Developing and Applying Machine Learning Techniques for Model-Agnostic Searches for New Physics at the LHC

Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Physik
der Universität Hamburg

vorgelegt von
Manuel Roland Sommerhalder

Hamburg
2024

Gutachter/innen der Dissertation:	Prof. Dr. Gregor Kasieczka Prof. Dr. Peter Schleper
Zusammensetzung der Prüfungskommission:	Prof. Dr. Johannes Haller Prof. Dr. Sarah Heim Prof. Dr. Gregor Kasieczka Prof. Dr. Daniela Pfannkuche Prof. Dr. Peter Schleper
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Daniela Pfannkuche
Datum der Disputation:	28.01.2025
Vorsitzender des Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Markus Drescher
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Wolfgang J. Parak
Dekan der Fakultät MIN:	Prof. Dr.-Ing. Norbert Ritter

Eidesstattliche Versicherung / Declaration on oath

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben.

Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated.

If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

Hamburg, den 30.09.2024

Unterschrift der Doktorandin/des Doktoranden

Abstract

The aim of this thesis is to search for physics beyond the Standard Model (SM) with minimal model dependence. Traditional searches for new physics at the Large Hadron Collider (LHC) target specific signatures based on well-motivated models. In contrast, model-agnostic approaches, typically depending on loose data selections or an accurate SM background description, lack high sensitivity for new physics signals. Recently, a new paradigm of model-agnostic searches has emerged based on anomaly detection, which is aimed at automatically identifying deviations from the background expectation in the data using machine learning. This work introduces CATHODE, a novel anomaly detection method, which combines generative models with classifiers, achieving state-of-the-art performance on a simulated benchmark dataset, particularly in addressing the more challenging setting of correlated observables. Further improvements to enhance the background estimation stability and the robustness against uninformative input features are proposed. CATHODE is applied to search for heavy resonances decaying to two large-radius jets with anomalous substructure in 13 TeV proton-proton collisions recorded by the CMS Experiment at the LHC. While no significant excess over the SM background is observed, it enables setting upper cross-section limits across a broad spectrum of signal models. In many cases, the sensitivity is shown to improve substantially over conventional generic search strategies. By demonstrating the potential of targeting many different signal models in a single analysis, this thesis lays the groundwork for more such data-driven, model-agnostic searches in the future.

Zusammenfassung

Das Ziel dieser Arbeit ist die Suche nach Physik jenseits des Standardmodells (SM) mit minimaler Modellabhängigkeit. Traditionelle Suchen nach neuer Physik am Large Hadron Collider (LHC) sind auf spezifische Signaturen ausgerichtet, die auf gut motivierten Modellen basieren. Im Gegensatz dazu sind modellunabhängige Ansätze, die typischerweise von einer losen Datenauswahl oder einer akkuraten SM-Untergrundbeschreibung abhängen, nicht besonders sensitiv für Signale neuer Physik. In jüngster Zeit hat sich ein neues Paradigma für modellagnostische Suchen herausgebildet, das auf der Erkennung von Anomalien basiert und darauf abzielt, mithilfe von maschinellem Lernen automatisch Abweichungen von der Untergrunderwartung in den Daten zu identifizieren. In dieser Arbeit wird CATHODE vorgestellt, ein neuartiges Verfahren zur Erkennung von Anomalien, das generative Modelle mit Klassifikationsverfahren kombiniert und auf einem simulierten Benchmark-Datensatz die beste Leistung erzielt, insbesondere bei der Bewältigung der schwierigeren Situation korrelierter Observablen. Es werden auch Verbesserungen vorgeschlagen, um die Stabilität der Untergrundabschätzung und die Robustheit gegenüber uninformativen Inputs zu erhöhen. CATHODE wird für die Suche nach massiven Resonanzen eingesetzt, die in zwei Jets mit großem Radius und anomaler Substruktur in 13 TeV Proton-Proton-Kollisionen zerfallen, die vom CMS-Experiment am LHC aufgezeichnet wurden. Es wird zwar kein signifikanter Exzess gegenüber dem SM-Untergrund beobachtet, aber es ermöglicht die Festlegung von Obergrenzen von Wirkungsquerschnitten für ein breites Spektrum an Signalmodellen. In vielen Fällen zeigt sich, dass die Sensitivität im Vergleich zu herkömmlichen generischen Suchstrategien erheblich verbessert wird. Durch die Demonstration des Potenzials, viele verschiedene Signalmodelle in einer einzigen Analyse zu berücksichtigen, legt diese Arbeit den Grundstein für weitere datengesteuerte, modellagnostische Suchen in der Zukunft.

Contents

Preface

The work presented in this thesis is the result of my research during my employment at the Institute of Experimental Physics at the University of Hamburg (UHH) from 2020 to 2024, in collaboration with other researchers. During this period, I have been a member of the CMS Collaboration at CERN. The results presented in this thesis are a more detailed account of the following publications:

- Anna Hallin, Joshua Isaacson, Gregor Kasieczka, Claudius Krause, Benjamin Nachman, Tobias Quadfasel, Matthias Schlaffer, David Shih, and Manuel Sommerhalder. "Classifying anomalies through outer density estimation". In: Phys. Rev. D 106, 055006 (Sept. 2022) [1].
- Anna Hallin, Gregor Kasieczka, Tobias Quadfasel, David Shih, and Manuel Sommerhalder. "Resonant anomaly detection without background sculpting". In: Phys. Rev. D 107, 114012 (June 2023) [2].
- Thorben Finke, Marie Hein, Gregor Kasieczka, Michael Krämer, Alexander Mück, Parada Prangchaikul, Tobias Quadfasel, David Shih, and Manuel Sommerhalder. "Tree-based algorithms for weakly supervised anomaly detection". In: Phys. Rev. D 109, 034033 (Feb. 2024) [3].
- Philip Harris, William Patrick McCormack, Sang Eon Park, Tobias Quadfasel, Manuel Sommerhalder, Louis Moureaux, Gregor Kasieczka, Oz Amram, Petar Maksimovic, Benedikt Maier, Maurizio Pierini, Kinga Anna Wozniak, Thea Klæboe Årrestad, Jennifer Ngadiuba, Irene Zoi, Samuel Kai Bright-Thonney, David Shih, and Aritra Bal. "Machine learning techniques for model-independent searches in dijet final states". CMS Note: CMS-NOTE-2023-013 (Nov. 2023) [4].
- CMS Collaboration. "Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV". CMS Physics Analysis Summary: CMS-PAS-EXO-22-026 (March 2024) [5].

In addition to the work discussed in this thesis, I have contributed to the following publications during my research period at UHH:

- CMS Collaboration. "Search for narrow resonances in the b-tagged dijet mass spectrum in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: Phys. Rev. D 108, 012009 (July 2023) [6].
- Tobias Golling, Gregor Kasieczka, Claudius Krause, Radha Mastandrea, Benjamin Nachman, John Andrew Raine, Debajyoti Sengupta, David Shih, and Manuel Sommerhalder. "The interplay of machine learning-based resonant anomaly detection methods". In: Eur. Phys. J. C 84 (Mar. 2024) [7].

Moreover, I contributed to the operation of the CMS Experiment by performing the following service tasks:

- MC generation contact for the Trigger Studies Group,
- MC production manager,
- Development of offline data quality monitoring tools.

I co-supervised the following Bachelor and Master thesis projects:

- Tore von Schwartz. “Semi-supervised learning as a new tool for particle physics analysis”. Bachelor Thesis at the University of Hamburg (2021) [8].
- Parada Prangchaikul. “Mitigating uninformative features in weak supervision”. Master Thesis at the University of Hamburg (2023) [9].

In the course of my research and writing this thesis, I have used the generative AI tools listed in the following, in accordance with the “declaration on oath”. In particular, none of the AI-generated suggestions were implemented without thorough manual review and modification.

- GitHub Copilot: for generating snippets of source code in analysis scripts and LaTeX documents. The latter includes suggestions for this thesis document.
- ChatGPT: for assisting in literature research and suggesting language refinement in the thesis text.

1 Introduction

The field of high-energy physics is concerned with the study of the most fundamental constituents of matter and their interactions. The established theoretical framework that describes these elementary particles and forces is the Standard Model (SM), and its predictions have been confirmed by numerous experiments. A major breakthrough in verifying the SM was the discovery of one of its key components, the Higgs boson, by the ATLAS [10] and CMS [11] Collaborations at the Large Hadron Collider (LHC) in 2012. However, the SM is known to be incomplete as it does not account for several observed phenomena, such as gravity, dark matter or neutrino masses. This incompleteness has motivated a plethora of theories beyond the SM (BSM) that predict new particles and interactions, which are being searched for with experiments.

The typical approach to searching for BSM physics with collider experiments at the LHC is the *dedicated search* paradigm: a selection is applied to collision event data, such that the sensitivity to a specific BSM signal model is maximized, i.e., the selection is optimized to reject as many events from SM background processes as possible while retaining events that could arise from the hypothesized BSM signal based on its predicted signature. This selection is often designed using machine learning (ML) techniques. Statistical tests are then performed to decide whether the data passing the selections are more compatible with the existence of the BSM signal or with the background-only hypothesis. While this approach is very powerful for finding a specific model of BSM physics, it is notable that despite the vast number of searches performed by the LHC experiments, no BSM process has been discovered so far. On the one hand, this lack of discovery could be due to the absence of measurable BSM effects at the targeted energy scale. On the other hand, it could be that many new particles have already been produced at the LHC, but they are overlooked because they do not match the targeted signature of any dedicated search so far. The latter option cannot be excluded because the dedicated searches are mainly sensitive to the specific signal model for which they were designed. Such a search would need to be performed individually for every possible BSM theory and measurable signature therein, which is unfeasible in terms of person power and computational resources. Moreover, a true BSM signal might not be found if the correct theory has not yet been conceived.

To circumvent this issue, several searches with more relaxed model assumptions have been performed. Both the ATLAS [12] and CMS [13] Collaborations have performed so-called *general searches* where generic deviations between the data and a simulated SM background expectation are tested for their significance. However, these searches come at the cost of being highly sensitive to the accuracy of the background modeling, and the statistical significance of any deviation is diluted by the large number of tests performed. Another approach is to make the relatively loose assumption of a smooth background distribution in a kinematic variable where a signal might appear as a resonant peak. This facilitates a data-driven background estimation rather than relying on a simulated background model, and it targets a broad range of well-motivated signal models. A prominent example of such a generic signature, which has been targeted by ATLAS [14, 15] and CMS [16, 17], is the decay of a heavy particle into hadrons resulting in two large-radius jets in the detector. However, the loose selections inevitably accept a large number of SM background events, which dilutes the statistical significance of any present signal.

Another class of model-agnostic searches is based on *anomaly detection* techniques, a subfield of ML concerned with identifying deviations from a concept of normality that is learned from data. If such an algorithm is calibrated to view the SM background as "normal", it could be used to define selections that retain the anomalous BSM signal events in the data while rejecting SM background. Combined with a data-driven background estimation, this effectively facilitates a search that is independent of both the signal and the background models. To date, this type of anomaly detection paradigm has been applied within a few BSM searches at the LHC by the ATLAS Collaboration [18–20].

In this thesis, I introduce a novel method for anomaly detection–based analyses that I developed with collaborators during my PhD studies. This method, named CATHODE, substantially improves upon previous methods in terms of signal sensitivity. In particular, previous similar methods relied on the absence of significant correlations between certain input features, and this limitation reduces the generality of the methods. It will be shown that the signal sensitivity of CATHODE is robust against such correlations.

Two more improvements to CATHODE and similar methods are discussed in this thesis. First, I examine the feasibility of data-driven background estimation techniques after anomaly detection–based selections are applied. It is shown how the earlier-mentioned type of input feature correlations can still obscure the background estimation, even if the signal sensitivity of an approach is robust against them. I discuss an improvement to CATHODE, named La-CATHODE, that addresses this issue. Second, I investigate the impact of uninformative input dimensions on the performance of anomaly detection methods and propose a modification to CATHODE, relying on tree-based ML methods rather than neural networks, that substantially increases the resilience of the method towards such uninformative features. As a consequence, the method can use more information in the anomaly detection task and thus cover a broader range of potential BSM signals.

Finally, the thesis describes in detail how I applied CATHODE to a real BSM anomaly detection search. Specifically, I performed a search for heavy resonances produced in proton-proton collisions with a center-of-mass energy of 13 TeV and decaying into final states of two large-radius jets with anomalous substructure in the CMS Experiment. The CATHODE method was used in the context of a larger analysis where multiple state-of-the-art anomaly detection methods were employed simultaneously on the same data and targeting the same final state. This thesis focuses on the deployment of CATHODE but also discusses the final results of the full analysis.

The document is structured as follows. Section 2 introduces the physics background and methodology, and Sec. 3 provides the ML basics for establishing the foundation of the anomaly detection methods discussed in this thesis. Section 4 gives an overview of various relevant ML anomaly detection methods, and Sec. 5 discusses previous model-agnostic searches at the LHC. The methodological studies that led to the development of the CATHODE method and its improvements are presented in Sec. 6. The anomaly detection–based BSM search in the CMS Experiment is discussed in Sec. 7. Finally, the thesis is concluded in Sec. 8.

2 Particle Physics Theory and Methods

This thesis is concerned with the field of *high-energy physics* (HEP), also referred to as particle physics, which studies the fundamental particles and forces of nature at the smallest scales. This section will provide an overview of the theoretical foundation and the experimental methods that are relevant for the studies presented in this thesis.

2.1 The Standard Model of Particle Physics

The established theory of particle physics is the Standard Model (SM), which describes the fundamental particles of matter and their interactions, and whose predictions have been confirmed experimentally with unprecedented precision. It is formulated in the framework of quantum field theory (QFT), which is a combination of classical field theory, special relativity and quantum mechanics. As a full treatment of QFT and the SM is beyond the scope of this thesis, only the most essential aspects in view of later sections will be covered. A more detailed introduction to these topics can be found in Refs. [21–25], which were used as basis for this section. The most up-to-date information on the SM can be found in the Particle Data Group (PDG) review [26]. It should be noted that this thesis will use the convention of *natural units* where the speed of light c and the reduced Planck constant $\hbar$ are set to unity, i.e., $c = \hbar = 1$.

2.1.1 The Lagrangian

The key object for formulating a QFT is the Lagrangian density $\mathcal{L}$, which is a function of the quantum fields and their derivatives. As is custom in particle physics, it will also simply be referred to as *Lagrangian* in the following. The dynamic behavior of the fields is described by the Euler-Lagrange equations, which are derived from the principle of least action:

$$\frac{\partial \mathcal{L}}{\partial \varphi} - \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)} \right) = 0, \tag{2.1}$$

where φ is a field described by the Lagrangian and ∂_μ is the four-gradient in relativistic index notation, in which the Greek indices run over the spacetime dimensions $0, 1, 2, 3$ and double indices imply summation over them. The individual terms in the Lagrangian can be associated with dynamics of the fields, such as kinetic terms, potential terms, and interaction terms.

A symmetry of the Lagrangian is a transformation of the fields that leaves the Lagrangian invariant. This is a key concept in multiple regards. The Noether theorem states that every continuous symmetry corresponds to a conserved quantity (e.g., translation symmetry implies conservation of momentum). Moreover, a symmetry under local gauge transformations ($\varphi \to e^{i\alpha(x)}\varphi$, where the operator $\alpha(x)$ is a function of spacetime coordinates) implies the existence of gauge bosons, which are the mediators of the fundamental forces. Lastly, the contemporary approach to stating a theory of particle physics is to state specific symmetries as a fundamental principle, and then construct the Lagrangian with all terms that are compatible with these symmetries.

The SM Lagrangian is based on local $U(1) \times SU(2) \times SU(3)$ gauge group symmetries, where $U(1)$ is the group of unitary 1×1 matrices, $SU(2)$ is the group of special unitary 2×2 matrices, and $SU(3)$ is the group of special unitary 3×3 matrices. The simple compressed notation, which finds its ways onto shirts and mugs of particle physicists is as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{SM}} = &-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \\ &+ i\bar{\psi}\not{D}\psi + \text{h.c.} \\ &+ \bar{\psi}_i y_{ij}\psi_j\phi + \text{h.c.} \\ &+ |D_\mu\phi|^2 - V(\phi).\end{aligned} \tag{2.2}$$

In the following paragraphs, the relevant terms will be explained in more detail. The abbreviation h.c. refers to the Hermitian conjugate of the preceding term.

The second line in Eq. 2.2 describe the dynamics of matter, which consists of fermions. These are spin-$\frac{1}{2}$ particles, and their fields are here represented in a shortened notation ψ, which relates to an array of multiple fermion fields. The full list of known fermions is summarized in Tab. 2.1. The notation $\not{D} \equiv \gamma^\mu D_\mu$ is the summation of Dirac gamma matrices γ^μ with the covariant derivative D_μ. The latter is defined as:

$$D_\mu = \partial_\mu - ig_1\frac{Y}{2}B_\mu - ig_2\frac{\tau^i}{2}W^i_\mu - ig_3\frac{\lambda^a}{2}G^a_\mu, \tag{2.3}$$

where g_1, g_2, g_3 are the coupling strengths of the $U(1)$, $SU(2)$, and $SU(3)$ gauge groups, respectively. The Y is the weak hypercharge operator, τ^i are the three weak isospin operators represented by Pauli matrices, and λ^a are the eight color charge operators represented by Gell-Mann matrices. The fields B_μ, W^i_μ, and G^a_μ are the gauge bosons of the $U(1)$, $SU(2)$, and $SU(3)$ groups, respectively. The appearance of these spin-1 gauge fields is a direct consequence of the local gauge symmetries of the SM, and they necessarily transform as $V^\mu \rightarrow -\frac{i}{g}(\partial^\mu U)U^{-1} + UV^\mu U^{-1}$ under gauge transformations, where V^μ is the respective gauge field, g is the coupling strength, and U is the gauge transformation. The gauge field terms in the covariant derivative result in interaction terms between the fermions and the respective gauge bosons, which are responsible for the fundamental forces. The G^a_μ directly relate to the eight gluons that mediate the strong interaction. Gluons couple to particles that carry color charge, which are both the quarks and the gluons themselves. The fields B_μ and the three W^i_μ mix as a linear combination governed by the weak mixing angle θ_W to form the fields corresponding to the physical photon and the $\mathrm{W}^\pm$ and Z bosons. Photons mediate the electromagnetic interaction and couple to all particles with electric charge, which is a combination of the weak hypercharge and the weak isospin. These electrically charged particles are the quarks, the charged leptons (electron, muon, tau) and the $\mathrm{W}^\pm$ bosons. The weak interaction is mediated by the $\mathrm{W}^\pm$ and Z bosons and couples only to right-handed chiral states of the fermions. A summary of the properties of the physical (vector) gauge bosons is provided in Tab. 2.2.

The first term in Eq. 2.2 describes the dynamics of the gauge bosons. The field strength tensor $F_{\mu\nu}$ is generally defined as:

$$F_{\mu\nu} = \partial_\mu V_\nu - \partial_\nu V_\mu - ig\left[V_\mu, V_\nu\right], \tag{2.4}$$

where the square brackets denote the commutator $[A, B] = AB - BA$. Equation 2.2 implies such a field tensor for every gauge field.

Table 2.2 shows the measured masses for Z and W bosons. If a gauge field V^μ was massive with mass m_V, the Lagrangian would necessarily involve a term of the form: $\frac{1}{2}m_V^2 V_\mu V^\mu$. However, this can be shown to break the gauge symmetry of the Lagrangian. The solution to this apparent contradiction is the Higgs mechanism [27–29], which introduces a complex scalar field ϕ, transforming as an SU(2) doublet, with a potential $V(\phi)$, as shown in the last term of Eq. 2.2.

Table 2.1: Summary of the Standard Model fermions, including their masses in natural units and their charges in units of the elementary charge e. Measured masses are taken from Ref. [26].

	Leptons		**Quarks**	
1st Generation	Electron (e)	Electron Neutrino (ν_e)	Up (u)	Down (d)
Mass	0.511 MeV	≈ 0	2.16 MeV	4.70 MeV
Electric Charge	-1	0	2/3	-1/3
2nd Generation	Muon (μ)	Muon Neutrino (ν_μ)	Charm (c)	Strange (s)
Mass	106 MeV	≈ 0	1.27 GeV	93.5 MeV
Electric Charge	-1	0	2/3	-1/3
3rd Generation	Tau (τ)	Tau Neutrino (ν_τ)	Top (t)	Bottom (b)
Mass	1.78 GeV	≈ 0	173 GeV	4.18 GeV
Electric Charge	-1	0	2/3	-1/3

Table 2.2: Summary of the Standard Model vector bosons, including their masses in natural units and their charges in units of the elementary charge e. Measured masses are taken from Ref. [26].

	Electromagnetic	**Weak**		**Strong**
	Photon (γ)	Z Boson (Z)	W Boson (W)	Gluon (g)
Mass	0	91.2 GeV	80.4 GeV	0
Electric Charge	0	0	±1	0

With an explicit form $V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2$, and $\mu^2 < 0$, the minimum of the potential is not at $|\phi| = 0$, but at a non-zero value $|\phi| = \sqrt{\frac{-\mu^2}{\lambda}} \equiv v$. Expanding this Higgs field ϕ around the minimum, one finds a massive real scalar field and three massless Goldstone bosons, where the latter can be absorbed in a gauge choice. This results in couplings of the gauge bosons to the vacuum expectation value v in the form of mass terms $\frac{1}{2}\left(\frac{vg}{2}\right)^2 V_\mu V^\mu$, without the need of adding explicit gauge boson mass terms to the Lagrangian. This process is known as *spontaneous symmetry breaking*: the symmetry of the Lagrangian is broken because the lowest-energy state of the system does not exhibit the full symmetry. The excitation of the real scalar field is the Higgs boson, which has been experimentally discovered at the LHC in 2012 [10, 11] with a mass around 125 GeV.

The third term in Eq. 2.2 describes the interaction of the fermions with the Higgs field. Similar to the heavy gauge bosons, this results in mass terms for the fermions, which are proportional to the vacuum expectation value v and the corresponding Yukawa coupling constants y_{ij}. These y_{ij} (and consequently the fermion masses) are not predicted by the SM, but are free parameters that must be measured. The masses of the fermions are summarized in Tab. 2.1.

2.1.2 Renormalization and Running Coupling

In order to compute predictions from the SM, one usually relies on *perturbation theory* where the transition amplitude leading from an initial state to a final state is expanded in a series of Feynman diagrams, whose vertices arise from interaction terms in the Lagrangian. This is performed at a finite order in the coupling parameter $\alpha = g^2/4\pi$, where g is the coupling strength of the physical interaction that appears in every vertex. This expansion is only valid at small values of α. Starting from next-to-leading order (quadratic in α), the individual diagrams involve integrals over the momenta of intermediate particles, which results in infinite values. This is known as the problem of *ultraviolet divergences.* The solution to this problem is the process of *renormalization*, where the divergences are absorbed in the definition of the parameters of the theory. A consequence, which is governed by the renormalization group equations, is that the parameters of the theory depend on the energy scale at which they are measured. Prominently, the coupling parameter α itself becomes “running”. For example in the case of the strong interaction, the coupling has the form:

$$\alpha_s(q^2) = \frac{\alpha_s(\mu^2)}{1 + \frac{\alpha_s(\mu^2)}{12\pi}(33 - 2n_f)\log\left(\frac{-q^2}{\mu^2}\right)}, \tag{2.5}$$

where q^2 is the probed energy scale, μ^2 is a reference scale at which α_s has been measured, and n_f is the number of active quark “flavors”, i.e., the individual species of quarks. With $n_f = 6$, the running in Eq. 2.5 implies decreasing coupling strength at higher energies, which is known as *asymptotic freedom.* This is a key feature of the strong interaction, also referred to as Quantum Chromodynamics (QCD). At a scale below 1 GeV, the coupling becomes so strong that quarks and gluons are confined into color-neutral hadrons, which is known as *confinement.* Perturbation theory is not applicable in this regime and other methods need to be chosen for predictions.

2.1.3 Shortcomings of the Standard Model

The predictions of the Standard Model have been confirmed by many experiments with remarkable precision. However, it exhibits several limitations that lead to the belief that it is not the final theory of particle physics. The following paragraphs briefly discuss an incomplete list of these issues.

A prominent shortcoming of the SM is the lack of a description of the gravitational force. Compared to the strong and electroweak interactions, the measurable effects of gravity at the length scales of HEP are much smaller, which makes probing the gravitational force at the quantum level challenging. The effects of a quantum theory of gravity are expected to become relevant at an energy scale of the order of 10^{19} GeV, known as the Planck scale.

However, at the much larger length scale of astrophysical observations, gravity is a dominant force. Observations of the rotational velocity of stars in spiral galaxies are inconsistent with the visible matter distribution. This led to the hypothesis that a significant fraction (roughly one quarter) of mass-energy in the universe is in the form of *dark matter*, which does not interact electromagnetically. The nature of this dark matter is unknown, and while it could consist of some form of particles, there are no dark matter candidates in the SM.

Table 2.1 describes the masses of the neutrinos as “almost zero”. While the SM Lagrangian does not include mass terms for neutrinos, there is experimental evidence for neutrino oscillations, i.e., a time-dependent mixing of neutrino flavors, which can only occur if at least two of the neutrino flavors have non-zero masses.

The SM treats the existence of three generations of both leptons and quarks as given, without a deeper explanation of their number and the symmetry between quarks and leptons. While their masses are attributed to the Yukawa coupling with the Higgs field, these coupling strengths are free parameters of the theory, and it is not clear why their values differ by multiple orders of magnitude.

Another open question concerning large variance in mass scales is the so-called *hierarchy problem.* The Higgs boson mass, which is measured to be around 125 GeV, is sensitive to quantum corrections that are proportional to the Planck scale. A very precise fine-tuning of the virtual corrections is required to keep the Higgs boson mass at the vastly lower value that is observed. In addition to this, the origin of the λ term in the Higgs potential and consequently the vacuum expectation value of the Higgs field, is not understood from a deeper theoretical perspective.

2.2 Physics Beyond the Standard Model

Finding a theory that extends the Standard Model to address the issues mentioned in Sec. 2.1.3 is an active field of research. Such theories are referred to as *beyond the Standard Model* (BSM) theories. They often predict new particles, and verifying their existence in an experimental setting is an important test of these theories. A few classes of such BSM theories will be briefly discussed in the following paragraphs.

2.2.1 Supersymmetry

Supersymmetry (SUSY), which is discussed extensively in Ref. [30], is a theoretical framework that introduces an additional symmetry between bosons and fermions, i.e., the SUSY Lagrangian is symmetric under an operator that changes between integer and half-integer spin particles, leaving the other quantum numbers unchanged. As a consequence, every SM particle has a superpartner with a spin that differs by $\frac{1}{2}$, which effectively results in at least twice the number of fundamental particles. Since no such superpartners of SM particles with equal masses have been measured, the supersymmetry needs to be spontaneously broken, resulting in superpartners with substantially higher masses.

The existence of these superpartners provides a solution to the hierarchy problem, because the virtual corrections to the Higgs boson mass from SM particles are canceled by the opposite-sign contributions of their superpartners. In addition, SUSY offers a natural dark matter candidate in the form of a stable lightest superpartner. It is also seen as a means of unifying the forces because at high energies, the running couplings computed from SUSY theories tend to become equal for the strong, weak, and electromagnetic interactions.

2.2.2 Composite Quarks and Leptons

Many theories predict that quarks and leptons are not fundamental particles, but are composed of more elementary constituents, often called “preons”, that appear at smaller scales. An overview of such theories is provided in Ref. [31]. Composite fermion models are aimed at explaining why three generations of both quarks and leptons exist, and in particular from where their highly hierarchical mass structure originates. Moreover, the fractional charges of quarks, as seen in Tab. 2.1, could arise as a natural consequence of a composite structure.

In composite models, the SM fermions are low-energy bound states of their constituents, thus one would expect to probe the preons at much higher energies. Furthermore, compositeness implies an inner dynamic system, and thus the existence of excited quark and lepton states that may decay into their SM counterparts.

2.2.3 Composite Higgs Bosons

Another class of theories revolves around the idea that the Higgs boson is a composite particle arising from the spontaneous symmetry breaking related to a new strong force. Examples of such models are discussed in Refs. [32, 33]. This would mitigate the hierarchy problem, as the Higgs boson mass would depend on the binding energy of its constituents rather than the Planck scale.

These models typically predict the existence of new gauge bosons associated with the extended symmetry group, for example particles called W′ and Z′ bosons that are analogous to the W and Z bosons of the SM. The new strong interaction also gives rise to a spectrum of bound states, such as “vector-like” fermions. These are heavier fermions whose left- and right-handed chiral components transform equally under the SM gauge group. The masses of the SM fermions can be explained by their mixing with these vector-like fermions.

2.2.4 Warped Extra Dimensions

A different angle of addressing the hierarchy problem is to hypothesize the existence of an additional spatial dimension that separates the Planck scale from the electroweak scale. The most prominent model of this kind is the Randall-Sundrum model [34], which introduces a compact extra dimension that is warped, i.e., the metric of the extra dimension is not flat but space-time and energy scales change exponentially along its direction. The two scales define separate lower-dimensional subspaces called “branes”. Some fields are confined to one of the branes, while others can propagate through the “bulk” space in between. The hierarchy problem is addressed via the exponential warp factor, which suppresses the Planck scale effects for the Higgs field that is confined to the electroweak brane. The gravity field, on the other hand, originates at the Planck brane and its effects on the electroweak brane is suppressed due to propagation through the bulk.

Extensions to the Randall-Sundrum model, such as the one discussed in Ref. [35], predict the existence of Kaluza-Klein (KK) excitations: excited states of the SM particles that arise from quantized momenta in the extra dimension. In the 4D effective theory seen from the electroweak brane, these states appear as massive particles with a mass spectrum that is determined by the geometry of the extra dimension. Moreover, the distance between the branes is subject to fluctuations, which can give rise to a scalar field called the “radion” that couples to SM particles.

2.3 Measurements at Particle Colliders

The predictions of both the SM and BSM theories are being tested at HEP experiments, which are often performed at particle colliders. This section relates the theory introduced in Sec. 2.1 to measurable quantities in collider experiments. A particular focus lies on proton-proton collision experiments, which will be the setting for the studies in this thesis. As in Sec. 2.1, the content of this section is based on the more detailed discussions in Refs. [21–25].

2.3.1 Decay Widths and Cross Sections

The transition probability between an initial state (i) and a final state (f) is described by the Lorentz-invariant matrix element $\mathcal{M}_{fi}$, which arises from the interaction terms in the Lagrangian and is computed in practice using Feynman diagrams in perturbation theory at fixed order. The differential transition rate per unit time, which corresponds to units of energy in natural units, is described by:

$$dw_{fi} = (2\pi)^4 V^{1-n_i} \delta^4 \left(\sum_{k=1}^{n_f} p_k^{(f)} - \sum_{l=1}^{n_i} p_l^{(i)} \right) |\mathcal{M}_{fi}|^2 \prod_{l=1}^{n_i} \frac{d^3 p_l^{(i)}}{2E_l^{(i)}} \prod_{k=1}^{n_f} \frac{d^3 p_k^{(f)}}{(2\pi)^3 2E_k^{(f)}}, \tag{2.6}$$

where V is the interaction volume, $n_{i/f}$ are the number of initial/final state particles, $E_k^{(i)/(f)}$ are the energies of the particles in the initial and final states, respectively, and $p_k^{(i)/(f)}$ are their momenta. The delta function enforces the conservation of energy and momentum in the transition. The full transition rate is obtained by integrating over the allowed phase space. Important special cases for experimental searches are the decay width and the cross section, discussed in the following.

In the case where the initial state is a single particle A decaying into a specific final state, the transition rate is referred to as the *partial decay width*: $\Gamma_{A\to f} = w_{fA}$. The sum over all possible final states is the *total decay width*: $\Gamma_A = \sum_f \Gamma_{A\to f}$. The *branching fraction* of a specific final state is defined as its partial decay width divided by the total decay width. In natural units, the total decay width is related to the lifetime of the particle by $\tau_A = 1/\Gamma_A$, i.e., the mean time interval before a particle decays. The lifetime is not a Lorentz-invariant quantity and is thus always stated in the rest frame of the particle.

In the context of scattering, where the initial state consists of two particles A and B, the transition rate is conventionally normalized by the flux, which is the number of incoming particles per area per time. The resulting quantity is referred to as the *cross section*, $\sigma_{A+B\to f} = w_{f(AB)}/\text{flux}$, which is Lorentz invariant. Its relationship to the number of interactions N (also referred to as *events*) in an experiment is described with the definition of *luminosity* L:

$$\frac{dN}{dt} = L\sigma, \tag{2.7}$$

which characterizes the performance of the particle accelerator beam. The amount of recorded data in an experiment is often stated in terms of the *integrated luminosity*, $L_{\text{int}} = \int L dt$, which is proportional to the number of events.

2.3.2 Resonances

In scattering interactions, it can occur that the two initial state particles form an unstable intermediate particle R that decays into a final state: $A + B \to R \to f$. These intermediate states are called *resonances*. The cross section for this process then follows a Breit-Wigner distribution:

$$\sigma_{A+B\to f} = c(A,B,R)\frac{\Gamma_{R\to AB}\Gamma_{R\to f}}{(E_R - E)^2 + \Gamma_R^2/4}, \tag{2.8}$$

where the factor $c(A,B,R)$ depends on the quantum numbers of the initial and intermediate particles, E_R is the energy of the intermediate particle, and E the energy of the system, i.e., the available energy in the initial state. Figure 2.1 illustrates the typical peak-like shape of such a resonance decay. While the partial widths $\Gamma_{R\to AB}$ and $\Gamma_{R\to f}$ determine the height, the resonance energy determines the location and the total resonance width Γ_R relates to the width of the peak. Varying the energy by $\pm\Gamma_R$ around E_R reduces the cross section by half, hence it is also referred to as the "full width at half maximum". Since in the center-of-mass frame the energy E_R is the mass of the resonance, measuring the location and width of such resonance peaks is a common method to infer the mass and lifetime of the intermediate particle.

2.3.3 Proton Scattering

While the SM predictions can be tested with a variety of experiments, the following sections will be concerned with the collisions of highly accelerated protons. The proton is a stable composite particle of two up quarks and one down quark, which are bound by the strong force. It has one positive charge in units of e and a mass of 938 MeV [26]. When colliding two protons at very high energy, the interaction can be approximately described by the interaction of two constituents, one from each proton, called *partons* because of the mechanism of asymptotic freedom (Sec. 2.1.2). The cross section $\hat{\sigma}_{i+j\to f}$ involving partons i and j can be computed with

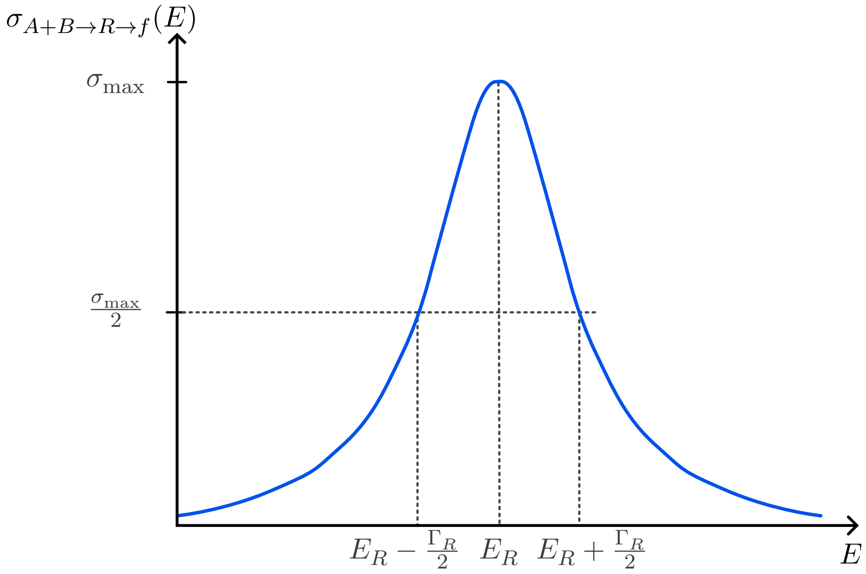

Figure 2.1: The typical Breit-Wigner resonance shape arising from the intermediate production of an unstable particle R between initial state $A + B$ and final state f.

perturbative calculations, and they carry only a fraction x of the respective proton momentum: $p_{\text{parton}} = xp_{\text{proton}}$.

It is important to note that the partons comprise not only the three bound quarks, but the inner dynamics of the proton involves virtual gluons and quark-antiquark pairs, which can all participate in the interaction. The probability of probing parton i with momentum fraction x is described by the *parton distribution function* (PDF), $f_i(x, \mu_F)$, which is a non-perturbative quantity that is determined from fits to experimental data and evolves as a function of the factorization scale μ_F according to the DGLAP evolution equations [36–38]. The value of μ_F is a free parameter that corresponds to the choice of energy scale that separates non-perturbative calculations (PDFs) from perturbative calculations (parton cross sections). The factorization theorem of QCD then facilitates approximating the proton-proton cross section as a convolution of the PDFs and the partonic cross sections, summed over all possible parton interactions:

$$\sigma_{pp\to f} = \sum_{i,j} \int_0^1 dx_1 \int_0^1 dx_2 f_i(x_1, \mu_F) f_j(x_2, \mu_F) \hat{\sigma}_{i+j\to f}(x_1, x_2, \mu_F, \mu_R). \tag{2.9}$$

Here, the dependence of both f_i and $\hat{\sigma}_{i+j\to f}$ on μ_F is made explicit. Moreover, the renormalization scale μ_R is introduced, which enters the cross section computation as briefly mentioned in Sec. 2.1.2. A complete computation to all orders would eliminate the dependence on μ_F and μ_R.

2.4 The CMS Detector

2.4.1 The Large Hadron Collider

Located at the border between France and Switzerland, the Large Hadron Collider (LHC) is the world's largest particle accelerator with a circumference of 27.6 km. It is operated by the European Organization for Nuclear Research (CERN) and is designed to accelerate and collide either protons or heavy ions. The LHC features two beam pipes, which cross at four interaction points where the particle beams are made to collide within the detectors ATLAS, ALICE, CMS, and LHCb, respectively. Superconducting magnets are used to guide the beams along the circular path, with dipoles for bending and quadrupoles for focusing [39].

The particles are not accelerated individually, but instead in *bunches* of the order of 10^{11} particles. These bunches are separated by a fixed time interval, which is typically 25 ns. The luminosity is given by:

$$L = \frac{N_1 N_2 n_b f_\text{rev}}{A}, \tag{2.10}$$

with N_1 and N_2 being the number of particles per bunch in each beam, n_b the number of bunches in the ring, f_rev the frequency with which a bunch completes one loop of the ring, and A the effective beam overlap cross section at the interaction point. The luminosity can be increased by increasing the number of bunches or decreasing the area by focusing the beams with strong quadrupole magnets [39].

So far, two data-taking periods have been completed at the LHC. The first run (Run 1) took place from 2011 to 2013, with a center-of-mass energy that was increased from 7 TeV in 2011 to 8 TeV in 2012. The second run (Run 2) started in 2015 and continued until 2018, with a center-of-mass energy of 13 TeV. A third run (Run 3) is ongoing since 2022 with a center-of-mass energy of 13.6 TeV and is planned to end in 2025 [40]. A common convention is to refer to the center-of-mass energy of the LHC as $\sqrt{s}$, which is the square root of the Mandelstam variable s that describes the total squared energy in the collision.

2.4.2 Detector Design

The Compact Muon Solenoid (CMS) detector is a general-purpose detector located at one of the interaction points of the LHC, designed to measure the properties of particles produced in proton-proton collisions with high precision. Its design is illustrated in Fig. 2.2. While the sketch and the following paragraphs provide a brief overview of the detector components, a more detailed description can be found in Ref. [41]. It should be noted that the detector has undergone several upgrades since its construction, and the following description reflects its state in 2017. The CMS detector has a cylindrical architecture, comprising concentric layers around the interaction point. Detector parts covering the lateral surface are referred to as the *barrel* section, and those in the direction of the beam axis are called *endcap*.

Coordinate System

The geometry of the CMS detector is conventionally described with a Cartesian coordinate system that has its origin at the nominal collision point. The x-axis points radially inward toward the center of the LHC, the y-axis points vertically upward, and the z-axis points along the beam direction. In spherical coordinates, the radius r corresponds to the distance from the origin in the x-y plane, the azimuthal angle ϕ measures the angle from the x-axis in the x-y plane, and the polar angle θ is measured from the z-axis. A preferred alternative to θ is the pseudorapidity η, defined as $\eta = -\log\tan(\theta/2)$, because particle production is roughly constant as a function of η, and differences in pseudorapidity $\Delta\eta$ are Lorentz invariant under boosts along the beam axis in the case of massless particles. The vector components in the x-y plane are referred to as *transverse*, such as the transverse momentum p_T and the transverse energy E_T.

