramér-Rao bound [51,52] states that the covariance matrix of an unbiased estimator, such as $\bar{\boldsymbol{\theta}}$, is bounded from below by the inverse of the Fisher information matrix, i.e.,

$$\text{Cov}(\bar{\boldsymbol{\theta}}) \geq I^{-1}(\boldsymbol{\theta}^{\text{true}}). \quad (49)$$

This inequality, understood in the sense that $\text{Cov}(\bar{\boldsymbol{\theta}}) - I^{-1}(\boldsymbol{\theta}^{\text{true}})$ is positive semidefinite, establishes a fundamental lower bound on the achievable variance and covariance for any unbiased estimator of the parameters. We evaluate the Fisher information in the Asimov data set by differentiating our parametric model using binned and unbinned components exactly as in Eq. (46). The details are provided in Appendix B.

We provide a visual representation of the eigensystems of the Fisher information in Fig. 11 for both the binned and unbinned cases. In these plots, the eigenvalues are shown on the vertical axis, indicating the sensitivity of the model to the combination of model parameters encoded as eigenvectors. On the horizontal axis, color-coded bars represent the absolute values of the corresponding eigenvector coefficients, illustrating the fractional contributions of each parameter. The uncertainty in a given eigendirection with eigenvalue EV is $\text{EV}^{-1/2}$ such that higher values are desirable. Notably, in the unbinned case, the largest eigenvalue is associated with an eigenvector that corresponds to a tight constraint on a combination of nuisance parameters—a feature that does not appear in the binned Fisher eigensystem. Moreover, the second largest eigenvalue in the unbinned case is linked to an eigenvector with a signal fraction of 30%, whereas the binned analysis yields two eigenvectors with a similar signal contribution but a lower combined fraction. This observation supports

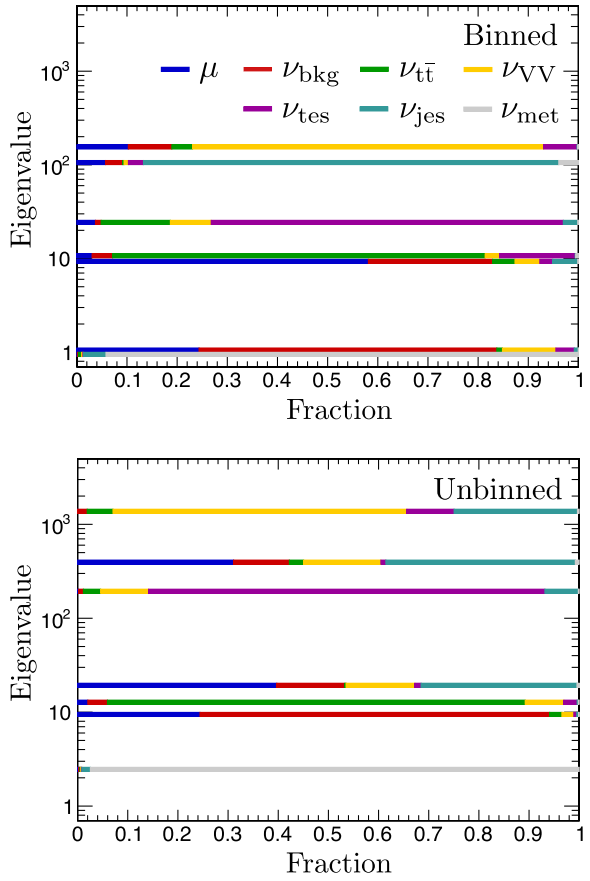

FIG. 11. A visual representation of the eigensystems of the Fisher information in the binned (top) and unbinned case (bottom). The eigenvalue gives the position on the y-axis, and the color-coded fractions in the horizontal direction represent the absolute values of the coefficients of the corresponding eigenvector.

our understanding that the unbinned analysis is more effective at disentangling nuisance parameters. Finally, the ν_{met} direction appears poorly constrained, reflecting its small linear impact, which dominates the Fisher information analysis.

D. Toy studies

To further assess the quality of the approach, we probe the unbinned model in Eq. (46a) with toy data samples.

We sample 5×10^4 model parameter configurations, with $\mu^{\text{true}} \sim U(0.1, 3)$ uniformly distributed, $\nu_{\text{met}}^{\text{true}} \sim \text{lnN}(0, 1)$ log-normally distributed, and the remaining five model parameters drawn from normal distributions. Samples with parameters outside the boundaries described in Sec. II C are removed. This reproduces the probability density function used to assess model performance in FAIR-HUC [22].

For each configuration, we sample from test data that was not used during surrogate training. The values of the signal strength and normalization uncertainties modify the expected event yields. The effects of the calibration-type

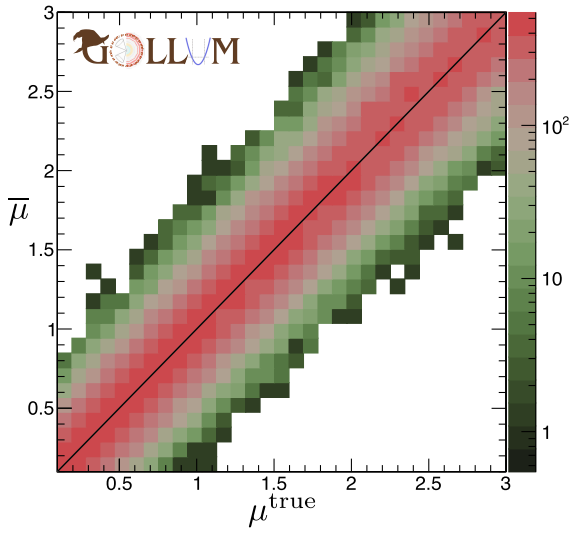

FIG. 12. Scatter plot of the true value of the $H \to \tau\tau$ signal strength parameter μ and the MLE $\bar{\mu}$ for 5×10^4 toys.

uncertainties `tes`, `jes`, and `met` are applied using the code snippets from Ref. [22], which adjust both primary and derived features. These modifications are computed on the fly for each toy dataset. The resulting datasets are used to compute the test statistic $q_\mu(\mathcal{D})$, the maximum likelihood estimate $(\bar{\mu}, \bar{\nu})$, the 68% confidence interval boundaries, μ_{16} and μ_{84}, and the 95% confidence interval boundaries, $\mu_{2.5}$ and $\mu_{97.5}$.

In Fig. 12, we show the scatter plot of μ^{true} versus $\bar{\mu}$. The MLE points lie along the diagonal with no visible bias. There is no population of outliers, indicating excellent stability of the MINUIT minimization and the unbinned surrogate model.

In Fig. 13, we show the postfit distribution of a randomly chosen toy fit for the observables $m_{\text{vis}}(\eta, \ell)$, $m^{j_1 j_2}$, and p_{T}^{H}. In the bottom panels we compare the prefit and postfit uncertainties. The much narrower postfit uncertainty bands visually demonstrate how the high-dimensional approach constraints the systematic uncertainties and extracts sensitivity. For this particular toy, we use $\mu^{\text{true}} = 3$ and found $\bar{\mu} = 3.24 \pm 0.25$.

Figures 14 and 15 show analogous scatter plots for the nuisance parameters. As discussed in Sec. V B, ν_{bkg} is only weakly constrained, and the MLE shows little correlation with the true value. In contrast, $\nu_{\text{t}\bar{\text{t}}}$ and ν_{VV} cluster tightly along the diagonal, although very small values of ν_{VV} are harder to reconstruct accurately. The values of ν_{tes} and ν_{jes} are precisely recovered, consistent with their tight postfit constraints. For ν_{met}, we observe a slight bias at both ends of the parameter range, but this has negligible impact given the small postfit impact of ν_{met}. Overall, we observe very few outliers, indicating stable reconstruction across all parameters.

In Fig. 16 (left), we show the mean interval widths at the 1σ and 2σ confidence levels, given by $\mu_{84} - \mu_{16}$ and

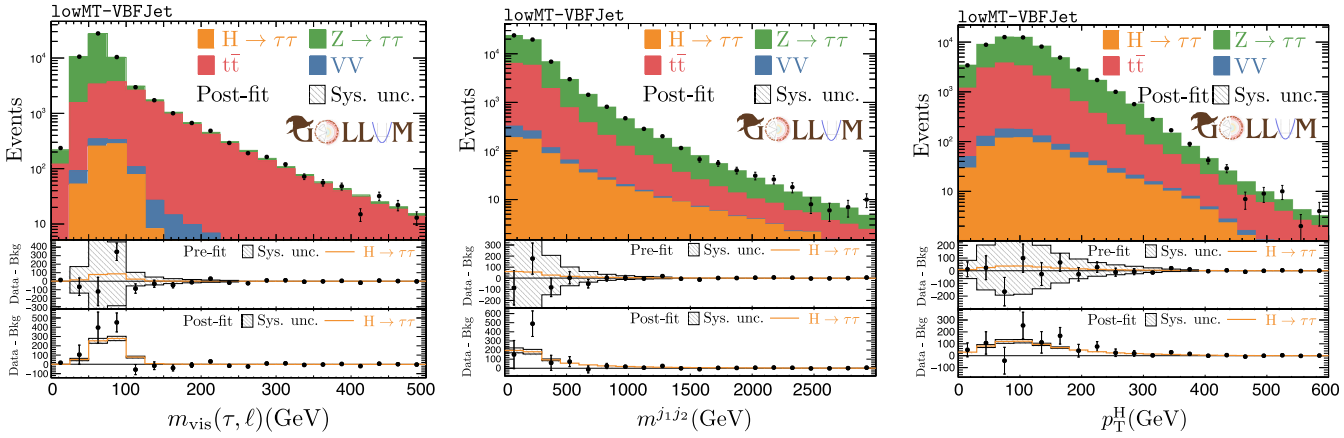

FIG. 13. Pre and postfit distributions in the `lowMT-VBFJet` region for the observables $m_{\mathrm{vis}}(\eta,\ell)$ (left), $m^{j_1 j_2}$ (middle), and $p_{\mathrm{T}}^{\mathrm{H}}$ (right) for a randomly chosen toy dataset with $\mu^{\mathrm{true}} = 3$. The MLE is $\bar{\mu} = 3.24 \pm 0.25$.