Magnetic Field

The key component is the superconducting solenoid, which generates a magnetic field of 3.8 T along the beam axis. The primary purpose of the magnetic field is to bend the trajectories

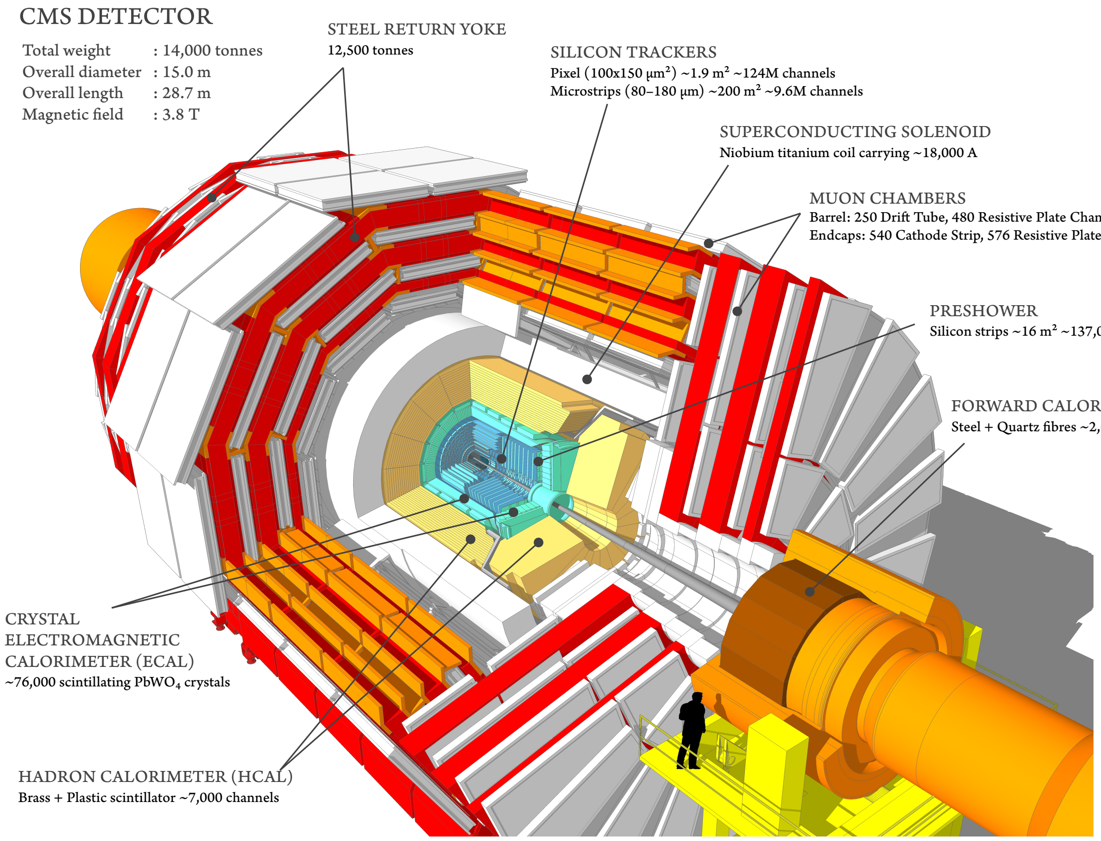
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm²) ~1.9 m² ~124M channels
Microstrips (80–180 μm) ~200 m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000 A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16 m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 2.2: Schematic illustration of the CMS detector with descriptions reflecting its state in 2017. Figure taken from Ref. [42].

of charged particles, which facilitates the measurement of their momenta and electric charges. The solenoid consists of niobium-titanium coils that are cooled down to 4.5 K. It has an internal diameter of 6 m and a length of 12.5 m. Outside the solenoid, a steel return yoke guides the magnetic field back to the solenoid, confining it to the detector volume and resulting in a uniform opposite-sign field in the outer barrel region for the muon system.

Tracking

The trajectories of charged particles are commonly referred to as *tracks*, and they are measured by the innermost silicon tracking system. These silicon detectors are subdivided into small units with either rectangular area, so-called pixels, or long strips. As charged particles pass through these units, they ionize the silicon atoms and the resulting free charges are collected by readout electronics. With the positional information of the pixels or strips, these signals are interpreted as localized hits, which are then reconstructed into tracks with geometric algorithms. Tracks are crucial to reconstruct *vertices*—the locations of particle decays into charged particles. The CMS tracking system consists of two sub-detectors.

First, the innermost pixel detector [43] comprises multiple layers of pixels with an area of $150 \times 100\,\mu\mathrm{m}$ each. The barrel section consists of four layers starting at a radius of around 3 cm and extending to a radius of 16 cm, and the endcap part has three disks on each side, ranging from a distance of around 3 to 5 cm along the beam axis. The pixel detector thus provides a fine-grained measurement of (r, ϕ, z) coordinates close to the interaction point, which is crucial for vertex reconstruction.

The second part of the tracking system is the strip detector. With ten barrel layers of silicon strips, it extends the precise measurement of (r, ϕ) coordinates to radii from 20 cm to 116 cm. The endcap part consists of 12 disks with radial strips on each side, which facilitate a ϕ measurement up to $|z|$ of 282 cm. The typical length of the strips is around 10 cm, and the distance between strip centers varies from 80 to 184 µm, depending on the layer.

Overall, the tracking system is designed to cover the pseudorapidity range $|\eta| < 2.5$ and to measure charged particles with a transverse momentum of at least 1 GeV. It is designed with a low material budget to minimize energy loss and multiple scattering of particles.

Calorimetry

The next layers after the tracker are the calorimeters. A general review of calorimetry in HEP can be found in Ref. [44]. These detectors consist of dense materials in which incident particles are fully absorbed, and their energy is inferred from measuring either the emitted light or the charge. Highly energetic electrons interact mostly by bremsstrahlung, which emits photons. High-energy photons interact primarily via electron-positron pair production. These emitted secondary particles again interact with the material, thus resulting in a shower of particles until the energy falls below a threshold where energy loss is dominated by ionization and excitation without producing secondary particles. The energy released by charged particles in the material from the latter processes is proportional to the incident particle energy and is often measured via their emission of light using photodetectors, such as photomultiplier tubes. In some active materials, such as semiconductors, the energy can also be inferred from the charge. Hadrons interact with the material mostly via the strong force, causing a more complex cascade of secondary particles. As the absorption of hadrons requires a denser material than electrons and photons, calorimeters are typically specialized for either electromagnetic (ECAL) or hadronic (HCAL) particles. Calorimeters can be further classified into *homogeneous* calorimeters, in which both the interaction and the measurement of the energy are performed in the same material, and *sampling* calorimeters, which consist of alternating layers of a dense absorber and an active medium for measuring the energy.

The CMS ECAL is a homogeneous calorimeter consisting of scintillating lead tungstate crystals. The barrel region of the ECAL starts at a radius of 1.29 m and covers the pseudorapidity range of approximately $|\eta| < 1.5$. The crystals have a cross-sectional area of $22 \times 22\,\mathrm{mm}^2$ in the front face and $26 \times 26\,\mathrm{mm}^2$ in the rear face, with a length of 230 mm. The crystals are of similar size in the endcap region, which covers the pseudorapidity range of $1.5 < |\eta| < 3.0$ and starts at a distance of 315 cm from the interaction point. The ECAL in the endcap region is complemented by a preshower detector within a pseudorapidity range of $1.7 < |\eta| < 2.6$. The preshower detector is designed to help distinguish neutral pions from single photons and consists of two layers of lead absorber and silicon strips.

The HCAL system in the CMS detector is a sampling calorimeter consisting of multiple sub-detectors. The barrel region of the HCAL covers the pseudorapidity range of $|\eta| < 1.3$ and starts at a radius of 1.77 m. It consists of 14 layers of brass plates and a steel plate in the front and back, respectively. These absorber layers have a typical thickness of around 50 mm and are interleaved with plastic scintillator tiles with a thickness of mostly 3.7 mm as the active material. The light emitted in the scintillator is collected by wavelength-shifting optical fibers and read out by photodetectors. The scintillators are divided into sectors to each cover a segment of $(\Delta\eta, \Delta\phi) = (0.087, 0.087)$. In this central region of $|\eta| < 1.3$, the material of the sampling calorimeter is not sufficient to fully contain hadronic showers. Thus, it is complemented by the *outer calorimeter*, consisting of 1–2 additional layers of scintillator tiles just outside the solenoid, treating the solenoid as an absorber. The endcap part of the HCAL covers the pseudorapidity range of $1.3 < |\eta| < 3.0$ and consists of 18 layers of brass plates with a thickness of 79 mm each, separated by 9 mm-thick scintillator tiles. Similar to the barrel, the HCAL endcap region has a component outside the solenoid: The *forward calorimeter* is located at 11.2 m from the interaction point at each side and has an inner radius of 12.5 cm from the beam line. It covers the pseudorapidity range of $3.0 < |\eta| < 5.0$ and is based on steel absorber layers with embedded quartz fibers.

Muon System

With their relatively high mass compared to electrons, muons only lose small amounts of energy when interacting electromagnetically with the detector material, and in contrast to hadrons, they do not interact via the strong force. As a consequence, their relatively long lifetime allows them to travel through the calorimeters and the solenoid. A dedicated muon system is thus built into the iron return yoke outside the solenoid, in order to measure muon tracks with high precision.

The barrel part of the muon system covers a pseudorapidity range of $|\Delta\eta| < 1.2$ and radii of around 4.0 to 7.3 m. It consists of four stations interspersed among the layers of the iron return yoke, each consisting of multiple chambers along the z-axis. The chambers consist of several layers of drift tubes, which are long tubes filled with gas that is ionized by passing muons. The time it takes for the ionization electrons to drift to a central wire is measured to determine the distance between the muon interaction and the wire. These drift tube cells are arranged in layers with offsets of half a cell to avoid blind spots.

The endcap part of the muon system covers the pseudorapidity range of $0.9 < |\Delta\eta| < 2.4$. Its four stations range from 5.7 to 10.6 m from the interaction point and consist of cathode strip chambers. These gas-filled chambers comprise anode wire planes interleaved with cathode panels. If a muon ionizes the gas, the electrons drift to the anode wires, and the position is determined with information from the positive charge signal on the cathode strips. Compared to drift tubes, the more costly cathode strip chambers have a finer segmentation, a faster response, and are more resilient to radiation damage. This is needed in the endcap region due to the non-uniform magnetic field and the higher particle flux.

Both the barrel and endcap muon stations are complemented by resistive plate chambers. With a strong electric field between the resistive plates, the ionization electrons ionize surrounding gas atoms, and the resulting avalanche of electrons is collected by readout strips. While the position measurement is less precise than the drift tubes or cathode strip chambers, the resistive plate chambers have a significantly faster response and are thus primarily used for the trigger system.

2.4.3 Trigger System

With a bunch separation of 25 ns, the rate of events at the LHC is 40 MHz, which is significantly more than can realistically be processed and stored. Moreover, only a small fraction of these events are of interest for physics analyses. The decision of which events are recorded is handled by the *trigger system*, which is described in more detail in Refs. [45, 46]. The CMS trigger system comprises two tiers.

The Level-1 (L1) trigger is the first stage and is based on custom hardware processors. Its target is to reduce the event rate to around 100 kHz with a latency of 4 µm. This is achieved by using coarsely segmented information from the calorimeters and the muon system. These so-called trigger primitives are processed separately and then combined into objects whose signatures are consistent with electrons, photons, muons, tau leptons or jets (see Sec. 2.5.1). A global trigger step ultimately evaluates these objects against up to 512 pre-defined selection criteria to decide whether to pass the event to the next stage. These criteria range from simple single-object requirements, such as passing a minimum transverse energy threshold, to more complex multi-object requirements, such as the computation of invariant masses.

The High-Level Trigger (HLT), whose state during Run 2 is described in Ref. [47], is the second stage, which further reduces the event rate to around 1 kHz. It is based on a farm of commercial processors that run a streamlined version of the offline reconstruction algorithms with information from all sub-detectors, including the tracker. Over 600 different HLT paths are employed, targeting a broad range of signatures that are deemed interesting for physics analyses. An HLT path is a sequence of reconstruction and filtering modules that are arranged to increase in computational cost. If a filter module rejects an event, the remainder of the path is skipped, thus saving computational resources. Many HLT paths are designed to target generally interesting signatures, such as the presence a pre-defined number of jets, muons, electrons or photons above a p_T threshold. Other paths target more specific signatures, such as the presence of displaced jets.

2.4.4 Particle Reconstruction

Only a small set of SM particles is stable enough to travel from the interaction point to the detector and leave measurable traces inside. These are electrons, muons, photons, and the lighter hadrons such as pions and kaons. These particles leave characteristic signatures in the detector components.

Electrons typically leave tracks in the inner tracking system and deposit energy in the ECAL. Muons result in tracks both in the inner tracking system and in the outer muon system, while depositing only small amounts of energy in the calorimeters through ionization. Photons are uncharged and thus invisible in the inner tracker, but they result in ECAL showers. Charged hadrons leave tracks in the inner tracker and deposit energy in the ECAL and HCAL. The signature of neutral hadrons is to deposit energy in the ECAL and HCAL without a track in the inner tracker. Neutrinos are not directly detectable because of their negligible interaction with matter in the detector. Instead, their presence is inferred from the imbalance in the transverse energy, since the colliding partons are expected to collide with approximately zero transverse momentum. This missing transverse component is referred to as missing transverse energy $E_\mathrm{T}^\mathrm{miss}$.

The CMS Collaboration uses a *particle-flow* (PF) algorithm to reconstruct the individual particles produced in the collisions. This algorithm aims at identifying and reconstructing all stable particles in the event, using the information from all sub-detectors simultaneously. It is described in full detail in Ref. [48], and it essentially follows the four-step procedure outlined in the following.

The first step of PF is to reconstruct the *PF elements*, which are the basic building blocks of the algorithm. In the inner tracking system, these are the tracks and their originating vertices. Their reconstruction is based on a combinatorial track finder algorithm, based on a Kalman filter [49], that uses the hits in the silicon detectors. In the calorimeters, the PF elements are energy clusters, which are reconstructed with a dedicated clustering algorithm. In the muon system, the PF elements are based on track segments, which are reconstructed from hits in the muon stations.

As a second step, a linking algorithm is applied to combine the PF elements across the various sub-detectors in pairs. Depending on the nature of the PF elements, specific linking criteria are applied. For example, a track in the inner tracker is linked to a calorimeter cluster if the extrapolated track position is within the cluster boundaries. Each link is assigned a distance metric, which is used to determine the quality of the link. The resulting chains of linked PF elements are referred to as *PF blocks*.

Once the PF blocks are reconstructed, the third step is to identify and reconstruct the particles from the PF blocks. This is based on criteria related to the particle signatures discussed earlier.

The final step is to apply a postprocessing step to improve the quality of the reconstructed particles. In particular, the presence of misreconstructed and/or misidentified muons can cause an overestimation of the $E_{\mathrm{T}}^{\mathrm{miss}}$ per event.

The output of the PF algorithm are the reconstructed particles, which are typically referred to as *PF candidates*. These are used as input to the physics analyses directly or as input to further reconstruction algorithms, such as jet clustering algorithms, which are discussed in Sec. 2.5.2. The stored properties of the PF candidates include the four-momentum vector and the type of particle. For hadrons the distinction is only made between charged and neutral ones—no further distinction is made into the specific type of hadron.

2.5 Jet Physics

If a high-energy particle collision results in the creation of quarks and gluons, they will not be detectable as free particles due to the nature of the strong force. Instead, they will be observable only as collimated sprays of color-less hadrons. These are referred to as *jets*, and they will be the main physics object of study throughout this thesis. The following subsections will briefly introduce the theoretical background of jets and then introduce how they are reconstructed in an experiment, such as CMS.

2.5.1 Jet Theory

The formation of jets is a direct consequence of the nature of QCD, where partons carrying color charge cannot be observed as free particles but instead are confined into color-neutral hadrons. For a more detailed discussion of the theoretical aspects of jets, the reader is thus referred to QCD literature, such as Ref. [50]. The description of the jet formation process is usually factorized into two parts based on the energy scale of the process: parton showering and hadronization.

When a high-energy parton is produced in a collision, it can emit additional partons in a process called *parton showering*. Specifically, a quark can radiate off a gluon, and gluons can either split into a quark-antiquark pair or another pair of gluons. The probability for a parton to split into two lower-energy partons is described by *splitting functions*, which can be derived from perturbative QCD calculations because of the high momentum scale. This process continues as a cascade of emissions until the energy of the partons drops into the non-perturbative regime. The radiation angles of subsequent emissions are limited by the angular emission cone of the previous one, a phenomenon known as *angular ordering*, which arises from the color coherence of the system. This results in the typical collimated shape of jets.

The lower-energy partons resulting from the parton shower subsequently undergo *hadronization*, during which they form color-neutral hadrons. As this process cannot be calculated via perturbative QCD, phenomenological models are used to describe it in practice. A common such model is the Lund String Model [51], which is implemented in the widely used PYTHIA event generator [52]. In this model, pairs of partons are connected by a one-dimensional string that represents the QCD potential, which increases linearly as a function of the parton separation due to the self-interaction of the color-carrying gluon mediator. As the partons separate, the energy in the potential increases until it becomes energetically more favorable to break into a quark-antiquark pair with opposite color charge. This process continues iteratively until the kinetic energy of the partons is low enough to bind them into hadrons.

2.5.2 Clustering Algorithms

The reconstruction of jets in an experiment is typically done with a *jet clustering algorithm*, which groups particle-like entities in the event into (usually a variable number of) jets. The underlying entities may be calorimeter clusters or reconstructed particles, such as PF candidates. The latter is the default approach in the CMS Experiment.

There are different classes of jet clustering algorithms. An important one comprises the *sequential recombination algorithms*. They rely on an angular difference metric Δ_{ij} between entities i and j, which is defined as:

$$\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \tag{2.11}$$

where ϕ is the azimuthal angle and y is the rapidity defined as $y = \frac{1}{2}\log\left(\frac{E+p_z}{E-p_z}\right)$. For massless particles, this rapidity definition is identical to the pseudorapidity η. The generalized distance metric d_{ij} between entities i and j and the distance of i with respect to the beam d_{iB} is then defined as:

$$d_{ij} = \min(k_{Ti}^{2p}, k_{Tj}^{2p})\frac{\Delta_{ij}^2}{R^2}, \tag{2.12}$$

$$d_{iB} = k_{Ti}^{2p}, \tag{2.13}$$

with k_{Ti} being the transverse momentum of entity i, R being a parameter that defines the size of the jet, and p being a parameter that defines the specific algorithm. The algorithm proceeds by identifying the smallest distance in the list of entities. If it is a d_{ij} pair, the entities i and j are combined into a new entity. If it is a d_{iB} pair, entity i is identified as a jet and removed from the list. This process is repeated until no entities are left.

The general formulation of sequential recombination algorithms above was identified by Ref. [53]. With a choice of $p = 1$, it recovers the k_T algorithm [54], with $p = 0$ it corresponds to the Cambridge-Aachen (CA) algorithm [55] and $p = -1$ is referred to as the anti-k_T algorithm [53]. This choice of specific algorithm effectively determines the order in which entities are merged. Roughly speaking, the k_T algorithm starts by merging low-momentum entities, the CA algorithm only considers angular separation, and the anti-k_T algorithm starts by adding low-momentum entities to high-momentum ones. More specifically, the anti-k_T algorithm builds jets around the highest-momentum particles if no other high-momentum particle is within an angular distance of $2R$, and these jets will be approximately circular with radius R in the y-ϕ plane. If two high-momentum particles are close with a distance of $R < \Delta_{ij} < 2R$, they will form two circular jets with a straight line boundary in between. If they are closer than R, they will first unify into a single complex jet and then accumulate low-momentum radiation around them. This centrality around high-momentum particles with an outer boundary that is insensitive to low-momentum radiation is the reason why the anti-k_T algorithm is nowadays the most commonly used jet clustering algorithm in HEP.

Any well-defined jet clustering algorithm is subject to theoretical constraints. In perturbative QCD calculations, divergences appear from integrals over intermediate states, which are cancelled by taking into account contributions of real emissions with momenta and angles below the measurement resolution. If the outcome of a jet clustering algorithm changes due to low-momentum emissions or collinear splitting, these cancellations could not occur, and no observables could be computed based on this jet definition. This property is referred to as *infrared* *and collinear (IRC) safety*, and all three clustering algorithms mentioned above can be shown to be IRC safe.

Once a jet is defined by a group of entities, global properties are derived from the entities within the jet. For example, the four-momentum of the jet is usually taken to be the vector sum of all the constituent momenta.

2.5.3 Pileup Mitigation

As discussed in Sec. 2.4.1, the LHC does not accelerate and collide single particles, but rather bunches of the order of 10^{11} protons. As a consequence, each collision event will involve additional low momentum transfer proton-proton interactions, referred to as *pileup* (PU) interactions. For example, the CMS Experiment recorded a mean number of 38 interactions per bunch crossing in 2017 [56]. In order to reconstruct the primary (high momentum transfer) interaction of interest, the effect of pileup needs to be mitigated. In the context of jet reconstruction, this is particularly important because the particles from pileup interactions will otherwise be clustered into jets of the primary interaction and distort the kinematic properties of the jets such that they are no longer representative of the primary interaction partons.

A common strategy, employed in the CMS Experiment [57], is to identify the primary interaction vertex and the PU vertices based on the reconstructed tracks. The vertex with the largest value of the summed p_{T}^2 of physics objects associated with it is selected as the *leading vertex* (or primary vertex). The rest of the vertices are considered PU vertices. This information can be used to filter out any PF candidates that are associated with PU vertices before the jet clustering algorithm is applied. This is done in the method called *charged hadron subtraction* (CHS). However, only charged hadrons leave tracks in the detector. The CHS approach corrects the jet four-momentum vector for neutral hadrons after the clustering steps based on the expected PU density within the jet area [48]. Since only global jet properties are corrected, the substructure (Sec. 2.5.5) and shape of the jet are left untreated for neutral hadron contributions.

The *pileup per particle identification* (PUPPI) algorithm [58] improves over CHS by assigning a weight between zero and one to each particle based on the information of surrounding particles. This weight describes the probability of the particle originating from the leading vertex. The particle (PF candidate) four-momenta are then rescaled based on these weights to correct the PU entirely at the particle level before jet clustering. Charged particles are assigned a weight of 1 if they are associated with the leading vertex and 0 if they are associated with a PU vertex. Neutral particles are assigned a weight based on the weights of the surrounding charged particles, with the proximity defined by the angular distance $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ in the η-ϕ plane.

2.5.4 Grooming Algorithms

In the context of jet physics, the term *jet grooming* refers to a wide class of postprocessing (after clustering) algorithms that are designed to systematically remove jet constituents that are thought to obscure the process of interest. This typically includes particles from PU interactions, initial-state radiation (ISR), and contributions of the interactions of the other partons in the interacting proton. Each of these processes tend to result in additional wide-angle low-momentum radiation. The reconstructed mass of the groomed jets is usually more compatible with the mass of the original particle that initiated the jet. This is particularly important in the context of boosted objects, where the decay products of a heavy particle are collimated into a single jet, discussed in Sec. 2.5.5. The most commonly used grooming algorithms are trimming [59], pruning [60], mass drop [61], and soft drop [62].

The *soft drop* algorithm [62] has become a popular choice in the realm of boosted jet physics. Its first step is to recluster the constituents of a jet (independently of the original choice of clustering algorithm) with the CA algorithm. This results in a pair-wise clustering tree with an angular-ordered structure. The algorithm then proceeds by undoing the last clustering step of jet j, which results in two subjets j_1 and j_2, and then testing whether these subjets pass the soft drop condition:

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{\text{cut}} \left(\frac{\Delta_{12}}{R_0}\right)^{\beta}, \tag{2.14}$$

where p_{T1} and p_{T2} are the transverse momenta of the subjets, Δ_{12} is the angular distance between the subjets in the y-ϕ plane, R_0 is the jet radius, and z_{cut} and β are parameters of the algorithm. If the subjets pass the condition, the jet j is considered the final soft-drop jet. If they do not pass, then the subjet with larger p_{T} becomes the new jet j while the other subjet is removed, as it corresponds to wide-angle and low-momentum radiation. This process is iterated until the condition is either satisfied or the jet is reduced to a single particle. In either case, the jet is considered the final soft-drop jet.

2.5.5 Jet Substructure

If a particle with high Lorentz boost decays, its products are collimated to an extent as they might fall within a single, typically large-radius, jet. The so-called *substructure* of this jet, consisting of overlapping smaller jets, is thus non-trivial and can be used to infer the nature of the original particle.

Several observables have been developed to quantify this substructure, such as the *N-subjettiness* [63], which measures the compatibility of a jet with a given number of subjets. The most common definition of N-subjettiness is

$$\tau_N = \frac{1}{d_0} \sum_k p_{\text{T},k} \min\left\{\Delta R_{1,k}, \Delta R_{2,k}, \cdots, \Delta R_{N,k}\right\}, \tag{2.15}$$

where the summation is performed over all particles in the jet, $p_{\text{T},k}$ are their transverse momenta, $\Delta R_{i,k}$ is the distance in the η-ϕ plane between the particle and the i-th subjet axis, and d_0 is a normalization factor:

$$d_0 = \sum_k p_{\text{T},k} R_0, \tag{2.16}$$

with R_0 being the characteristic jet radius of the original jet clustering algorithm. The definition assumes that a set of N candidate subjets is present. These are commonly found by re-clustering the jet with a jet clustering algorithm enforcing the desired number of N subjets. A more optimal choice would be to compute the minimum τ_N over every possible set of N candidate subjet axes, but in practice this is computationally intensive.

Small values of τ_N imply that the particles within the jet align well with the N subjet axes, i.e., the jet is likely to have N or fewer subjets. Large values, on the other hand, arise when a large fraction of energy is distributed at larger angles from the candidate subjet axes and thus the large-radius jet is more compatible with at least $N+1$ subjets. While a basis consisting of τ_N with multiple values of N can be used to identify the number of subjets with multivariate analysis techniques, the more interpretable ratios τ_N/τ_{N-1} are often used to quantify the substructure of a jet, as low values in this ratio indicate that the jet is well described by exactly N subjets. In the later sections, such N-subjettiness ratios will be denoted as $\tau_{XY} = \tau_X/\tau_Y$, and it should be clear from context whether the subscript denotes a ratio with X and Y being single-digit subjet numbers or a single XY-subjettiness value with XY being one two-digit subjet number.

Reference [63] also introduces a more general definition of N-subjettiness:

$$\tau_N^{\alpha,\beta} = \frac{1}{d_0} \sum_k \min_J \left\{ p_{T,k} \left(p_{T,J}\right)^{\alpha} \left(\Delta R_{J,k}\right)^{\beta} \right\}, \tag{2.17}$$

where the normalization factor d_0 becomes:

$$d_0 = \max_J \left\{ \left(p_{T,J}\right)^{\alpha} \right\} \left(R_0\right)^{\beta} \sum_k p_{T,k}. \tag{2.18}$$

Here, the index J runs over the N candidate subjets, and the parameters α and β can be used to weight the transverse momentum and angular separation of the particles with respect to the subjet axes, respectively. The original definition of Eq. 2.15 is recovered for $\alpha = 0$ and $\beta = 1$ and is the definition used within this thesis unless stated otherwise.

2.5.6 Heavy Flavor Tagging

Several algorithms have been developed to distinguish jets initiated by heavy quarks, either b or c, from so-called *light-flavor jets*, which are initiated by the u, d, and s quarks and the gluons. These algorithms are referred to as *heavy flavor tagging* algorithms, or *b tagging* and *c tagging* for the specific identification of the respective quark flavor jets. They facilitate the selection of more specific hadronic final states, e.g., for the measurement of $\mathrm{H} \to \mathrm{b\overline{b}}$ decays.

An overview of the heavy-flavor tagging algorithms used in Run 2 analyses of the CMS Experiment can be found in Ref. [64], which all rely on the different properties of the heavy-flavor hadrons compared to the light-flavor ones. These properties include the longer lifetimes of b and c hadrons, resulting in decay lengths of several millimeters up to a centimeter in the detector. This gives rise to the presence of displaced *secondary vertices* that are reconstructed in the tracking system. Moreover, the decay products of heavy-flavor hadrons tend to have larger transverse momenta relative to the jet axis compared to the other jet constituents, and in around 20% (10%) of the b (c) hadron decay chains, an electron or muon is present. One of these algorithms introduced in Ref. [64] is the DeepCSV tagger, which is a neural network classifier (see Secs. 3.1.1 and 3.3.1) that uses input features based on tracks that are clustered in the jet and secondary vertex properties, such as the distance to the primary vertex and the reconstructed secondary vertex mass.

The type of heavy-flavor tagging discussed above is applicable to the case where the jets are initiated by single partons. As mentioned in Sec. 2.5.5, jets may also arise from collimated decay products of a highly boosted even heavier particle, such as a W, Z or H boson, or a top quark. An overview of algorithms for this specific regime of *boosted* heavy flavor tagging is provided in Ref. [65] for the CMS Experiment, and multiple of them make use of substructure observables as discussed in Sec. 2.5.5.

2.6 Statistical Analysis Methods

This section will provide an introduction to the common statistical methods in HEP, used in the context of this thesis. A more detailed discussion of these concepts can be found in Refs. [26, 66].

In searches for new physics at the LHC, results are most commonly reported in terms of a frequentist interpretation. In this paradigm, one estimates the probability of observing the data under a specific hypothesis, where probability is identified with the relative frequency of a measurement in an ensemble of repeated equivalent experiments. This is in contrast to a

Bayesian interpretation, where the probability is identified with a degree of belief in a hypothesis, and the aim is to update the prior belief in a hypothesis based on the data.

The two main outputs of a frequentist statistical analysis in the context of new physics searches are the *significance* of an excess, and *exclusion limits* on parameters of a new physics model. They are both based on the *p-value*, which is the probability of observing an outcome that is at least as extreme as the one in the data, under the assumption of a specific hypothesis. The notion of what constitutes as "extreme" depends on further assumptions, such as an alternative hypothesis.

The significance of an excess is a measure of how unlikely it is to observe the data under the hypothesis of only known physics processes (referred to as *background*). It relates to the p-value of the data under the background-only hypothesis. In a more narrow sense, the significance in HEP usually refers to the Z-score, which is the number Z of Gaussian standard deviations above the median outcome under the background-only hypothesis. It can be computed from the p-value p as:

$$Z = \Phi^{-1}(1 - p), \tag{2.19}$$

where Φ^{-1} is the quantile function (the inverse of the cumulative distribution function) of the standard Gaussian distribution. The convention in HEP is to regard a significance of $Z \geq 5$ ($p \leq 2.87 \times 10^{-7}$) as a "discovery" of new physics, and any such result would be widely reported as a breakthrough in the scientific community. A significance of $Z \geq 3$ is often more weakly referred to as "evidence" for new physics. In addition to the *observed* significance in the data, the *expected* significance can be calculated as a measure of the experimental sensitivity for a given new physics (*signal*) model. It corresponds to the median (background-only) p-value one would find if the data were following the signal+background hypothesis.

Even if no significant excess is found in an analysis, exclusion limits can be set on parameters of a given new physics model. The most-targeted parameter is the signal strength, i.e., the expected number of signal events, which translates to the cross section. The upper limit on the signal strength corresponds to the value at which the p-value under the corresponding signal+background hypothesis becomes lower than a certain threshold value α. Signal strengths beyond this limit are considered excluded at a confidence level (CL) of $1-\alpha$. A common threshold is $\alpha = 0.05$, corresponding to a 95% CL. Even more often than in the case of significance, the observed limit is conventionally complemented by the expected limit, quantifying the sensitivity of the analysis to the new physics model. The expected limit is the parameter value for which the median p-value under the signal+background hypothesis and with (artificial) data arising exclusively from background is equal to α. In addition, one commonly reports the $\pm 1\sigma$ and $\pm 2\sigma$ bands around the expected limit, i.e., the p-values corresponding to the 2.3, 16, 84 and 97.7 percentiles in addition to the median. This provides context for the fluctuations of the observed limit, i.e., deviations of the observed limit from the expected limit by more than 2σ upwards or downwards are hints for an excess or deficit, respectively.

Figure 2.3 shows a typical example of how the significance (bottom) and exclusion limits (top) are reported in LHC searches, here for the discovery of the SM Higgs boson [10] by the ATLAS Collaboration. Section 2.6.1 will discuss how the underlying p-values for significance and limits are computed.

2.6.1 Hypothesis Testing

As discussed earlier, the statistical analysis revolves around excluding one hypothesis in favor of another one. The two hypotheses are commonly labeled as the *null hypothesis* H_0 and the *alternative hypothesis* H_1, respectively. In the context of computing the significance of excesses, H_0 is the background-only hypothesis and H_1 the presence of signal. For computing exclusion

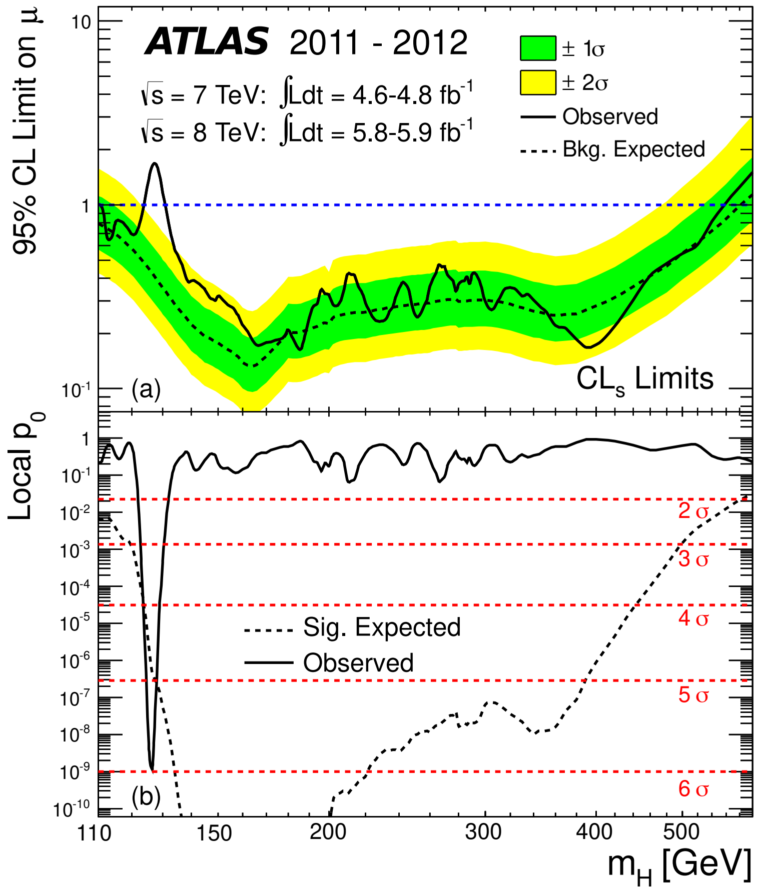

Figure 2.3: Adjusted figure from Ref. [10], showing the exclusion limit (a) and significance (b) as a function of the particle mass, resulting from the search for the SM Higgs boson by the ATLAS Collaboration. The 95% CL limit is set on a signal strength parameter μ and the significance is shown in terms of the local p-value under the background-only hypothesis p_0. The red dashed lines indicate the p-values corresponding to specific thresholds in terms of Z-score. The observed limit is complemented with the background-only expectation, and the observed significance is compared to the expected significance arising from the presence of the SM Higgs boson at the given mass. The presence of an SM Higgs boson signal at a mass of 125 GeV is visible both via the low p-value and the observed limit exceeding the expected bands.

limits, H_0 is the presence of signal with a certain parameter choice and H_1 the background-only hypothesis. The main quantity to measure the compatibility of the observed data x with either H_0 or H_1 is the so-called test statistic. The Neyman-Pearson lemma [67] states that the ratio of the likelihoods of the data under either hypothesis $L(x|H)$ is a most powerful test statistic for distinguishing between the two hypotheses:

$$\lambda(x) = \frac{L(x|H_1)}{L(x|H_0)}. \tag{2.20}$$

The condition is that the hypotheses are simple, i.e., they must not depend on undetermined parameters. The following paragraphs will discuss how to construct likelihoods and thus test statistics for typical HEP searches, and how to subsequently derive p-values for significance and limits.

A typical measurement in HEP consists of a histogram of an observable, $n = (n_1, n_2, ..., n_N)$, where the entries correspond to the number of collision events passing the analysis selections in each of the N bins. The expectation value for the bin number i is

$$\mathbb{E}(n_i) = \mu s_i + b_i, \tag{2.21}$$

where $b_i \in \mathbb{N}$ is the number of expected background events in the bin, $s_i \in \mathbb{N}$ is the number of signal events of some nominal total signal amount falling into bin i, and $\mu \in \mathbb{R}$ is the signal strength parameter. The parameter μ is the quantity of interest in the analysis, as it is related to the cross section of the signal process. The bins of the histogram are assumed to be statistically independent, and the likelihood function of the data given the signal strength μ is thus the product of Poisson probabilities for each bin j:

$$L_{\text{poisson}}(n, \mu) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)}. \tag{2.22}$$

The values s_i and b_i typically depend on signal and background models, respectively, which themselves depend on a set of additional parameters θ that are not entirely fixed. These are referred to as *nuisance parameters* and are often constrained by auxiliary measurements m. These m could be another histogram $m = (m_1, m_2, ..., m_M)$ with M bins and expectation value in bin i:

$$\mathbb{E}(m_i) = u_i(\theta), \tag{2.23}$$

where u_i are computable quantities depending on θ. In that case, one can introduce another Poisson likelihood term for the auxiliary measurement:

$$\pi(m, \theta) = \prod_{k=1}^{M} \frac{u_k(\theta)^{m_k}}{m_k!} e^{-u_k(\theta)}. \tag{2.24}$$

Another common choice for m are direct estimates of θ from other experiments or theoretical calculations, resulting in a prior distribution for θ. This could, for example, be a Gaussian constraint:

$$\pi(m, \theta) = \prod_{l=1}^{L} \frac{1}{\sqrt{2\pi}\sigma_l} e^{-\frac{(\theta_l - \theta_l^0)^2}{2\sigma_l^2}}, \tag{2.25}$$

where $m = (\theta_1^0, \theta_2^0, ..., \theta_L^0, \sigma_1, \sigma_2, ..., \sigma_L)$ are the measured values and uncertainties of the L nuisance parameters.

The full likelihood function of the data given the signal strength μ and the nuisance parameters θ is the product of the Poisson likelihoods for the data and the auxiliary measurements:

$$L(n, m, \mu, \theta) = L_{\text{poisson}}(n, \mu, \theta)\pi(m, \theta). \tag{2.26}$$

The explicit likelihood function of Eq. 2.26 relates to $L(x|H)$ in Eq. 2.20 with $x = \{n, m\}$ and $H = \{\mu, \theta\}$. One choice of H_0 and H_1 is to set H_0 to $\mu = 0$ and H_1 to a specific $\mu > 0$. However, with the introduction of nuisance parameters θ, they are no longer simple hypotheses, as required by the Neyman-Pearson lemma. The common frequentist strategy to circumvent this issue is to define the *profile likelihood ratio*:

$$\lambda_p(\mu) = \frac{L(n, m, \mu, \hat{\hat{\theta}}(\mu))}{L(n, m, \hat{\mu}, \hat{\theta})}, \tag{2.27}$$

where $\hat{\hat{\theta}}(\mu)$ are the values of θ that maximize $L(n, m, \mu, \theta)$ for a given μ, and $\hat{\mu}$ and $\hat{\theta}$ are the values that unconditionally maximize $L(n, m, \mu, \theta)$ over all μ and θ. The effect of including free parameters θ is to broaden the profile likelihood ratio for a given value of μ, which represents the loss of information due to systematic uncertainties.

With the profile likelihood ratio, one can define a test statistic:

$$t_\mu = -2 \log \lambda_p(\mu), \tag{2.28}$$

which is monotonically related to the optimal Neyman-Pearson test statistic, but has computational advantages due to the logarithm. High values of t_μ indicate increasing incompatibility between the data and the hypothesized μ. A p-value under the μ hypothesis can be obtained by integrating the probability density function of t_μ on data following a given μ, $f(t_\mu|\mu)$, from the observed value $t_{\mu,\text{obs}}$ to infinity:

$$p_\mu = \int_{t_{\mu,\text{obs}}}^{\infty} f(t_\mu|\mu) dt_\mu. \tag{2.29}$$

Obtaining the distribution $f(t_\mu|\mu)$ is in general non-trivial and could be either approached by running Monte Carlo (MC) simulation of data under the μ hypothesis, or by asymptotic formulae, derived in Ref. [66], that approximate the distribution of t_μ for large sample sizes. The latter is often used in practice, as it is computationally more efficient.

A common assumption in many (but not all) new physics searches, is that a signal process only adds more events, i.e., $\mu \geq 0$. An alternative test statistic, which has the highest agreement for the obtained estimator $\hat{\mu} < 0$ when $\mu = 0$, is defined as:

$$\tilde{t}_\mu = \begin{cases} -2 \log \frac{L(n,m,\mu,\hat{\hat{\theta}}(\mu))}{L(n,m,\hat{\mu},\hat{\theta})} & \hat{\mu} \geq 0 \ , \\ -2 \log \frac{L(n,m,\mu,\hat{\hat{\theta}}(\mu))}{L(n,m,0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0 \ . \end{cases} \tag{2.30}$$

An important special case, which is often used for discovery tests, is:

$$q_0 = \tilde{t}_0 = \begin{cases} -2 \log \lambda_p(0) & \hat{\mu} \geq 0 \ , \\ 0 & \hat{\mu} < 0 \ . \end{cases} \tag{2.31}$$

This results in a p-value under the background-only null hypothesis, p_0, which is used to quantify the significance of an excess in the data:

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f(q_0|0) dq_0. \tag{2.32}$$

In order to compute the expected significance of an experiment if the data were to actually arise from a signal+background hypothesis with signal strength μ', one would replace the $q_{0,\text{obs}}$ in the integration range of Eq. 2.32 by the median of the distribution of q_0 under the μ' hypothesis: $\text{med}\left(f(q_0|\mu')\right)$. The distribution $f(q_0|\mu')$ can again be numerically obtained via MC simulation, or via asymptotic formulae. Reference [66] derives a simple analytic approximation involving the so-called “Asimov dataset”, which is an artificial dataset constructed such that the estimated parameter values reproduce the true values.

For setting upper limits on the signal strength parameter μ, one needs a test statistic that results in lower compatibility between data and μ if the estimator $\hat{\mu}$ is smaller than μ. This is achieved by the test statistic:

$$q_\mu = \begin{cases} -2 \log \lambda_p(\mu) & \hat{\mu} \leq \mu \ , \\ 0 & \hat{\mu} > \mu \ . \end{cases} \tag{2.33}$$

In this notation, q_0 from Eq. 2.31 is not the same as q_μ with $\mu = 0$, but its own distinct definition. The definition Eq. 2.33 can be further restricted to the case where $\mu \geq 0$, leading to the alternative test statistic:

$$\tilde{q}_\mu = \begin{cases} -2\log \frac{L(n,m,\mu,\hat{\hat{\theta}}(\mu))}{L(n,m,0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0 \;, \\ -2\log \frac{L(n,m,\mu,\hat{\hat{\theta}}(\mu))}{L(n,m,\hat{\mu},\hat{\theta})} & 0 \leq \hat{\mu} \leq \mu \;, \\ 0 & \hat{\mu} > \mu \;. \end{cases} \tag{2.34}$$

With either test statistic from Eq. 2.33 or 2.34 (both will be implied when using q_μ in the following), one can compute the p-value for a given μ hypothesis by integrating the respective probability density function:

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)dq_\mu. \tag{2.35}$$

In order to find a $1-\alpha$ CL upper exclusion limit on μ, one may set Eq. 2.35 to α and, if possible, solve analytically for μ, or scan numerically for the μ that satisfies the equality.

Similarly to the case of expected significance, one can compute the expected limit, i.e., the upper exclusion limit achieved when the data are only arising from background, $\mu' = 0$, by replacing the $q_{\mu,\mathrm{obs}}$ in the integration range of Eq. 2.35 by the median of the distribution of q_μ under the $\mu' = 0$ hypothesis: med $\big(f(q_\mu|0)\big)$. Analogously to the observed limit, the expected limit is computed by setting this p-value to α and solving for μ. The same procedure applies for the $\pm 1\sigma$ and $\pm 2\sigma$ bands around the expected limit, replacing the lower integration range by the respective quantiles of $f(q_\mu|0)$. In practice, this is again mostly solved by asymptotic formulae involving an Asimov dataset. Figure 2.4 illustrates the quantities needed for setting exclusion limits.

2.6.2 CLs Method

The $1 - \alpha$ CL exclusion limits presented in Sec. 2.6.1 were based on computing the p-value under the signal+background hypothesis, p_μ, and finding the value of μ for which p_μ is equal to a threshold value α. This approach, which is sometimes referred to as CL_{s+b} limit, has the correct coverage under the frequentist interpretation, i.e., the probability of excluding a true signal is α.

The shortcoming of the CL_{s+b} exclusion limits is their behavior in the case of low experimental sensitivity, i.e., when predictions of the signal+background hypothesis are almost indistinguishable from the background-only prediction. In that case, there remains a probability of α to exclude the entire range of signal strengths down to zero. This is due to downward fluctuations of the data with respect to the background expectation. However, it is commonly regarded as unfavorable to report an exclusion limit in this case, as the data do not provide enough information to make a statement about the presence of a signal.

The most common remedy of this in HEP experiments is the CL_s test statistic [68, 69], defined as:

$$\mathrm{CL}_s = \frac{p_\mu}{1 - p_b}, \tag{2.36}$$

where p_b is the background-only p-value from test statistic q_μ:

$$p_b = \int_{0}^{q_{\mu,\mathrm{obs}}} f(q_\mu|0)dq_\mu. \tag{2.37}$$

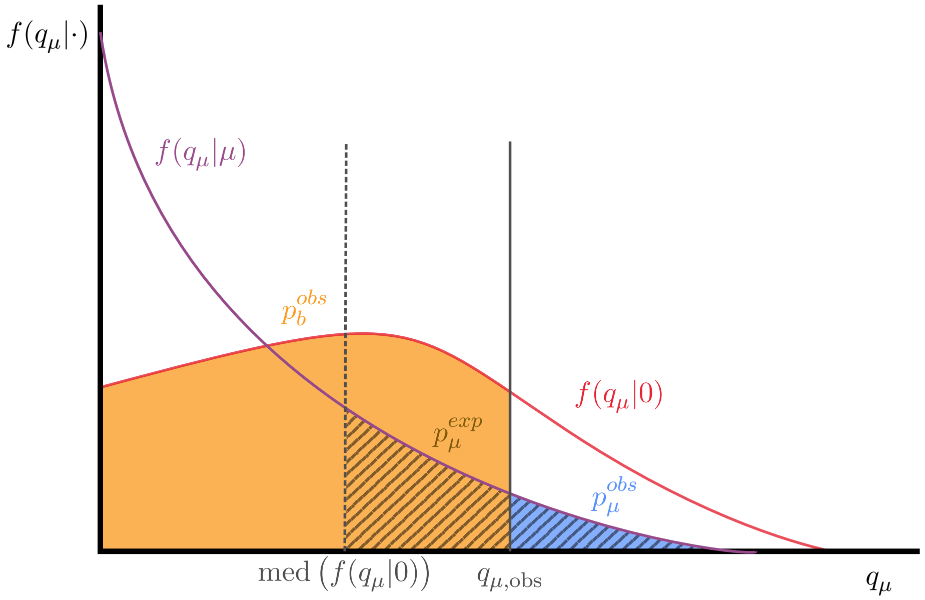

Figure 2.4: Illustration of statistical quantities needed for setting exclusion limits. Two distributions of the test statistic q_μ (signal+background hypothesis) are shown: with the nominal signal strength $\mu' = \mu$ in the data (purple curve) and with only background in the data $\mu' = 0$ (red curve). The blue area on the right of the observed value $q_{\mu,\mathrm{obs}}$ corresponds to the observed signal+background p-value p_μ, whereas the orange area is the observed p_b, which is needed for setting CL$_s$ limits. The corresponding quantities for the *expected* limits depend on the threshold of the median of the background-only distribution of q_μ, which are the expected p_μ (hatched area) and p_b (orange area left of $\mathrm{med}\left(f(q_\mu|0)\right)$).

The integral for p_b is performed over the left end of $f(q_\mu|0)$, as the test statistic q_μ is defined such that higher values indicate higher compatibility with the background-only hypothesis. By setting CL$_s = \alpha$ and solving for μ, one thus obtains the CL$_s$ exclusion limit. The quantities p_μ and p_b are illustrated in Fig. 2.4.

The CL$_s$ limits are more conservative in the sense that the coverage probability is strictly higher than the nominal confidence level $1 - \alpha$ because the denominator of Eq. 2.36 is smaller or equal to one. In the case of low sensitivity, a decrease in p_μ will be accompanied by a decrease in $1 - p_b$, and the exclusion of low μ is suppressed. In the case of high sensitivity, the CL$_s$ limit approaches the CL$_{s+b}$ limit. The conceptual disadvantage is the departure from a strictly frequentist interpretation of the result.

The same asymptotic approximations to the test statistic distributions $f(q_\mu|\mu')$ from Ref. [66] can be used to compute the observed CL$_s$ limit, as well as the expected limits with their $\pm 1\sigma$ and $\pm 2\sigma$ bands.

2.6.3 Background Estimation With Bump Hunts

The analysis procedure discussed in Sec. 2.6.1 relies on an expected background b_i in each bin of the histogram. This *background estimation* can be based on MC simulation following the known description of the background processes, and the nuisance parameters would account for systematic uncertainties in the background prediction. However, in many cases the modeling of the known background processes and the corresponding systematic uncertainties is not reliable enough. It might then be more feasible to resort to a *data-driven* background estimation, i.e., to estimate it directly from data. The *bump hunt* approach is a widely used technique for this purpose and has been adopted, e.g., in the discovery of the ρ meson [70] or the Higgs boson [10, 11]. While the term "bump hunt" is not fully standardized, it has often been used for this context in the recent literature. This section thus also establishes the terminology for the remainder of this thesis and in particular resolves potential confusion with the BumpHunter algorithm [71]. Other sources use the term to refer to the full statistical analysis procedure that involves the type of background estimation discussed here, but also the subsequent hypothesis testing. This thesis will use the term exclusively for the background estimation part.

The underlying assumption of a bump hunt background estimation is that the background b_i is only smoothly varying across the histogram bins i, and that any localized excess in only few bins is indicative of a signal process. The latter is thus an implicit restriction in s_i as well. This is well motivated in the case of resonance searches (Sec. 2.3.2) in invariant mass spectra, where known SM processes typically do not exhibit sharp peaks in contrast to signal models predicting the decay of massive new particles.

The background is then estimated by fitting a smooth parametric function to the data. This can be performed directly within the profile likelihood of Eq. 2.27 by introducing the background shape parameters among the nuisance parameters θ. The parameters can also be estimated in a separate control region, such as a complementary sideband region to the signal region, i.e., where the narrow signal hypothesis is tested.

The choice of suitable parametric fit function depends on the preconceptions about the background shape. Typical choices are polynomials, exponential functions, or power laws. The function should be chosen such that it is flexible enough to describe the background well, but not too flexible such that it would absorb signal-like features. Rather than deciding for a specific function in advance, on can use a more data-driven approach to choose one function from a set of increasingly flexible shapes by comparing the improvement in goodness of fit on a control region with the increase in complexity.

Usually, the statistical procedure discussed in this section is applied after a series of well-motivated selections on additional observables, aimed at reducing the background to a level where the signal can be observed with higher significance. However, another concern arises when the additional observables are correlated with the feature in which the statistical analysis is performed. In that case, the selection might change the shape of the background in the feature of interest, a process referred to as *background sculpting*. Depending on the resulting background shape, this can violate the smoothness assumption of the bump hunt and, if undetected, lead to interpreting bump-like structures in the background as signal and thus overestimating the significance. Not all forms of background sculpting are inherently problematic for a bump hunt. A scaling of linear dependence in the feature of interest, i.e., a pure change of slope, can be absorbed by the data-driven background estimation. However, once this scaling has at least quadratic dependence, i.e., a change of curvature, the background estimation will be biased. It can either be mitigated by modeling the potential change of background shape as a systematic uncertainty, or it can be prevented by decorrelating the selection on the additional observables from the feature of interest.

The similar (and thus often confused) term BumpHunter [71], on the other hand, does not relate to a specific type of background estimation. Instead, it is a statistical hypertest algorithm to quantify global p-values for localized excesses. Essentially, the feature of interest is scanned by constructing various possible windows from the histogram bins. In each window, the local p-value under the background-only null hypothesis is computed, in line with the procedure from Sec. 2.6.1, by treating the entire window as a histogram with a single bin. By repeating a statistical test over a large phase space, the probability for finding a low p-value solely from fluctuations becomes increasingly high, which is referred to as the *look-elsewhere effect*. In order to interpret the result as a *global p-value*, one selects the lowest p-value found in the scan as the observed value of a new test statistic, and estimates the distribution of this test statistic by repeating this procedure many times on pseudo-experiments, sampled from the background-only hypothesis. The global p-value is then the fraction of pseudo-experiments that yield a lower minimum p-value than the observed one. Reference [71] also discusses additional data-driven quality criteria based on sideband regions.

3 Machine Learning Methods

The field of machine learning (ML) is generally concerned with the development of algorithms that can learn from and make predictions based on data. In this section, I introduce the most relevant concepts and methods from the literature that are used in this thesis. A more detailed discussion of these topics can be found in text books, such as Ref. [72].

An ML problem can be formulated as a one of function approximation where some (often unknown and/or intractable) function f is to be approximated by some generic model $\hat{f}_\theta$ with a number D_model of free parameters $\theta \in \mathbb{R}^{D_\text{model}}$. The specific learning task is encoded in a loss function $\mathcal{L}$, designed to quantify the difference between $\hat{f}_\theta$ and f (or at least some proxy for f) on a set of D-dimensional input data $\mathcal{D} = \{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^D$ and N is the number of data points. The aim is to find the set of parameters θ^* that minimizes the loss function:

$$\theta^* \equiv \arg\min_\theta \mathcal{L}(\hat{f}_\theta, f, \mathcal{D}), \tag{3.1}$$

in order to obtain the best function approximation $\hat{f}_{\theta^*}$. This process of finding the optimal parameters is called training or fitting, and is usually done with the help of numeric minimization algorithms.

This section briefly introduces the core ML concepts used in this thesis, starting with an overview of the relevant learning tasks in Sec. 3.1. The minimization algorithms for training ML models are covered in Sec. 3.2. Section 3.3 then introduces the most prevalent types of models, namely neural networks and boosted decision trees, which are the basis of many modern ML methods, such as some of the more specific approaches for unsupervised learning discussed in Sec. 3.4. Finally, Sec. 3.5 discusses common evaluation metrics, with a focus on HEP applications.

3.1 Common Machine Learning Tasks

Learning tasks are commonly classified into one of three subfields:

- Supervised learning: the algorithm is trained on a labeled dataset, i.e., for each data point the correct output is known. The aim is to learn a mapping from input to output that generalizes well to unseen data.
- Unsupervised learning: the algorithm is trained without labels. The learning task is to learn the inherent structure of the data.
- Reinforcement learning: the algorithm learns to make decisions by interacting with an environment. Based on its actions, it receives feedback in the form of rewards, which it uses to improve its decision-making process.

This distinction is not entirely unambiguous as some learning tasks can be seen as a combination of the above. For example, the subfield of semi-supervised learning is concerned with supervised tasks that make use of unlabeled data to improve the learning process. Moreover, many unsupervised learning tasks are approached by phrasing the problem as a supervised one, thus making use of the well-studied supervised learning methodology.

This is also the case for the work presented in this thesis, where the inherently unsupervised learning task of anomaly detection is partially approached by rephrasing it as supervised learning method. Thus, the following subsections are covering relevant examples of both supervised and unsupervised learning.

3.1.1 Supervised Learning

The subfield of supervised machine learning is commonly segregated into two types of tasks: regression and classification.

In *regression*, the output is a continuous variable, i.e., the aim is to learn a function f:

$$f : \mathbb{R}^{D_{\text{in}}} \to \mathbb{R}^{D_{\text{out}}}, \tag{3.2}$$

where D_{in} and D_{out} are the dimensions of the input and output, respectively. A common choice of loss function for regression is the mean squared error (MSE), defined as:

$$\mathcal{L}(\hat{f}_\theta, f, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{f}_\theta(x_i) - f(x_i) \right)^2 , \tag{3.3}$$

where $f(x_i)$ is the known target output of the model for input x_i, often denoted as y_i.

In a *classification* task, on the other hand, each data point is assigned to one of $K \in \mathbb{N}$ classes. The aim is to learn a function f such that:

$$f : \mathbb{R}^{D_{\text{in}}} \to \Delta^{K-1} \tag{3.4}$$

where Δ^{K-1} is the $(K-1)$-dimensional standard simplex, i.e., the set of all vectors $t \in \mathbb{R}^K$ with $\sum_{k=0}^{K-1} t_k = 1$ and $t_k \geq 0$. The output of the model can then be interpreted as a vector of probabilities, where the k-th element is the probability of the data point belonging to class k. A common choice of loss function for classification is the cross-entropy loss, defined as:

$$\mathcal{L}(\hat{f}_\theta, f, \mathcal{D}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log \hat{f}_\theta(x_i)_k, \tag{3.5}$$

where $y_{i,k}$ is the k-th element of the so-called one-hot encoded target output y_i, which is 1 if the i-th data point belongs to class k, and 0 otherwise. An important special case of classification is binary classification, where $K = 2$. In this case, the output is sufficiently described by the probability of the data point belonging to class $k = 1$, and so the label is often encoded as $y_i \in \{0, 1\}$, also referred to as negative and positive labels, respectively. The cross-entropy loss then simplifies to:

$$\mathcal{L}(\hat{f}_\theta, f, \mathcal{D}) = -\frac{1}{N} \sum_{i=1}^{N} \Big(y_i \log \hat{f}_\theta(x_i) + (1 - y_i) \log(1 - \hat{f}_\theta(x_i)) \Big) . \tag{3.6}$$

A more detailed discussion of the cross-entropy loss and its properties can be found in Appendix A.3.

3.1.2 Unsupervised Learning

Common types of unsupervised machine learning tasks are representation learning, density estimation, and generation.

A *representation learning* task is concerned with mapping the input data to a space where some relevant structure is preserved and/or more manifest. The target space is typically lower in dimension than the input space.

Density estimation is the task of learning a tractable probability density function $\hat{p}_\theta(x_i)$ that approximates the true data distribution $p(x_i)$. As the latter is often unknown, a common proxy task is to find the choice of θ that maximizes the likelihood of the data under the model, which can be achieved by minimizing the negative log-likelihood (NLL). This defines a loss function:

$$\mathcal{L}(\hat{p}_\theta, \mathcal{D}) = -\log\left(\prod_{i=1}^{N} \hat{p}_\theta(x_i)\right) = -\sum_{i=1}^{N} \log \hat{p}_\theta(x_i). \tag{3.7}$$

The relationship between NLL and approximating the true density is elaborated on in Appendix A.2.

Finally, the field of *generative* ML is concerned with learning a stochastic model that can generate new data points that resemble the training data. It typically constitutes tuning a function $\hat{f}_\theta$ such that it maps noise drawn from a simple distribution (e.g., a standard normal distribution) to the data distribution, often under the condition of additional variables:

$$\hat{f}_\theta : \mathbb{R}^{D_\text{noise}} \times \mathbb{R}^{D_\text{cond}} \to \mathbb{R}^{D_\text{in}}, \tag{3.8}$$

where D_noise is the dimension of the noise space, and D_cond is the dimension of the condition space.

3.2 Model Training

Once a parametric model $\hat{f}_\theta$ and a loss function $\mathcal{L}$ have been chosen, the main task is to find optimal parameters such that the loss function is minimized, as shown in Eq. 3.1. Only in a few cases can this be done analytically. In most cases, the parameters are learned by first randomly initializing their values and then optimizing them via numeric minimization algorithms.

3.2.1 Numeric Minimization Algorithms

The basis for most modern numeric minimization algorithms is *gradient descent*, which iteratively updates the parameters in the direction of the negative gradient of the loss function with respect to the parameters:

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta \mathcal{L}, \tag{3.9}$$

where t denotes discrete time steps during the learning process and α is the *learning rate*, which is a hyperparameter[1] that controls the size of the update step. The gradient $\nabla_\theta \mathcal{L}$ is here implied to be computed over the entire dataset $\mathcal{D}$. In practice, however, this is often infeasible due to the size of the dataset. Instead, one typically divides $\mathcal{D}$ randomly into smaller fixed-size subsets called *mini batches* (or simply “batches”) $\mathcal{D}_i$ and performs the update steps from Eq. 3.9 sequentially for each mini batch. This batch-wise approach is known as *stochastic gradient descent* (SGD). Iterating through the entire dataset is called an *epoch*, and training is usually performed for multiple epochs.

The choice of learning rate α in Eq. 3.9 is crucial for convergence, as a learning rate that is too large can lead to divergence when the parameter updates become larger than the width of local optima. On the other hand, a learning rate that is too small can slow down convergence and result in globally suboptimal local minima. Thus, it is common to use adaptive learning rates, which change during training based on the history of the parameter updates, in order to yield large parameter updates far away from global optima and small ones in proximity to them. One popular choice is the *Adam* optimizer [73], which modifies the parameter update step from

[1] While the term “parameter” usually refers to those being updated during training, “hyperparameters” are pre-defined constants that relate to explicit design choices.

Eq. 3.9 to be an approximated rescaling of first moment $\hat{m}$ (the mean) of the gradient by the square root of the second moment $\hat{v}$ (the variance):

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_{t+1}}{\sqrt{\hat{v}_{t+1}}}. \tag{3.10}$$

The subscript t denotes that the values of the first and second moment estimates are computed as a moving averages at each step t, along with the parameter updates. Algorithm 1 shows the full pseudocode. The bias correction terms shown there are introduced to account for the initialization of first and second moments to zero, which can lead to biased estimates in the early stages of training. *Adam* has the advantage of being invariant to rescaling of the gradient, approximately bounding the parameter update magnitude by the hyperparameter α, working with sparse gradients, and having automatic step size annealing.

Algorithm 1 *Adam* optimization algorithm. All operations on vectors are performed element wise, including the square of gradients $g_t^2 = g_t \odot g_t$. The hyperparameter ϵ is introduced to avoid zero-division issues and is set to a small number, such as 10^{-8}. Copied from Ref. [73].

```
Require: $\alpha$: Stepsize
Require: $\beta_1, \beta_2 \in [0,1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters $\theta$
Require: $\theta_0$: Initial parameter vector
  $m_0 \leftarrow 0$ (Initialize 1$^{\text{st}}$ moment vector)
  $v_0 \leftarrow 0$ (Initialize 2$^{\text{nd}}$ moment vector)
  $t \leftarrow 0$ (Initialize timestep)
  while $\theta_t$ not converged do
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep $t$)
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
    $\hat{m}_t \leftarrow m_t/(1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
    $\hat{v}_t \leftarrow v_t/(1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$ (Update parameters)
  end while
  return $\theta_t$ (Resulting parameters)
```

3.2.2 Overfitting Mitigation

A model $\hat{f}_\theta$ with more free parameters has more capacity to fit the training data. This can lead to *overfitting*, where the model learns the statistical noise in the training data. This reduces its ability to generalize to unseen data.

A standard approach to reduce overfitting effects is to monitor the loss during training on a separate *validation dataset*, which is not used for training. If the validation loss starts to increase while the training loss continues to decrease, the model is likely overfitting. Thus, it is common to stop training when the validation loss starts to increase, a technique known as *early stopping*. As the validation loss is subject to statistical fluctuations, a grace period of multiple epochs, the so-called *patience*, is usually introduced before stopping the training. By recording the model state at each epoch, one can use the model with the lowest validation loss as final model for the inference on unseen data. When comparing the performance of different machine learning models, it is important to use another separate *test dataset* that has neither been used for training nor for validation. This ensures that the performance metrics are not biased by the model selection process.

In a realistic application, one might want to simultaneously use all data points for both training (for better model performance) and testing (or inference). In order to retain separate training and test sets, one can use *cross-validation* techniques. A common choice is k-fold cross-validation, illustrated in Fig. 3.1, where the data are randomly partitioned into a number k of equally sized subsets. The machine learning model is trained on $k-1$ of these subsets and tested on the remaining one. This process is repeated k times, each time using a different subset as the test set. This can be generalized to a rotating validation set as well, by splitting the $k-1$ training subsets further into l sub-subsets. The ML model is then trained on $l-1$ of these and the remaining one serves as validation set. This is permuted over all options. As this results in l trained models per data point in the test set, one can average the predictions of these models to obtain the final prediction. A convenient choice is $l = k - 1$, such that the data only need to be partitioned once.

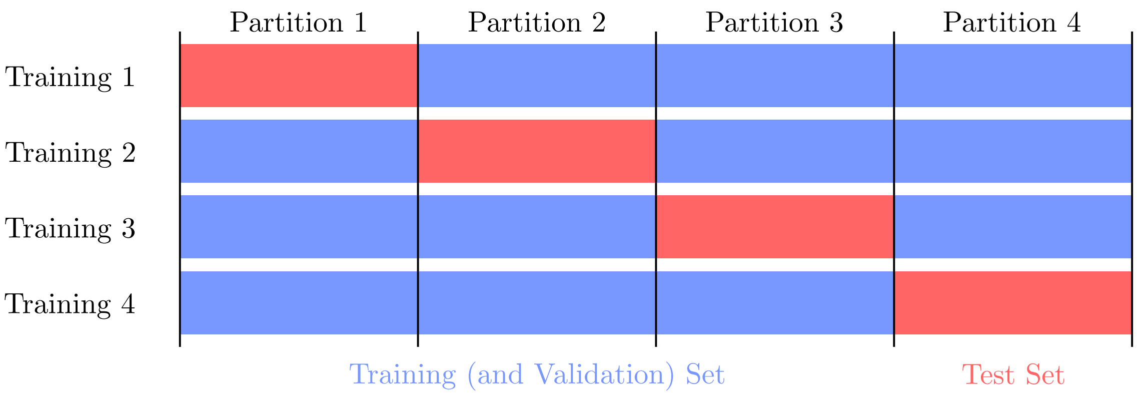

Figure 3.1: Illustration of k-fold cross-validation with $k = 4$. After randomly partitioning the data into k parts, the ML algorithm is trained k times, on each iteration with a different partition as the test set (red). The validation set is an implicit subset of the training set (blue), but could also be rotated by adding another layer of l-fold cross-validation inside each training set.

Another common approach to reduce overfitting is to add so-called regularization terms to the loss function. These terms penalize large parameter values, which can lead to a smoother model that generalizes better. A common choice is the L_p regularization, which adds the sum of the p-norms of all D_{model} free parameters to the loss function:

$$\mathcal{L}_{\text{reg}}(\hat{f}_\theta, f, \mathcal{D}) = \mathcal{L}(\hat{f}_\theta, f, \mathcal{D}) + \lambda \sum_{i=1}^{D_{\text{model}}} |\theta_i|^p, \tag{3.11}$$

where λ is a hyperparameter that controls the regularization strength and p is the order of the norm. The most common choices are $p = 1$ (L_1 regularization) and $p = 2$ (L_2 regularization). The former is known to induce sparsity in the model, i.e., many parameters are driven to zero, which can be useful for feature selection. L_2 regularization, on the other hand, is known to induce smoothness in the model.

3.3 Basic Model Architectures

While the previous sections have generally covered the basic principles of fitting any parametric model to data, this section will focus on two specific classes of models for $\hat{f}_\theta$: neural networks and boosted decision trees.

3.3.1 Neural Networks

Neural networks [74–76] have been particularly successful in the past and are still being widely used as building blocks in more sophisticated modern ML applications.

The base unit of a neural network is the *neuron*, which is a simple function consisting of a weighted sum of the inputs, a bias term $b \in \mathbb{R}$ and some non-linear activation function $\phi : \mathbb{R} \to \mathbb{R}$:

$$z = \phi\left(\sum_{i=1}^{D_\text{in}} w_i x_i + b\right), \tag{3.12}$$

where $w_i \in \mathbb{R}$ are the weights and x_i are the inputs. This is illustrated in pictorial form on the left-hand side of Fig. 3.2. By defining w_i as elements of a weight vector $w \in \mathbb{R}^{D_\text{in}}$ (and again x_i as elements of an input vector $x \in \mathbb{R}^{D_\text{in}}$), Eq. 3.12 can be written in a more compact vectorized form:

$$z = \phi\left(w \cdot x + b\right). \tag{3.13}$$

This alone could be used as a simple linear ansatz for $\hat{f}_\theta$ where the free parameters are $\theta = \{w, b\}$, mapping the input $x \in \mathbb{R}^{D_\text{in}}$ to the scalar output $z \in \mathbb{R}$. In order to achieve higher output dimensionality, one can stack them in *layers* of parallel neurons:

$$z = l_\theta^{D_{out},D_{in}}(x) = \sum_{j=1}^{D_\text{out}} \phi\left(\sum_{i=1}^{D_\text{in}} w_{ij} x_i + b_j\right) = \phi\left(Wx + b\right), \tag{3.14}$$

where $z \in \mathbb{R}^{D_\text{out}}$ is now implied to be a vector of outputs, w_{ij} are elements of a weight matrix $W \in \mathbb{R}^{D_\text{in} \times D_\text{out}}$, and b_j are elements of a bias vector $b \in \mathbb{R}^{D_\text{out}}$. On the right-hand side of Eq. 3.14, ϕ is applied element wise to the vector of weighted sums of inputs, which is now a matrix-vector product.

By stacking N such layers, each with their own parameter vector θ_i, one obtains a *neural network*:

$$\hat{f}_\theta = l_{\theta_N}^{D_\text{out},D_{N-1}} \circ l_{\theta_{N-1}}^{D_{N-1},D_{N-2}} \circ \ldots \circ l_{\theta_2}^{D_2,D_1} \circ l_{\theta_1}^{D_1,D_\text{in}}. \tag{3.15}$$

Specifically, this architecture is referred to as a *fully connected* neural network, as every neuron in a given layer takes all outputs from the previous layer as input. The input and output dimensionality, D_in and D_out, are determined by the task at hand, while the intermediate, so-called *hidden layer* dimensions D_i are hyperparameters that can be tuned. An example of such a fully connected neural network is illustrated on the right-hand side of Fig. 3.2.

By stacking many hidden layers with many neurons, the resulting $\hat{f}_\theta$ can approximate any continuous function to arbitrary precision, given enough neurons. This is known as the *universal approximation theorem* [77]. The non-linear activation function ϕ is crucial for this property, as it allows the network to learn non-linear relationships between the input and output. Common choices for ϕ are the rectified linear unit (ReLU) function $\phi = \max(0, x)$, and the sigmoid function $\phi = 1/(1 + \exp(-x))$. The activation function in the output layer needs to match the desired co-domain of the output, e.g., the identity function for regression tasks or the sigmoid function for binary classification. In the case of multi-class classification with K classes, the

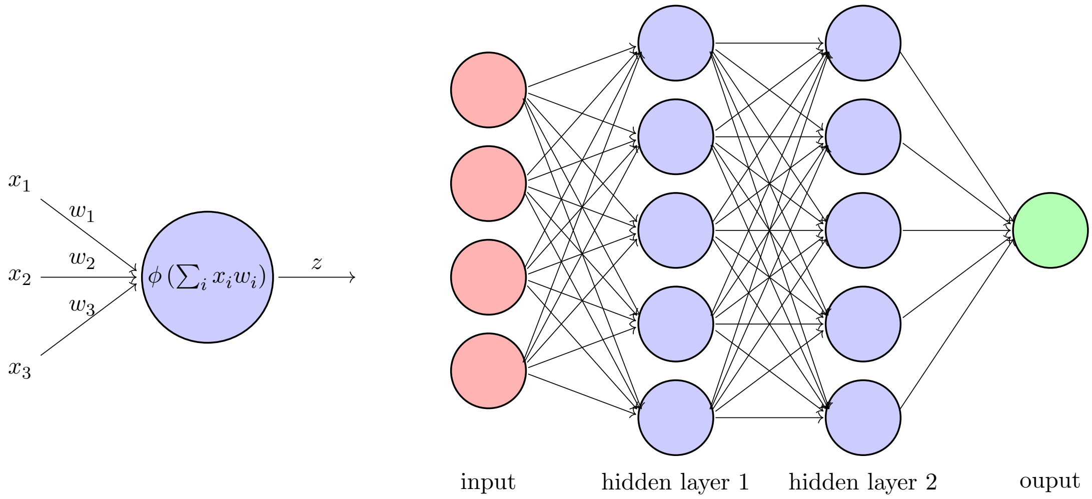

Figure 3.2: Left: Illustration of a single neuron with input dimension $D_{\text{in}} = 3$. Right: Illustration of a simple fully connected neural network with two hidden layers. The input layer is simply a representation of the input vector x and has dimension $D_{\text{in}} = 4$, the hidden layers have dimensions $D_1 = D_2 = 5$, and the output layer has dimension $D_{\text{out}} = 1$.

softmax function $\phi = \exp(x)/\sum_{k=1}^{K}\exp(x_k)$ is often applied, which normalizes the output to a probability distribution over the classes.

One key advantage of neural networks is that the gradient $\nabla_\theta \hat{f}_\theta$ of the output (and thus ultimately the loss function) with respect to the parameters can be computed via the backpropagation algorithm [78–80]. This algorithm is essentially an efficient application of the chain rule of calculus, where at each layer the gradient of the loss with respect to the output of the previous layer is computed, and then propagated backwards through the network to compute the gradient with respect to the parameters. Both the forward pass as in Eq. 3.15 and the gradient computation are based on matrix-vector operations, which can be efficiently parallelized on modern hardware, in particular on graphics processing units (GPUs).

3.3.2 Boosted Decision Trees

Another class of models that has been widely used for regression and classification tasks are boosted decision trees: the application of so-called boosting techniques to decision trees.

A decision tree is a simple model that recursively partitions the input space into regions, assigning a constant value to each region. It can be illustrated as a flowchart, as in Fig. 3.3, where each internal node represents a decision based on some feature, each branch represents the outcome of that decision, and each final "leaf" node represents the final prediction. The decision at each node corresponds to a threshold on a single feature. The tree is constructed iteratively by selecting the feature and threshold that best splits the data into two subsets, according to some criterion, e.g., the Gini impurity or information gain. It is grown until some stopping criterion is triggered, for example a maximum depth.

A single complex decision tree might not generalize well to unseen data, as it could overfit the training data. To mitigate this, one can instead train an ensemble of decision trees, where each tree is trained on a subset of the data, which is randomly sampled with replacement for each iteration. This is known as *bagging* (bootstrap aggregating) [81] and can be applied for other models than trees as well. The ensemble prediction is obtained by aggregating over the ensemble members, e.g., via taking the mean of all individual model predictions. In classification,

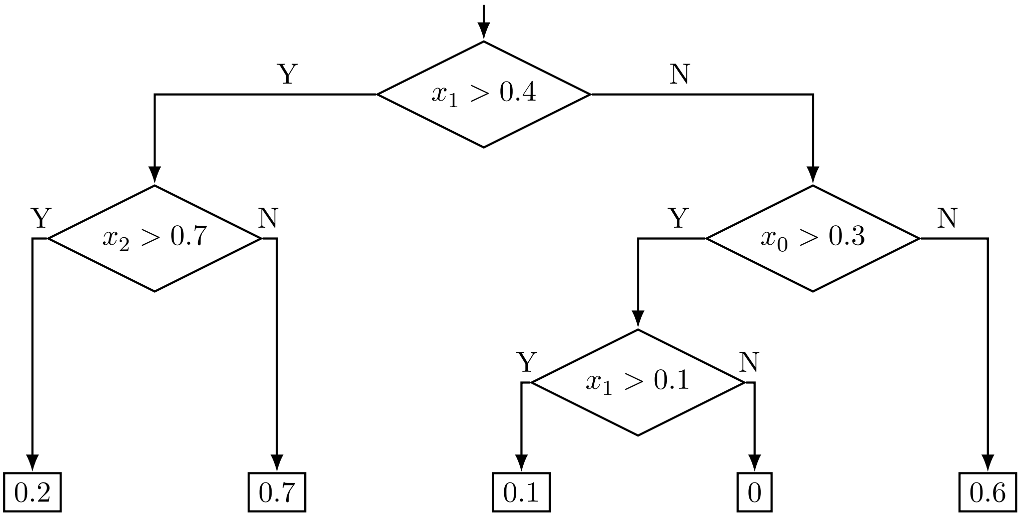

Figure 3.3: Example of a simple decision tree with three input features x_0, x_1, and x_2, four internal nodes and five leaf nodes. The leaf nodes denote the prediction associated with the data point, in this case discrete numbers between zero and one. The labels “Y” and “N” denote for each internal node whether the corresponding condition is fulfilled or not fulfilled, respectively.

a majority vote is a popular choice of aggregation. Another layer of model-by-model randomness can be introduced by only considering a random subset of input features at each split. Decision tree ensembles with sample and feature bagging are known as *random forests* [82].

Instead of training the ensemble members independently, one may also train them sequentially, where each new tree is trained on the residuals of the ensemble prediction so far. This is known as *boosting*. A popular choice of boosting algorithm is *AdaBoost* [83], which assigns weights to the training samples based on the performance of the previous ensemble members, i.e., the next iteration puts more emphasis on those samples where the last iteration performed poorly. The final ensemble prediction is the weighted average of the individual model predictions. Another popular class of boosting algorithms is *gradient boosting* [84], which allows minimizing an arbitrary differentiable loss function. The idea is to fit each new model to the negative gradient of the loss function with respect to the ensemble prediction. The ensemble prediction is then updated by adding the new model with a small learning rate. It is thus more in line with the gradient descent training paradigm, discussed in Section 3.2.1.

There are multiple popular frameworks for gradient boosting–based algorithms, such as *XGBoost* [85] and *LightGBM* [86]. They provide many additional features for improving the performance and stability of the models. One such feature is to discretize the input features into histograms, which can speed up the training process by significantly reducing the number of unique feature values that need to be considered at each split.

3.4 Concrete Unsupervised Learning Approaches

In this thesis, unsupervised ML techniques will serve as an important tool for anomaly detection and eventually the model-agnostic search strategy for new physics. The following subsections introduce the most relevant methods that form the basis of the anomaly detection techniques used in this thesis.

3.4.1 Autoencoders

An important example of unsupervised machine learning models are *autoencoders* (AEs). These are (variants of) neural networks consisting of two parts: an *encoder* and a *decoder*. The encoder maps the input data to a typically lower-dimensional latent space representation, while the decoder maps the latent space back to the input space. This can, for example, be realized simply by using a fully connected neural network with fewer nodes in one of the hidden layers than there are in the input and output layer. A common choice of loss function is the MSE between the input and the output of the autoencoder:

$$\mathcal{L}(\hat{f}_\theta, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{f}_\theta(x_i) - x_i \right)^2 . \tag{3.16}$$

Once trained, the encoder part can be used to map new data points to the latent space representation, which can be useful for tasks, such as dimensionality reduction.

If the probability distribution of the data within the latent space $p(z)$ were known, one could sample from it and transform the samples back to the input space via the decoder. This would then be a generative model, with the ability to sample data points that resemble the ones in the training data. However, such a tractable density model would need to be enforced during training. The *variational autoencoder* [87] (VAE) is one way to achieve this. Rather than learning a deterministic mapping from input to latent space, the encoder learns the latent space distribution under the condition of the input values $q(z|x)$. This posterior distribution is commonly chosen to be a multi-variate Gaussian distribution with diagonal covariance matrix, where the mean μ and variance σ^2 are the outputs of a neural network:

$$q(z|x) = \mathcal{N}(\mu(x), \sigma^2(x)\mathbf{I}). \tag{3.17}$$

Following the derivation in Ref. [87] based on maximizing the so-called *evidence lower bound*, this results in an additional loss term $\mathcal{L}_{\mathrm{KL}}$ that encourages the posterior distribution to match a tractable prior distribution $p(z)$, e.g., a standard normal distribution, by introducing the Kullback-Leibler divergence (see Appendix A.1) between the two distributions. If one further weights the KL term by a hyperparameter β, introduced in Ref. [88], the VAE loss function becomes:

$$\mathcal{L}(\hat{f}_\theta, \mathcal{D}) = \mathcal{L}_{\mathrm{reco}} + \beta \mathcal{L}_{\mathrm{KL}} = \mathcal{L}_{\mathrm{reco}} + \beta \sum_{i=1}^{N} D_{\mathrm{KL}}(q(z|x_i)||p(z)), \tag{3.18}$$

where $\mathcal{L}_{\mathrm{reco}}$ is the reconstruction loss, as used for the AE. In fact, the same MSE loss as in Eq. 3.16 can be inserted, implying that the input features can be described as a product of Gaussian functions with unit variance.

3.4.2 Kernel Density Estimation

A very simple method for the task of density estimation is *kernel density estimation* (KDE) [89]. The idea is to approximate the data distribution $p(x)$ by placing a kernel function $K : \mathbb{R}^D \to \mathbb{R}$ on each data point x_i in the training set[2] and summing them up:

$$\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h} \right), \tag{3.19}$$

[2] The procedure here does not involve a training in the sense of minimizing a loss function, but the “training set” here rather refers to the data points that are used to estimate the density, which are potentially different from the inference data point x.

where $h \in \mathbb{R}$ is the so-called *bandwidth* of the kernel. A larger choice of h results in a smoother density estimate. Various options for kernel functions are mentioned in the literature, such as a simple normal distribution. This approach is not a machine learning method in the sense of optimizing parameters based on a loss function, as it depends only on pre-defined hyperparameters (h and the choice of K), but it is fully data driven in nature.

A major disadvantage of the KDE is that its expressiveness is limited as it assumes the same local smoothness around each data point, rather than capturing more complex structures in the data. In particular, dependencies between multiple dimensions are generally not well described by a static kernel. Moreover, the computational complexity of inferring the density at a new data point scales linearly with the number of training data points, which all need to be stored in memory.

In addition to its simplicity and lack of assumptions on the (global) functional form of the data distribution, the KDE has the advantage that it can be used for sample generation as well. This is in practice achieved by uniformly sampling a data point from the training set and then sampling from the kernel function centered around this point.

3.4.3 Normalizing Flows

A more sophisticated class of generative models that can also be used for density estimation, are *normalizing flows* [90, 91]. Their core idea is to learn an invertible mapping between a simple and analytically tractable distribution to the more complex data distribution.

When transforming a random variable x via an invertible function $h : \mathbb{R}^D \to \mathbb{R}^D$ to a new variable $z = h(x)$, the probability density function of z transforms according to the change of variables formula:

$$p_x(x) = p_z(h(x)) \left| \det \left(\frac{\partial h}{\partial x} \right) \right| , \tag{3.20}$$

where p_x and p_z are the probability density functions of x and z, respectively. In this context, $\frac{\partial h}{\partial x}$ is the Jacobian matrix of h with respect to x. Finding such a function h that transforms to arbitrarily complex data distributions while being invertible and yielding tractable Jacobian determinants can be challenging, but is simplified by restricting the form of h to a composition of $T \in \mathbb{N}$ simple, invertible functions, $h = h_1 \circ \ldots \circ h_T$. Its inverse then corresponds to the composition of the inverses in reverse order, $h^{-1} = h_T^{-1} \circ \ldots \circ h_1^{-1}$. The advantage of this compositional form is that the Jacobian determinant of the composition is simply the product of the Jacobian determinants of the individual functions:

$$\left| \det \left(\frac{\partial h}{\partial x} \right) \right| = \prod_{t=1}^{T} \left| \det \left(\frac{\partial h_t}{\partial z_{t-1}} \right) \right| , \tag{3.21}$$

where $z_t \equiv h_t^{-1}(z_{t-1})$, $z_0 \equiv z$, and $z_T \equiv x$. As usual in machine learning applications, these functions h_t are parametrized by a set of learnable parameters θ_t, in order to approximate the (unknown) true data distribution $p(x)$ with a model distribution $p_x(x|\theta)$. One can show (see Appendix A.2) that this is equivalent to minimizing the NLL of the data under the model, which is thus a suitable loss function for training a normalizing flow. Using the relations in Eq. 3.20 and Eq. 3.21 yields the explicit form:

$$\mathcal{L}(h_\theta, \mathcal{D}) = \sum_{i=1}^{N} \left(-\log p_x(x|\theta) \right) = -\sum_{i=1}^{N} \left(\log p_z(h_\theta(x)) + \sum_{t=1}^{T} \log \left| \det \left(\frac{\partial h_{\theta_t}}{\partial z_{t-1}} \right) \right| \right) . \tag{3.22}$$

In addition to modeling the data density using a composition of simpler functions h_{θ_t}, one can generate new data points by sampling from the simple distribution $z_{\text{new}} \sim p_z$ and transforming them via the inverse $x_{\text{new}} = h_\theta^{-1}(z_{\text{new}})$.