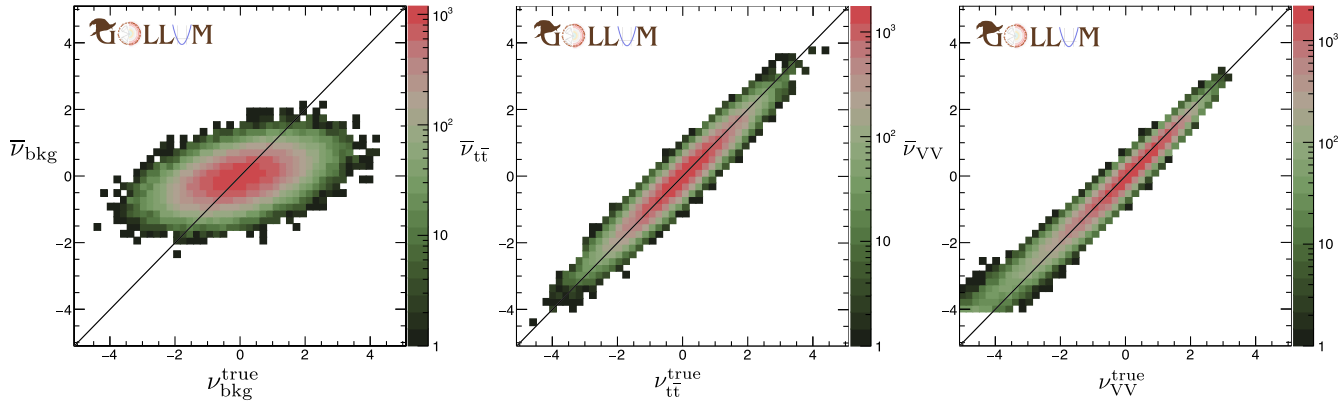

FIG. 14. Same as Fig. 12, but for ν_{bkg} (left), $\nu_{\mathrm{t\bar{t}}}$ (middle), and ν_{VV} (right).

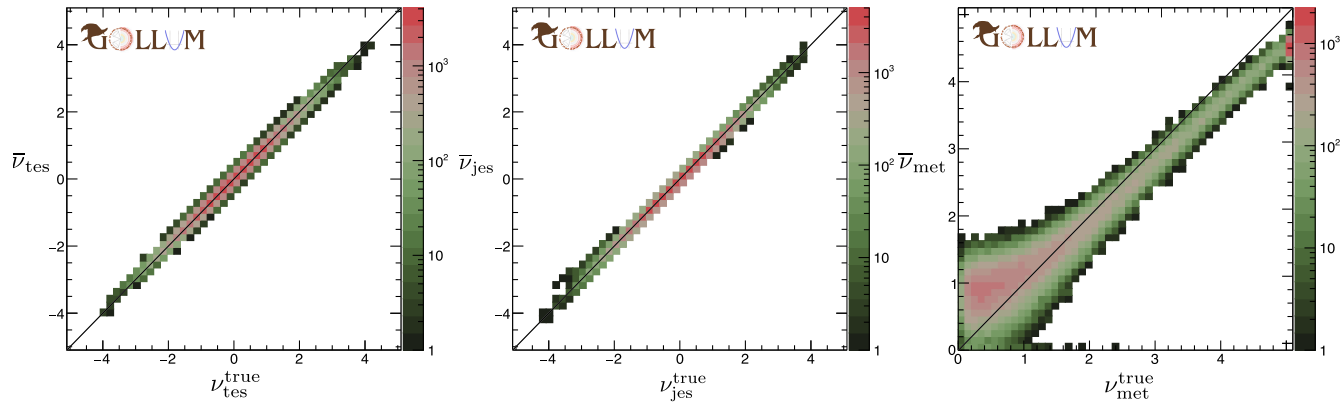

FIG. 15. Same as Fig. 12, but for ν_{tes} (left), ν_{jes} (middle), and ν_{met} (right).

$\mu_{97.5} - \mu_{2.5}$, respectively. Both increase slightly with μ^{true} across the probed range.

The final evaluation, which also tests the quality of the learned parametrization, is the coverage test shown in Fig. 16 (right). As a function of μ^{true}, we show the fraction of toy fits for which μ^{true} lies within the reconstructed 1σ and 2σ intervals. We observe excellent agreement, indicating that the unbinned model provides reliable coverage.

Our code guaranteed optimal likelihood-based unbinned method (GOLLUM) for ML and inference is publicly available at Ref. [23].

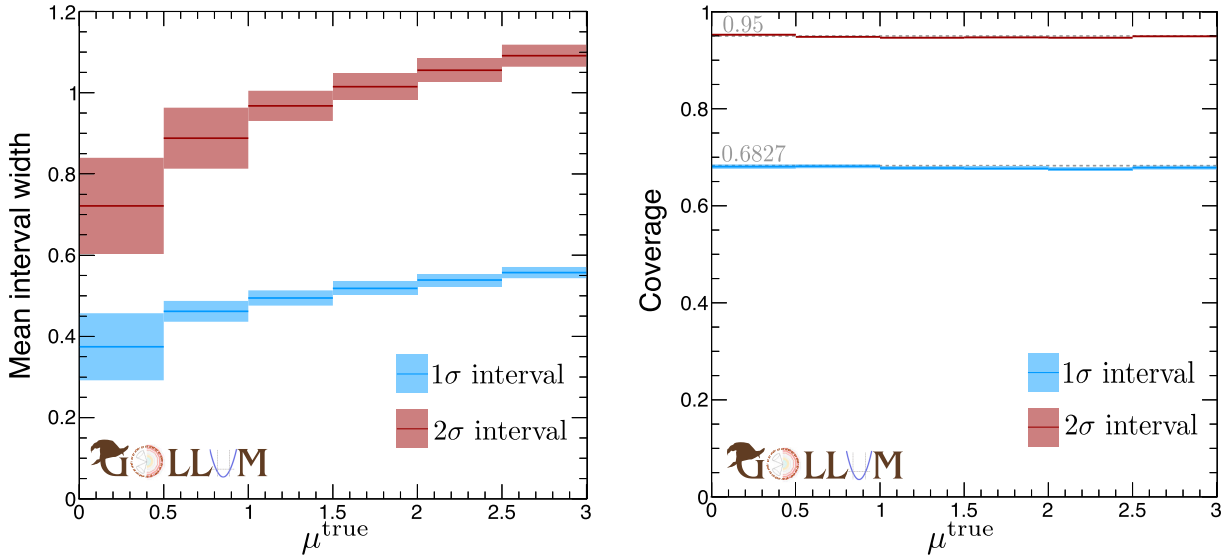

FIG. 16. Mean interval width of the signal strength parameter μ (left) and coverage (right) as a function of μ^{true} obtained from 5×10^4 toy datasets.