The main challenge is to find building blocks for h_{θ_t} (or equivalently $h_{\theta_t}^{-1}$) that are expressive, invertible and have tractable Jacobian determinants. A popular choice is the *coupling layer* [92], which splits the input into two (at least roughly) equal-sized parts $x = \{x_1, x_2\}$, where x_1 is left unchanged and x_2 is modified by an affine transformation, i.e., scaled and shifted, depending on x_1:

$$h_{\theta_t}(x) = \begin{bmatrix} x_1 \\ e^{\alpha_{\theta_t}(x_1)} \odot x_2 + \mu_{\theta_t}(x_1) \end{bmatrix} . \tag{3.23}$$

The scaling and shift strengths α_{θ_t} and μ_{θ_t} are neural networks that take x_1 as input. The inverse of this transformation is simply:

$$h_{\theta_t}^{-1}(z) = \begin{bmatrix} z_1 \\ e^{-\alpha_{\theta_t}(z_1)} \odot (z_2 - \mu_{\theta_t}(z_1)) \end{bmatrix} , \tag{3.24}$$

and the Jacobian of the transformation is

$$\frac{\partial h_{\theta_t}}{\partial x} = \begin{bmatrix} \mathbf{I} & 0 \\ \frac{\partial h_{2,\theta_t}}{\partial x_1} & \mathrm{diag}(e^{\alpha_{\theta_t}(x_1)}) \end{bmatrix} . \tag{3.25}$$

Because of the diagonal structure of the Jacobian, the determinant is simply the product of the diagonal elements:

$$\left| \det \left(\frac{\partial h_{\theta_t}}{\partial x} \right) \right| = \prod_{i=1}^{D} e^{\alpha_{\theta_t}(x_1)} = \exp \left(\sum_{i=1}^{D} \alpha_{\theta_t,i}(x_1) \right) . \tag{3.26}$$

As only half of the input is transformed, the coupling layer is usually alternated with a *permutation layer* that changes the order of the input dimensions, such that all dimensions are transformed at some point.

Another popular class of normalizing flows are *autoregressive flows*, which can be seen as a more flexible generalization of the coupling layer flows, in particular the *inverse autoregressive flow* (IAF) [93] and the *masked autoregressive flow* (MAF) [94]. The IAF models a step in the generative direction[3] $x = g_{\theta_t}(z) = h_{\theta_t}^{-1}(z)$ as an affine transformation, where the scaling and shifting strengths are computed per for each dimension i based on input dimensions with lower index $j < i$:

$$g_{\theta_t}(z) = \begin{bmatrix} z_1 \\ z_2 e^{\alpha_{\theta_t}(z_1)} + \mu_{\theta_t}(z_1) \\ z_3 e^{\alpha_{\theta_t}(z_1,z_2)} + \mu_{\theta_t}(z_1, z_2) \\ \vdots \\ z_D e^{\alpha_{\theta_t}(z_1,\ldots,z_{D-1})} + \mu_{\theta_t}(z_1, \ldots, z_{D-1}) \end{bmatrix} . \tag{3.27}$$

Similarly to the coupling layer, the inverse of this transformation is straightforward to compute, and the Jacobian determinant reduces to the exponential sum of scaling strengths (Eq. 3.26), because of its triangular structure.

Since the inputs of the neural networks α_{θ_t} and μ_{θ_t} are values z along the generative direction $x = g_{\theta_t}(z)$, the simple latent space samples can be transformed to data-like samples in one forward pass. On the other hand, in order to compute the likelihood of a data point x one needs to map the first dimension x_1 to z_1, then use the z_1 to map the second dimension x_2

[3] By the inverse function theorem, the log Jacobian determinant of the inverse transformation is simply the negative of the log Jacobian determinant of the forward transformation.

to z_2, and so on. In total, this requires D forward passes through the neural networks, which can be computationally expensive for high-dimensional data. This is particularly important as the likelihood computation is needed for training with the loss function in Eq. 3.22. The MAF addresses this issue by the scaling and shifting strengths depending on the more data-like direction x:

$$g_{\theta_t}(z) = \begin{bmatrix} z_1 \\ z_2 e^{\alpha_{\theta_t}(x_1)} + \mu_{\theta_t}(x_1) \\ z_3 e^{\alpha_{\theta_t}(x_1,x_2)} + \mu_{\theta_t}(x_1,x_2) \\ \vdots \\ z_D e^{\alpha_{\theta_t}(x_1,\ldots,x_{D-1})} + \mu_{\theta_t}(x_1,\ldots,x_{D-1}) \end{bmatrix}. \tag{3.28}$$

In this case, the likelihood computation is more efficient, as it is realized in a single forward pass, but the generative direction requires D forward passes through the neural networks.

The scaling and shifting strengths α_{θ_t} and μ_{θ_t} of IAD and MAF can be realized with a single neural network per transformation t with D input dimensions and $2D-2$ outputs, corresponding to the scaling and shifting strengths for each dimension except the first. The weight matrices in each neural network layer are multiplied element wise with a binary mask (values of 0 and 1) that enforces the autoregressive property by construction, i.e., the scaling and shifting strengths for a given dimension only depend on the previous dimensions. This is inspired by the masked autoencoder for density estimation [95].

The formalism of normalizing flows is easily extended to *conditional* density estimation and generation, i.e., modeling $p(x|y)$ where $y \in \mathbb{R}^{D_{\text{cond}}}$ is a set of D_{cond} additional context variables. The original proposal [96] of the conditional likelihood obtained from a normalizing flow h_θ is to modify Eq. 3.20 as follows:

$$p(x|y,\theta) = p_z(h_\theta(x,y)|y) \left| \det \left(\frac{\partial h_\theta(x,y)}{\partial x} \right) \right|, \tag{3.29}$$

i.e., the context variables appear both as input to the transformation h_θ and as condition in the latent space prior p_z. In practice, the former can be achieved by including them as input to the neural networks that parametrize each of the transformations h_{θ_t}. The latter is often omitted, i.e., the latent space prior is chosen to be independent of the context variables, as the conditioning can already be modeled by a sufficiently expressive h_θ. In the generative direction, these context variables need to be provided to $g(z_{\text{new}}|y)$ along with the latent space samples $z_{\text{new}} \sim p_z$.

Compared to the VAE, normalizing flows have the advantage of explicitly modeling the likelihood. A disadvantage is that the latent space dimensionality is fixed to the data dimensionality, whereas a VAE is able to learn a lower-dimensional latent space representation.

3.5 Evaluation Metrics

Before using a trained ML model for inference, it is crucial to have an estimate on the expected performance. This is in particular important when comparing different models or hyperparameters, in order to decide which one to use for the final application. This is commonly performed on a separate test dataset, which has been used neither for training nor validation. The optimal choice of metric depends on the nature of the task.

A straightforward choice of evaluation metric is the loss function used during training, computed on the test set. This results in a single scalar number and thus allows for unambiguous ordering of models by performance. The loss function also only depends on information that has been available for the training set. However, the loss function often only encodes a proxy task for what is ultimately demanded from the model. Moreover, it would not be trivially possible to compare between different choices of loss functions.

In the case of classification, the most popular choices of evaluation metrics are based on counting how many data points have been correctly classified. This depends on knowing the true labels for the test set data points, which might not always be available if the classification task is trained with unsupervised learning techniques. In this case, a representative set of labels can be available for the sake of testing, with the caveat that the labels might not be fully generalizable to every case.

A widely used binary classification metric is the *accuracy*, which counts the number of correctly classified test set examples, divided by the size of the test set. It is straightforward to interpret, as 100% is the maximum and 50% corresponds to mere random guessing. However, the conventional definition of accuracy implies a pre-defined decision threshold, i.e., when the output of the model $\hat{f}_\theta(x)$ is above a certain value, the data point is classified as positive, otherwise as negative. This decision threshold is conventionally set to 0.5, since the model output is often interpreted as a probability. This might not be the optimal choice, as the relative importance of true and false positives depends on the specific application.

The *receiver operating characteristic* (ROC) curve is a graphical representation of the trade-off between *true positive rate* (TPR), the fraction of positive-label examples that were classified as such, and *false positive rate* (FPR), the fraction of wrongly classified negative-label examples, for all possible decision thresholds. A ROC curve is illustrated in Fig. 3.4, showing the TPR as function of the FPR. The diagonal line, on which the two values are equal, corresponds to random guessing. In HEP, the ROC curves are often drawn as the *background rejection*, which is the reciprocal of FPR, as function of TPR. The reason for this lies in the high rate of expected background compared to small new physics signal rates. It is also easier to compare different models, as the numeric difference between models in the vertical axis increases from the reciprocal of FPR, which only ranges from 0 to 1.

While the ROC curve provides a full overview of the model capability over all possible decision thresholds, it does not yield a strict ordering between models. The ROC curve is thus often summarized by measuring the *area under the curve* (AUC), in the TPR vs FPR plane. This can easily be interpreted, as the maximum value of 1 corresponds to perfect classification, while 0.5 corresponds to random guessing. However, the AUC implicitly assumes that the relative importance of TPR and FPR is equal, which might not always be the case. Especially in HEP, rejecting background is usually more important than retaining signal. Another ROC curve–based metric is the FPR (or background rejection) at a fixed value of TPR, which assumes an a-priori choice of working point.

In HEP, the focus of many ML classifiers is to enhance the fraction of collision events originating from new physics processes within an overwhelming amount of background events. In the limit of large datasets, the statistical significance of a new physics process can be approximated as:

$$Z = \frac{N_S}{\sqrt{N_B}}, \tag{3.30}$$

with N_S and N_B being the number of signal and background events, respectively. This implies a specific relative importance of TPR and FPR, which is reflected in the *significance improvement characteristic* (SIC) [97], defined as:

$$\mathrm{SIC} = \frac{Z_\mathrm{cut}}{Z_0} = \frac{N_{S,\mathrm{cut}}}{\sqrt{N_{B,\mathrm{cut}}}} \Bigg/ \frac{N_{S,0}}{\sqrt{N_{B,0}}} = \frac{N_{S,\mathrm{cut}}}{N_{S,0}} \Bigg/ \sqrt{\frac{N_{B,\mathrm{cut}}}{N_{B,0}}} = \frac{\mathrm{TPR}}{\sqrt{\mathrm{FPR}}}, \tag{3.31}$$

where the subscripts “0” and “cut” refer to the quantities before and after applying a selection on the classifier output, respectively. Equation 3.31 shows the correspondence between improving

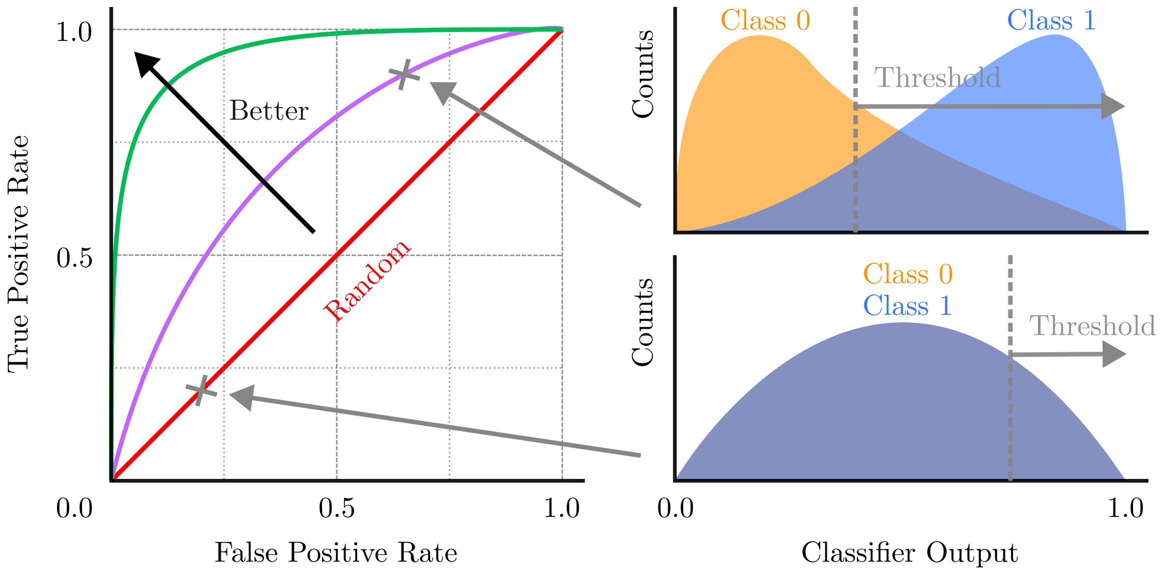

Figure 3.4: The left panel illustrates three receiver operating characteristic (ROC) curves. The red diagonal line corresponds to random guessing, and the purple and green curves arise from non-trivial separation between the two classes. The curves arise from scanning decision thresholds through the classifier output, as illustrated in the right panel. The upper output distribution corresponds to the purple ROC curve and the lower one to a random classifier. Each show an example decision threshold, which maps to a single point on each ROC curve.

the significance and the quantity in terms of TPR and FPR. A SIC curve can be constructed by plotting the SIC values as a function of anything related to the decision threshold, such as TPR or FPR. A value of SIC > 1 corresponds to a non-trivial improvement in sensitivity, which is a stricter requirement than improving over random guessing.

4 Anomaly Detection Methods

This section provides an introduction to the specific ML subfield of anomaly detection, with a particular focus on the application in model-agnostic searches for BSM physics. The following paragraphs give an overview of the core concepts, including a categorization of different types of anomalies. Dedicated subsections later review specific methods in the literature along each of these categories that have been developed for model-agnostic searches in HEP. The focus and the level of detail is chosen in light of what will be relevant in the remainder of this thesis. For a broader review of anomaly detection outside of HEP, the reader is referred to Ref. [98]. Previous reviews of anomaly detection in HEP can be found, for example, in Refs. [99, 100].

An anomaly is by definition a deviation from the normal, i.e., the expected behavior of a system. Anomaly detection is thus the task of identifying such deviations (usually directly) in data using machine learning. It is useful to distinguish three types of anomalies, based on the necessary degree of preconception: outliers, overdensities, and anomalies according to alternative priors. They are illustrated in Fig. 4.1.

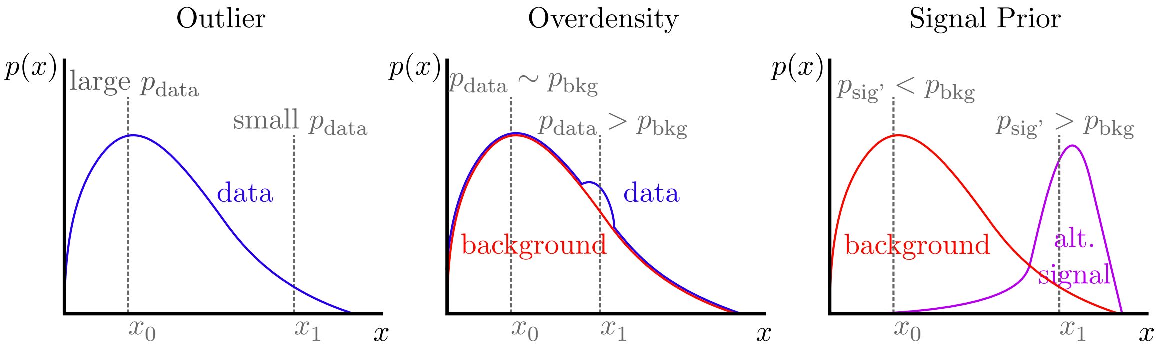

Figure 4.1: Illustration of the three types of anomalies: outliers (left), overdensities (center), and anomalies according to alternative signal priors (right). The points x_0 are non-anomalous in each case, while the points x_1 are likely associated with a high anomaly score.

Outliers are by far the most commonly targeted type of anomaly. In the broad literature (outside of HEP), the terms outlier and anomaly are often used interchangeably. They are data points that are in some sense substantially different from the bulk of the data. From a statistical point of view, they correspond to samples x that lie in low-density regions of the data distribution $p(x)$. Such outliers can be identified without any assumptions on the expectation, and thus it is often referred to as fully unsupervised anomaly detection. The definition of a low-density region is however not unique as it depends on the choice of coordinate system for x, i.e., the same data point can be an outlier in one coordinate system but not in another [101, 102].

Overdensities, on the other hand, are regions of the data distribution that are more populated than one would expect from some a-priori model of what is considered normal. This so-called background model can also be obtained from the data themselves, however this construction then relies on further assumptions, such as a localization of the anomaly in certain obervables. The definition of an overdensity is often phrased in terms of a likelihood ratio of a resulting signal (capturing the anomalous contribution) and background model:

$$R_{\text{sig,bkg}}(x) = \frac{p_{\text{sig}}(x)}{p_{\text{bkg}}(x)}, \tag{4.1}$$

which is independent of the coordinate system because the Jacobian determinants of any transformation would cancel out between numerator and denominator.

Finally, if one not only has some preconception of the background, but also some (limited) prior knowledge about the type of anomaly that could be expected, one may define anomalies according to how well data points match this *alternative prior*. It can also be phrased as a likelihood ratio as in Eq. 4.1 but with the signal model being replaced by some alternative signal hypothesis. Thus, it does not depend on the coordinate system either. As the outcome depends on the choice of alternative signal priors, it is the least unsupervised paradigm of the three types of anomalies.

With each definition above, one is able to construct an *anomaly score*, i.e., a continuous number related to the probability of a data point being an anomaly. As in binary classification tasks, one can then define a threshold on the anomaly score to flag data points as either anomalous or non-anomalous.

4.1 Model-Agnostic Searches in High-Energy Physics

In the context of high-energy physics analyses, anomaly detection is usually linked to *model-agnostic searches* for new physics.

The usual setting of a model-agnostic new physics search is that the bulk of the recorded data is assumed to follow a known background, the Standard Model, and there might be relatively few data points resulting from new physics processes. The aim is to identify these new physics signal events if they exist, and ideally to do so with no assumptions on the nature of the signal. This is being referred to as a *signal model–agnostic* search. Ideally, such a search would also not rely on an exact description of the background, making it *background model agnostic* as well. In practice, neither of these two "agnosticisms" can be fully achieved, and the goal is to minimize the assumptions made.

Machine learning anomaly detection has become a promising toolkit for such model-agnostic searches. Applying anomaly detection in this context implies that the potential new physics signal is an anomaly as defined earlier. This further assumes that there exist measured observables, in which a signal would manifest itself as either an outlier or a localized excess (overdensity) in the data distribution. Otherwise, it cannot be detected, even if it is present.

It is possible to construct model-agnostic searches for new physics without machine learning. Important examples of such analyses are discussed in Secs. 5.1 and 5.2. However, recent developments in anomaly detection show a promising potential to improve the sensitivity of such searches. The following sections will discuss some of these methods, grouped into the three types of anomalies as categorized in the beginning of this section.

4.2 Unsupervised Anomaly Detection

As introduced earlier, outlier detection is the most unsupervised type of anomaly detection. In the context of machine learning, this corresponds to learning the data distribution $p_{\text{data}}(x)$ or a proxy for it, and then selecting data points with low probability.

One of the most established unsupervised anomaly detection approaches is to train an autoencoder (Section 3.4.1) to compress the data into a lower-dimensional latent representation and then reconstruct the input. The AE has limited capacity and will learn to effectively reconstruct the most common patterns in the data. Anomalies, which by definition only occur infrequently in the training data, will be reconstructed worse. The reconstruction error, which is also the typical choice of loss function for training AEs (e.g., Eq. 3.16), is thus a proxy for

$p_{\text{data}}(x)$ and can be used as the anomaly score. Early examples of demonstrating the potential of AEs for anomaly detection in HEP include Refs. [103–105].

Rather than learning $p_{\text{data}}(x)$, one can also learn a proxy for $p_{\text{bkg}}(x)$ directly by incorporating additional assumptions. This is commonly done by either motivating control regions in the data where new physics signals are expected to be more depleted than in the analysis region, or by training the model on MC simulation of the background.

As discussed in Section 3.4.1, VAEs are a generalization of AEs that learn a probabilistic model of the data distribution in the latent space. Reference [106] is an early example in HEP where the reconstruction error of a VAE was tested as an anomaly score with promising performance. A key advantage of the VAE is that the regularization term $\mathcal{L}_{\text{KL}}$ in the loss function encourages the latent space data distribution to approximate a known prior $p(z)$. This regularization term (or the full loss, consisting of both reconstruction and regularization terms in Eq. 3.18) can thus also be used to define an anomaly score, as it is large for data points that are unlikely under the prior. This has been shown to exceed the sensitivity to certain types of anomalies over a pure reconstruction loss in Ref. [107]. The latent space prior $p(z)$ can also be used in other ways to define an anomaly score. In particular, the sensitivity to certain types of anomalies can be enhanced by encoding additional assumptions into the prior, such as a bimodal structure [108].

A key advantage of (V)AE-based anomaly detection is that it tends to scale well to a large number of input features. In HEP, this is particularly useful for incorporating the full collision event information, which is typically described by a high-dimensional "low-level" feature space (e.g., the full space of PF candidates from Sec. 2.4.4). By constructing a limited number of "high-level" features that are motivated by expert knowledge (e.g., N-subjettiness from Sec. 2.5.5), one might lose sensitivity to new physics signals that are not well captured by these features. However, this also means that the resulting anomaly score is prone to be correlated with certain features of interest, such as invariant masses, that are used in the statistical analysis after selecting the most anomalous data points. Selecting on this anomaly score thus results in a change of the background shape in these features, which might be undesirable. Early studies of VAEs with low-level input features, such as Ref. [107], addressed this issue by explicitly decorrelating the anomaly score from such features that are aimed to be orthogonal.

Another known caveat of (V)AE-based anomaly detection is the so-called complexity bias. The loss is not an exact proxy for the data likelihood, but tends to be higher for more complex data. Abundant data points that are intrinsically more difficult to reconstruct are often more likely to be flagged as anomalies than rare but simple ones. One approach to mitigate this is to use a normalization scheme that decreases the probability of rare data points, regardless of their complexity [109].

Another approach to learn the data (or background) distribution more directly would be to use a normalizing flow. This is explored for HEP anomaly detection in Refs. [110, 111]. A downside of this approach is that normalizing flows usually come with higher model complexity, especially for modeling many dimensions, which would be necessary to incorporate more low-level features, as is common in VAE-based anomaly detection in HEP. Furthermore, it is a known issue in the broader ML literature outside HEP that normalizing flows tend to associate high density to out-of-distribution data points [112]. One potential reason is the mismatch between typicality and density in high-dimensional spaces, which is a problem for outlier detection in general [113]. Another explanation is attributed to the inductive bias that is necessary to generalize to data points outside the training set [114], but these conclusions are drawn from works that focus on image data, and it remains to be shown how well they translate to HEP data.

4.3 Weakly Supervised Anomaly Detection

Compared to purely model-agnostic outlier searches, overdensity detection–based applications rely on some additional assumptions. In the context of HEP, they are usually encoded by rephrasing the learning problem as a *weakly supervised* classification task.

4.3.1 Weak Supervision and CWoLa

Weak supervision is the subfield of ML that is concerned with learning tasks that fall in-between the fully supervised case, described in Sec. 3.1.1, and unsupervised learning, described in Sec. 3.1.2. Reference [115] aims at providing a clear taxonomy of different branches of weak supervision. In particular interest for this work is the problem setting of *multi-instance learning* (MIL) [116], where the individual training data (referred to as *instances*) are grouped into so-called *bags* and the labels are only available per bag and not per instance. An important special case of MIL is the *noisy label* setting, where the label of a bag is not correctly reflecting the per-instance labels that should be inferred by the model. A broad mathematical formulation and conditions for a solution to this problem are discussed in Ref. [117]. In the context of HEP applications, this was formalized within the framework of *Classification Without Labels* (CWoLa) [118].

CWoLa can be understood by noting that a binary parametric classifier model (e.g., a neural network) trained to separate two classes, signal and background, approximates (a monotonic rescaling of) the likelihood ratio, as shown in Eq. 4.1, between the two classes[4]. This is a consequence of using a suitable loss function, such as the binary cross-entropy, and is discussed in more detail in the Appendix A.3. For simple hypothesis tests, in which the densities $p_{\text{sig}}(x)$ and $p_{\text{bkg}}(x)$ are uniquely defined without depending on extra model parameters, the likelihood ratio is a uniformly most powerful test statistic according to the Neyman-Pearson lemma [67]. This means that for every fixed false positive rate, the likelihood ratio test has the highest true positive rate.

Assuming there are two *mixed sets*, M_1 and M_2, that are composed of both signal and background data points with signal fractions f_1 and f_2, respectively, then their respective probability densities are given by

$$p_{M_1}(x) = f_1 p_{\text{sig}}(x) + (1 - f_1) p_{\text{bkg}}(x) \tag{4.2}$$

$$p_{M_2}(x) = f_2 p_{\text{sig}}(x) + (1 - f_2) p_{\text{bkg}}(x). \tag{4.3}$$

If one further assumes that $f_1 > f_2$, then the core theorem of CWoLa [118] is that an optimal classifier trained to separate M_1 and M_2 will simultaneously approximate an optimal classifier for distinguishing signal from background directly. This is because the mixed-set likelihood ratio $R_{M1,2}(x) = p_{M_1}(x)/p_{M_2}(x)$, which is approximated by the mixed-set training, is a monotonically increasing rescaling of the signal-background likelihood ratio $R_{\text{sig,bkg}}(x) = p_{\text{sig}}(x)/p_{\text{bkg}}(x)$:

$$R_{M1,2}(x) = \frac{p_{M_1}(x)}{p_{M_2}(x)} = \frac{f_1 p_{\text{sig}}(x) + (1 - f_1) p_{\text{bkg}}(x)}{f_2 p_{\text{sig}}(x) + (1 - f_2) p_{\text{bkg}}(x)} = \frac{f_1 R_{\text{sig,bkg}}(x) + (1 - f_1)}{f_2 R_{\text{sig,bkg}}(x) + (1 - f_2)}. \tag{4.4}$$

The theorem holds as long as $f_1 > f_2$, since then the derivative of the mixed-set likelihood ratio with respect to the signal-background likelihood ratio is strictly positive:

$$\left.\frac{\partial R_{M1,2}}{\partial R_{\text{sig,bkg}}}\right|_x = \frac{f_1 - f_2}{\left(f_2 R_{\text{sig,bkg}}(x) + (1 - f_2)\right)^2} > 0. \tag{4.5}$$

[4] It should be noted, however, that this asymptotically holds in the limit of infinite training data, an infinitely flexible model and perfect training convergence. In practice, the classifier is an approximate estimate.

In the reverse case, $f_1 < f_2$, the two likelihood ratios would be anti-correlated, and a reverse signal-from-background classifier would be learned. If $f_1 = f_2$, i.e., the mixed sets follow the same distribution, the theorem breaks down and no information about the signal-background likelihood ratio can be inferred.

4.3.2 Resonant Anomaly Detection and the Idealized Anomaly Detector

The CWoLa paradigm inspired a class of methods for detecting overdensities in HEP data, often referred to as *resonant anomaly detection*. The core setup consists of one so-called *resonant feature* m, in which a small localized signal presence is hypothesized, and a set of *auxiliary features* x, which might capture potential differences of such a new physics signal and the background. The spectrum of m is divided into two parts: the *signal region* (SR), in which a potential signal is hypothesized[5], and the complementary *sidebands* (SB). This is illustrated in Fig. 4.2. If one were to have a sample of perfect SR background data, thus following $p_{\text{bkg}}(x|m \in \text{SR})$, one could train a classifier with input features x to distinguish this from the SR data, following $p_{\text{data}}(x|m \in \text{SR}) = f_{\text{sig}} p_{\text{sig}}(x|m \in \text{SR}) + (1 - f_{\text{sig}}) p_{\text{bkg}}(x|m \in \text{SR})$ with the unknown (potentially zero) signal fraction f_{sig}. This would correspond to a classifier learning the data-to-background likelihood ratio:

$$R^{\text{SR}}_{\text{data,bkg}}(x) = \frac{p_{\text{data}}(x|m \in \text{SR})}{p_{\text{bkg}}(x|m \in \text{SR})}, \tag{4.6}$$

and the CWoLa theorem implies that this classifier would also be optimal for distinguishing signal from background in the SR, since the pure background set corresponds simply to a mixed set with signal fraction of 0.

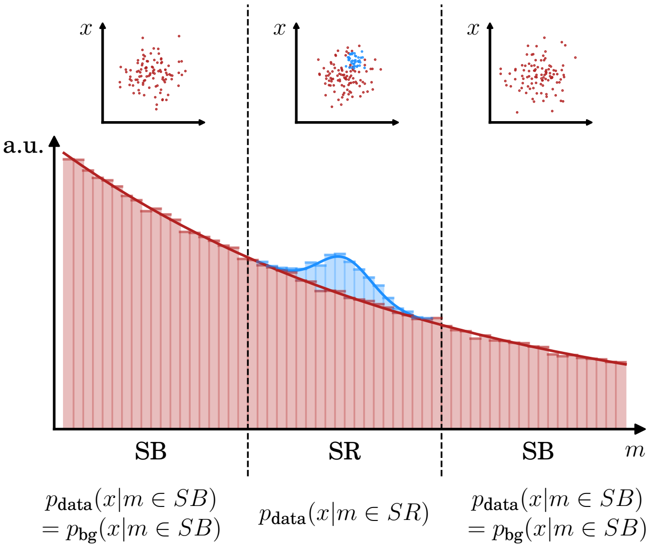
x
x
x
a.u.
SB
SR
SB
m
$p_{\text{data}}(x|m \in SB)$
$= p_{\text{bg}}(x|m \in SB)$
$p_{\text{data}}(x|m \in SR)$
$p_{\text{data}}(x|m \in SB)$
$= p_{\text{bg}}(x|m \in SB)$

Figure 4.2: Setup of a typical resonant anomaly detection scenario with a resonant feature m and auxiliary features x. The m spectrum is divided into signal region (SR) and sidebands (SB). The presence of signal is exaggerated for illustration. Figure taken from Ref. [1].

[5] As there is usually not an a-priori optimal choice of SR in a model-agnostic search, the procedure discussed here will be repeated for multiple SR choices covering a fiducial region of m with steps motivated by the detector resolution.

This type of approximating a signal-sensitive classifier by training a classifier to distinguish between signal+background and a perfect background sample is referred to as an *idealized anomaly detector* (IAD), introduced as such in Ref. [1][6]. It is idealized in the sense that the availability of a perfect background sample is unlikely in a real-world scenario. On the other hand, the training of such an IAD classifier on limited training data and numeric optimization is a departure from learning the exact likelihood ratio. It thus serves as an upper performance boundary when comparing different methods for estimating this background sample. An imperfect background sample implies different p_{bkg} in Eqs. 4.2 and 4.3, which breaks the CWoLa theorem to some extent, such that the equality in Eq. 4.4 is only approximate.

4.3.3 Approximating the Background Sample

A conceptually straightforward strategy for obtaining the pure background sample would be to use MC simulation. The disadvantage is that this introduces an additional degree of background model dependence, further relying on the background being well modeled in the simulation. Otherwise, the resulting classifier learns to associate systematically mismodeled phase space regions with being more signal like. To address this issue, several methods have been proposed that learn a faithful background sample in a data-driven manner from the SB.

The earliest approach for a data-driven background sample approximation via the SB is referred to as *CWoLa Hunting* [120, 121]. Here, the background sample is simply constructed by selecting data points from the SB region. The method is valid only in the case where there are no substantial differences between the SR and SB background: $p_{\text{bkg}}(x|m \in \text{SR}) \approx p_{\text{bkg}}(x|m \in \text{SB})$. This is in general not the case and needs to be ensured by a suitable choice of auxiliary features, such that there are no notable correlations between m and x. As this is often not possible to ensure over the full spectrum of m, it is common to select background-like data points only from an adjacent region in m to the SR, where the background is expected to be similar. This region is sometimes referred to as *short sidebands* (SSB). While a narrower SSB choice reduces the correlation between m and x, it also reduces the size of the background sample, thus increasing the statistical uncertainty and consequently leading to a less optimal classifier training. In order to account for the different number of data points in SR and SSB, the loss function of the classifier is usually weighted, such that left and right SSB contribute equally to the loss, and together contribute as much as the SR.

A hybrid approach between using MC simulation and CWoLa Hunting is the *simulation-assisted likelihood-free anomaly detection* (SALAD) method [122]. An imperfect MC background simulation sample is corrected within the SR by deriving a correction function from the SB data. This is achieved by training a binary classifier, usually a neural network, with a cross-entropy loss to distinguish between SB data and SB simulation, using the auxiliary features x as well as the resonant feature m as input. The output of this classifier $h(x, m)$ can be used to derive a sample weight corresponding to an explicit likelihood ratio:

$$w(x,m) = \frac{h(x,m)}{1-h(x,m)} \approx \frac{p_{\text{data}}(x,m)}{p_{\text{MC}}(x,m)} = \frac{p_{\text{data}}(x|m)}{p_{\text{MC}}(x|m)} \frac{p_{\text{data}}(m)}{p_{\text{MC}}(m)} \tag{4.7}$$

where the very last factor is an overall constant, corresponding to the ratio of the total number of data and MC simulation samples. The first equality is derived explicitly in the Appendix A.3. The parameterization with m allows the model, trained on the SB only, to be interpolated into the SR. This learning of a reweighting function via neural networks has been explored in a HEP context under the term *deep neural networks using classification for tuning and reweighting*

[6] It has been used for comparison earlier, under the name of *optimal CWoLa* in Ref. [119].

(DCTR) [123]. These weights are then used to reweight the MC background simulation samples in the binary cross-entropy loss when training the SR classifier $\hat{f}_\theta$ to distinguish data from the background:

$$\mathcal{L}_{\text{SALAD}}(\hat{f}_\theta, f, \mathcal{D}) = -\sum_{m_i \in \text{SR data}} \log \hat{f}_\theta(x_i) - \sum_{m_i \in \text{SR MC}} w(x_i, m_i) \log(1 - \hat{f}_\theta(x_i)). \tag{4.8}$$

This classifier is not subject to the same limitations of CWoLa Hunting mentioned above. The disadvantage is that it relies on some prior MC background simulation sample as starting point, and it assumes that the discrepancy between data and simulation in the SR can be smoothly interpolated from the SB. In addition, the implicit interpolation of the correction function (querying values within a region of m that was not shown during training) relies on a smoothness assumption in the inductive bias of the classifier architecture.

A similar path is taken by the *simulation-augmented CWoLa* (SA-CWoLa) method [119]. Instead of using the SB to correct the SR MC background simulation, it modifies the CWoLa Hunting training to penalize the learning of differences between the SR and SB (other than a potential signal presence). This is achieved by adding an opposite-sign regularization term to the loss function that captures whether the classifier is learning differences between SR and SB in the MC background simulation:

$$\begin{aligned}\mathcal{L}_{\text{SA-CWoLa}}(\hat{f}_\theta, f, \mathcal{D}) = &-\sum_{m_i \in \text{SR data}} \log \hat{f}_\theta(x_i) - \sum_{m_i \in \text{SB data}} \log(1 - \hat{f}_\theta(x_i)) \\ &+ \lambda \left(\sum_{m_i \in \text{SR MC}} \log \hat{f}_\theta(x_i) + \sum_{m_i \in \text{SB MC}} \log(1 - \hat{f}_\theta(x_i)) \right).\end{aligned} \tag{4.9}$$

Here, λ is a positive hyperparameter that controls the strength of the regularization, where the limit of $\lambda \to 0$ recovers the default CWoLa Hunting loss. The MC simulation is thus only used as a decorrelation tool, rather than a direct background sample. On the other hand, the approach still assumes that the MC simulation differs from the data similarly in SR and SB.

The first fully data-driven attempt at learning the background sample directly from the SB, and generalizing to x and m being arbitrarily correlated, is *anomaly detection with density estimation* (ANODE) [124]. Rather than learning the data-to-background likelihood ratio directly via a binary classifier, it involves an explicit construction of the ratio using two normalizing flows. One conditional normalizing flow is trained to approximate the conditional likelihood of x given m in the SR data $p_{\text{data}}(x|m)$. A second flow learns the background likelihood by only training on SB data $p_{\text{bkg}}(x|m)$. Because of the conditioning on m, the latter model can be implicitly interpolated by querying it with m values in the SR, similar to the SALAD correction function. For each data point, the likelihood ratio can then be constructed by taking the explicit ratio of the likelihoods from the two flows. With this construction, the method is not limited to the case where x and m are uncorrelated. The downside is that normalizing flows are computationally much more expensive than training a simple binary classifier such as a neural network. Moreover, the crucial task of correctly modeling the small overdensity on top of the overwhelming background is inherently more difficult with a normalizing flow, based on smooth mappings, than with a binary classifier that learns the likelihood ratio directly.

A notable extension of the CWoLa Hunting protocol is *Tag N' Train* (TNT) [125]. This method does not address the issue of needing x and m to be uncorrelated, but rather focuses on increasing the signal fraction during training via well-motivated physics assumptions. Specifically, it assumes that the anomaly manifests in at least two objects, O1 and O2. In HEP this could, for example, correspond to a heavy BSM resonance decaying into two lighter BSM particles. This means that the features can be decomposed into those capturing properties of O1 (x_{O1}) and O2 (x_{O2}). The procedure then consists of training an unsupervised anomaly detector, as described in Sec. 4.2, on x_{O1} only, then dividing the data points into a signal-enriched and a background-enriched set based on the resulting anomaly score (e.g., using the highest and lowest anomaly score values, respectively). A CWoLa classifier is then trained to distinguish signal-enriched from background-enriched data based on x_{O2}. The same procedure is repeated with the roles of x_{O1} and x_{O2} reversed. The final two classifiers, one using x_{O1} and the other using x_{O2}, approximate the signal-to-background likelihood ratio if the two objects are uncorrelated and the unsupervised anomaly detector learns at least weak discrimination between signal and background. They can then either be combined to form the final TNT classifier, or the procedure is iterated to further increase the signal fraction. This is illustrated in Fig. 4.3. The feature m and its distinction between SR and SB is not strictly needed in TNT, but can be used to further enhance the a-priori signal fraction. The method is thus a hybrid between weakly supervised and unsupervised anomaly detection, with the additional assumption that the signal results in at least two uncorrelated objects.

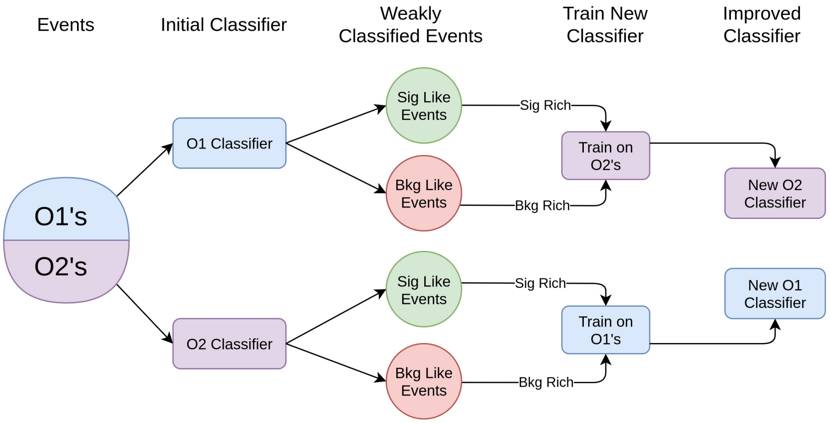

Figure 4.3: Schematic illustration of the Tag N' Train method. Figure taken from Ref. [125].

While ANODE benefits from the explicit construction of $p_{\mathrm{bkg}}(x|m)$ in the SR, by interpolating a flow trained on the SB, its direct likelihood ratio construction is less performant than the binary classifier training of CWoLa Hunting. The *classifying anomalies through outer density estimation* (CATHODE) [1] algorithm combines the advantages of both approaches. After training (as in ANODE) a conditional normalizing flow to learn $p_{\mathrm{bkg}}(x|m)$ from the SB, it involves generating new samples from this flow with m drawn from a simple one-dimensional approximation to the SR distribution $p_{\mathrm{bkg}}(m)$. By the invertible nature of normalizing flows, these samples are distributed according to $p_{\mathrm{bkg}}(x, m)$, and can be used to train a binary classifier to distinguish data from these samples. It thus learns the likelihood ratio in the same manner as CWoLa Hunting, but with a background sample whose correlations between x and m are directly modeled rather than ignored. An additional advantage is that once the flow is trained, one can draw arbitrarily many background samples from it. This facilitates an improved classifier training, as the background sample size is not limited by the size of the (S)SB. CATHODE can thus approach the performance of the IAD in a fully data-driven manner, in the general case of correlated x and m. It relies on the assumption that the background shape can be well interpolated into the SR via the smooth inductive bias of the normalizing flow architecture. Furthermore, it requires all features x to be well modeled by the flow, which is based on a sequence of smooth transformations. This can be challenging for general HEP data, which might consist not only of smooth distributions, but also of non-trivial structures, such as sharp peaks or discrete observables. The CATHODE method will be discussed in more detail in Sec. 6.2, as it has been developed in the context of this thesis.

Another approach to constructing the background sample is *constructing unobserved regions by transforming adjacent intervals* (CURTAINs) [126]. Instead of learning the full p_{bkg} from the SB, it uses the invertible transformations directly, i.e., the normalizing flow without the prior distribution, to transport data from the SB to the SR using the difference in m between target and source value, Δm, as conditional. This mapping is originally trained via an optimal transport loss that quantifies how well two batches of SB data points are mapped into each other. A later version combines two flows: one trained to learn the base distribution of the SB and the other to learn the transformation between SB points [127]. With the latter approach, it is possible to train with the conventional NLL loss. The conceptual advantage of CURTAINS over CATHODE is that an a-priori simpler task is learned: the invertible map between SR and SB x, which are expected to be somewhat similar. However, in order to learn it, either a batch-wise optimal transport loss (Ref. [126]) or an additional base density flow (Ref. [127]) is needed during training. The former is known to be suboptimal compared to an exact NLL loss. The latter approach is an additional layer of complexity (and approximation) compared to CATHODE where only the base density flow is trained. The conditioning on Δm also comes with advantages and disadvantages. On the one hand, the transformation can be performed using data pairs within a single SB (left or right SB exclusively) as well as between left and right SB, so the evaluation does not rely on querying conditional values unseen during training. On the other hand, this choice limits the expressiveness of the transport function, as it potentially ignores more complex dependence of x on the actual value of m.

The feature transporting property of flows has also been explored in an approach similar to SALAD. The method *flow-enhanced transportation for anomaly detection* (FETA) [128] uses a flow to map an imperfect MC background prior to a more faithful distribution of the expected background in data. It employs the same double-flow construction as Ref. [127] to learn the transformation from MC background simulation to data in the SB conditioned on m. Once trained, this flow is applied to map MC background simulation to the expected data background in the SR, which is then used to train a binary classifier against the SR data. Compared to the binary classifier reweighting in SALAD, the flow-based transfer function is expected to generalize better for the case where data and simulation x distributions have non-overlapping support, i.e., a pure weighting scheme cannot learn to populate an empty region of x space in simulation to non-zero data density in the same region. A disadvantage is the higher complexity of flow models compared to classifiers based on simple neural networks. Otherwise, the method exhibits similar advantages and limitations as SALAD.

While CWoLa Hunting only works in the case where x and m are approximately uncorrelated, multiple of the aforementioned methods extend the background sample construction to the more general case of finite correlations. However, once a binary classifier $\hat{f}_\theta$ is trained on this correlated x to learn the likelihood ratio, it also learns the implicit m dependence: $\hat{f}_\theta(x(m))$. Applying a selection on this anomaly score thus results in an m-dependent selection, which alters the background distribution in m. This can be detrimental during the background estimation needed to quantify the deviation from the background-only hypothesis, as discussed in Sec. 2.6.3. The CATHODE protocol can be modified slightly in order to prevent the classifier from learning the m dependence. Rather than training it to distinguish SR data from a sample drawn from the normalizing flow approximation to $p_{\mathrm{bkg}}(x|m)$, the underlying invertible transformation of the flow $h_\theta(x, m)$ is first used to map the SR data to the latent space, and the background sample is drawn from the simple prior $p_z(z)$[7]. The classification task between the two samples in the latent space is similar to the original task, but with the m dependence removed as background has been trained to be mapped to a simple prior distribution without m dependence. This method is referred to as *latent CATHODE* (LaCATHODE) [2] and, apart from mitigating the background shaping in m, shares most of the advantages and limitations of the original CATHODE method. While the mapping of SR background data to the prior distribution has been trained from learning the task in the SB, the lack of control over where in the latent space the potential signal is mapped to presents a potential limitation compared to CATHODE. It implies that the overdensity location might change from one randomly initialized flow training to another, introducing additional variance, and it might be mapped to regions of high background density, where the classifier is less sensitive to anomalies. The problem of the m dependence for background estimation and the solution of LaCATHODE will be discussed in more detail in Sec. 6.3, as it has been developed in the context of this thesis.

4.4 Semi-Supervised Anomaly Detection

Section 4.2 discussed improving outlier detection methods by incorporating additional assumptions on the background, thus learning a proxy for $p_{\mathrm{bkg}}(x)$ instead of $p_{\mathrm{data}}(x)$. The Quasi Anomalous Knowledge (QUAK) [129] strategy extends this idea by incorporating prior assumptions on the new physics signal as well, i.e., learning a proxy for both $p_{\mathrm{bkg}}(x)$ and $p_{\mathrm{sig}}(x)$. The core reasoning is that even in a model-agnostic search, there exist assumptions motivated by physics. For example, that the decay products of a massive particle will be contained within a cone determined by the particle energy and Lorentz invariance. The difference with respect to a classical supervised search is that the targeted $p_{\mathrm{sig}}(x)$ proxy is not specific to one single signal hypothesis, but should rather guide towards a broad class of signal-like patterns.

Instead of training a single VAE to reconstruct examples from a background MC simulation, QUAK involves training N independent VAEs[8], one on the background MC simulation and the remaining ones on $N-1$ additionally provided signal MC simulation samples. The orthogonal losses of each VAE then define an N-dimensional space, the QUAK space, where background-like data are expected to cluster in the corner of low background loss and high signal loss. Points with both high background and high signal loss, which would be classified as anomalous in the typical one-dimensional VAE approach, are more likely to arise from detector glitches. The region of interesting anomalies is located at high background loss and low signal losses. One can thus select a subset of well-motivated anomalies by constructing an N-dimensional window in this QUAK space region.

An advantage of the QUAK paradigm is that one expects higher sensitivity for signal processes close to the provided priors. The downside is a higher dependence on the types of signal MC simulation samples that are used during training. Naively one would include as many signal hypotheses as possible, but the higher dimensionality poses a higher risk of diluting any real signal in a larger space, making it more difficult to find.

[7] This procedure assumes the design choice of a conditional normalizing flow with an unconditional latent space prior $p_z(z)$.

[8] While the explanation here follows the original implementation of QUAK, which was based on VAEs, it can also be realized with other density approximating ML models, such as normalizing flows.

5 Existing Model-Agnostic Searches

This section summarizes a few previous attempts at model-agnostic searches for BSM physics, with a focus on experiments at the LHC. They are sorted into three paradigms. Section 5.1 discusses a class of analyses that make no assumption about the signal, but rely on an accurate SM background description. It is followed by Sec 5.2, discussing a type of search that allows a data-driven background estimation by restricting the targeted signals to the well-motivated case of resonances. Lastly, Sec. 5.3 reviews previous analyses at the LHC that follow the approach of using anomaly detection methods to enhance sensitivity for BSM signals in a data-driven manner. The latter category corresponds to the same paradigm that was followed in the search performed in the context of this thesis and described in detail in Sec. 7.

5.1 "General Searches"

An established class of searches for new physics with substantially less model dependence than a dedicated search comprises so-called *general searches*. These searches aim to be unspecific with respect to any new physics signal, while still assuming a known description of the SM background. Such general searches have previously been conducted by the DØ [130–133] and CDF [134, 135] Experiments at the Tevatron, and by the H1 Experiment [136, 137] at HERA. More recently, they have also been employed at the LHC: the ATLAS general search [12] and the model-unspecific search in CMS (abbreviated as MUSiC) [13].

Both the ATLAS and CMS general searches were based on the same principle: they sorted events into classes based on the multiplicity of various reconstructed final state particles, and an automated algorithm measured the deviations of the observed data from the expected SM background, with the background being estimated using MC simulation. The search algorithms used in these analyses scanned histograms in a few kinematic variables with promising sensitivity for new signal processes (e.g., invariant masses of final state particle combinations). Various connected regions of interest (RoI) were constructed in these categories and features, and the local p-value of the observed data under the SM-only hypothesis was measured in each RoI. In line with the BumpHunter algorithm [71], which was briefly discussed in the end of Sec. 2.6.3, the RoI with the smallest local p-value (the largest disagreement) was selected for every distribution and subsequently a global p-value was derived from running pseudo-experiments, in order to reduce the look-elsewhere effect. A global p-value below a cut-off value, e.g., a value of 5%, was interpreted as a significant disagreement between the measured data and the SM expectation. However, such a disagreement cannot directly be interpreted as a discovery of new physics, as it might have been caused by mismodeling the SM background. The primary use of the general searches was therefore to trigger dedicated searches within the given RoI in case such deviations had been found, ideally on a statistically independent dataset and with a more sophisticated background estimation scheme.

There are some differences between the search strategies of the ATLAS and the CMS general searches. ATLAS only analyzed the $3.2\,\mathrm{fb}^{-1}$ of 13 TeV data collected in 2015, reserving the remaining Run 2 dataset from 2016 to 2018 for statistically independent dedicated searches, that might be triggered by the general search. The MUSiC analysis was performed on $35.9\,\mathrm{fb}^{-1}$ of 13 TeV data collected in 2016. Moreover, the choice of event classes and kinematic features differed between the two.

In both analyses, no significant deviation between observed data and the SM expectation was found. Thus, no dedicated analyses were triggered to be conducted on their respective RoIs.

5.2 Inclusive Dijet Searches

Another paradigm for model-agnostic analyses are inclusive dijet resonance searches, typically applying only loose analysis selections other than targeting a final state of two high-momentum jets. By restricting the search to heavy particles decaying into two particles with hadronic decays, they are more specific in terms of the targeted type of new physics signal than the general searches. Nevertheless, it is a topology that is motivated by many BSM scenarios, as discussed in Sec. 2.3.2. On the other hand, they tend to make fewer assumptions on the exact shape of the background than the general searches. Rather than estimating the background directly via SM MC simulation, they may follow the looser assumption that the SM background has a smoothly falling dijet invariant mass distribution in the high-mass region of interest, and thus the background shape can be estimated directly from data via the bump hunt technique described in Sec. 2.6.3.

Dijet resonance searches have been performed by several experiments in the past, such as UA1 [138], UA2 [139], CDF [140], and DØ [141]. Their background shapes were primarily based on MC simulation, whereas the CDF Collaboration eventually introduced a data-driven background estimation [142] by fitting a smooth function to the dijet invariant mass distribution. This method was also adopted, among others, by the ATLAS [14] and CMS [16] Collaborations in their inclusive dijet resonances searches at $\sqrt{s} = 7\,\mathrm{TeV}$, and later at $\sqrt{s} = 13\,\mathrm{TeV}$ [15, 17].

These searches target a relatively wide range of BSM models, and thus the significance scans and 95% CL exclusion limits, in terms of the product of cross section, acceptance, and branching fraction, can be interpreted for more than a few dedicated signal processes. However, the overwhelming number of QCD background events passing their loose selections makes it challenging to achieve a good signal-to-background ratio, and thus the sensitivity to new physics is substantially lower than in more dedicated searches. In order to achieve higher sensitivity, the ATLAS and CMS Run 2 searches have been extended to searches with additional requirements, such as the presence of b-tagged jets [6, 143], thus sacrificing some model independence. Another extension, retaining more model independence, is achieved via machine learning anomaly detection algorithms. These are discussed in Sec. 5.3.1 for ATLAS and Sec. 7 for CMS.

5.3 Machine Learning Anomaly Searches

To date, three full analyses [18–20] have been performed at the LHC using machine learning anomaly detection methods of the type described in Sec. 4. They have all been performed by the ATLAS Collaboration using their Run 2 dataset with a center-of-mass energy of 13 TeV and are briefly summarized in the following.

5.3.1 ATLAS Dijet Resonance Search with Weak Supervision

The first such search [18] employed the CWoLa Hunting method, as described in Sec. 4.3.3. The target topology was any BSM particle A that decays into particles B and C, which can potentially also be BSM, that then decay hadronically. The final state under consideration consisted of two large-radius jets.

The invariant mass of the two leading jets was used as the resonant feature to distinguish between signal region and sidebands as needed in the CWoLa Hunting protocol. Specifically, the invariant mass spectrum was divided into varying-width bins, with widths motivated considering the detector resolution. Then, for each bin the CWoLa Hunting procedure was applied using the current bin as signal region and the neighboring bins as sidebands.

The neural network classifier used the two leading jet masses as input features to detect differences between signal region and sideband events, and was trained multiple times in a k-

fold cross-validation scheme in order to use the full dataset for both training and evaluation. The most anomalous 1% and 10% of the data points according to these classifier models were then selected and used to perform a bump hunt background estimation via a heuristic functional form.

This scan over invariant mass bins yielded local p-values for excesses as a function of the mass, both for the 1% and 10% selection efficiency. The largest local excess in the analysis was observed with a significance of 3σ, and corresponded to a global 1.5σ excess after taking into account the look-elsewhere effect.

In the absence of a significant excess, the search was used to set 95% CL exclusion limits on a generic heavy-vector triplet [144] W' signal with varying masses, m_{B} and m_{C}, of the two decay products. As the signal efficiency of weakly supervised anomaly detection methods depends on the presence of signal in the training data, the classifiers were trained with varying amounts of signal MC simulation events injected into the data. The point where the exclusion limit matches the injected signal cross section was interpreted as the 95% CL exclusion limit on the signal. The reported limits outperform those of a generic inclusive dijet resonance search (Sec. 5.2).

5.3.2 ATLAS H + X Anomaly Search

The second LHC anomaly search [19] targeted a heavy resonance Y decaying into a Standard Model Higgs boson and a new particle X, again considering only fully hadronic final states. The search was performed in the regime where the Y mass is high and thus both H and X have a high Lorentz boost, resulting in two high-momentum large-radius jets in the detector.

The two leading-p_{T} jets were first classified into an H candidate, $\mathrm{J_H}$, and an X candidate, $\mathrm{J_X}$. This was achieved through a neural network, trained to distinguish between jets arising from an $\mathrm{H} \to \mathrm{b\overline{b}}$ process and top or QCD jets. The resulting jet-level discriminant quantifying the probability of originating from an $\mathrm{H} \to \mathrm{b\overline{b}}$ decay was denoted as $D_{\mathrm{H_{bb}}}$. The jet with higher $D_{\mathrm{H_{bb}}}$ was classified as $\mathrm{J_H}$, and the other one as $\mathrm{J_X}$.

Three analysis regions were defined via selections on $\mathrm{J_X}$, as illustrated in Fig. 5.1. The first region, the anomaly region, was defined via a jet-level anomaly score (S_{A}) [145] based on a variational recurrent neural network (VRNN) [146]. Each jet was modeled as an ordered sequence of up to 20 constituent 4-vectors. The order of the constituents was deduced from transverse momenta. A VAE was trained to reconstruct these sequences via a recurrent neural network, conditioned on high-level jet features. The conditioning aimed at revealing correlations between the constituents and the jet substructure. The VAE loss at each time step consisted of a reconstruction loss and the KL divergence D_{KL} between the encoded posterior distribution and a Gaussian one. The average loss over all time steps per sequence was used for the parameter update, while the average D_{KL} term per jet was used as the anomaly score: $S_{\mathrm{A}} = 1 - \exp(-D_{\mathrm{KL}})$. The anomaly region was defined by selecting events where $S_{\mathrm{A}} > 0.5$ for $\mathrm{J_X}$.

Moreover, two additional analysis regions were defined to provide a benchmark for the anomaly detection approach, optimized for an $\mathrm{X} \to \mathrm{q\overline{q}}$ decay. The merged two-prong region covered the case where both quarks are within a single large-radius jet. It was selected by requiring $D_2^{\mathrm{trk}} < 1.2$, where this energy-correlator substructure variable [147] is lower for two-prong jets. The complementary region $D_2^{\mathrm{trk}} > 1.2$ was the resolved two-prong region. Since the two subjets are not well contained within a single large-radius jet in this case, they were reconstructed as two individual small-radius jets. These two analysis regions were constructed to be orthogonal to each other, but could have an overlap with the anomaly region.

The analysis procedure was performed three times, once in each of the analysis regions. Each analysis region was divided into six subregions, based on a selection on $D_{\mathrm{H_{bb}}}$ (where the signal region is above a cut capturing 60% of preselected events) and one on the H jet mass m_{H} (where

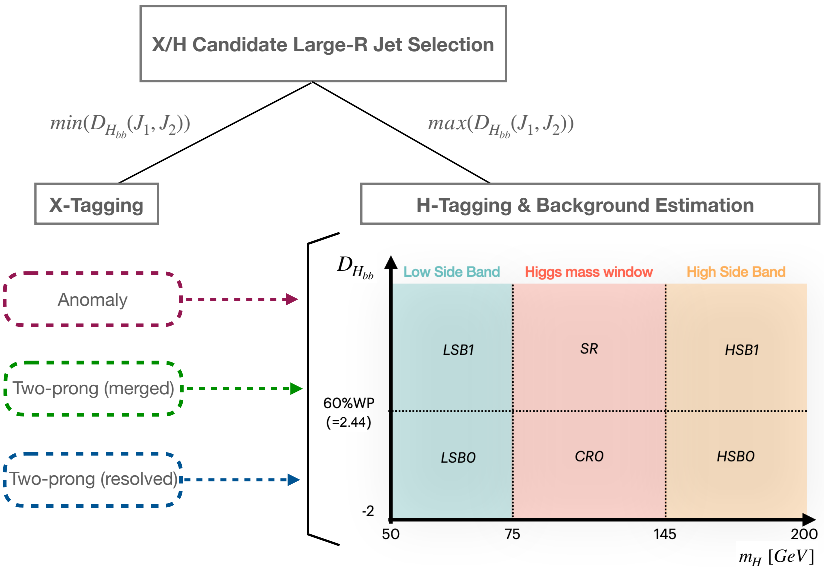

Figure 5.1: Illustration of the ATLAS H + X anomaly search signal regions. The H → $b\bar{b}$ jet classifier score $D_{\mathrm{H_{bb}}}$ is used to sort the leading jets J_1 and J_2 into an X and H jet candidate. The X jet candidate is used to classify the event into one of the three analysis regions: anomaly, two-prong (merged) and two-prong (resolved). The H jet candidate is used to divide the events further into signal region (SR) and sideband regions, based on $D_{\mathrm{H_{bb}}}$ and the H jet mass m_{H}. Figure taken from Ref. [19].

the signal region is defined by a window around the nominal Higgs boson mass), illustrated in Fig. 5.1. A small neural network was trained to reweight the dijet mass distribution in the high-mass sideband of the low $D_{\mathrm{H_{bb}}}$ region to match the high $D_{\mathrm{H_{bb}}}$ one. The reweighting function was then validated in the low sideband region and finally used to construct a background template in the signal region by reweighting the low $D_{\mathrm{H_{bb}}}$ control region. This background template was then used to perform a bump hunt in the dijet invariant mass, in overlapping bins in the X mass. This thus resulted in three 2-dimensional local p-value scans in the m_{X}-m_{Y} plane. The smallest observed p-value corresponded to a global significance of 1.43σ.

As no significant excess was seen, 95% CL exclusion limits were set on the production cross section of several benchmark signal models for m_{Y} between 1.5 and 6 TeV and m_{X} between 65 and 3000 GeV. The limits from the anomaly region were generally competitive with the ones from the two-prong regions, and in one case more stringent than the more dedicated benchmarks.

5.3.3 ATLAS Unsupervised Two-Body Invariant Mass Search

The third published anomaly search by the ATLAS Collaboration [20] generally targeted any resonances in two-body final states of a jet and object Y. The jet could either be a light jet or a b jet, and these two define separate categories. The Y could be either an electron, a muon, a photon, a light jet, or a b jet. These combinations resulted in nine different categories, where the highest-p_{T} candidate of each object was selected. The same event could thus contribute to multiple categories. The events required the presence of an isolated lepton during online selection, as this allows probing invariant mass regions below 1 TeV, which is more challenging with jet-based trigger algorithms.

The primary analysis selection was done through a data-driven anomaly detection algorithm, in this case an autoencoder that was trained to reconstruct the input through an information bottleneck, as described in Sec. 3.4.1. The input features were the entries of the rapidity mass matrix (RMM) [148, 149], which contains scaled invariant mass and rapidity information of all pairs of final state objects, as well as the missing transverse energy. For the creation of the RMM, up to ten light jets, ten b jets, five electrons, five muons and five photons in the final state were considered. Missing values were filled with zeros, and the nine invariant mass combinations defining the analysis categories were removed in order to reduce bias. This autoencoder was trained 50 times on a random 1% subset of the preselected data and the median model, in terms of validation loss, was selected. The reconstruction loss served as the anomaly score. The cross section passing a selection on this anomaly score was used to define three overlapping anomaly regions: 10, 1, and 0.1 pb.

Bump hunts were performed independently in the respective invariant mass of each of the nine categories, once for each of the anomaly regions. The strongest local excesses were found to be 2.8 and 2.9σ respectively, which were not deemed globally significant.

As in the other analyses, the scan for excesses was followed by setting 95% CL exclusion limits, in this case on the product of cross section, acceptance, efficiency and branching fraction of generic Gaussian-shaped signals. Only the 10 pb anomaly region was used for this purpose, as it was the most sensitive one.

6 Developing and Improving Anomaly Detection Methods

While the previous sections introduced and reviewed concepts from the literature, this section presents the results of my own research with collaborators on developing and improving anomaly detection methods in the context of BSM searches. All the studies presented in this section will use the LHCO R&D dataset, introduced in Sec. 6.1, for establishing performance benchmarks. Section 6.2 presents the CATHODE method, which is followed by a modification that facilitates a more stable background estimation, called LaCATHODE and covered in Sec. 6.3. Section 6.4 discusses the issue of signal sensitivity degradation in the presence of uninformative features in weak supervision, and proposes a strategy for substantially improving the signal sensitivity in such cases.

6.1 Benchmark Dataset

For many of the anomaly detection methods that are designed for HEP, performance benchmarks are established using the LHC Olympics (LHCO) R&D dataset [150]. It was initially released for developing and testing methods in the context of the LHCO 2020 anomaly detection challenge [151]. The motivation behind this dataset is that conventional, dedicated searches are expected to have low sensitivity to the type of signal that is present in the dataset, because both the primary resonance and its decay products are BSM particles.

It contains simulated proton-proton collision events at an LHC-like collider, consisting of one million QCD dijet background events and 100,000 signal events of a heavy BSM particle W$'$ decaying into two lighter BSM particles X and Y, which then each decay into quark-antiquark pairs[9]. The masses of these three BSM particles are $m_{\mathrm{W}'} = 3500\,\mathrm{GeV}$, $m_{\mathrm{X}} = 500\,\mathrm{GeV}$ and $m_{\mathrm{Y}} = 100\,\mathrm{GeV}$, respectively. The events were generated with PYTHIA 8.219 [52] and the detector response was modeled with DELPHES 3.4.1 [152] using the default settings. The simulation does not include pileup or multiparton interactions. The events were required to pass a trigger based on the presence of at least one anti-k_{T} [53] jet with $R = 1.0$ and $p_{\mathrm{T}} > 1.2\,\mathrm{TeV}$. Up to 700 reconstructed particles are stored per event, where each particle is represented by its transverse momentum p_{T}, pseudorapidity η and azimuthal angle ϕ, and setting their mass to zero.

A high-level feature representation was also released, where the particles are clustered into anti-k_{T} jets with $R = 1.0$ using FastJet [153], and only features for the two highest-p_{T} jets are stored: the Cartesian 3-momenta, their masses, and the three subjettiness variables τ_1, τ_2 and τ_3 (as defined in Sec. 2.5.5).

Several studies, including the following sections of this thesis, sort the two leading-p_{T} jets by ascending mass, such that $m_{j1} < m_{j2}$, and extract the dijet invariant mass m_{jj}, the mass of the lighter jet m_{j1}, the mass difference between the two jets $\Delta m_j = m_{j2} - m_{j1}$, and the subjettiness ratios $\tau_{21} = \tau_2/\tau_1$ of the two jets. Another feature that will be used is the angular separation between the two jets, $\Delta R_j = \sqrt{(\Delta\eta_j)^2 + (\Delta\phi_j)^2}$.

In the course of this work, an additional set of 612,858 QCD dijet background events were generated [154] with the same settings as used for the original LHCO R&D dataset, but only within the dijet invariant mass window of $3.3\,\mathrm{TeV} < m_{jj} < 3.7\,\mathrm{TeV}$. The motivation was to reduce the statistical uncertainty in the performance evaluation of methods using this mass region as the signal region, and to allow the training of benchmark methods with a higher degree of idealization.

[9] Some sources in the literature, including the original LHCO paper [151], refer to the signal particle as a Z$'$ particle, which is conventionally a neutral heavy boson whereas W$'$ is its charged counterpart. However, the configuration file in the dataset repository [150] refers to the particle as a W$'$.

One focus of investigation will be the behavior of anomaly detection methods when certain input features are correlated. The Pearson correlation coefficients, measuring linear correlation of the auxiliary features $(m_{j1}, \Delta m_j, \tau_{21,1}, \tau_{21,2})$ with m_{jj}, are shown in Fig. 6.1. One can see that within the full region the jet mass–based features are correlated with m_{jj} by up to 11%. However, within the relatively thin slice that will in the following be used as the primary choice of “signal region”, the correlations are only around 1%, and they increase roughly by a factor of two when additionally considering the adjacent 200 GeV–wide SSBs on each side.

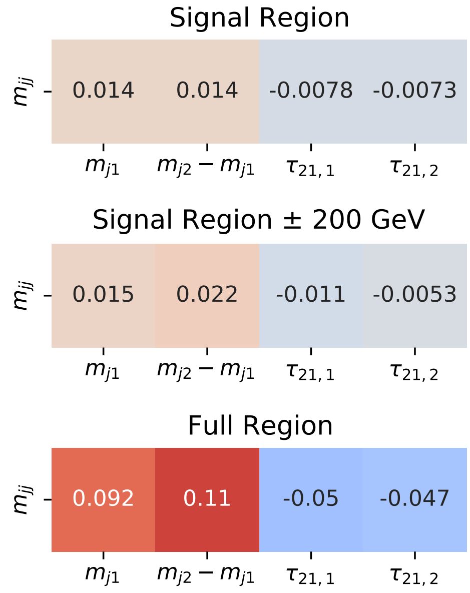

Figure 6.1: Pearson correlation coefficients showing the linear correlations between m_{jj} and the auxiliary features m_{j1}, Δm_j, $\tau_{21,1}$ and $\tau_{21,2}$. Three overlapping m_{jj} regions are considered: the signal region $3.3\,\text{TeV} < m_{jj} < 3.7\,\text{TeV}$ (SR), the SR and adjacent short sidebands (SSB) $m_{jj} \in [3.1, 3.9]$ TeV, and the full region.

6.2 CATHODE

This section is based on work presented in the publication in Ref. [1], which was developed with the co-authors Anna Hallin, Joshua Isaacson, Gregor Kasieczka, Claudius Krause, Benjamin Nachman, Tobias Quadfasel, Matthias Schlaffer, and David Shih. The figures and results thus largely overlap with the publication. I made key contributions to the work by implementing the CATHODE algorithm prototype consistently along with the benchmark methods, proposing major algorithmic improvements, such as preprocessing schemes and the conditional feature sampling, obtaining and analyzing the majority of the results, assisting in writing the article, and submitting its preprint.

Section 4.3.2 discussed the concept of resonant anomaly detection, where the data-over-background likelihood ratio of Eq. 4.6 is approximated by a machine learning classifier, trained to distinguish observed data from a background-only simulation. This background template is typically constructed with the use of real data. Section 4.3.3 provides an overview of such approaches, where most rely on the distinction into signal region (SR) and sideband (SB) regions. The likelihood ratio is estimated within the SR, while the background template is constructed from the SB. The algorithm *classifying anomalies through outer density estimation* (CATHODE) [1] was the first to construct this background template in a purely data-driven manner using machine learning. This section describes the technical details of the method and evaluates its performance.

6.2.1 Algorithm and Implementation

CATHODE requires the core setup, described in Sec. 4.3.2, with one resonant feature m and a set of auxiliary features x. Similarly, the spectrum of m is split into the SR, where the presence of a signal overdensity is tested, and the complementary SB regions. The purpose of x is to capture potential deviations between said signal process and the background. The algorithm consists of three main components:

1. Training an ML model to learn the conditional density of SB data: $p_{\mathrm{data}}(x|m \in \mathrm{SB})$. Under the assumption that a signal, if it exists, is mostly contained within the SR, this corresponds to a background model: $p_{\mathrm{data}}(x|m \in \mathrm{SB}) \approx p_{\mathrm{bkg}}(x|m \in \mathrm{SB})$.

2. Sampling synthetic background data from the interpolated model. This is achieved by first sampling m values within the SR $m \sim p_{\mathrm{bkg,SR}}(m)$ and then sampling x using the SR m as condition: $x \sim p_{\mathrm{bkg,SR}}(x, m) = p_{\mathrm{bkg}}(x|m)p_{\mathrm{bkg,SR}}(m)$. The m distribution can either be fitted from the SB or approximated by a one-dimensional density estimator to the SR data, assuming the signal contribution in m is negligible in the setup of anomaly detection: $p_{\mathrm{data,SR}}(m) \approx p_{\mathrm{bkg,SR}}(m)$.

3. Training an ML classifier to distinguish between observed SR data $\mathcal{D}_{\mathrm{data}} \sim p_{\mathrm{data,SR}}(x)$ and the synthetic background $\mathcal{D}_{\mathrm{bkg}} \sim p_{\mathrm{bkg,SR}}(x)$.

The following paragraphs describe the implementation of the CATHODE proof of concept from Ref. [1] to gain signal sensitivity in the LHCO R&D dataset, introduced in Sec. 6.1. The code to reproduce the results is available in a public repository [155].

For this test case, the high-level feature prescription described in Sec. 6.1 is used, i.e., the two highest-p_{T} jets are ordered by their masses, such that $m_{j1} < m_{j2}$. The auxiliary features are chosen to be $x = \left(m_{j1}, \Delta m_j, \tau_{21,j1}, \tau_{21,j1}\right)$. The resonant feature is the invariant mass of the resulting dijet system, which is split into the SR $m_{jj} \in [3.3, 3.7]$ TeV and the SB $m_{jj} \notin [3.3, 3.7]$ TeV. This choice of features and SR/SB separation follow closely Ref. [124] and are illustrated in Fig. 6.2. The m_{jj} distribution shows that the choice of SR conveniently aligns with the actual resonance peak, capturing $\sim 77\%$ of the signal events. This is an idealization for the proof of concept that focuses on the signal sensitivity in case of a well-chosen SR. In a real analysis, where the signal is unknown, the SR is scanned over various positions of the mass spectrum. In addition to providing signal sensitivity when the SR captures the signal peak, the method should also not produce false positives in the absence of signal. The latter will be briefly discussed later in this section and addressed more systematically in Sec. 6.3. Another degree of idealization is that all four auxiliary features exhibit differences between the signal and the background. The bulk of signal events peak narrowly at the Y mass of 100 GeV in m_{j1}, whereas the QCD background shows no such peak. Similarly, for the signal the distribution of Δm_j peaks at 400 GeV, the mass difference between X and Y, with a small secondary peak towards zero corresponding to events that are not well reconstructed with the two leading jets. The subjettiness ratios τ_{21} are constructed such that the events with two well-separated subjets have low values, which is more common in the signal than in the background, as both the X and Y each decay into a quark-antiquark pair. In practice, it is unlikely that all auxiliary features have such high discrimination power for a real signal. Departures of this idealization will be discussed in Sec. 6.4.

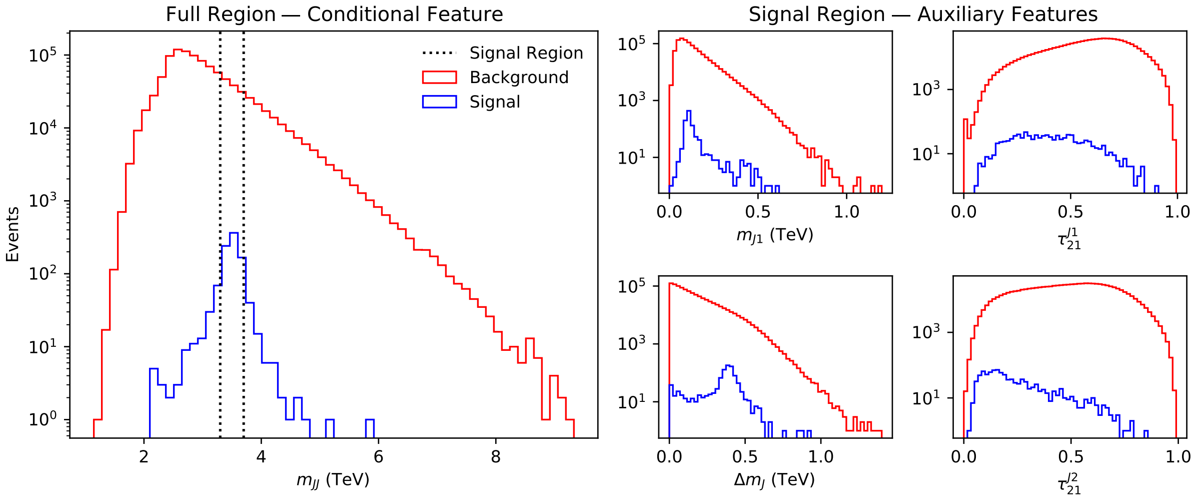

Figure 6.2: Illustration of the features used in the CATHODE proof of concept using the LHCO R&D dataset. The resonant feature is the dijet invariant mass m_{jj}, which is split into the signal region (SR) and sideband (SB). The auxiliary features are the mass of the lighter jet m_{j1}, the mass difference between the two jets Δm_j, and the subjettiness ratios τ_{21} of the two jets. The full region contains one million background and one thousand signal events.

In order to create a proxy for real detector data, the one million QCD dijet background events from the LHCO R&D dataset are mixed with one thousand signal events. This choice, which is also in line with Ref. [124], corresponds to a realistic setting where the signal presence is too small to be observed directly with a typical bump hunt. The approximate significance over the full m_{jj} range is $N_S/\sqrt{N_B} \approx 1$. If the range is restricted to the SR, it corresponds to $N_S/\sqrt{N_B} \approx 2.2$ and a direct signal-to-background ratio of $N_S/N_B \approx 6 \times 10^{-3}$. The performance with respect to smaller signal fractions is studied later in Sec. 6.2.3. In the SB region, 500,000 of the proxy data events are used for training and the remaining (378,876) events are reserved for the validation set. In the SR, the proxy data are split into two equal parts for training and validation sets, resulting in approximately 60,000 events in each. The performance evaluation in Sec. 6.2.3 will be performed on a separate test set. In a real application, one would rather implement a k-fold cross-validation procedure, as described in Sec. 3.2.2, in order to use all data for both training and inference.

Various ML models would technically qualify for learning $p_{\text{data}}(x|m \in \text{SB})$. The requirements are that the model can learn to generate samples of x conditioned on m following the same distribution as the training data, and that the conditioning interpolates well into an unseen region. The latter requires a smooth inductive bias, which is in general the case for neural networks. As the explicit density is not necessary, the conditional generative model could be realized with a generative adversarial network (GAN) [156] or a conditional VAE (Sec. 3.4.1). Another option would be a mixture density network (MDN) [157], which explicitly models the density but has only limited expressive power. Normalizing flows (Sec. 3.4.3) are a particularly attractive choice for this task, as both the prior distribution in the latent space and the conditional mapping, parametrized by a neural network, are inherently smooth. Moreover, they are very expressive and the associated NLL minimization is a well-defined training objective.

A known downside of normalizing flows is that sharp boundaries in the data distribution can be difficult to model with smooth distortions from a Gaussian prior. An example for such a sudden discontinuity in this dataset is the left edge of the Δm_j distribution in Fig. 6.2, where the background shape starts immediately with a peak at zero. This is mitigated by preprocessing the features. First, all x features are scaled to lie within a range of $[0, 1]$. Then a logit function $\log(x/(1-x))$ is applied, which stretches the points near the boundaries to $(-\infty, +\infty)$. Lastly, the mean is subtracted, and the result is divided by the standard deviation, which centers the data around zero and scales it to unit variance. For an exact description of the likelihood, this would need to be accounted for in the Jacobian of the transformation. For use as a generative model, the transformations are applied in reverse order during sampling. The feature processing is omitted for m, as it is already naturally of $\mathcal{O}(1)$ as units of TeV were chosen for m_{jj}.

In this study, a masked autoregressive flow (MAF) with affine transformations (as described in Sec. 3.4.3) was used, in exact analogy to Ref. [124]. The code implementation and hyperparameters are the same as in Ref. [158]. The MAF was thus realized in PyTorch [159], with 15 MADE blocks, each consisting of a single hidden layer with 128 nodes. Between each MADE block, a permutation of the feature order is performed. The prior latent space density is chosen to be a four-dimensional Gaussian distribution with zero mean and unit variance, independent of the conditional variable. The batch size during training with the Adam optimizer was set to 256 and the learning rate to 10^{-4}. In order to improve the numerical stability of the training process, the output of each MADE block hidden layer was scaled to a mean of zero and variance of one with respect to the entire training set before each new epoch. This significantly reduced the risk of training divergence, however it also involved mean and variance computations on the entire training set, which is computationally expensive. The model was trained for 100 epochs, after each recording the validation loss. The minimum of the monitored validation loss was observed to be well before the final epoch, as can be seen in an example loss curve in Fig. 6.3.

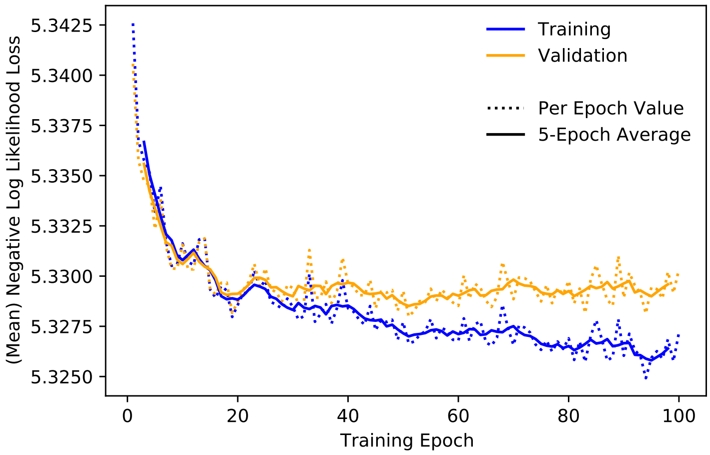

Figure 6.3: Training and validation losses per epoch (dotted lines) during one training of the MAF model for CATHODE. The values are also averaged over 5 epochs (solid lines) to improve visibility of overall trends. The training loss in the plot is defined by evaluating the NLL of the training set after each epoch, and thus is not necessarily monotonically decreasing. Figure taken from Ref. [1].

Once the $p_{\text{bkg}}(x|m)$, parametrized as a normalizing flow, is learned from the SB, it can be used to generate SR background by sampling SR values of m and inserting these into the sampling function of the MAF. Multiple approaches for obtaining $p_{\text{bkg,SR}}(m)$ were considered. As a first prototype, m was sampled uniformly within the SR of 3.3 to 3.7 TeV. Despite not being an accurate description of the real background, it still resulted in a relatively faithful sample of x, i.e., a classifier trained to distinguish $\mathcal{D}_{\text{data}}$ from $\mathcal{D}_{\text{bkg}}$ purely based on values in x (but not m) still learned to discriminate signal from background. This is because the correlations between x and m are relatively small over the SR in the LHCO R&D dataset. One alternative that was considered is to fit a smooth heuristic function from the SB m values and then sample from the function inside the SR. While this would reduce the risk of signal contamination in $p_{\text{bkg,SR}}(m)$, its additional degrees of freedom decrease the stability of an automated method, and it requires an a-priori choice of functional form rather than being entirely data driven. A more straightforward approach is to instead fit a one-dimensional kernel density estimator (Sec. 3.4.2) to the data m values in the SR. As anomaly detection is usually applied in the case where a signal peak in m is too small to detect without additional selections (e.g., on x), it is reasonable to assume that its contribution to $p_{\text{data,SR}}(m)$ is negligible. Moreover, even if the background template were to model a dominant peak in m from the signal, its values in x, which are modeled by $p_{\text{bkg}}(x|m)$, may still differ from the signal. The KDE was implemented with the Scikit-learn library [160] using a Gaussian kernel and a bandwidth of 0.01. Instead of fitting the m distribution directly, it was first rescaled to the range of $[0, 1]$ and then transformed with a logit function, as described for the MAF preprocessing. Otherwise, the Gaussian kernel would visibly smoothen out the sharp SR boundaries. The inverse transformation is applied after sampling m values. An illustration of the KDE samples with and without the logit preprocessing, and in comparison to naive uniform sampling, is shown in Fig. 6.4.

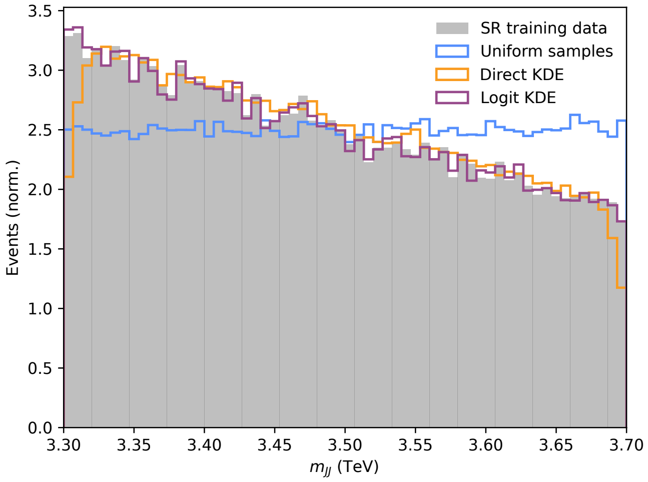

Figure 6.4: Comparison of methods for sampling m values in the SR of the LHCO R&D dataset. A simple uniform sampling is compared to a KDE with and without the logit preprocessing. The boundaries are not well described by the KDE without logit preprocessing.