VI. CONCLUSIONS

We have presented a novel unbinned methodology for inclusive cross-section measurements that systematically incorporates machine-learned approximations of systematic uncertainties. By combining a calibrated multiclass classifier with a parametric neural network surrogate for the differential cross-section ratio, we reconstruct a high-dimensional likelihood ratio test statistic that captures both known analytic and unknown data-driven dependencies on nuisance parameters. Using the FAIR-HUC benchmark dataset, we demonstrate significant improvements over traditional binned analyses. In Asimov data and large-scale toy studies, the unbinned model outperforms the binned reference by reducing the uncertainty in the signal strength by 20% and achieving higher significance for signal exclusion. The Fisher information matrix further confirms that the unbinned likelihood extracts more information from the nuisance parameters. The method is fully refinable and compatible with established workflows, enabling stage-wise improvements without retraining the entire analysis. Our implementation is provided in the public GOLLUM codebase [23], offering a flexible and robust toolkit for unbinned likelihood-based inference at the LHC and beyond.

ACKNOWLEDGMENTS

We thank the organizers of the FAIR-HUC challenge for the organization and technical support. C. G. and M. S. received support from the Austrian Science fund (FWF) under Grant No. I5687. D. S. received support under Grant No. P33771. The computational results were obtained using the Vienna Bio Center and the CLIP computing cluster of the Austrian Academy of Sciences at [53].

DATA AVAILABILITY

The data that support the findings of this article are openly available [23,54]. The underlying FAIR-HUC data [55] is described in [22].

APPENDIX A: BINNED AND UNBINNED ASIMOV DATA

Asimov data [2] can be used to compute sampling-free expected results within continuous parametric models. In the well-known binned case, the Asimov expectation amounts to replacing the integer-count observation in, e.g., Eq. (44), by the expected yield under the (assumed) true hypothesis, which we denote by $\mu'\boldsymbol{\nu}'$, that is

$$\langle N_{\text{obs}} \rangle_{\mu',\boldsymbol{\nu}'} \to \mathcal{L}\sigma(\mu'\boldsymbol{\nu}') \tag{A1}$$

for every bin.

The unbinned generalization, first obtained in Ref. [16], is the continuum limit of this procedure and amounts to replacing the sum over observed events in the last term in Eq. (30) by a weighted sum over simulated events, concretely

$$\left\langle \sum_{i=1}^{N_{\text{obs}}} (\cdots) \right\rangle_{\mu',\boldsymbol{\nu}'} \to \sum_{\{w_i,\boldsymbol{x}_i\} \in \mathcal{D}_{\mu',\boldsymbol{\nu}'}} w_i \times (\cdots). \tag{A2}$$

For obtaining Asimov limits with arbitrary μ' it thus suffices to compute the log-term in the unbinned test statistic as a weighted sum and scale the signal weights by μ'.

We can also use the learned model instead, that is,

$$\left\langle \sum_{i=1}^{N_{\text{obs}}} (\cdots) \right\rangle_{\mu',\boldsymbol{\nu}'} \to \sum_{\{w_i,\boldsymbol{x}_i\} \in \mathcal{D}_{1,\mathbf{0}}} w_i \hat{R}(\boldsymbol{x}_i|\mu'\boldsymbol{\nu}') \times (\cdots). \tag{A3}$$

The approximate expectation of the log-likelihood ratio then becomes

$$\left\langle -\frac{1}{2}u_{\mathrm{UB}}(\mathcal{D}|\mu,\boldsymbol{\nu})\right\rangle_{\mu',\boldsymbol{\nu}'} \approx \sum_{r=1}^{N_r}\sum_{\substack{\{w_i,\boldsymbol{x}_i\}\in\\ \mathcal{D}_{1,\mathbf{0}}\cap r}} w_i\big[1-\hat{R}_r(\boldsymbol{x}_i|\mu,\boldsymbol{\nu}) + \hat{R}_r(\boldsymbol{x}_i|\mu',\boldsymbol{\nu}')\log\hat{R}_r(\boldsymbol{x}_i|\mu,\boldsymbol{\nu})\big], \tag{A4}$$

which we can evaluate parametrically.

Because we defined u in Eq. (4) with respect to $(\mu,\boldsymbol{\nu})=(1,\mathbf{0})$, we need not subtract the reference log-likelihood to reobtain the general case. The Asimov expectation for generic $(\mu',\boldsymbol{\nu}')$ is

$$\begin{aligned}-\frac{1}{2}\Lambda &\equiv \left\langle\frac{L(\mathcal{D}|\mu,\boldsymbol{\nu})}{L(\mathcal{D}|\mu',\boldsymbol{\nu}')}\right\rangle_{\mu',\boldsymbol{\nu}'}\\ &= -\frac{1}{2}\langle u_{\mathrm{UB}}(\mathcal{D}|\mu,\nu)-u_{\mathrm{UB}}(\mathcal{D}|\mu',\nu')\rangle_{\mu',\boldsymbol{\nu}'}\\ &= \sum_{r=1}^{N_r}\sum_{\substack{\{w_i,\boldsymbol{x}_i\}\in\\ \mathcal{D}_{1,\mathbf{0}}\cap r}} w_i\left[\hat{R}_r(\boldsymbol{x}_i|\mu',\boldsymbol{\nu}')-\hat{R}_r(\boldsymbol{x}_i|\mu,\boldsymbol{\nu}) + \hat{R}_r(\boldsymbol{x}_i|\mu',\boldsymbol{\nu}')\log\frac{\hat{R}_r(\boldsymbol{x}_i|\mu,\boldsymbol{\nu})}{\hat{R}_r(\boldsymbol{x}_i|\mu',\boldsymbol{\nu}')}\right],\end{aligned} \tag{A5}$$