With the logit-preprocessed KDE sampling, a pre-defined number N_{samples} of samples in m is drawn from $p_{\text{bkg,SR}}(m)$. Unless otherwise stated, $N_{\text{samples}} = 400{,}000$, which corresponds to roughly 3.3 times as many synthetic background events as there are proxy data. This feature of *oversampling* background is a strength of the CATHODE method over other approaches, such as CWoLa Hunting, that are limited by the number of available data points to directly build the background template samples. The effect of various degrees of oversampling, i.e., different choices of N_{samples}, on the signal extraction performance is discussed later in Sec. 6.2.3. Figure 6.3 shows that the validation loss of the MAF training oscillates visibly from epoch to epoch. In order to increase the stability of the sampling, the MAF model is not only used at the epoch with the lowest validation loss, but instead an ensemble of the ten model states with the lowest validation losses are used. Specifically, the N_{samples} values in m are split into ten equal parts. Each of these parts is then used as conditional to sample $x \sim p_{\text{bkg}}(x|m)$ at each model state, and these are merged into a single set and shuffled. Finally, they are split proportionally into training and validation sets. Figure 6.5 compares the resulting background template with the actual background in the proxy data, showing that the in-situ background simulation by the MAF approximates the true background well.

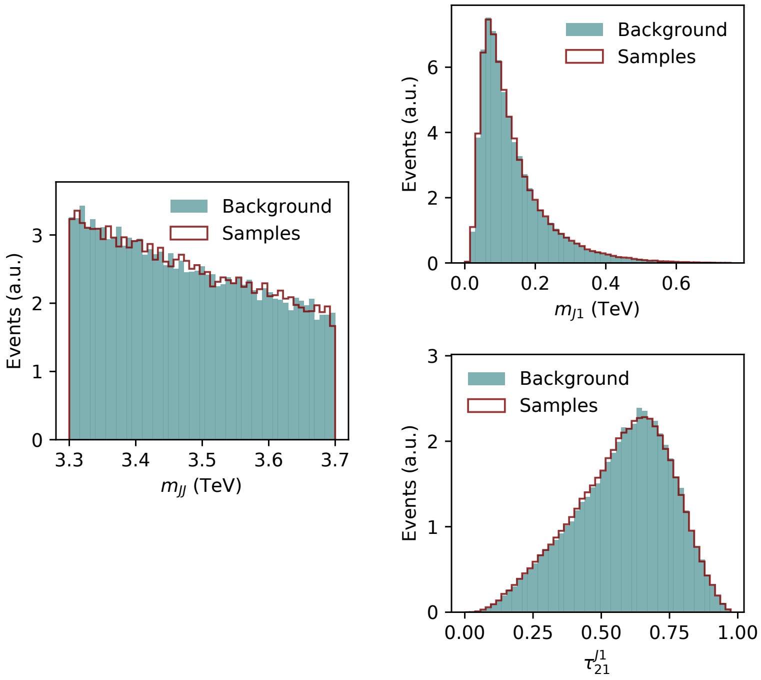

Figure 6.5: Comparison of the background template generated by the MAF with the actual SR background in the LHCO R&D dataset. Figure taken from Ref. [1].

The binary classifier, tasked to distinguish between $\mathcal{D}_{\text{data}}$ and $\mathcal{D}_{\text{bkg}}$, was implemented via PyTorch [159] as a fully connected neural network with three hidden layers of 64 nodes each and ReLU activation. The output was realized as a single node with sigmoid activation. The Adam optimizer was used to minimize the binary cross-entropy between proxy data and background samples with a learning rate of 10^{-3} and a batch size of 128. Again, the training was performed with 100 epochs, in order to generously surpass the validation loss minimum. Figure 6.6 shows the training and validation losses during one training of the classifier, which already converges after 15 epochs. In order to account for the different sizes of the $\mathcal{D}_{\text{data}}$ and $\mathcal{D}_{\text{bkg}}$, a class weight is applied to the loss function, such that the two datasets contribute equally. This is applied both to the training and the validation loss computation. The absence of such reweighting would bias the classifier towards predicting more background-like labels more often. The features are scaled to zero mean and unit variance as a preprocessing step to increase the numerical stability of the training. A logit transformation, as used for the flow training, is not applied here as it consistently decreased the performance. Only values of x are used as input, whereas m could technically be included as well. A study on the effect of including m as an input feature is discussed later in Sec. 6.2.3. Again, the learning task is subject to epoch-by-epoch noise, in particular because the discrimination between almost identical samples is inherently noisy. To increase the stability of the classifier prediction, an ensemble of the ten lowest-validation loss epochs is used and implemented via a mean of the model predictions per data point.

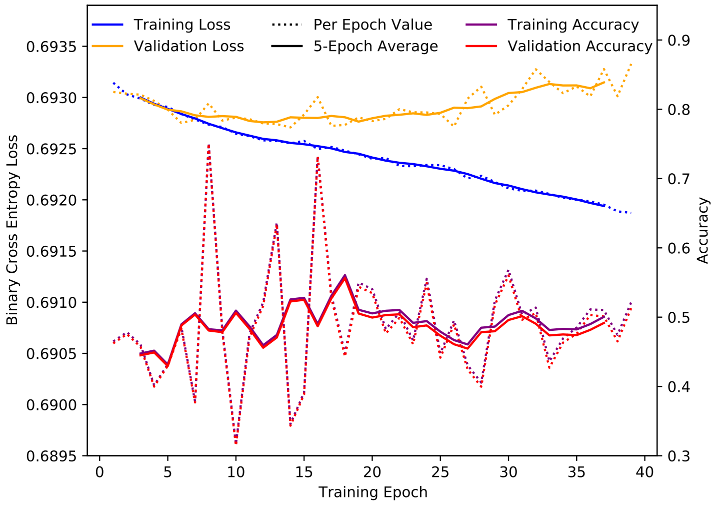

Figure 6.6: Binary cross-entropy loss, as well as the accuracy of validation and test set, respectively, per epoch (dotted lines) during one CATHODE classifier training. The values are also averaged over 5 epochs (solid lines) to improve visibility of overall trends. An accuracy of 50% corresponds to finding no difference between the two datasets, which is the desired result given the small signal contribution in the proxy data. Figure taken from Ref. [1].

6.2.2 Benchmark Method Implementation

The performance of CATHODE is evaluated in comparison to several benchmark methods, which provides context for the results.

On the one hand, the performance of CATHODE is assessed in comparison to CWoLa Hunting and ANODE (see Sec. 4.3.3), since they are similar data-driven methods. CWoLa Hunting uses an ML classifier for the likelihood ratio estimation, but has no generalizable background template construction as it directly relies on SB data values of x. ANODE estimates the background likelihood explicitly from the SB region with a conditional density estimator but relies on a second such model for constructing the likelihood ratio in the SR. Other methods that existed before CATHODE either depend on a prior MC background simulation sample (SA-CWoLa, SALAD) or make stronger assumptions on the type of signal (TNT).

On the other hand, the performance is also compared to two more idealized benchmarks. The first one is an IAD, where the flow-sampled background template is replaced by a background sample following the same distribution as the background in the SR proxy data. This yields an upper performance boundary of what could be expected from CATHODE if the background template was perfect. The second one is a fully supervised classifier, trained directly to distinguish signal from background events. This quantifies what signal sensitivity could be achieved by a typical dedicated search, where the target signal (and the SM background) is known and well modeled with MC simulation. The next paragraphs describe the implementation of these benchmark methods, as shown in the figures of Section 6.2.3.

For CWoLa Hunting, the same binary classifier architecture and training scheme was used as for CATHODE, including the ten-epoch ensembling of the anomaly score. The only difference is the background template, which is drawn directly from the SB. Specifically, a short SB (SSB) region was defined as the range of $m_{jj} \in [3.1, 3.3] \cup [3.7, 3.9]$ TeV, following Ref. [124]. Across these 200 GeV–wide SSBs, only marginal correlations between m and x are present. The same 50/50 training/validation set split is applied to this SSB background sample as to the proxy data. Since the left SSB includes more events than the right one because of the exponentially falling nature of the m_{jj} distribution, a sample weight is applied such that the two SSBs contribute equally to the loss function. This is multiplied with a class weight that ensures equal contributions between SR data and the background template.

Similarly, the ANODE implementation is based on the same conditional normalizing flow architecture and training scheme as in CATHODE. In addition to the flow trained in the SB, another flow with the same architecture is trained in the SR. The anomaly score is then computed as the ratio of the likelihoods of the SR and SB flows. Similarly to other methods, a ten-epoch ensemble is used to increase the stability. For ANODE, this is obtained by averaging the likelihood values of the ten epochs with the lowest validation loss per data point, in each of the two trainings, before their ratio is computed.

The IAD classifier is fully analogous to the CATHODE case. The only difference is that the background template is realized with 272,000 events from the additionally produced SR background events that follow the same distribution as the original LHCO R&D background. They are split evenly into training and validation set. A more optimal comparison would have been to use 400,000 events, just as in the default CATHODE background sample. Since the rest of the additional background events was already set aside for the test set, it would have been necessary to produce more and this was not possible due to time constraints. In the later studies, which are discussed in Sec. 6.3, this shortcoming was resolved by applying a different dataset split but obtaining equivalent results.

The fully supervised classifier architecture, training and inference are equivalent to CATHODE, but it uses signal and background labels directly. The same 272,000 additional background events as for the IAD are used as background, while 55,000 SR signal events that are not used for the data proxy injection constitute the signal sample. The use of only complementary events to the proxy data, as well as the relatively large signal sample, are chosen to imitate a scenario where an abundance of both signal and background MC simulation events are available. These sets are split evenly into training and validation sets, and class weights are applied to the loss function, such that the signal and background samples contribute equally.

The amount and type of data used for training and validation of each method are summarized in Tab. 6.1.

Table 6.1: Numbers of events in units of 1000 ≡ 1k used for training, validation, and testing for each method. In this context, validation refers to selecting model epochs. The test set is the same for all methods. DE and CLS refer to density estimators and classifiers, respectively. “Bkg” and “sig” refer to background and signal, respectively. The term “bkg samples” here indicates that they originate from the synthetic background template.

Method	Type	Train	Validation	Testing
CATHODE	DE	500k SB data	380k SB data	340k SR bkg 20k SR sig
	CLS	200k SR bkg samples 60k SR data	200k SR bkg samples 60k SR data	
ANODE	DE	500k SB data 60k SR data	380k SB data 60k SR data	
CWoLa Hunting	CLS	65k SSB data 60k SR data	65k SSB data 60k SR data	
Idealized AD	CLS	136k SR bkg 60k SR data	136k SR bkg 60k SR data	
Fully Supervised	CLS	136k SR bkg 27k SR sig	136k SR bkg 27k SR sig	

6.2.3 Results

The following paragraphs discuss the performance of CATHODE, in particular with respect to the benchmark methods. The discussions include the signal sensitivity (the discrimination power between signal and background), the robustness of this sensitivity with a smaller signal presence and correlated input features, and exploring the background estimation potential.

The test set, used to compute all performance metrics, consists of 340,000 SR background events and 20,000 SR signal events. This is a different relative signal fraction than in the training data, however all signal sensitivity quantities are based on TPR and FPR, which are independent of the signal fraction. The results are expected to vary among different trainings due to the stochastic nature of ML trainings, in particular in this case where the number of signal events to detect during the training is small. Therefore, the ML models (classifiers and/or normalizing flows) were trained ten times, each with a different random weight initialization. The performance is then reported in terms of the median of these ten trainings and the variance is quantified by the central 68% of the results around the median, i.e., the range from the 16^{th} to the 84^{th} percentile. The training, validation and test sets are kept fixed for all ten trainings, except when hatched uncertainty bands are shown. In that case (Fig. 6.8) this is also varied, i.e., the sizes and signal fractions remain the same but are randomly shuffled, in order to test the variance with respect to different realizations of the same signal-to-background ratio.

ROC and SIC curves depend on computing the TPR and FPR at all possible decision thresholds on the anomaly score. In the case of very high thresholds, only few individual events will pass the selection and the performance metrics are thus more prone to statistical fluctuations. In order to remove statistically unstable regions from ROC and SIC curves, a lower limit is imposed on the FPR values, corresponding to a Poisson uncertainty in the remaining background events of 40%. The median of the ten trainings is displayed if at least five trainings satisfy this criterion. The same criterion is imposed whenever the maximum SIC value is quoted for a method.

Signal Sensitivity

Figure 6.7 shows ROC and SIC curves of CATHODE in comparison to the benchmarks discussed in Sec. 6.2.2. CATHODE outperforms the other anomaly detection methods significantly over a wide range of signal efficiency values. The CATHODE SIC improves by roughly a factor two with respect to ANODE and a factor of 1.3–2 with respect to CWoLa Hunting. CATHODE, ANODE and CWoLa Hunting reach a maximum SIC of 14, 6.5 and 11, respectively.

The improvement over ANODE, which uses the same SB-trained model of $p_{\text{bkg}}(x|m)$, shows that a simple fully connected neural network classifier is significantly more capable of approximating the likelihood ratio than training another normalizing flow to estimate the SR likelihood first and then computing the ratio explicitly. This likely stems from the ANODE task of training a smooth normalizing flow to accurately model the small, generally sharply peaking, signal contribution on top of the large background.

The CWoLa Hunting benchmark uses the same classifier architecture, but suffers from a less accurate background template. This is likely due to the percent-level correlation between x and m (shown in Fig. 6.1), which is small enough for CWoLa Hunting to perform well but does not model the true background perfectly. Moreover, the training set size of CWoLa Hunting is limited by the number of events in the SSB (here approximately 65,000), whereas CATHODE could sample arbitrarily many background events (here chosen to be 200,000).

The comparison between CATHODE and the IAD shows that their ROC and SIC curves are overlapping almost everywhere. The learned background template seems to model the true

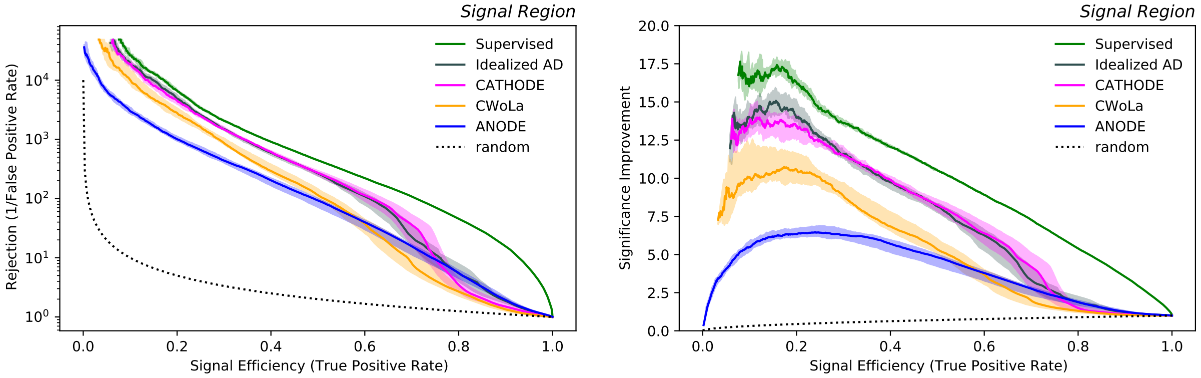

Figure 6.7: Comparison of background rejection (left) and significance improvement (right) of different classification methods as a function of the signal efficiency. The solid lines represent the median of ten trainings with different random weight initialization on the same training, validation and test set. The uncertainty bands contain the central 68% of these runs around the median. Figure taken from Ref. [1].

background very well, and thus CATHODE approximates the optimal sensitivity that can be expected from resonant anomaly detection on this dataset.

Finally, the IAD and CATHODE are outperformed by the fully supervised classifier, which shows that a dedicated search will still outperform resonant anomaly detection in this case of limited training data. It reaches a maximum SIC value of 17.5, which is only 25% higher than CATHODE.

Lower Signal Strengths

While Fig. 6.7 demonstrates excellent signal sensitivity of CATHODE in the LHCO R&D dataset with a signal-to-background fraction of 0.6% in the SR of the training data, a full test of the method should involve a performance estimate in the presence of smaller signal contributions. An ideal anomaly detection approach would detect arbitrarily few signal events in the data. Such a scan is performed in Fig. 6.8 for CATHODE and all benchmarks methods, i.e., the signal fraction in the training and validation set is decreased in steps. Here the median and variance cover the aforementioned reshuffling of train/validation/test splits. Otherwise, the SIC curves at low signal injections would depend on which signal events are randomly chosen for the training set.

The left plot of Fig. 6.8 shows the maximum SIC at decreasing signal amounts. CATHODE follows the IAD over the whole range and therefore retains its near optimality as a weakly supervised anomaly detection method. Above a signal-to-background ratio of 0.25%, it outperforms ANODE and CWoLa Hunting. The right plot shows the achieved maximum significance, i.e., the product of the significance without any anomaly score–based selection and the maximum SIC of each method (the left-hand side of Fig. 6.8). It is visible that in the region below the signal-to-background ratio of 0.25%, no anomaly detection method achieves a 3σ significance with their maximum SIC, including the IAD. This means that CATHODE retains optimalilty in the region where a non-trivial significance is achievable. The same plot also shows that at a signal-to-background ratio of 0.3%, CATHODE still has the potential to improve an initial significance of 1.02 to a discovery significance of 5σ.

The fully supervised classifier maximum SIC is independent of the signal fraction, as it is trained on an independent simulation dataset. The maximum achieved significance thus

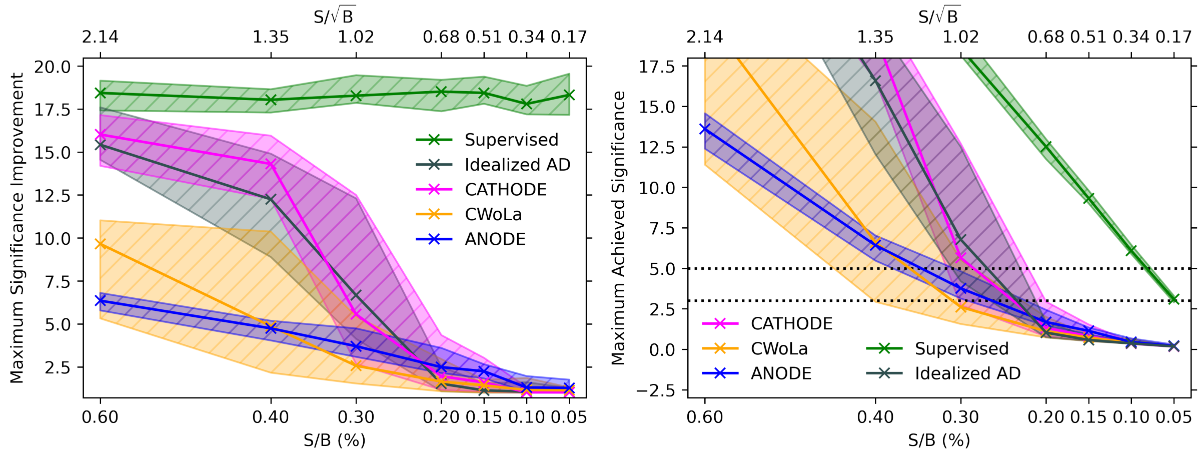

Figure 6.8: Left: Maximum significance improvement of each classification method as a function of the signal-to-background ratio. The solid lines represent the median of ten trainings with both different random weight initialization and different training/validation/test splits. The hatched bands represent the central 68% of these runs around the median, thus quantifying the variance from retrainings as well as from different realizations of the (increasingly few) injected signal events. Right: The achieved maximum significance, computed by multiplying the initial significance without any anomaly score–based selection with the maximum significance improvement of each method. The dotted lines correspond to significance values of 3 and 5σ. The initial significance is displayed on the upper horizontal axis of both plots. Figure taken from Ref. [1].

decreases linearly with the initial significance, and still manages to reach a 3σ significance down to a signal-to-background ratio of 0.05%. This illustrates the inherent gap between resonant anomaly detection and a dedicated search at limited classifier training capabilities.

Stability Under Correlations

The main shortcoming of CWoLa Hunting is its dependence on x and m being approximately uncorrelated over SSB and SR. In the default feature choice of the LHCO R&D dataset, this is the case as can be seen both by the small degree of linear correlations in Fig. 6.1 and by the high performance of the CWoLa Hunting method in Fig. 6.7. In order to test a highly correlated setting, Ref. [124] proposed to artificially increase correlation between m_{jj} and auxiliary features by adding 10% of the m_{jj} to both the lower jet mass m_{j1} and the jet mass difference Δm_j in every event:

$$\begin{aligned} m_{j1} &\rightarrow m_{j1} + 0.1 \times m_{jj} \\ \Delta m_j &\rightarrow \Delta m_j + 0.1 \times m_{jj} \end{aligned} \tag{6.1}$$

This shifts the lower SSB mass variables on average by 0.32 TeV, the SR by 0.35 TeV, and the upper SSB by 0.38 TeV. This dominates the m_{j1} and Δm_j distributions, and the CWoLa Hunting classifier learns obvious differences between SSB and SR instead of gaining sensitivity to the signal presence.

Figure 6.9 (left) demonstrates this sensitivity loss in terms of SIC curves, trained and evaluated on this shifted dataset. As can be seen, the discrimination power of CWoLa Hunting reduces to a mere random classifier. The sensitivity of ANODE also decreases due to the mass

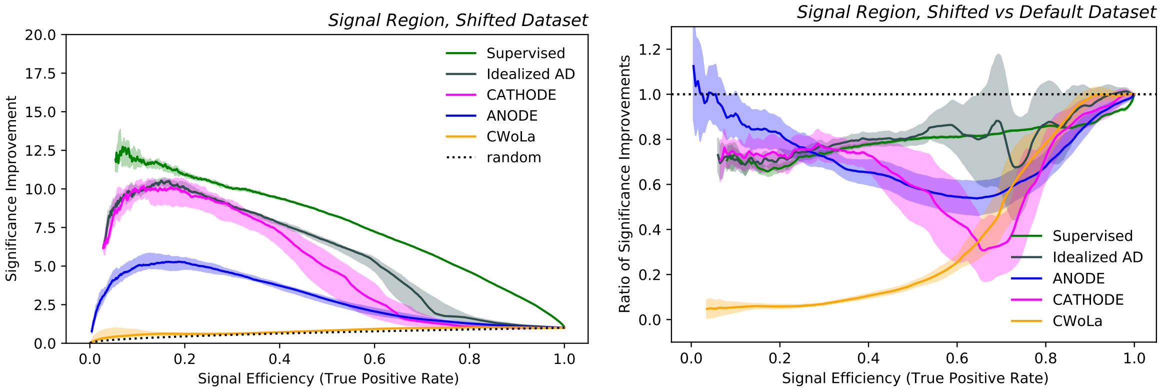

Figure 6.9: Left: Significance improvement of different classification methods as a function of the signal efficiency on the mass-shifted dataset. The solid lines and uncertainty bands are defined analogously to Fig. 6.7. Right: The significance improvement on the mass-shifted dataset divided by the default dataset values in Fig. 6.7. Figure taken from Ref. [1].

shift, which can be seen from Fig. 6.9 (right). It shows the reduction in SIC from the shift as a function of the signal efficiency. This is likely caused by the smearing of the originally well-peaking signal in m_{j1} and Δm_j by the more dominant m_{jj} shift. Moreover, the task of training and interpolating a conditional normalizing flow might be more difficult in this smeared setting. Even the performance of the supervised classifier and the IAD deteriorate from this smearing of the signal, both by the same amount. This can be expected because the smearing is caused by an independent random variable that is not known to the classifier. The behavior when this variable is known, is discussed in the next paragraph.

CATHODE is conceptually a mixture between ANODE and the IAD, as it involves the learning of the conditional flow in addition to the classification task. This also seems reflected in the decline of performance due to the shift in Fig. 6.9. It remains comparable to the IAD at lower signal efficiency, but in the medium higher signal efficiency range, where the ANODE performance decreases more substantially, CATHODE also becomes worse than the IAD. This study demonstrates that the explicit construction of the m-dependent background template renders CATHODE robust against correlations between x and m, in contrast to CWoLa Hunting.

Correlated vs Noisy Features

The above study of smearing features with m_{jj} has the shortcoming that the smearing by an independent random variable itself reduces the signal sensitivity of CATHODE and all benchmarks. It is thus convolved with the behavior under correlations between x and m. A different approach to investigate the robustness with respect to correlations is simply to include m_{jj} as an auxiliary feature for training the classifier, i.e., m becomes also part of x and thus this dimension of x is trivially maximally correlated with m. The result is shown in Fig. 6.10 for the fully supervised classifier, the IAD and CATHODE, each once on the default dataset and once with the mass shift applied.

The supervised classifier performs slightly better in Fig. 6.10 when m_{jj} is included among the four default features, as it is able to use the extra information. With the mass smearing applied, the inclusion of m_{jj} as an input feature fully recovers the performance that was previously lost on the smeared dataset. This shows that a neural network classifier with all information available is able to learn to reverse the smearing transformation.

The IAD in Fig. 6.10, on the other hand, performs worse on the default dataset when m_{jj} is

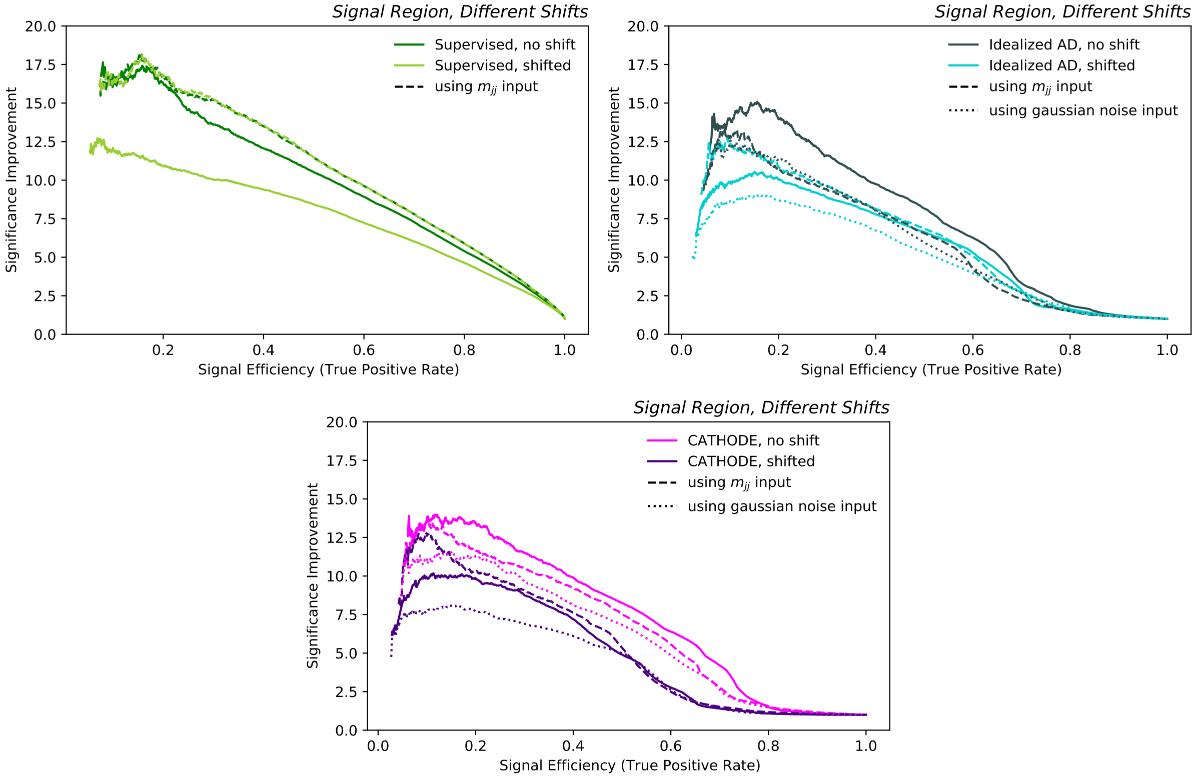

Figure 6.10: Significance improvement of different classification methods as a function of the signal efficiency: fully supervised (upper left), idealized anomaly detector (upper right), and CATHODE (lower). All lines represent the median of ten trainings with different random weight initialization on the same training, validation and test set, while the uncertainty bands are omitted for clarity. Two datasets are compared: the LHCO R&D dataset in the default and shifted setting. Each are trained with only the four default features (solid line), with m_{jj} as a fifth input feature (dashed line), and with random Gaussian noise as a fifth input feature (dotted line). Figure taken from Ref. [1].

used for classification. This seems to contradict the expected behavior from the fully supervised case, which was able to benefit from the small discrimination power of m_{jj} within the SR. To understand why this additional feature degrades the performance, the plot includes another configuration where instead of m_{jj} a random Gaussian noise feature is included for the training, i.e., a feature where signal and background follow the same distribution. The resulting SIC curve aligns with the one from the m_{jj}-included training, which indicates that the reason for the performance decline is not the physical property of the conditional feature m, but rather that a weakly supervised classifier loses signal sensitivity if more input features are included in the training without contributing substantial discrimination between signal and background. While the signal is strongly peaking in m_{jj} when considering the entire spectrum, the difference between signal and background is relatively small within the thin SR window around the signal peak. This issue of performance degradation due to uninformative features will be discussed more thoroughly in Sec. 6.4. On the shifted dataset, the IAD also yields strictly worse performance with a pure noise feature added. With the physical m_{jj} input feature, which is not uncorrelated noise now, the SIC is similar to without. In fact, it overlaps mostly with the SIC of the IAD of the unsmeared case with m_{jj} added. Thus, it seems that also the IAD learns to revert the smearing transformation when given the full necessary information, but is still impacted by the higher total number of input features.

Finally, CATHODE exhibits qualitatively similar behavior in Fig. 6.10 as the IAD. Just as the mass shifting study was convolved with the effect of signal smearing, this study cannot be entirely disentangled from the effect of uninformative features in weak supervision. But in combination, it confirms that CATHODE, just like an IAD, is robust against correlations between x and m.

Background Template Size

One benefit of CATHODE over CWoLa Hunting is that arbitrarily many SR background template samples can be generated once the normalizing flow has been trained on the SB. Figure 6.11 explores the signal sensitivity of CATHODE with different choices of background sample size for the classifier training. Aside from the default choice of 200,000 background samples in the training set, the choices of 60,000 and 800,000 events are also explored. The former is approximately as many events as there are SR proxy data events. The left plot, which is deduced on the default LHCO R&D dataset, shows that this small background sample with 60,000 events results in a strictly lower, although still relatively high, SIC curve. Moreover, it still outperforms CWoLa Hunting, which has slightly more background training samples (65,000 events). This demonstrates that the explicit modeling of the background template is a measurable factor why CATHODE outperforms CWoLa Hunting on this dataset. The background sample size of 800,000 events does not achieve higher performance than with 200,000 events. Because of this observed saturation, 200,000 has been chosen as the default on this dataset.

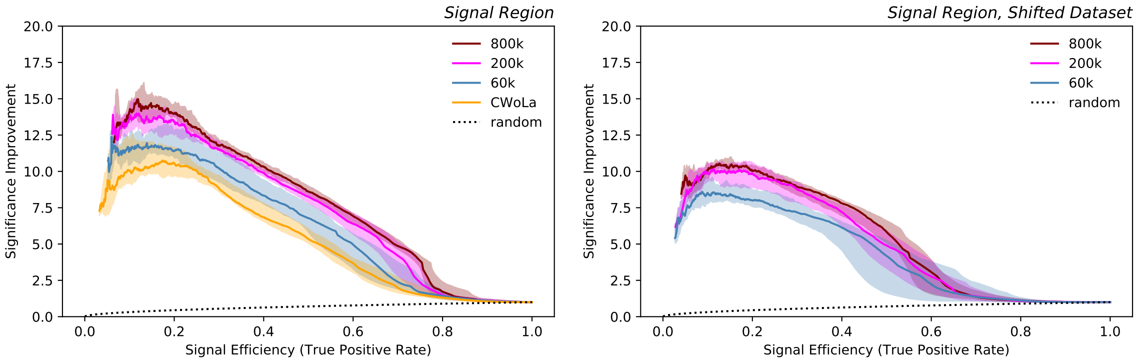

Figure 6.11: Comparison of CATHODE classifiers trained with different numbers of sampled background events in the training set, whereas the training proxy data are fixed at approximately 60,000 events. The solid lines and uncertainty bands are defined analogously to Fig. 6.7. The left plot shows the performance on the default dataset. This includes the CWoLa Hunting benchmark, which uses approximately 65,000 background events sampled from the SSB for training. The right plot shows the performance on the mass-shifted dataset, without CWoLa Hunting as it has no sensitivity due to the correlation between m_{jj} and the auxiliary features. Figure taken from Ref. [1].

For completeness, the right plot of Fig. 6.11 shows the performance in the mass-shifted setting, which mostly confirms the conclusions from the default dataset. In addition, it shows more clearly that a larger background sample also decreases the performance variance between independent model trainings.

In summary, Fig. 6.11 demonstrates that the adjustable background template sample size of CATHODE contributes an additional handle to increase the signal sensitivity and reduces its variance until a saturation point is reached. The improvement beyond the actual number of events in the proxy data might be in contrast to an intuition that the number of statistically independent data points is limited to what is available during training. A counterpoint to this intuition is that the smooth inductive bias of neural networks allows them to generalize, given that the underlying distribution is smooth, such that they can generate more data points than available during training. This is discussed more generally in Ref. [161]. In fact, the smooth inductive bias is already responsible for the capability of sampling events from the SR, which is unseen during training.

Background Estimation

The results in this section have so far focused on the signal sensitivity based on ROC and SIC curves. While these are useful metrics to measure the discrimination power between the background and a target signal based on respective labels, they do not yet directly relate to the ultimate sensitivity of the method on real observed data. For this, a statistical analysis, as described in Sec. 2.6, must be applied to quantify a p-value of the background-only hypothesis in favor of a signal+background hypothesis and without relying on unknown information, such as signal labels. A large class of these methods relies on selecting the most anomalous events, according to the anomaly score, as signal candidates and then estimating to what fraction they comprise background events. The bump hunt, a data-driven method that is more thoroughly discussed in Sec. 2.6.3, is a common example for such a background estimation technique. The underlying assumption of the bump hunt is that the background distribution in m is smooth and that any bump-like structure in the data is due to a signal. Thus, it is crucial that a selection on the anomaly score does not artificially introduce bumps into the background distribution.

The smoothness after applying a CATHODE-based selection is tested in Fig. 6.12 (left) by training on the same default LHCO R&D data configuration as previously, but without any signal (and thus expecting no bump-like structure in m) present in the proxy data. The plot shows the background shape before any selection and with the most anomalous 20% and 5% of events selected. The classifier has been evaluated not only in the SR, but also in the adjacent SB region. This is possible because the classifier only depends on inputs x, and not on m. The plot demonstrates that the background shape is not distorted by the selection on the anomaly score and thus a bump hunt is not expected to be biased towards yielding an excess without the presence of a signal. A more thorough study on background sculpting and the limitations of evaluating the classifier beyond the SR is provided in Sec. 6.3.

An alternative background estimation technique is to directly utilize the background template learned from the SB. Reference [124] proposed this in the context of ANODE in their Figure 8. After a selection on the anomaly score, one counts the fraction of background template events passing the decision threshold and then identifies the same fraction as background in the SR proxy data. The m_{jj} range is chosen such that the initial background is both smoothly falling and not subject to high statistical uncertainty.

The same study is shown in Fig. 6.12 (right) for the background-only trained CATHODE, where for every possible decision threshold the number of passing background template events is divided by the number of passing proxy data background events. Because the background template sample size has been scaled to match the number of proxy data background events, a value of one corresponds to an unbiased estimate of the true SR background. For the most part, the obtained ratio is compatible with unity within the statistical uncertainty. ANODE, on the other hand, has been demonstrated in Ref. [124] to suffer from a systematic shift towards values below one, i.e., the ANODE likelihood ratio is biased towards finding background events

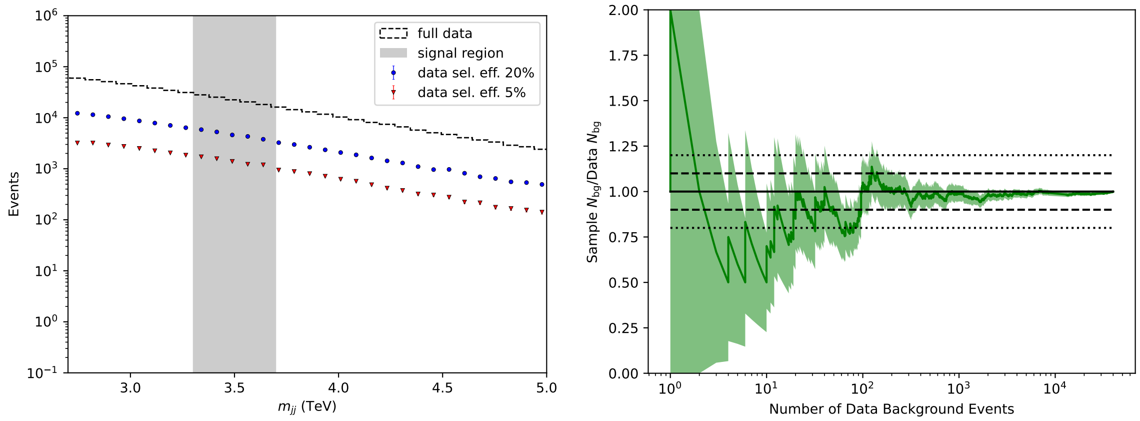

Figure 6.12: Tests of the background estimation in CATHODE using a training without signal events. The left plot shows the m_{jj} distributions of proxy data background after selecting the 20% and 5% most anomalous events according to the classifier, respectively. The right plot shows the ratio of sampled background events and proxy data background events passing the same selection on the anomaly score, as a function of the respective number of background events. The uncertainty band corresponds to the statistical uncertainty of the background samples, propagated to the ratio. Figure taken from Ref. [1].

more signal like. The classifier-based likelihood ratio of CATHODE thus seems to reflect the true likelihood ratio more accurately than the flow-based ratio of ANODE, and the direct background estimation via the SB-trained background template is a promising avenue for more studies.

6.2.4 Conclusions

CATHODE is a weakly supervised anomaly detection method that arises as a hybrid between ANODE and CWoLa Hunting. The classifier-based learning of the likelihood ratio, similar to CWoLa Hunting, results in a more accurate estimate than the ratio of two separately trained conditional density estimators. The SB-trained conditional density estimator, as is trained in ANODE, achieves a more faithful background template than the direct SB sampling of CWoLa Hunting.

CATHODE was shown to outperform the two predecessor methods in various aspects on the LHCO R&D dataset. The signal sensitivity in terms of SIC curves is the highest, and it approximates the maximally expected sensitivity of an idealized anomaly detector with a perfect background template. This holds also at lower signal fractions. A key advantage over CWoLa Hunting is that its signal sensitivity is robust against correlations between the auxiliary features and the mass variable. Lifting this former limitation on the choice of auxiliary features is a significant step towards the inclusion of more information and thus increases the sensitivity for a broader range of potential signal processes in the data. The adjustable classifier training sample size provides another advantage, as it measurably increases the signal sensitivity and decreases the variance from random weight initialization. The proof-of-concept study of the background estimation potential shows satisfactory compatibility with a bump hunt, and it opens the door for the method-intrinsic background estimation via the SB-trained background template.

This work has sparked further questions. A more in-depth discussion of the background estimation potential of CATHODE is provided in Sec. 6.3. In particular, it is shown that the results of the background estimation test in Fig. 6.12 will change significantly when considering correlated input features. The reasons behind the loss of signal sensitivity due to the presence of uninformative features is investigated more thoroughly in Sec. 6.4, also outlining potential solutions. Moreover, this proof-of-concept study lacks the full statistical analysis procedure to quote the true signal sensitivity in terms of a p-value. Section 7 will provide a complete CATHODE-driven analysis on more realistic MC simulation as well as real detector data.

More issues are to be addressed in future work. For example, the conditional density estimation based on normalizing flows implicitly relies on auxiliary feature distributions that can be learned by smoothly deforming a simple prior distribution. If some auxiliary features exhibit highly discrete distributions, this modeling might not be successful. Moreover, it needs to be shown under which circumstances the inductive bias of the conditional density estimator can guarantee an accurate interpolation from the SB into the SR that has been blinded during training. Related to this is also the question of how this interpolation works if a large fraction of the signal is outside the SR. This could be the case if the signal manifests itself as a resonance peak that is wider than the SR. Ideally it would *not* correctly interpolate the signal contribution in x, but only the background.

6.3 LaCATHODE

The work in this section is based on the publication in Ref. [2], developed in cooperation with Anna Hallin, Gregor Kasieczka, Tobias Quadfasel, and David Shih. Accordingly, the results and figures presented here are the same as in the publication. My contribution to the work consists of significant parts of the conceptual development of the algorithm, as well as the full implementation of both the algorithm and the benchmark methods. I have also performed all studies involving the LHCO R&D data, developed the metrics for testing background sculpting, contributed to writing the article, and handled its submission to the preprint server and journal, along with the review process.

As discussed in Sec. 6.2, CATHODE achieves near-optimal signal sensitivity in the test case of the LHCO R&D dataset. This could also be seen in the case where the auxiliary features x and the resonant feature m are highly correlated. Moreover, the background estimation study showed that a selection on the CATHODE anomaly score does not result in artificial bump-like structures in the passing background distribution. Such so-called background sculpting would potentially lead to a false discovery of a signal when using background estimation techniques, such as a bump hunt (Sec. 2.6.3), if it is not properly reflected in the uncertainty estimation, which can be difficult to model. However, the studies in Sec. 6.2.3 did not combine the background estimation study with the correlated feature scenario. It was only discovered later that strong correlations between x and m actually introduce undesired background sculpting in resonant anomaly detection methods, such as CATHODE. This is not related to the data-driven learning of the background template—in fact it occurs even in the case of an idealized anomaly detector.