which has the same form as Eq. (B.15) in Ref. [16]. The number Λ on the lhs is the noncentrality parameter of a noncentral χ^2 distribution which, according to Wald's theorem [56], is the asymptotical limit of the distribution of the test-statistic. Using the learned model in Eq. (A4), we have Λ parametrically as a function of μ' and $\boldsymbol{\nu}'$.

APPENDIX B: ENTRIES OF THE FISHER INFORMATION MATRIX

In this section, we compute the entries of the seven-parameter Fisher information matrix for binned and unbinned datasets. Combining the components of the model for binned yields in Sec. III B and abbreviating $\boldsymbol{\theta}=(\mu,\boldsymbol{\nu})$, we have

$$\begin{aligned}\lambda(\boldsymbol{\theta}) = \mathcal{L}\big[&\mu\sigma_{\mathrm{H}}\hat{S}_{\mathrm{H}}(\boldsymbol{\nu}) + (1+\alpha_{\mathrm{bkg}})^{\nu_{\mathrm{bkg}}}\\ &\times\big(\sigma_{\mathrm{Z}}\hat{S}_{\mathrm{Z}}(\boldsymbol{\nu}) + (1+\alpha_{\mathrm{t\bar{t}}})^{\nu_{\mathrm{t\bar{t}}}}\sigma_{\mathrm{t\bar{t}}}\hat{S}_{\mathrm{t\bar{t}}}(\boldsymbol{\nu})\\ &+ (1+\alpha_{\mathrm{VV}})^{\nu_{\mathrm{VV}}}\sigma_{\mathrm{VV}}\hat{S}_{\mathrm{VV}}(\boldsymbol{\nu})\big)\big]\end{aligned} \tag{B1}$$

where $\hat{S}_p=\exp(\nu_A\hat{\Delta}_{A,p})$.

The Fisher information is additive, so if r labels all bins contributing to the likelihood, we find from Eq. (48)

$$\begin{aligned}I^{(\mathrm{binned})}_{ab}(\boldsymbol{\theta}) &= \sum_r\left\langle\frac{\partial}{\partial\theta_a}\log\mathrm{P}(N_{r,\mathrm{obs}}|\lambda_r(\boldsymbol{\theta}))\frac{\partial}{\partial\theta_b}\log\mathrm{P}(N_{r,\mathrm{obs}}|\lambda_r(\boldsymbol{\theta}))\right\rangle_{\boldsymbol{\theta}}\\ &= \sum_r\frac{1}{\lambda_r(\boldsymbol{\theta})}\frac{\partial\lambda_r(\boldsymbol{\theta})}{\partial\theta_a}\frac{\partial\lambda_r(\boldsymbol{\theta})}{\partial\theta_b}\end{aligned} \tag{B2}$$

where we label the seven parameters in $\boldsymbol{\theta}$ with indices a, b. The derivatives, evaluated at $\boldsymbol{\theta}=(1,\mathbf{0})$, are

$$\frac{\partial}{\partial\mu}\lambda(\boldsymbol{\theta})\Big|_{1,\mathbf{0}} = \mathcal{L}\sigma_{\mathrm{H}} \tag{B3a}$$

$$\frac{\partial}{\partial\nu_{\mathrm{bkg}}}\lambda(\boldsymbol{\theta})\Big|_{1,\mathbf{0}} = \mathcal{L}(\sigma_{\mathrm{Z}}+\sigma_{\mathrm{t\bar{t}}}+\sigma_{\mathrm{VV}})\log(1+\alpha_{\mathrm{bkg}}) \tag{B3b}$$

$$\frac{\partial}{\partial\nu_{\mathrm{t\bar{t}}}}\lambda(\boldsymbol{\theta})\Big|_{1,\mathbf{0}} = \mathcal{L}\sigma_{\mathrm{t\bar{t}}}\log(1+\alpha_{\mathrm{t\bar{t}}}) \tag{B3c}$$

$$\frac{\partial}{\partial\nu_{\mathrm{VV}}}\lambda(\boldsymbol{\theta})\Big|_{1,\mathbf{0}} = \mathcal{L}\sigma_{\mathrm{VV}}\log(1+\alpha_{\mathrm{VV}}) \tag{B3d}$$

$$\frac{\partial}{\partial\nu_{\mathrm{tes}}}\lambda(\boldsymbol{\theta})\Big|_{1,\mathbf{0}} = \mathcal{L}\big(\sigma_{\mathrm{H}}\hat{\Delta}_{\mathrm{tes,H}}+\sigma_{\mathrm{Z}}\hat{\Delta}_{\mathrm{tes,Z}} + \sigma_{\mathrm{t\bar{t}}}\hat{\Delta}_{\mathrm{jes,t\bar{t}}}+\sigma_{\mathrm{VV}}\hat{\Delta}_{\mathrm{jes,VV}}\big) \tag{B3e}$$

$$\frac{\partial}{\partial\nu_{\mathrm{jes}}}\lambda(\boldsymbol{\theta})\Big|_{1,\mathbf{0}} = \mathcal{L}\big(\sigma_{\mathrm{H}}\hat{\Delta}_{\mathrm{jes,H}}+\sigma_{\mathrm{Z}}\hat{\Delta}_{\mathrm{jes,Z}} + \sigma_{\mathrm{t\bar{t}}}\hat{\Delta}_{\mathrm{jes,t\bar{t}}}+\sigma_{\mathrm{VV}}\hat{\Delta}_{\mathrm{jes,VV}}\big) \tag{B3f}$$

$$\frac{\partial}{\partial\nu_{\mathrm{met}}}\lambda(\boldsymbol{\theta})\Big|_{1,\mathbf{0}} = \mathcal{L}\big(\sigma_{\mathrm{H}}\hat{\Delta}_{\mathrm{met,H}}+\sigma_{\mathrm{Z}}\hat{\Delta}_{\mathrm{met,Z}} + \sigma_{\mathrm{t\bar{t}}}\hat{\Delta}_{\mathrm{met,t\bar{t}}}+\sigma_{\mathrm{VV}}\hat{\Delta}_{\mathrm{met,VV}}\big) \tag{B3g}$$

where we suppressed the index r on all yields λ on the lhs and on all per-bin cross sections σ on the rhs From these equations, we can compute $I^{(\mathrm{binned})}$. While Eq. (B3) is not particularly illuminating, it implies that in the absence of nuisance parameters, i.e., for an idealized single-parameter measurement of μ in the presence of backgrounds,

$$I^{(\mathrm{binned})}_{\mu\mu}(1,\mathbf{0}) = \frac{(\mathcal{L}\sigma_H)^2}{\mathcal{L}(\sigma_H+\sigma_Z+\sigma_{\mathrm{t\bar{t}}}+\sigma_{\mathrm{VV}})} = \rho\mathcal{L}\sigma_H \tag{B4}$$

where ρ is the signal purity. The Cramér-Rao bound then states that

$$\sqrt{C_{\mu\mu}} \geq \frac{1}{\sqrt{\rho\mathcal{L}\sigma_H}} \tag{B5}$$

i.e., the uncertainty of an unbiased estimator is bounded by the reciprocal square root of the expected signal yield, as usual, but it is also penalized by the purity in the bin.

Next, we compute the Fisher information for the unbinned extended likelihood where we drop the sum over the unbinned regions for simplicity. The general expression then is

$$I_{ab}(\boldsymbol{\theta}) = \int \mathrm{d}\mathcal{D} p(\mathcal{D}|\boldsymbol{\theta}) \partial_a \log p(\mathcal{D}|\boldsymbol{\theta}) \partial_b \log p(\mathcal{D}|\boldsymbol{\theta}) = -\int \mathrm{d}\mathcal{D} p(\mathcal{D}|\boldsymbol{\theta}) \partial_a \partial_b \log p(\mathcal{D}|\boldsymbol{\theta}), \tag{B6}$$

where the last line follows from partial integration. The integration measure over the space of the observed variable-length datasets is $\mathrm{d}\mathcal{D} = \mathrm{d}N \prod_{i=1}^{n} \mathrm{d}\boldsymbol{x}_i$ where $\mathrm{d}N$ is the counting measure on $\mathbb{N}_0$ and the $\mathrm{d}\boldsymbol{x}_i$ are standard Lebesgue measures on $\mathbb{R}^{N_f}$ with, in our case, $N_f = 28$. Inserting Eq. (2), we find after a few steps

$$I_{ab}^{(\mathrm{unbinned})}(\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{x} \frac{\partial_a(\mathcal{L}\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta})) \partial_b(\mathcal{L}\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}))}{\mathcal{L}\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta})} \tag{B7}$$

where we have used Eq. (3) to convert $p(\boldsymbol{x}|\boldsymbol{\theta})$ to DCRs. Equation (B7) is the continuum limit of Eq. (B2) and we evaluate it for $\boldsymbol{\theta} \equiv (\mu, \boldsymbol{\nu}) = (1, \mathbf{0})$. Using a simulated sample $\mathcal{D}_{1,\mathbf{0}}$, we get

$$I_{ab}^{(\mathrm{unbinned})}(1, \mathbf{0}) \approx \sum_{\substack{\{w_i, \boldsymbol{x}_i\} \\ \in \mathcal{D}_{1,\mathbf{0}}}} w_i \partial_a \hat{R}(\boldsymbol{x}_i|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} \partial_b \hat{R}(\boldsymbol{x}_i|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} \tag{B8}$$

with $\hat{R}(\boldsymbol{x}|\boldsymbol{\theta})$ again taken from Eq. (22).