The reason for this background sculpting in the presence of correlated features is twofold. First, if the features x are a (noisy) function of m, then applying a selection on these $x(m)$, and thus on any function $f(x(m))$, corresponds to a selection on m. Secondly, if the phase space of x varies significantly from one region of m to another, the ML classifier trained in the SR cannot be trivially extrapolated to the SB regions anymore, as was done in Fig. 6.12 (left). Instead, it will rely on the inductive bias and random initialization of the classifier model when evaluated on x values outside the training domain, and is subject to uncontrolled behavior from this extrapolation.

This section studies the background sculpting due to correlations between x and m using a toy experiment as well as the LHCO R&D dataset. Furthermore, we will propose a modification to CATHODE in which the classification task is performed in the latent space of the conditional normalizing flow that has been trained on the SB data. This modification, called *latent CATHODE* (LaCATHODE), benefits from the decorrelating property of the conditional invertible mapping of data from $p_{\mathrm{bkg}}(x|m)$ to the simple prior $p_z(z)$ where the m dependence has been removed. Applying a selection based on these latent space values (and thus a classifier function with input z) thus does not translate to a selection on m. The prior distribution is the same in SR and SB, thus a classifier trained in one region can be directly applied to the other region.

6.3.1 Algorithm and Implementation

LaCATHODE can be seen as a modification to the CATHODE algorithm in the specific case when a normalizing flow was used for learning the conditional background density $p_{\mathrm{bkg}}(x|m)$. In order to underline the differences to LaCATHODE, the CATHODE algorithm with a normalizing flow is again quickly summarized as follows:

1. The conditional background density $p_{\mathrm{bkg}}(x|m)$ is learned by training a normalizing flow h on the SB data to map the auxiliary features x to the latent space $z = h(x;m)$ following a simple prior $p_z(z)$ distribution, continuously for every corresponding value in m: $p_{\mathrm{bkg}}(x|m) = p_z(h(x;m)) \left|\det\left(\frac{\partial h}{\partial x}\right)\right|$. The common choice of $p_z(z)$ is a standard Gaussian distribution $\mathcal{N}(0,1)^D$ with the data dimensionality D.

2. The background density of m in the SR, $p_{m,\mathrm{SR}}(m)$, is estimated, e.g., with a one-dimensional KDE fitted to the observed SR m values.

3. A background-like set of x values is created by first independently sampling z and m values, then passing them backwards through the normalizing flow: $\mathcal{D}^x_{\mathrm{bkg,SR}} = \{h^{-1}(z_i;m_i)\}$ where $z_i \sim p_z$ and $m_i \sim p_{m,\mathrm{SR}}$.

4. A binary ML classifier is trained to distinguish the real data $\mathcal{D}^x_{\mathrm{data,SR}}$, as represented in the feature space x, from the background sample $\mathcal{D}^x_{\mathrm{bkg,SR}}$.

The LaCATHODE algorithm, which is illustrated pictorially in Fig. 6.13, follows step 1 above, but then continues as follows:

2. The signal region data $\mathcal{D}^x_{\mathrm{data,SR}}$ are mapped to the latent space z using the conditional normalizing flow from step 1: $\mathcal{D}^z_{\mathrm{data,SR}} = \{h(x_i;m_i)\}$.

3. The background-like set of z values is created by sampling from the simple prior distribution: $\mathcal{D}^z_{\mathrm{bkg,SR}} = \{z_i \sim p_z\}$.

4. A binary ML classifier is trained to distinguish the latent data $\mathcal{D}^z_{\mathrm{data,SR}}$ from the background sample $\mathcal{D}^z_{\mathrm{bkg,SR}}$.

Therefore, the basic principle is the same for both methods, but the LaCATHODE classification task is performed in the latent space of the normalizing flow instead of the physical feature space. The motivation behind this transformation is that the background in the latent space has been trained to follow a simple prior distribution where the dependence on m has been removed. This means that the conditional flow mapping effectively decorrelates $z = h(x;m)$ from m because it learns to map all background data in the SB to the same simple prior distribution, regardless of the value of m. A selection on the classifier score based on latent input values z should thus not translate to a change in background shape in m. This also eliminates the need for an estimation of $p_{m,\mathrm{SR}}(m)$ in Step 2 of CATHODE above, because the real m values are available when mapping data to the latent space along the forward direction of h, and the latent background template is sampled without any m dependence.

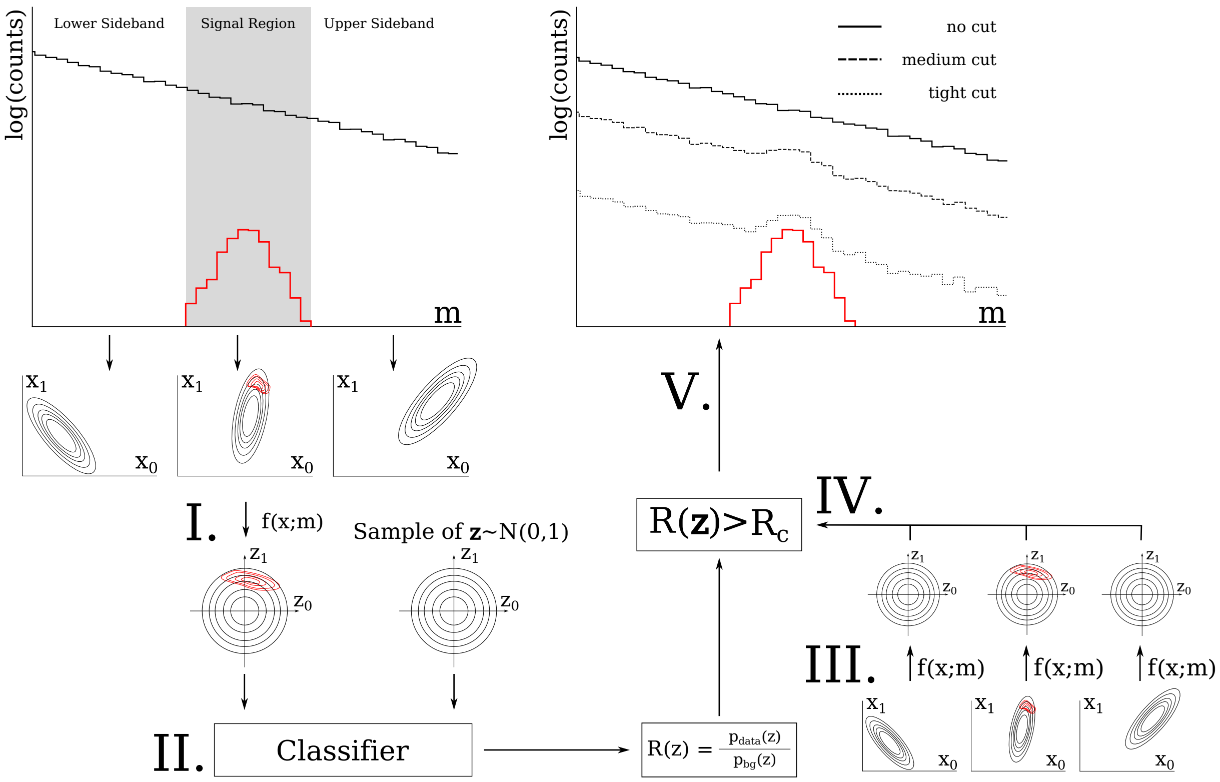

Figure 6.13: Illustration of the LaCATHODE algorithm. The signal region data are mapped to the latent space representation with the conditional normalizing flow $f(x; m)$, where the background is expected to be distributed according to the prior of the flow, usually a standard Gaussian (I). A binary ML classifier is trained to distinguish the latent data from a pure standard Gaussian sample (II). Once trained, the test data from both SR and SB are mapped to the latent space where the background should in each result in a standard Gaussian distribution (III). These values are passed through the classifier to obtain the anomaly score (IV). The shape of the background in the conditional feature m will remain unchanged if a selection is applied based on this anomaly score (V). Figure taken from Ref. [2].

The next paragraphs will specify the implementation details of the LaCATHODE proof-of-concept study. The code to reproduce the results is available in a public repository [162]. While Sec. 6.3.2 will illustrate the working principle of LaCATHODE with a simple toy experiment, the main results of Sec. 6.3.4 will again rely on the LHCO R&D dataset.

The default dataset is defined via the same high-level feature representation as described in Sec. 6.2.1, i.e., the four auxiliary features are $x = (m_{j1}, \Delta m_j, \tau_{21,1}, \tau_{21,2})$ and the resonant feature is $m = m_{jj}$. The SR/SB choices and signal-to-background ratio are analogous as well. The main dataset-related difference to the studies in Sec. 6.2 is a simpler separation into training/validation/test datasets. Here, half of the full proxy data are selected for training. The remaining half is split into a validation set (1/6) and test set (1/3). The resulting SR and SB datasets thus follow the same split.

In order to test the robustness with respect to correlations, two additional datasets featuring stronger correlations are defined. The first of these datasets comprises the same m_{jj}-dependent shift applied to the two jet mass features as in Eq. 6.1. As discussed in Sec. 6.2, this introduces a dominant correlation with x but this random smearing also decreases the discrimination power of the features. The second of these datasets is instead constructed by including the angular separation $\Delta R_j = \sqrt{(\Delta\eta_j)^2 + (\Delta\phi_j)^2}$ between the two leading jets as a fifth auxiliary feature. This feature is known to be strongly correlated with m_{jj}, as can be seen in Fig. 6.14. Adding a correlated feature rather than smearing existing ones is similar to the study in Sec. 6.2.3 where m_{jj} has been included among the auxiliary features, but with a realistic correlation between random distributions instead of a trivial equality in one dimension. This is also crucial when a conditional normalizing flow is trained on the dataset, as including m_{jj} itself would reintroduce the exact same feature in the condition as in the modeled data: $p(x, m|m)$. This manner of testing the robustness with respect to correlations was first suggested in Ref. [126].

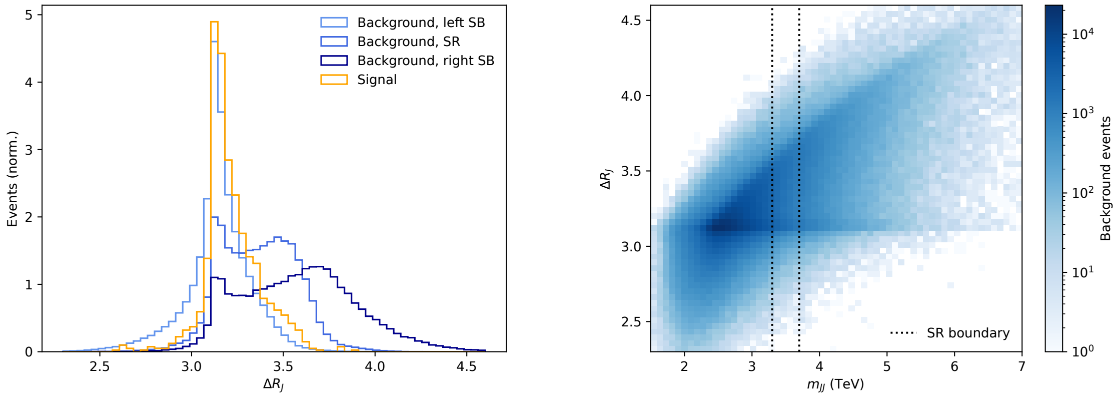

Figure 6.14: Left: normalized marginal distributions of the angular separation ΔR_j between the two leading jets in the LHCO R&D dataset, separated into left sideband, signal region, right sideband and signal. Right: two-dimensional background distribution of ΔR_j and m_{jj}. Both plots show that there are significant correlations between ΔR_j and m_{jj} in the dataset.

The normalizing flow implementation and training is fully analogous to Sec. 6.2.1. When using the flow for mapping SR data to the latent space, the model state at the epoch with the lowest validation loss is used. Hence, no ten-epoch ensembling scheme, as used for CATHODE in Sec. 6.2.1, is applied. While such a multi-epoch ensemble could be realized with the mean of the transformed values from each model state, this degree of freedom is not further explored in this study.

Once the SR training data are mapped to the latent space, the background template is generated by sampling values from the standard Gaussian prior, using the random number generator implemented in the NumPy library [163]. The size of this background template is chosen to be 267,000 events. By splitting them into training and validation sets with equal proportions as the proxy data, this results in 200,000 for training and 67,000 for validation. This matches the same oversampling factor of approximately 3.3 as in Sec. 6.2.1. Once the two sets $\mathcal{D}^z_{\text{data,SR}}$ and $\mathcal{D}^z_{\text{bkg,SR}}$ are created, the subsequent binary classifier training is performed in full analogy to Sec. 6.2.1.

6.3.2 Toy Experiment

In order to illustrate the relationship between x-m correlations and background sculpting in resonant anomaly detection methods, and how LaCATHODE mitigates this issue, a toy experiment is constructed with simple known distributions.

The setup consists of a one-dimensional feature m, which is sampled from a uniform distribution with a range of $[-10, 10]$. The signal region is defined as the narrow subset $m \in [-0.3, 0.3]$.

The two auxiliary features $x = (x_1, x_2)$ are governed by the conditional density $p_c(x|m) = \mathcal{N}(\mu = cm, \sigma = 1)^2$, where $c \in \mathbb{R}$ is a parameter controlling the correlation strength between x and m. For every considered choice of c, two independent but identically distributed samples of x and m are drawn: one representing the (background-only) proxy data and one representing the background template.

A binary ML classifier is trained to distinguish the SR events in the two samples based on their values in x. Each sample consists of one million events, which are divided into training, validation and test sets with fractions of 50%, 16.7% and 33.3%, respectively. The classifier is implemented as a fully connected neural network with three hidden layers of 64 nodes, with ReLU activation function, and a single output node with sigmoid activation function. It is implemented with the Keras library [164] and the TensorFlow backend [165]. The training is performed by minimizing the binary cross-entropy loss with the Adam optimizer [73], using a batch size of 128 and a learning rate of 10^{-3}. From the total of 50 training epochs, the five with the lowest validation loss are ensembled in the inference step by computing their mean prediction per event.

Once trained in the SR, the classifier is evaluated on the test set of the proxy data in the full region of m (SR and SB). Based on the resulting anomaly score, the 1% most anomalous events are selected. The magnitude of how much the resulting classifier sculpts the background shape in m can be visually inspected. Deviation from a uniform distribution indicates the presence of sculpting.

Figure 6.15 shows the toy experiment with the two choices of $c \in \{0.001, 0.1\}$. The left column displays the two-dimensional distribution of x, where only the larger choice of $c = 0.1$ results in a visible departure from a standard Gaussian distribution. The elongated shape towards the diagonal is due to the m-dependent shift in the x values and results in a domain shift between the classifier training region (SR) and the evaluation region (SR and SB). The right column shows the distribution of m for the selected most anomalous percentile of events, based on three independent classifier trainings. The $c = 0.001$ case is compatible with a uniform distribution of m for every classifier training, indicating no sculpting due to this small degree of correlation. In contrast, the $c = 0.1$ case is subject to a significant change in shape for all three classifiers. Moreover, there is a visible disagreement between the shapes resulting from the three independent classifiers. This illustrates non-deterministic behavior once the neural networks are extrapolated into phase space that is not accessed during training.

While the above test corresponds to a simplified CATHODE-like scenario (or more specifically an IAD), one can now also imagine a simplified LaCATHODE application. For this, one would use a normalizing flow that maps the conditional background likelihood $p(x|m; c)$ to a simple prior distribution $p_z(z)$, where the m dependence has been removed. In this simple setup, a perfect normalizing flow, mapping x to the latent space following a two-dimensional standard Gaussian distribution, is analytically obtained as

$$z = x - cm. \tag{6.2}$$

The resulting z distribution and the corresponding m distribution, originally sampled uniformly, for the selected 1% most anomalous events are shown in Fig. 6.16. The z distribution coincides in this toy experiment with $x \sim p(x|m; c)$ for $c = 0$. The right plot shows that a selection on this uncorrelated z does not at all result in a change in background m shape, as expected.

This simple toy experiment shows that correlations between x and m cause a classifier, trained on values in x, to learn an implicit function in m and thus a selection on this classifier changes the shape of the passing m distribution. This effect is negligible at sufficiently small correlations but becomes dominant with more significant correlations. Due to the extrapolation

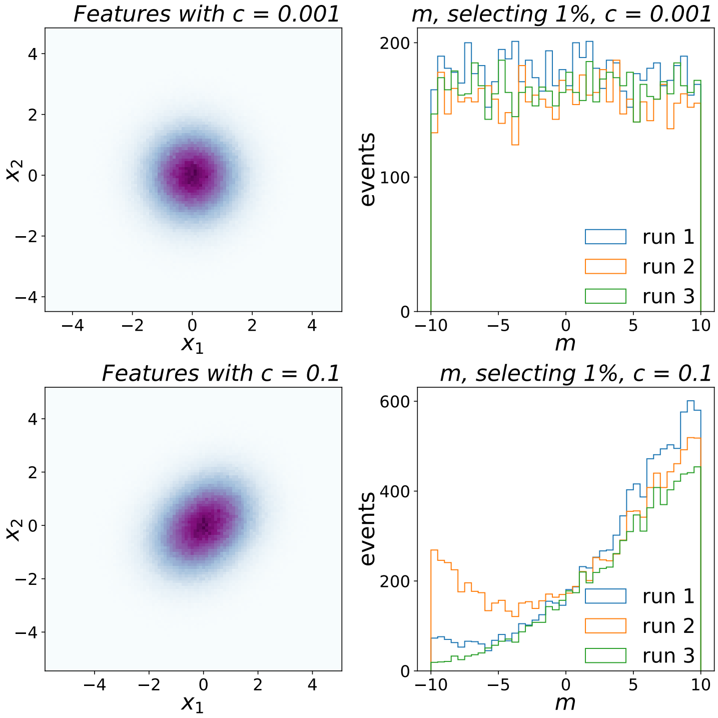

Figure 6.15: Two-dimensional toy experiment auxiliary feature x distribution (left column) and the resulting distribution in the conditional feature m, originally sampled uniformly, when a binary classifier is trained to distinguish equivalent samples in x and selecting the 1% data points with the highest anomaly score (right column). This scenario is illustrated with an x-to-m correlation parameter of $c = 0.001$ (top row) and $c = 0.1$ (bottom row). Three independent classifiers were trained for each case. Figure taken from Ref. [2].

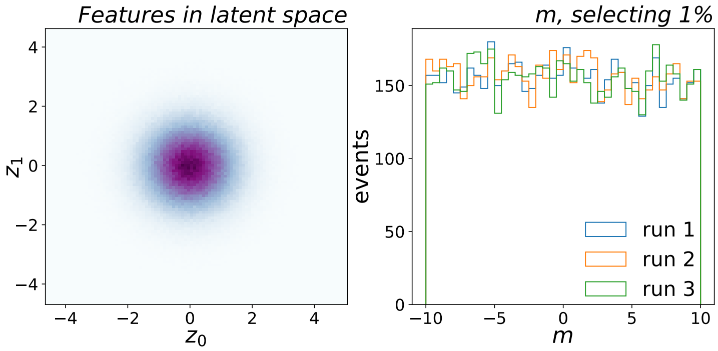

Figure 6.16: Two-dimensional toy experiment latent space z distribution of a perfect normalizing flow (left) and the resulting distribution in the conditional feature m, originally sampled uniformly, when a binary classifier is trained to distinguish equivalent samples in z and selecting the 1% data points with the highest anomaly score (right). Three independent classifiers were trained for each case. Figure taken from Ref. [2].

from an SR-only training to a broader domain in m, the behavior of a neural network classifier on correlated x depends increasingly on the specific initialization of the weights, thus leading to more uncontrolled behavior between repeated trainings. Furthermore, it shows how decorrelating x and m before classifier training mitigates the sculpting issue. It thus illustrates the motivation behind modifying CATHODE to LaCATHODE, where the latent space of a normalizing flow is used for the classification task.

6.3.3 Benchmark Method Implementation

Section 6.3.4 will compare the LaCATHODE algorithm to CATHODE, as they are directly related. In addition, the results are compared to an IAD, which corresponds to CATHODE with a perfect background template. Comparing to both thus helps to disentangle the behavior due to mismodeling the background template from the more general challenges arising from feature space classification.

The CATHODE implementation closely follows the one in Sec. 6.2.1 with only a few departures. The data split and the sizes of the background templates used for the training and validation sets are the ones described earlier in Sec. 6.3.1. In addition, the ten-epoch ensembling of the normalizing flow is omitted, in order to compare more directly to the LaCATHODE case where this is also not performed.

The IAD implementation is also analogous to Sec. 6.2.2 except for the data split. The proxy data are the same as for (La)CATHODE, and the background template is created by selecting 200,000 events from the additional LHCO R&D background samples for the training set and 67,000 for the validation set. Thus, this idealized background template now has the exact same size as the one used for (La)CATHODE.

Table 6.2 summarizes the amount and type of data used for training and validation of each method.

Table 6.2: Numbers of events in units of 1000 ≡ 1k used for training, validation, and testing for each method. In this context, validation refers to selecting model epochs. The test set is the same for all methods. DE and CLS refer to density estimators and classifiers, respectively. "Bkg" and "sig" refer to background and signal, respectively. The term "bkg samples" here indicates that they originate from the synthetic background template.

Method	Type	Training	Validation	Testing
CATHODE	DE	439k SB data	147k SB data	For sculpting studies: 333k bkg (SB+SR) For SIC: 386k SR bkg 75k SR sig
	CLS	61k SR data 200k SR bkg samples	20k SR data 67k SR bkg samples	
LaCATHODE	DE	439k SB data	147k SB data	
	CLS	61k SR data 200k SR bkg samples	20k SR data 67k SR bkg samples	
Idealized AD	CLS	61k SR data 200k SR bkg	20k SR data 67k SR bkg	

6.3.4 Results

The performance of LaCATHODE, CATHODE and an IAD will be discussed in the next paragraphs in terms of both the signal sensitivity and robustness against background sculpting. In order to study the signal sensitivity, the same SR signal-to-background ratio of approximately

0.6% as in Sec. 6.2.3 is chosen in the proxy data. No signal events were present in the training data for studying the background sculpting because the main focus is to avoid reporting false discoveries in the absence of true signal. The same metrics were also derived with the signal-injected trainings, without leading to substantial differences.

Signal Sensitivity

The main performance metric for signal sensitivity is the SIC in the SR. To ensure relatively small statistical uncertainty in the SIC measurement, the test set is enhanced with events from the additional SR background events following the LHCO R&D dataset background simulation, as well as all SR signal events that are not used for training and validation. This thus comprises approximately 386,000 background events and 75,000 signal events. Again, the results are presented in terms of the median and variance of ten independent retrainings of the full method on the same dataset, as in Sec. 6.2.3.

Figure 6.17 shows the SIC as a function of the background efficiency for the three methods on the default dataset (top left), the mass-shifted dataset (top right), and the dataset with the additional angular separation feature ΔR_j (bottom). The background efficiency is chosen here to quantify the position of the decision threshold of the anomaly score because, compared to the signal efficiency, it relates more closely to the data selection efficiency that one can control in a real analysis. Since the signal content in an anomaly search is likely small, the data selection efficiency and the background efficiency are approximately equal, and the SIC curves in Fig. 6.17 allow an estimate of how tight a selection needs to be in a real analysis in order to reach a given signal sensitivity.

The SIC curves for CATHODE and the IAD confirm the behavior discussed in Sec. 6.2.3, where CATHODE saturates the IAD performance in the default case, and is more affected by the mass smearing. The bottom plot of Fig. 6.17 also indicates that the inclusion of the highly correlated ΔR_j feature leads to a greater mismodeling by the flow, and thus the CATHODE performance is still high but lower than for the IAD.

LaCATHODE, on the other hand, shows similar but slightly lower SIC with higher variance in the default case. In the mass-shifted case, LaCATHODE is similar to CATHODE with a larger downward variance. On the dataset with ΔR_j, LaCATHODE is both lower in median and with larger variance than CATHODE. One reason for this might be the less advantageous feature representation in the latent space for discriminating signal from background in a weakly supervised manner. Moreover, while the background is trained to be mapped to a standard Gaussian prior, the signal is not well controlled in the flow and might be mapped to more or less populated phase space regions. Hence, the larger variance. Besides the performance loss with respect to CATHODE, the SIC is still relatively high, in particular compared to other anomaly detection methods, such as CWoLa Hunting or ANODE.

Background Sculpting

While the study above shows that the signal sensitivity when using LaCATHODE persists to be high, although slightly lower than for CATHODE, the main focus of the method is to retain an unsculpted background distribution of m after a selection on the anomaly score is applied. Similar to the toy experiment in Sec. 6.3.2, this can be tested by selecting the most anomalous 1% events according to each classifier and plotting their m distribution. The selection efficiency of 1% is to some degree arbitrary, but has been chosen to be tight enough for non-trivial signal sensitivity (according to the SIC curves in Fig. 6.17), but loose enough for the resulting m distribution to not be dominated by statistical uncertainties. The result for a single training of

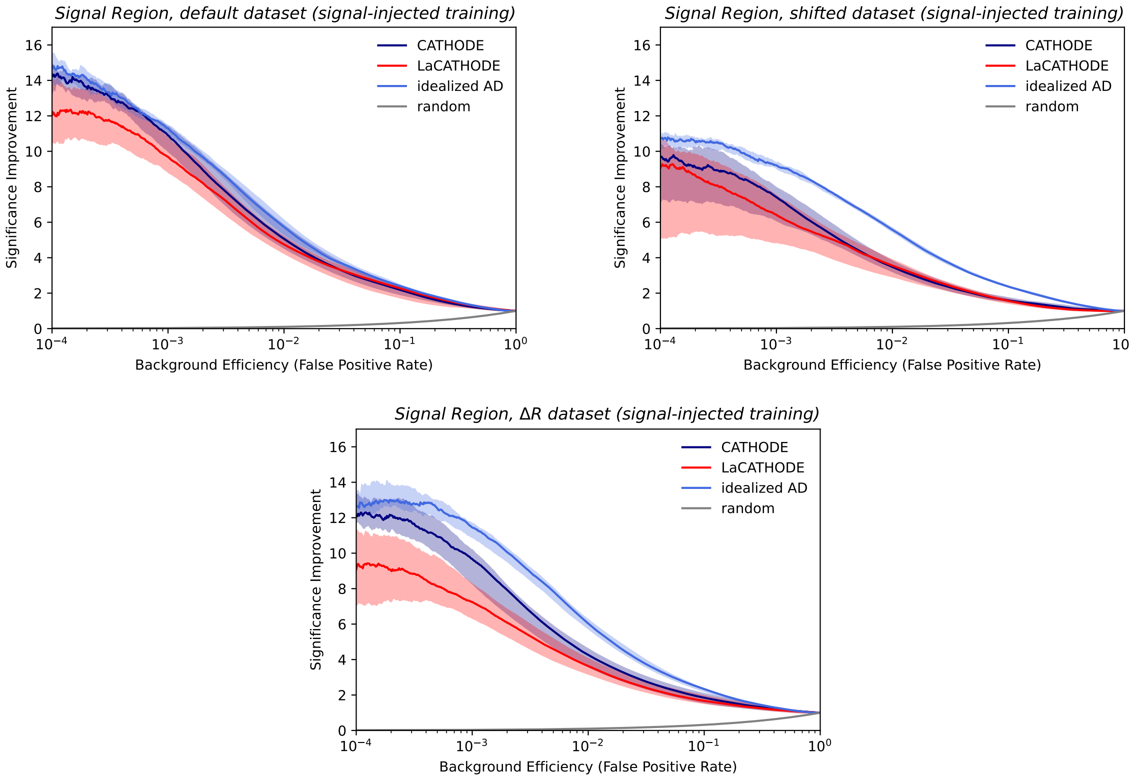

Figure 6.17: Significance improvement characteristic (SIC) as a function of the background efficiency for LaCATHODE in comparison to CATHODE and an idealized anomaly detector applied to the default LHCO R&D high-level feature dataset (top left), the mass-shifted representation (top right), and including the angular separation ΔR_j as an additional auxiliary feature (bottom). The signal-to-background ratio in the signal region is approximately 0.6%. The solid lines represent the median of ten trainings with different random weight initialization on the same training, validation and test set. The uncertainty bands contain the central 68% of these runs around the median. Figure taken from Ref. [2].

each of the three anomaly detectors on each of the three datasets is summarized in Fig. 6.18.

In the default case with only minor correlations, all three lines appear similar to the unselected background. Some noticeable deviation is present for the IAD, which is slightly shifted to the left between 1.6 and 2.8 TeV. The difference becomes more striking in the two strongly correlated datasets. There, both CATHODE and the IAD result in significantly different m distributions after selection, confirming the expectation from the toy experiment. The similarity between CATHODE and the IAD in terms of the magnitude of sculpting also confirms that the issue is not caused by the data-driven learning of the background template, but by the feature space classification itself. LaCATHODE, on the other hand, shows a very similar m distribution before and after selection in all three cases, confirming that the decorrelating mapping to the flow latent space is successful in mitigating the background sculpting issue.

While the test in Fig. 6.18 is illustrative of the background sculpting of feature space classifiers and the lack thereof in the latent space, they are based only on one training and a single selection efficiency choice. We thus want to construct a more quantitative measure of the change of background shape after a selection on the anomaly score. For this, a histogram of the test data m_{jj} distribution is computed with $n_{\text{bins}} = 300$ bins and boundaries such that all bins are equally

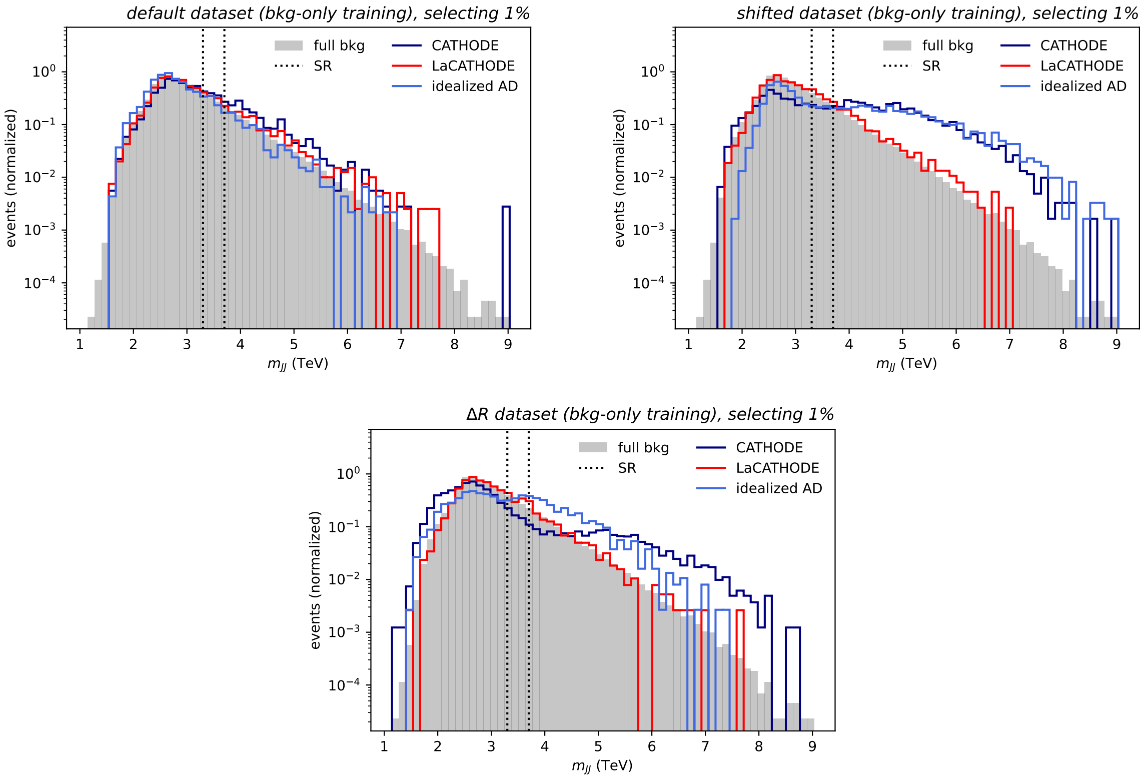

Figure 6.18: The normalized distribution of m_{jj} after selecting the 1% most anomalous events after a single training of LaCATHODE, CATHODE and an idealized anomaly detector applied to the default LHCO R&D high-level feature dataset (top left), the mass-shifted representation (top right), and including the angular separation ΔR_j as an additional auxiliary feature (bottom). Only background events were used in the training and evaluation. The gray area represents the full m_{jj} shape of the background before any anomaly score selection. Figure taken from Ref. [2].

populated: $N_i^{\text{full}} = N^{\text{tot}}/n_{\text{bins}}$, where N_i^{full} is the content of bin number i and $N^{\text{tot}} = 333{,}000$ is the total number of background events. The choice of n_{bins} was made such that a 1% selection efficiency would still yield at least ten events per bin. An area-normalized histogram with bin content $n_i^{\text{full}} = a_i N_i^{\text{full}}$ is then created, where a_i is the respective normalization factor. After training an anomaly detection method and applying a selection with efficiency ε_{sel}, a histogram of the passing m_{jj} values N_i^{sel} is constructed with the same binning as derived for the inclusive case, and separately area normalized $n_i^{\text{sel}} = b_i N_i^{\text{sel}}$ with respective normalization factor b_i. This then allows computing the χ^2 distance metric between the two normalized histograms:

$$\chi^2 = \sum_{i=1}^{n_{\text{bins}}} \frac{\left(n_i^{\text{full}} - n_i^{\text{sel}}\right)^2}{\sigma_i^2} = \sum_{i=1}^{n_{\text{bins}}} \frac{\left(n_i^{\text{full}} - n_i^{\text{sel}}\right)^2}{a_i^2 \varepsilon_{\text{sel}} N_i^{\text{full}}}. \tag{6.3}$$

It should be noted that this corresponds to the one-sided χ^2 test, where the full distribution scaled down by ε_{sel} is considered the true target distribution. The expected statistical uncertainty per bin σ_i is thus the product of the Poisson error $\sqrt{\varepsilon_{\text{sel}} N_i^{\text{full}}}$ and the normalization factor a_i. The advantage of this metric over, e.g., a two-sided χ^2 test or the Jensen-Shannon divergence, is that its value approximates the number of degrees of freedom $n_{\text{dof}} = n_{\text{bins}} - 1$ in the case of a perfect agreement between the two histograms. We thus divide the χ^2 value by

n_{dof} to obtain the χ^2/n_{dof} value with an expected value of one. This metric is plotted for each anomaly detection method as a function of decreasing $\varepsilon_{\mathrm{sel}}$, ranging from 20% down to 1%, in Fig. 6.19. As in the SIC curves, the solid lines and bands correspond to the median and central 68% of ten independent trainings, respectively. As a reference, the gray line and bands denote the median and variance of 100 randomly drawn subsets at each choice of $\varepsilon_{\mathrm{sel}}$, confirming that the χ^2/n_{dof} metric converges towards unity with our choice of n_{bins}.

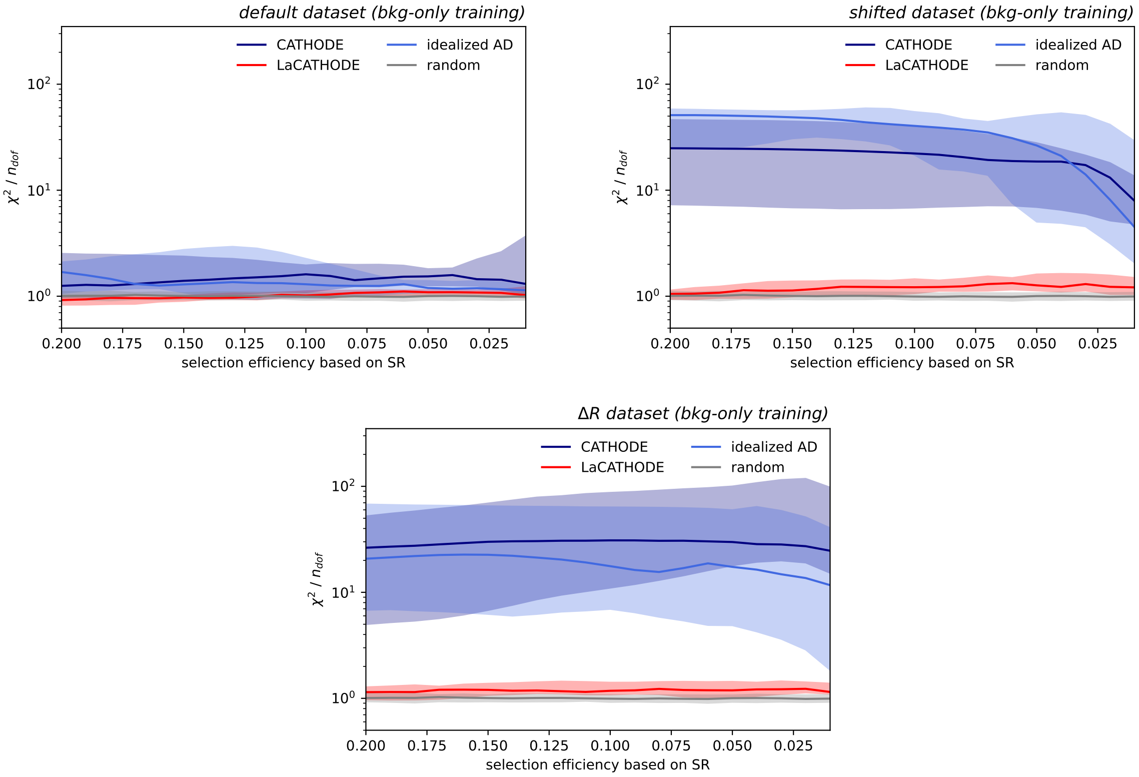

Figure 6.19: The χ^2/n_{dof} between the m_{jj} distribution after applying a selection on the anomaly score and the full m_{jj} shape of the proxy data before any anomaly score selection, as a function of the respective selection efficiency in the signal region. The result is shown for LaCATHODE, CATHODE and an idealized anomaly detector applied to the default LHCO R&D high-level feature dataset (top left), the mass-shifted representation (top right), and including the angular separation ΔR_j as an additional auxiliary feature (bottom). Only background events were used in the training and evaluation. The solid lines and bands are defined as in Fig. 6.17. The gray band denotes the χ^2/n_{dof} of 100 randomly drawn subsets at the respective efficiency. Figure taken from Ref. [2].

The resulting figure shows that the change in background shape is noticeable but remains relatively limited for CATHODE and the IAD with the default feature choice. However, in the two correlated datasets, their χ^2/n_{dof} values increase dramatically and are subject to large variance. Again, the similarity between CATHODE and the IAD confirm that this behavior is intrinsic to training a classifier on physical features with correlations to m. LaCATHODE, on the other hand, remains strikingly close to unity in all cases with only small variance. It thus confirms that the background shape remains largely unchanged after a selection on the LaCATHODE anomaly score, for all considered choices of $\varepsilon_{\mathrm{sel}}$ and across multiple retrainings, even with heavily correlated x and m.

6.3.5 Conclusions

LaCATHODE is a modification to the CATHODE algorithm that involves using the already trained normalizing flow for decorrelating the physical auxiliary features from the resonant feature m before training a binary ML classifier. With this modification, the classifier does not learn an implicit function of m and thus does not change the background shape in m once a selection based on the anomaly score is applied. It was explicitly shown that LaCATHODE results in an unchanged background shape even in a highly correlated setting while retaining much of the high signal sensitivity of CATHODE, without adding more complexity to the method. In fact, it even eliminates the need for estimating the SR background distribution of m.

An unchanged background shape is not strictly necessary to perform a bump hunt, as linear modifications to the shape without introducing bump-like (at least quadratic dependencies on m) structures would also not result in false discoveries. Moreover, if one manages to model the sculpting and/or the resulting systematic uncertainties, it could be accounted for during the statistical procedure. However, this type of recalibration is particularly challenging if the data-driven nature of the analysis should be preserved. The strictly unmodified background shape of LaCATHODE automatically facilitates a more robust statistical analysis, in particular given that its loss of signal sensitivity compared to CATHODE is relatively small.

The limitations of the method are comparable to the ones outlined for CATHODE, with the addition that the latent feature representation empirically reduces the signal sensitivity, measured in terms of SIC, and increases the variance.

Appendix B.5 provides another illustration of the mechanism behind background sculpting under correlated auxiliary and resonant features. This study is performed on a larger and more realistic dataset, confirming that background sculpting in m is an inherent consequence of selecting based on a classifier with input x if x is correlated with m, and that explicit decorrelation methods, such as LaCATHODE, are necessary if these correlations are substantial.

6.4 Noise Issue in Weak Supervision

This section is based on the associated publication in Ref. [3], which was prepared together with the co-authors Thorben Finke, Marie Hein, Gregor Kasieczka, Michael Krämer, Alexander Mück, Parada Prangchaikul, Tobias Quadfasel, and David Shih. The results and figures are only slightly modified for the context of this thesis. I contributed to the work by performing initial studies identifying the underlying issue of uninformative features in weak supervision, closely supervising a student who conducted more in-depth noise studies, implementing the TMVA BDT benchmark, discussing the results in the publication, as well as assisting in writing and proofreading of the article.

In the studies of Secs. 6.2 and 6.3, it was observed that a weakly supervised classifier tends to lose sensitivity to a signal when more input features without any additional discrimination power between the background and the respective signal are included. An example is shown in Fig. 6.10, where both m_{jj} and a pure Gaussian noise feature similarly decrease the signal extraction performance of an IAD classifier. In contrast, a fully supervised classifier was able to exploit the small additional information contained in m_{jj}. This section studies the impact of uninformative features in weak supervision in more detail, and proposes replacing the neural network classifier with a boosted decision tree ensemble, which is shown to be substantially more robust to noisy features. The code for reproducing the results is available in a public repository [166].