The required derivatives are

$$\begin{aligned}
\frac{\partial}{\partial \mu} \hat{R}(\boldsymbol{x}|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} &= \hat{g}_{\mathrm{H}}(\boldsymbol{x}) \\
\frac{\partial}{\partial \nu_{\mathrm{bkg}}} \hat{R}(\boldsymbol{x}|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} &= (1 - \hat{g}_{\mathrm{H}}(\boldsymbol{x})) \log(1 + \alpha_{\mathrm{bkg}}) \\
\frac{\partial}{\partial \nu_{\mathrm{t\bar{t}}}} \hat{R}(\boldsymbol{x}|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} &= \hat{g}_{\mathrm{t\bar{t}}}(\boldsymbol{x}) \log(1 + \alpha_{\mathrm{t\bar{t}}}) \\
\frac{\partial}{\partial \nu_{\mathrm{VV}}} \hat{R}(\boldsymbol{x}|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} &= \hat{g}_{\mathrm{VV}}(\boldsymbol{x}) \log(1 + \alpha_{\mathrm{VV}}) \\
\frac{\partial}{\partial \nu_{\mathrm{tes}}} \hat{R}(\boldsymbol{x}|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} &= \left(\hat{g}_{\mathrm{H}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{tes,H}}(\boldsymbol{x}) + \hat{g}_{\mathrm{Z}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{tes,Z}}(\boldsymbol{x}) + \hat{g}_{\mathrm{t\bar{t}}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{jes,t\bar{t}}}(\boldsymbol{x}) + \hat{g}_{\mathrm{VV}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{jes,VV}}(\boldsymbol{x})\right) \\
\frac{\partial}{\partial \nu_{\mathrm{jes}}} \hat{R}(\boldsymbol{x}|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} &= \left(\hat{g}_{\mathrm{H}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{jes,H}}(\boldsymbol{x}) + \hat{g}_{\mathrm{Z}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{jes,Z}}(\boldsymbol{x}) + \hat{g}_{\mathrm{t\bar{t}}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{jes,t\bar{t}}}(\boldsymbol{x}) + \hat{g}_{\mathrm{VV}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{jes,VV}}(\boldsymbol{x})\right) \\
\frac{\partial}{\partial \nu_{\mathrm{met}}} \hat{R}(\boldsymbol{x}|\mu, \boldsymbol{\nu})\Big|_{1,\mathbf{0}} &= \left(\hat{g}_{\mathrm{H}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{met,H}}(\boldsymbol{x}) + \hat{g}_{\mathrm{Z}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{met,Z}}(\boldsymbol{x}) + \hat{g}_{\mathrm{t\bar{t}}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{met,t\bar{t}}}(\boldsymbol{x}) + \hat{g}_{\mathrm{VV}}(\boldsymbol{x}) \hat{\Delta}_{\mathrm{met,VV}}(\boldsymbol{x})\right).
\end{aligned} \tag{B9}$$

For example, the systematics-free Fisher information of the signal strength parameter is

$$I_{\mu\mu}^{(\mathrm{unbinned})}(1, \mathbf{0}) = \sum_{\{w_i, \boldsymbol{x}_i\} \in \mathcal{D}_{1,\mathbf{0}}} w_i \hat{g}_{\mathrm{H}}^2(\boldsymbol{x}_i), \tag{B10}$$

a quantity that could be visualized in one- or two-dimensional distributions to identify regions that contribute more substantially to the measurement.

[1] Lyndon Evans and Philip Bryant, LHC machine, J. Instrum. **3**, S08001 (2008).

[2] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur. Phys. J. C **71**, 1554 (2011); **73**, 2501(E) (2013).

[3] The ATLAS and CMS Collaborations and The LHC Higgs Combination Group, Procedure for the LHC Higgs boson search combination in Summer 2011, https://cds.cern.ch/record/1379837.

[4] Kyle Cranmer, Practical statistics for the LHC, in *Proceedings of the 2011 European School of High-Energy Physics* (2014), p. 267, arXiv:1503.07622.

[5] CMS Collaboration, The CMS statistical analysis and combination tool: COMBINE, Comput. Software Big Sci. **8**, 19 (2024).

[6] Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez, Constraining effective field theories with machine learning, Phys. Rev. Lett. **121**, 111801 (2018).
[7] Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez, A guide to constraining effective field theories with machine learning, Phys. Rev. D **98**, 052004 (2018).
[8] Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer, Mining gold from implicit models to improve likelihood-free inference, Proc. Natl. Acad. Sci. U.S.A. **117**, 5242 (2020).
[9] Johann Brehmer, Felix Kling, Irina Espejo, and Kyle Cranmer, MadMiner: Machine learning-based inference for particle physics, Comput. Software Big Sci. **4**, 3 (2020).
[10] Kyoungchul Kong, Konstantin T. Matchev, Stephen Mrenna, and Prasanth Shyamsundar, New machine learning techniques for simulation-based inference: InferoStatic nets, kernel score estimation, and kernel likelihood ratio estimation, arXiv:2210.01680.
[11] Kyle Cranmer, Johann Brehmer, and Gilles Louppe, The frontier of simulation-based inference, Proc. Natl. Acad. Sci. U.S.A. **117**, 30055 (2020).
[12] Henning Bahl, Victor Bresó, Giovanni De Crescenzo, and Tilman Plehn, Advancing tools for simulation-based inference, arXiv:2410.07315.
[13] Johann Brehmer, Sally Dawson, Samuel Homiller, Felix Kling, and Tilman Plehn, Benchmarking simplified template cross sections in WH production, J. High Energy Phys. 11 (2019) 034.
[14] Rahool Kumar Barman, Dorival Gonçalves, and Felix Kling, Machine learning the Higgs boson-top quark *CP* phase, Phys. Rev. D **105**, 035023 (2022).
[15] Henning Bahl and Simon Brass, Constraining $\mathcal{CP}$-violation in the Higgs-top-quark interaction using machine-learning-based inference, J. High Energy Phys. 03 (2022) 017.
[16] Raquel Gomez Ambrosio, Jaco ter Hoeve, Maeve Madigan, Juan Rojo, and Veronica Sanz, Unbinned multivariate observables for global SMEFT analyses from machine learning, J. High Energy Phys. 03 (2023) 033.
[17] Ricardo Barrué, Patricia Conde-Muíño, Valerio Dao, and Rui Santos, Simulation-based inference in the search for *CP* violation in leptonic WH production, J. High Energy Phys. 04 (2024) 014.
[18] Radha Mastandrea, Benjamin Nachman, and Tilman Plehn, Constraining the Higgs potential with neural simulation-based inference for di-Higgs production, Phys. Rev. D **110**, 056004 (2024).
[19] Chi Lung Cheng, Gup Singh, and Benjamin Nachman, Incorporating physical priors into weakly supervised anomaly detection, Phys. Rev. Lett. **135**, 021801 (2025).
[20] Chi Lung Cheng, Ranit Das, Runze Li, Radha Mastandrea, Vinicius Mikuni, Benjamin Nachman, David Shih, and Gup Singh, Generator based inference (GBI), arXiv:2506.00119.
[21] Robert Schöfbeck, Refinable modeling for unbinned SMEFT analyses, Mach. Learn. Sci. Tech. **6**, 015007 (2025).
[22] Wahid Bhimji *et al.*, FAIR universe HiggsML uncertainty challenge competition, arXiv:2410.02867.
[23] Lisa Benato, Cristina Giordano, Claudius Krause, Ang Li, Robert Schöfbeck, Dennis Schwarz, Maryam Shooshtari, and Daohan Wang, GOLLUM Code repository, https://github.com/HephyAnalysisSW/GOLLUM (2025).
[24] ATLAS Collaboration, The ATLAS experiment at the CERN large hadron collider, J. Instrum. **3**, S08003 (2008).
[25] ATLAS Collaboration, An implementation of neural simulation-based inference for parameter estimation in ATLAS, Rep. Prog. Phys. **88**, 067801 (2025).
[26] Georges Aad *et al.*, Measurement of off-shell Higgs boson production in the $H^* \to ZZ \to 4\ell$ decay channel using a neural simulation-based inference technique in 13 TeV pp collisions with the ATLAS detector, Rep. Prog. Phys. **88**, 057803 (2025).
[27] CMS Collaboration, The CMS experiment at the CERN LHC, J. Instrum. **3**, S08004 (2008).
[28] CMS Collaboration, Search for neutral Higgs bosons decaying to tau pairs in pp collisions at $\sqrt{s} = 7$ TeV, Phys. Lett. B **713**, 68 (2012).
[29] ATLAS Collaboration, Search for the standard model Higgs boson in the H to $\tau^+\tau^-$ decay mode in $\sqrt{s} = 7$ TeV pp collisions with ATLAS, J. High Energy Phys. 09 (2012) 070.
[30] CMS Collaboration, Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons, J. High Energy Phys. 05 (2014) 104.
[31] ATLAS Collaboration, Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector, J. High Energy Phys. 04 (2015) 117.
[32] CMS Collaboration, Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector, Phys. Lett. B **779**, 283 (2018).
[33] ATLAS Collaboration, Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Rev. D **99**, 072001 (2019).
[34] CMS Collaboration, Measurements of Higgs boson production in the decay channel with a pair of τ leptons in proton–proton collisions at $\sqrt{s} = 13$ TeV, Eur. Phys. J. C **83**, 562 (2023).
[35] ATLAS Collaboration, Measurements of Higgs boson production cross sections in the $H \to \tau^+\tau^-$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, J. High Energy Phys. 08 (2022) 175.
[36] CMS Collaboration, Measurement of the inclusive and differential Higgs boson production cross sections in the decay mode to a pair of τ leptons in pp collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. Lett. **128**, 081805 (2022).
[37] CMS Collaboration, Measurement of the production cross section of a Higgs boson with large transverse momentum in its decays to a pair of τ leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Lett. B **857**, 138964 (2024).
[38] ATLAS Collaboration, Differential cross-section measurements of Higgs boson production in the H $\tau^+\tau^-$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, J. High Energy Phys. 03 (2025) 010.
[39] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands, An introduction to PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015).

[40] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, DELPHES 3, a modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.
[41] Jerzy Neyman and Egon Sharpe Pearson, On the problem of the most efficient tests of statistical hypotheses, Phil. Trans. R. Soc. A **231**, 289 (1933).
[42] S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. **9**, 60 (1938).
[43] Roger J. Barlow, Extended maximum likelihood, Nucl. Instrum. Methods Phys. Res., Sect. A **297**, 496 (1990).
[44] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, and David Cournapeau *et al.*, Array programming with NumPy, Nature (London) **585**, 357 (2020).
[45] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, and Tyler Reddy *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).
[46] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, and Zhifeng Chen *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015, Software available from https://www.tensorflow.org/.
[47] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, On calibration of modern neural networks, arXiv:1706.04599.
[48] Kyle Cranmer, Juan Pavez, and Gilles Louppe, Approximating likelihood ratios with calibrated discriminative classifiers, arXiv:1506.02169.
[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.*, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).
[50] F. James and M. Roos, MINUIT: A system for function minimization and analysis of the parameter errors and correlations, Comput. Phys. Commun. **10**, 343 (1975).
[51] C. R. Rao, Information and the accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc. **37**, 81 (1945).
[52] H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, NJ, 1946).
[53] https://www.clip.science/.
[54] L. Benato, C. Giordano, C. Krause, A. Li, R. Schoefbeck, D. Schwarz, M. Shooshtari, and D. Wang, Training Data for Gollum in Higgs Uncertainty Challenge (Zenodo, 2025), 10.5281/zenodo.15322773.
[55] W. Bhimji *et al.*, FAIR Universe—HiggsML Uncertainty Challenge Public Dataset (Zenodo, 2025), 10.5281/zenodo.15131565.
[56] Abraham Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Am. Math. Soc. **54**, 426 (1943).