The task of a weakly supervised anomaly detection classifier, as employed in CATHODE, is essentially to classify between two almost identical distributions: both consist primarily of the same overwhelmingly large background, and they differ only by the small additional signal contribution in one sample. The discrimination power between the underlying signal and background can only be learned if the classifier finds these small differences between the two mixed samples. This is inherently more challenging than in a fully supervised setting, where pure samples of signal and background are learned to be separated directly. The task becomes particularly difficult if the signal contribution is relatively small compared to the statistical fluctuation of the samples. Adding more dimensions in which signal and background differ only based on statistical fluctuations, in the following referred to as "noise", hides the signal even more. Therefore, the more noisy dimensions are present, the more difficult it becomes for the classifier to find a decision boundary through the phase space, such that it separates the underlying signal from the background.

The aim of anomaly detection searches is to be sensitive to a wide range of potential new physics signals. If, however, the classifier training is constrained to only few input features, which need to be discriminative between the background and a signal, the model agnosticism of the search is severely limited. One would want to include sufficiently many input features to cover many potential signals, but then for any specific signal, most of the input features will inevitably be uninformative. Thus, overcoming the limitation of only using a few hand-crafted features, which has been the case in all weakly supervised anomaly detection methods so far, is a crucial step towards a more model agnostic search.

The studies in this section aim to investigate the impact of including many uninformative features in the training of a weakly supervised classifier. In order to factor out the challenge of simulating a background in a data-driven manner, only an IAD setting is considered, leaving the generalization to CATHODE to future work. The BDT and neural network classifiers are exposed to various types of noisy feature sets, different numbers of signal and background events during training, as well as different types of signal processes.

6.4.1 Neural Networks vs Boosted Decision Trees

For a long time, BDTs used to be the standard method for classifying new physics signal against SM background in HEP. They can be seen as a natural progression from manually designing selections based on few high-level features towards a more automated approach. Eventually, deep neural networks were identified to outperform BDTs on these tasks, in particular as they are able to exploit lower-level information rather than relying on hand-crafted features [167]. However, it is known in the field of machine learning that BDTs still outperform neural networks on *tabular data*, in particular on *small- to medium-sized datasets* [168, 169]. In this context, the term "tabular data" refers to a fixed number of features per sample where the values of features can be both categorical and continuous, and there is no positional information between the features, i.e., their order is irrelevant. Reference [168] defines medium-sized datasets as up to an order of 10,000 samples. In their empirical study, the authors highlight three key findings why tree-based models outperform deep neural networks in this case:

1. **Smoothness bias:** while models based on decision trees learn piece-wise constant functions, neural networks are inherently smooth. Thus, the latter need more data to learn irregular decision boundaries.

2. **Uninformative features:** tabular data features tend to be less correlated than structured ones, such as images or sequences. Thus, many features are likely uninformative. BDTs are seen to be less affected by uninformative features than neural networks.

3. **Rotation invariance:** neural network trainings are rotationally invariant, i.e., if each feature vector is multiplied by a unitary matrix before training, the model will learn to

associate the same output as for the original feature vector. This is a more complex learning task, and the reason why neural networks are more sensitive to uninformative features. BDTs, on the other hand, are not rotationally invariant as they consider each feature individually in its original form.

The superiority of neural networks over BDTs in most HEP applications seems to be due to the typically large dataset sizes, as well as the high-dimensional and structured nature of collider data. The case of weakly supervised anomaly detection, however, is characterized by a relatively small *effective dataset size*: the small signal contribution is the only difference between the two (usually large) samples. Moreover, the common feature representation for weakly supervised anomaly detection, such as the high-level representation of the LHCO R&D dataset, is of tabular nature[10]. Therefore, BDTs are expected to be more robust than neural networks in the case where many of these features are uninformative.

Another key advantage of BDTs over deep neural networks is that they are computationally less expensive to train. This is in particular the case for implementations that use histogramming techniques (as briefly mentioned in Sec. 3.3.2). This makes them more attractive for ensembling many individual models. It is common that the average prediction of an ensemble of models is more accurate than the prediction of any individual model. This can, for example, be seen in the comparison of top taggers in Ref. [170]. As the variance of individual model predictions tends to be large in the case of a small signal in weak supervision, ensembling techniques are expected to significantly stabilize the classifier performance. In order to factor out the variance of not only the intrinsic randomness of the model (random weights initialization for the neural network and subsampling schemes for the BDTs), but also the underlying randomness of the training data, the ensemble is built from models trained on different training/validation splits of the data. This can be regarded as a randomized type of cross-validation.

6.4.2 Implementation Details

As in the previous methodological studies, the LHCO R&D dataset in the high-level feature presentation is used to benchmark different approaches, with all definitions analogous to Sec. 6.2.1, unless stated otherwise. Only events within the SR are used, as the studies are limited to training IAD and fully supervised classifiers. This results in a default composition of 120,000 background and 772 signal events in the SR proxy data, as well as a background template of 272,000 events. The dataset is split into 50% training and 50% validation data. In addition to the default two-prong $\mathrm{W}' \to \mathrm{X}(\to \mathrm{q\bar{q}})\mathrm{Y}(\to \mathrm{q\bar{q}})$ signal process, a three-prong signal process $\mathrm{W}' \to \mathrm{X}(\to \mathrm{qqq})\mathrm{Y}(\to \mathrm{qqq})$ with otherwise equal masses will also be considered later.

The main focus of the study is to investigate the behavior of IAD classifiers in the presence of many uninformative features. To this end, we modify the baseline feature set $x = \{m_{J_1}, \Delta m_J, \tau_{21}^{\beta=1,J_1}, \tau_{21}^{\beta=1,J_2}\}$ in two ways[11]. First, we add pure noise dimensions by sampling from a standard normal distribution, indifferently to the signal label of the event. This will be denoted by nG where n is the number of noise dimensions. In a realistic scenario, there will likely not just be fully uncorrelated noise features, but rather a small amount of information will be spread over many physical features. To mimic this, we define three physical extended feature sets, which are summarized in Tab. 6.3. Extended Set 1 adds additional subjettiness ratios

[10] The restriction to hand-crafted high-level features is already made to mitigate the noise issue, thus the promise of BDTs is primarily to use substantially more of them. It would be even more ideal, if this restriction could be lifted entirely. In this case, the tabular nature of the data would not hold anymore and BDTs might no longer be the best choice.

[11] The β parameter in the general N-subjettiness definition of Eq. 2.17 is made explicit in this section, as different choices will be explored. The parameter α remains at the default value of zero.

$\tau_{N,N-1} = \tau_N/\tau_{N-1}$ for $3 \leq N \leq 5$, thus including subjettiness ratios that are not expected to be relevant for the signal in the proxy data. Extended Set 2 replaces the (more human-readable) subjettiness ratios by individual (more correlated) subjettiness variables τ_N for $N \leq 5$. At last, Extended Set 3 also removes the subjettiness ratios from the baseline set, and instead includes all direct subjettiness variables τ_N^β for $N \leq 9$ and $\beta \in \{0.5, 1, 2\}$, thus covering a large phase space with plenty of uninformative and redundant attributes.

Table 6.3: Overview of the four physical feature representations of the LHCO R&D dataset used in the studies of Sec. 6.4.3. The baseline feature set corresponds to the high-level feature choice in the CATHODE studies. The Extended Sets 1–3 add various subjettiness variables either as ratios $\tau_{N,N-1} = \tau_N/\tau_{N-1}$, individual τ_N or all τ_N with different β parameters.

Name	Number of Features	Features
Baseline	4	$\{m_{J_1},\, \Delta m_J,\, \tau_{21}^{\beta=1,J_1},\, \tau_{21}^{\beta=1,J_2}\}$
Extended Set 1	10	$\{m_{J_1},\, \Delta m_J,\, \tau_{N,N-1}^{\beta=1,J_1},\, \tau_{N,N-1}^{\beta=1,J_2}\}$ for $2 \leq N \leq 5$
Extended Set 2	12	$\{m_{J_1},\, \Delta m_J,\, \tau_N^{\beta=1,J_1},\, \tau_N^{\beta=1,J_2}\}$ for $N \leq 5$
Extended Set 3	56	$\{m_{J_1},\, \Delta m_J,\, \tau_N^{\beta,J_1},\, \tau_N^{\beta,J_2}\}$ for $N \leq 9$ and $\beta \in \{0.5, 1, 2\}$

The neural network architecture is the same as in Sec. 6.2.1, with the same hyperparameters and training setup. The only difference is the choice of implementation via Keras [164] and TensorFlow [165].

Multiple choices of BDT architectures were tested and compared (partially shown in Fig. 6.20), out of which the histogram-based gradient boosting (HGB) classifier was found to be the most performant. The HGB classifier was implemented with the `HistGradientBoostingClassifier` from the Scikit-learn library [160]. The default parameters were used, i.e., a learning rate of 0.1, a maximum of 31 leaf nodes per tree, and a maximum of 255 bins for the histogramming. Early stopping was employed with a patience of 10 iterations, and the maximum number of iterations was increased from the default 100 to 200 to ensure that the training stops at the minimum validation loss and not at the maximum number of trees. Similar to the epoch selection procedure for the neural networks, the iteration with the lowest loss on the validation dataset is selected for inference. Subsampling was used for individual trees, such that the predictions can vary between different runs.

Other BDT architectures that were considered and compared are as follows, in line with the concepts introduced in Sec. 3.3.2:

- Random forests (RFs): a decision tree ensemble with sample and feature bagging. Implemented with the `RandomForestClassifier` from Scikit-learn [160].
- AdaBoost: an iterative decision tree ensemble where each tree samples training data based on weights that depend on the performance of the previous trees. Implemented with the `AdaBoostClassifier` from Scikit-learn [160].
- TMVA BDT: a frequently used BDT implementation in HEP based on the Toolkit for Multivariate Analysis (TMVA) [171], which is part of the ROOT framework [172]. We used version 6.28.4, with the BDT default settings, which are based on AdaBoost.

Except for the computationally expensive TMVA BDT, all BDT models were optimized with an automatic hyperparameter search on the baseline feature set, using the Optuna library [173]. Because of the near-optimality of the HGB default settings, we resorted to those for simplicity.

As mentioned before, ensembles of classifier models were built to increase the stability and performance of the predictions. For this, the mean predictions of N_{ens} independent models are computed, where each model is trained with a different randomized training/validation split of the data. Unless stated otherwise, $N_{\mathrm{ens}} = 50$ independent models were used for the ensemble, which was seen to saturate the performance in most settings.

6.4.3 Results

All results in this section will be based on the SIC, either as curves as a function of the signal efficiency, or condensed into a single number as the maximum SIC. This will be evaluated on a separate test set, comprising 340,000 background and 20,000 signal events. In order to ensure that the SIC at tight selections (and thus the maximum) is not dominated by statistical fluctuations, it is only considered in the range where enough background events remain to ensure a relative Poisson uncertainty below 20%. For the BDTs the results will be shown as before in terms of a median and central 68% confidence interval of ten independent retrainings. For the neural networks, a single run is shown because of their high computational cost, especially when ensembling over 50 models. However, it was verified that the variance of the neural network ensemble predictions is low, by examining 10 independent runs for the smaller ensemble size of $N = 10$ on every plot.

BDT Architecture Comparison

In order to decide the BDT architecture for the following studies, a comparison of the HGB, RFs, AdaBoost, and TMVA BDT was performed on the baseline and ten Gaussian noise dimension feature sets. The results are shown in Fig. 6.20. The HGB classifier was found to be the most performant, especially in the presence of noise. This is the reason why the following studies will be performed with the HGB as the choice of BDT. A main factor might be the histogramming technique, which not only accelerates the training, but also reduces the impact of statistical fluctuations in the training data by first aggregating the data into bins. The other approaches perform worse than the HGB and overall relatively similarly with respect to each other.

Baseline Feature Set

With the BDT architecture decided, the main focus is the comparison between the HGB and a neural network. Figure 6.21 compares the SIC curves of BDT and NN ensembles on the baseline feature set in a fully supervised and a weakly supervised setting. While the two classifier approaches have strikingly similar performance in the fully supervised case, the BDTs strictly outperform the neural networks in the weakly supervised setting. It thus seems that even without the presence of uninformative features, the BDT has a more suitable inductive bias for the tabular dataset and the small signal in the IAD setup.

Adding Pure Noise Features

The difference between BDT and NN ensembles becomes more significant when adding uncorrelated Gaussian noise dimensions to the baseline feature set. Figure 6.22 compares the respective SIC curves with various numbers of noise dimensions up to 50. A single noise dimension reduces the neural network IAD performance visibly, while a second one reduces the performance by approximately half across most of the signal efficiency range. With five uninformative features,

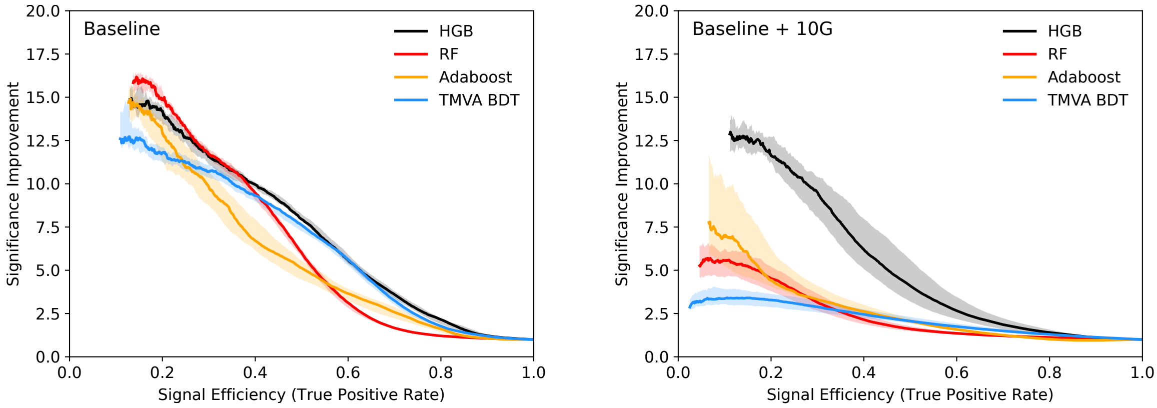

Figure 6.20: Significance improvement of (ensembles of) different boosted decision tree architectures as a function of the signal efficiency using the baseline LHCO R&D dataset (left) and ten Gaussian noise dimension feature set (right). The lines and bands correspond to the median and the 68% confidence interval of ten independent runs. Figure adapted from Ref. [3].

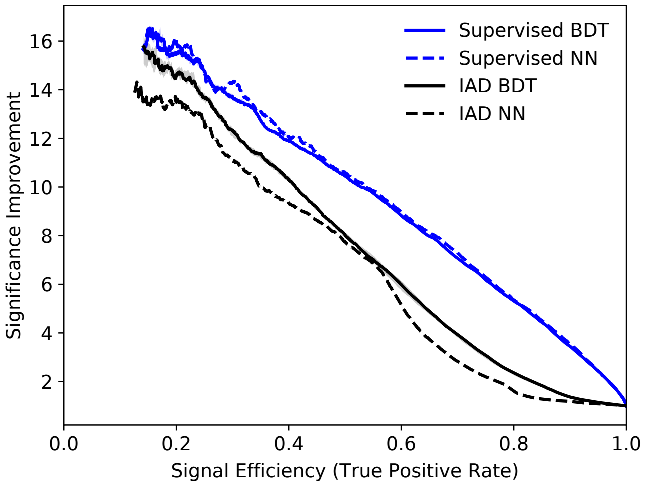

Figure 6.21: Significance improvement of different classifiers as a function of the signal efficiency using the four baseline LHCO R&D dataset features. The lines correspond to either an ensemble of boosted decision trees (solid) or neural networks (dashed) in a fully supervised (blue) or a weakly supervised setting (black). The BDT lines and bands corresponds to the median and the 68% confidence interval of ten independent runs. Figure adapted from Ref. [3].

the SIC is already too low to translate the initial significance of 2.2σ into a discovery significance of 5σ.

The BDT ensemble, on the other hand, retains high performance with many noise dimensions. The performance drop from ten noise dimensions is only marginal. Adding 30 or 50 uninformative features does reduce the performance visibly, but it still remains in a non-trivial range, and retains a maximum SIC around 12. This already demonstrates a substantial improvement in the robustness of the BDT ensemble to uninformative features compared to the neural network ensemble.

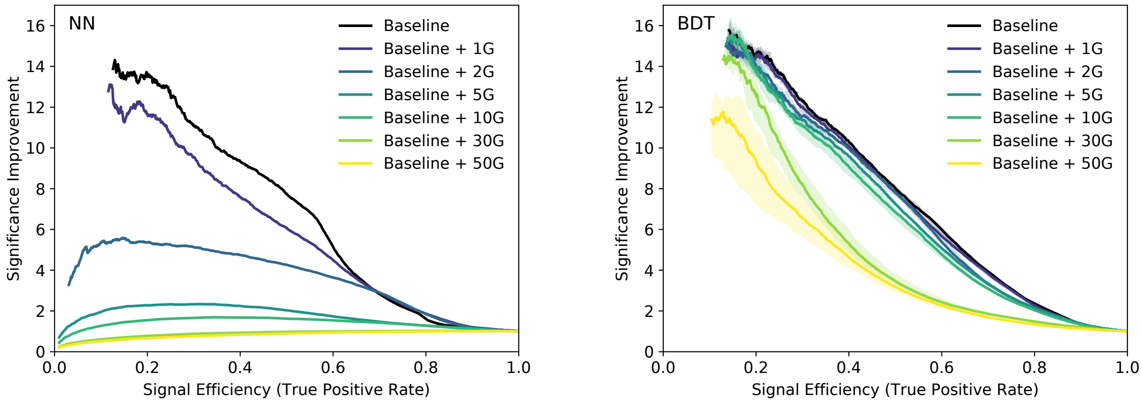

Figure 6.22: Significance improvement of neural network (left) and boosted decision tree (right) classifier ensembles as a function of the signal efficiency when varying numbers of pure Gaussian noise dimensions are added to the baseline LHCO R&D dataset features. The number n of Gaussian noise dimensions is denoted as nG in the legend. The BDT lines and bands correspond to the median and the 68% confidence interval of ten independent runs. Figure adapted from Ref. [3].

Physical Extended Feature Sets

While the uncorrelated noise is relatively easily filtered out by the BDT, the more realistic test is to add additional (generally correlated) physical features with only limited discrimination power for the present signal. The comparison of both classifier architectures is shown for all extended feature sets in Fig. 6.23. The NN ensemble performance decreases visibly for most of the signal efficiency range on Extended Set 1, which includes higher subjettiness ratios. The true signal is a two-prong process and the additional subjettiness ratios τ_{32}, τ_{43}, and τ_{54} are expected to be only discriminative for three-prong, four-prong, and five-prong signals, respectively. The NN is thus exposed to six mostly uninformative dimensions, which reduce the performance but not as much as the fully uncorrelated noise features have done, comparing to the 5G line in Fig. 6.22. The other feature sets, Extended Set 2 and Extended Set 3 improve the NN performance with respect to the baseline, which means that the direct subjettiness features contain significantly more discriminative information than the ratios.

The BDT ensemble, on the other hand, performs strictly better than the NN counterpart on all extended feature sets. It thus also manages to exploit the additional information contained in the large feature sets. The most striking difference lies in Extended Set 1, which for the BDT results in a slightly better performance than the baseline feature set. The mostly uninformative higher subjettiness ratios are thus not only non-detrimental, but even slightly beneficial for the BDT.

Different Signal Strengths

The maximum SIC of the BDT in Fig. 6.23 reaches 15 in the baseline feature set and 35 with Extended Set 3. Multiplied by the initial significance of 2.2σ, this corresponds to a maximum achieved significance of 33σ and 77σ, respectively. Since both are well beyond the conventional 5σ discovery threshold, this difference is not of practical relevance. The more crucial metric in practice is how the methods scale to rarer signals, i.e., what is the smallest signal strength for which the methods achieve a SIC that would result in a discovery. To investigate this, Fig. 6.24

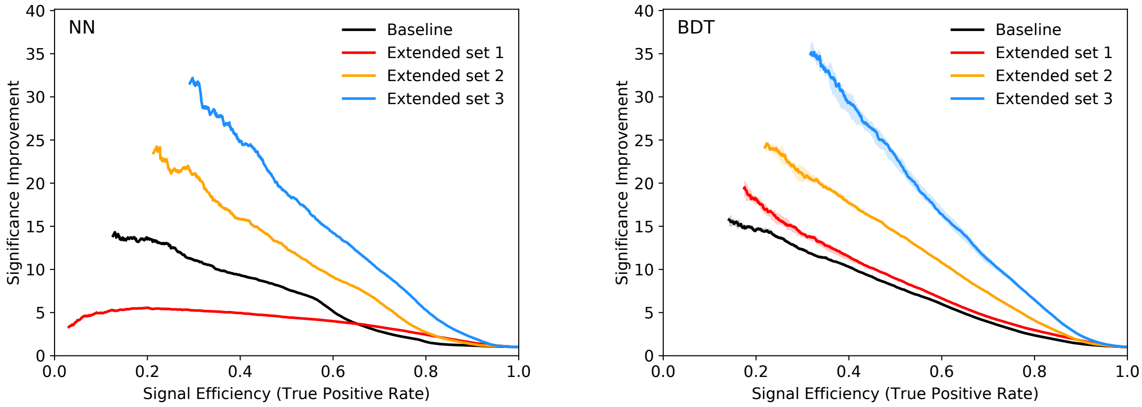

Figure 6.23: Significance improvement of neural network (left) and boosted decision tree (right) classifier ensembles as a function of the signal efficiency using the baseline LHCO R&D dataset features and the three extended feature sets. The BDT lines and bands correspond to the median and the 68% confidence interval of ten independent runs. Figure adapted from Ref. [3].

shows the maximum SIC as a function of the number of SR signal events in the proxy data, while keeping the same number of background events. The BDT achieves a plateau around a SIC of 15 at approximately 500 signal events in the baseline feature set, corresponding to a signal-to-background ratio of 0.4%. The NN performs overall similarly, but it reaches a plateau later. The Extended Set 3 BDT achieves a similar performance at around half the number of signal events. This demonstrates that the sensitivity for rare BSM phenomena can be increased by using more input features and a suitable model (BDTs rather than NNs) that is able to incorporate them.

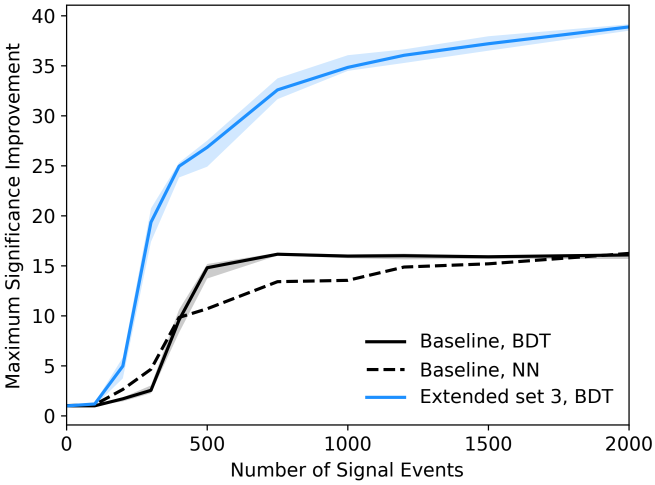

Figure 6.24: Maximum significance improvement as a function of the number of SR signal events during training in the LHCO R&D dataset for the baseline (black) and the Extended Set 3 (blue) feature sets. The BDT lines and bands correspond to the median and the 68% confidence interval of ten independent runs. The neural network results are shown for the baseline case a dashed line. Figure adapted from Ref. [3].

The study in Fig 6.24 can be generalized by calculating the maximum SIC for different combinations of the number of background events and the initial significance. This is shown in Fig. 6.25, for the BDT ensemble with baseline features and Extended Set 3, where the horizontal axis denotes the number of background events and the vertical axis measures the signal presence in terms of the initial significance, $N_{\mathrm{sig}}/\sqrt{N_{\mathrm{bkg}}}$, for the respective number of background events. The classifiers in this case were all trained with the background template size set equal to the respective proxy data background size. This is done for the practical reason that it increases the number of available simulated background events and thus more choices of proxy data background sizes can be scanned. It can be seen that for both feature sets the maximum SIC values only depend weakly on N_{bkg} for fixed values of $N_{\mathrm{sig}}/\sqrt{N_{\mathrm{bkg}}}$. This means that with increasing dataset sizes, one can achieve the same sensitivity with lower signal-to-background ratios. Comparing the two feature sets, the Extended Set 3 BDT achieves a significantly higher maximum SIC for the same $N_{\mathrm{sig}}/\sqrt{N_{\mathrm{bkg}}}$, which is consistent with the results in Fig. 6.23.

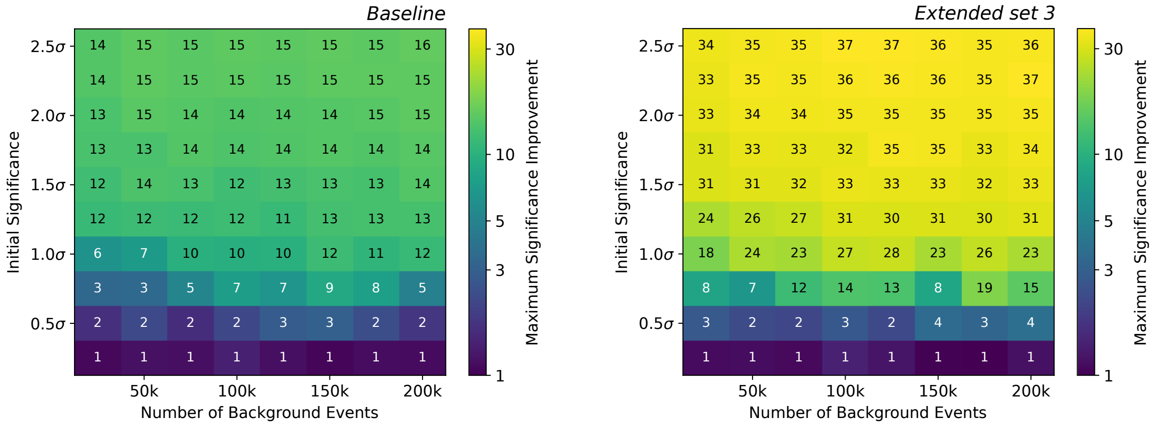

Figure 6.25: Maximum significance improvement as a function of the approximate initial significance $N_{\mathrm{sig}}/\sqrt{N_{\mathrm{bkg}}}$ and the number of background events in the LHCO R&D dataset during training for the baseline (left) and the Extended Set 3 (right) feature sets. Figure adapted from Ref. [3].

Different Signal Process

In order to check how well the findings above generalize to other signal processes, we repeated the feature set comparison study, shown in Fig. 6.23, for the three-prong signal, introduced in Sec. 6.4.2, with the same number of events. The results are shown in Fig. 6.26. It can be seen that the SIC curve for the NN using the baseline feature set is below one for a wide range of signal efficiencies, indicating that the NN is not able to identify the three-prong signal. A possible explanation is that the τ_{21} feature is not discriminative for the three-prong jets, and thus half of the four input features serve as noise. Once the more discriminative subjettiness ratio, τ_{32}, is added in Extended Set 1, the NN performance becomes non-trivial, but remains low. By adding direct subjettiness values, the performance becomes significantly better but lower than in the case of the two-prong signal.

The BDT ensemble, on the other hand, is more performant than the neural network on every feature set, thus showing that its inductive bias is better able to ignore uninformative features, such as τ_{21}. Moreover, the BDT already achieves non-trivial sensitivity with the baseline features, to an extent that the performance is even higher than the NN with Extended

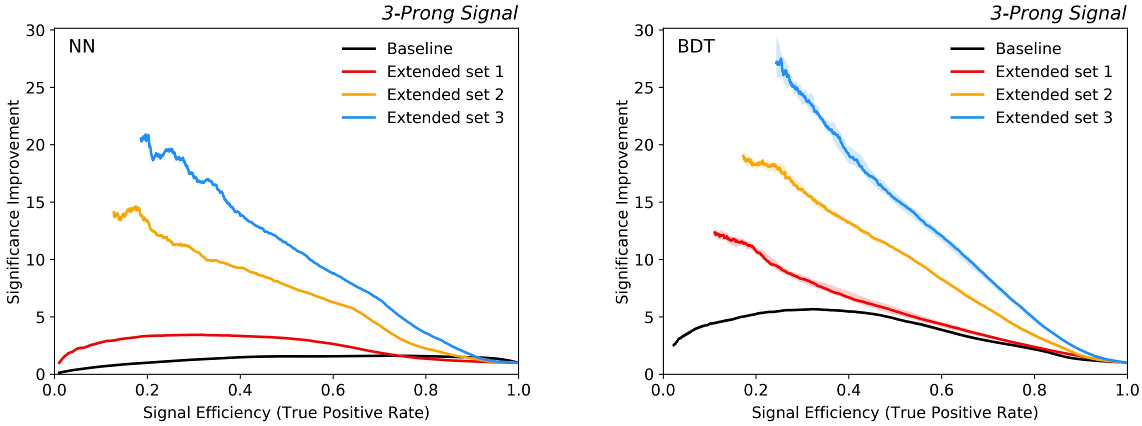

Figure 6.26: Significance improvement of neural network (left) and boosted decision tree (right) classifier ensembles as a function of the signal efficiency using the baseline LHCO R&D dataset features and the three extended feature sets in the presence of a three-prong signal. The BDT lines and bands correspond to the median and the 68% confidence interval of ten independent runs. Figure adapted from Ref. [3].

Set 1. The observation that the BDT achieves a maximum SIC of 5, even with the limited information of the mass features and the noise-like τ_{21}, demonstrates that it is more model agnostic and better able to exploit the information contained in the data.

Ensembling

Lastly, the effect of using ensembles of models instead of individual classifiers is investigated. Figure 6.27 shows the ensemble classifier SIC curves as green lines for neural networks (left) and BDTs (right) on the ten Gaussian noise dimension feature set. The gray lines show the SIC curves that one would obtain by evaluating each of the ensemble members individually, and the black line shows the median SIC curves of these individual ones. The improvement from averaging predictions versus the median performance using individual predictions is striking, both for the NN and the BDT. In the BDT case, the ensemble SIC curve is also significantly better than each individual model at every point in signal efficiency. For the NN, some individual SIC curves are slightly higher than the ensemble one in specific regions in signal efficiency, but lower in others.

6.4.4 Conclusions

Deep neural networks and their more recent variants have replaced boosted decision trees in many high-energy physics tasks because they are well suited for the high-dimensional nature and typically large sizes of collision event data. However, BDTs are known to outperform deep learning approaches on tabular small- to medium-sized datasets, in particular when many uninformative input features are present. We demonstrated in our studies that weakly supervised anomaly detection, with its small effective dataset size and the common choice of tabular input features, remains a HEP use case where BDTs are the superior choice. This becomes especially apparent in the presence of many uninformative features, which naturally arises when targeting a broad range of signal models. Ensembles of BDTs were shown to be resilient to the presence of many noisy features, both in the case of uncorrelated Gaussian noise and in the case of unnecessary physical features. By adding more physical features and replacing the neural network

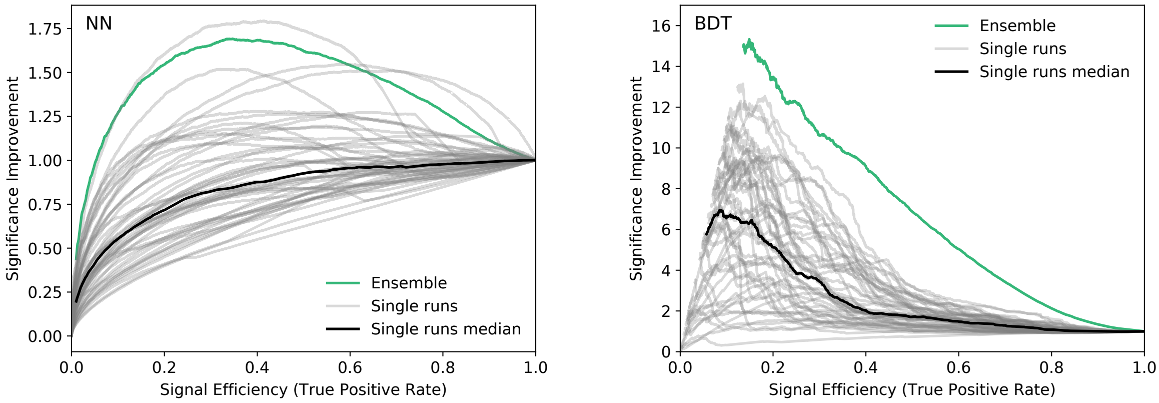

Figure 6.27: Significance improvement of neural network (left) and boosted decision tree (right) classifiers as a function of the signal efficiency on the ten Gaussian noise dimension LHCO R&D dataset features. The gray lines denote single classifiers, whereas the green line denotes the ensemble built from their mean prediction. The median significance improvement of the individual classifiers is shown as black lines. Figure adapted from Ref. [3].

classifier of the idealized anomaly detector by an ensemble of BDTs, the threshold sensitivity for rare signals can be significantly reduced. This is a promising result for the application of weakly supervised anomaly detection in the search for new physics at the LHC.

This proof-of-concept study leaves several avenues for more research. Background sculpting checks, as presented for neural network classifiers in Sec. 6.3, should be performed thoroughly for BDT ensembles as well. Moreover, the tests in this section were limited to the case where a perfect background template is available. Thus, it still needs to be explicitly shown that the results translate to the use of an approximated background template, such as with CATHODE. Whether BDTs retain their superiority in LaCATHODE, where the input features have been rotated away from their physical representation, is another interesting question. While the lower computational cost of histogram-based gradient boosting decision trees was striking and justified training ensembles of 50 models, a more detailed study of the computational cost of the two approaches would be beneficial, as the computational cost of training many classifiers for an analysis, such as the one discussed in Sec. 7, can become a bottleneck and limit the scope of a new physics search. Finally, it has been observed during the studies that the BDT iteration with the lowest validation loss does not necessarily coincide with the highest SIC, which motivates more research into a reliable model selection scheme for anomaly detection in general.

During the final stage of this study, we became aware of other work investigating the benefits of BDTs in weakly supervised anomaly detection, and it was published later under Ref. [174]. Their findings confirm the results presented in this section.

7 CMS Anomalous Dijet Search

This section presents an analysis performed by the CMS Collaboration that is in the final stages of the internal review process at the time of writing. Preliminary results derived using simulated data are publically available as a "CMS Note" in Ref. [4] and using real data as "Physics Analysis Summary" in Ref. [5]. Accordingly, many figures in this section are taken from these documents, or adapted from the collaboration-internal "Analysis Note" [175]. Some of the results differ slightly from the ones shown in Ref. [5] because the treatment of systematic uncertainties was improved during later stages of the review process, and the results might be subject to a few more changes by the time of final publication.

I was one of the main contributors, along with Thea Åarrestad, Oz Amram, Aritra Bal, Samuel Bright-Thonney, Nadya Chernyavskaya, Phil Harris, Gregor Kasieczka, Benedikt Maier, Petar Maksimovic, Patrick McCormack, Louis Moureaux, Jennifer Ngadiuba, Sang Eon Park, Maurizio Pierini, Tobias Quadfasel, David Shih, Tore von Schwartz, Kinga Wozniak, Chitrakshee Yede, and Irene Zoi.

The analysis comprises an application of five different anomaly detection methods within a common statistical framework. This thesis focuses on the application of the CATHODE method, which I implemented together with Gregor Kasieczka, Louis Moureaux, and Tobias Quadfasel. The specifics of the other anomaly detection methods are not discussed in detail.

My contributions to the common analysis framework consist of improvements of the common preselection implementation, studying their trigger efficiency, developments of the statistical procedure, conceptual contributions to the weakly supervised limit setting procedure, major work in defining the data control region for validation studies, investigating the intrinsic bias of the overall statistical procedure, optimizing the selection of suitable signal models for limit setting, investigating the impact of novel systematic uncertainties, and iterating with conveners during the collaboration-internal review process. In the CATHODE application, my contributions include major R&D on the implementation of the method on realistic detector simulation and real collision data, a fully automated limit setting procedure including systematic uncertainties, in-depth studies of the background sculpting and fit function bias specific to CATHODE, building the excess interpretation framework for CATHODE, and performing many of the significance scans, limits, and expected significance estimation on simulation and real data.

7.1 Introduction

The analysis presented in this section probes the dijet final state in proton-proton collision events, recorded by the CMS Experiment with a center-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$, for the existence of new physics via an anomalous jet substructure. The target topology is some heavy new particle A, with a mass of the order of a few TeV and a narrow decay width, decaying into two other generic particles, B and C, with hadronic final states, i.e., these particles could be either single quarks or gluons, or massive resonances ultimately decaying to quarks and gluons. In the latter case, the focus is on masses of A such that B and C are highly Lorentz boosted, resulting in their decay products being merged into two large-radius jets.

This topology is motivated by a relatively broad array of theoretical signal models that predict massive resonances, such as the ones introduced in Sec. 2.2. The substructure of large-radius jets provides a handle to capture anomalous features in the decay of B and C, which can be exploited in a fully data-driven manner by ML-based anomaly detection techniques to reduce the overwhelmingly large SM background by multiple orders of magnitude.

This section specifically focuses on the application of the CATHODE algorithm, whose development is detailed in Sec. 6.2, on CMS data recorded during the years 2016–2018. CATHODE is applied on top of well-motivated triggers and analysis preselections, and followed by a data-driven background estimation procedure via a bump hunt. As CATHODE requires the explicit choice of a signal region, the analysis strategy is to scan multiple choices of overlapping dijet mass bins, repeating the full analysis in each bin. The subsequent statistical analysis yields the significance of potential excesses in the data as a function of the resonance mass, as well as CL_s exclusion limits on a range of explicit signal models at their respective hypothetical mass. The latter are complicated by the signal efficiency of CATHODE depending non-trivially on the signal cross section in the data. The analysis strategy is validated in both MC simulation and data sidebands, before the final results are presented. This CATHODE-based analysis is part of an overarching anomaly detection search described in Refs. [4, 5], where multiple other anomaly detection methods are applied to the same dataset using the same preselections. The performance of all these methods is compared in terms of expected significance and exclusion limits.

The outline of the section is as follows. Section 7.2 first describes the datasets used, both the data recorded by the CMS Experiment and simulated ones, as well as the reconstruction and preselection of events. Section 7.3 discusses the implementation of the CATHODE method in this analysis. The subsequent statistical analysis, including background and signal modeling, is detailed in Sec. 7.4. This is followed by Sec. 7.5, which discusses the systematic uncertainties that are considered in the analysis. The subtleties arising in the computation of upper cross section limits when applying weak supervision–based anomaly detection are discussed in Sec. 7.6. The entire analysis is validated in MC simulation and a data control region in Sec. 7.7. Section 7.8 describes a procedure to interpret potential excesses if they arise in data. The implementation of the other anomaly detection methods is briefly discussed in Sec. 7.9, along with additional benchmarks for comparison. Finally, Sec. 7.10 presents the results of the analysis with real collision data, followed by a conclusion in Sec. 7.11.

7.2 Data and MC Simulation Samples

This section covers the details of the used samples, both real data and MC simulation, how events were reconstructed, and the selections that are applied before they are used as input to the anomaly detection algorithms. The selections consist of generic data quality filters, high-level triggers during data-taking, and analysis-specific preselections.

7.2.1 Data Samples

The data comprise the proton-proton collision events recorded in the CMS detector at a center-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ during the Run 2 data-taking period from 2016 to 2018. The total integrated luminosity of the dataset is $138\,\mathrm{fb}^{-1}$. The quality of the data is ensured by requiring certain criteria were fulfilled during the recording of the events, i.e., all sub-detectors were considered properly working during data taking. All luminosity sections that fulfill these criteria are listed in so-called “golden JSON” files, which are listed in Tab. 7.1 together with the respective integrated luminosity of each data-taking year.

All events recorded by the CMS Experiment need to pass one or more HLT paths (see Sec. 2.4.3) and are then sorted in associated *primary datasets*. The events used in this analysis are stored in the so-called `JetHT` dataset, which contains events passing trigger requirements focusing on highly energetic jets. The data were processed within the CMS Ultra Legacy (UL) campaign in the CMS Software (CMSSW) releases `CMSSW_8_X`, `CMSSW_9_X`, `CMSSW_10_X` depending on the year. The *global tag*, which is a label that uniquely identifies a set of detector configurations needed for the correct analysis of the data [176], is `106X_dataRun2_v32`. The samples are

Table 7.1: "Golden JSON" files used in the analysis, i.e., record of events fulfilling data quality monitoring criteria, along with the corresponding integrated luminosities.

Year	**Golden JSON file name**	**Integrated luminosity (fb^{-1})**
2016	Cert_271036-284044_13TeV_Legacy2016_Collisions16_JSON.txt	36.4
2017	Cert_294927-306462_13TeV_UL2017_Collisions17_GoldenJSON.txt	41.5
2018	Cert_314472-325175_13TeV_Legacy2018_Collisions18_JSON.txt	59.8
Total		138

Table 7.2: CMS data sample identifiers and their corresponding integrated luminosities. None of the events of the first sample passed the quality criteria set in the "golden JSON" files, Tab. 7.1.

Sample name	**Integrated luminosity (fb^{-1})**
/JetHT/Run2016B-21Feb2020_ver1_UL2016_HIPM-v1/MINIAOD	0
/JetHT/Run2016B-21Feb2020_ver2_UL2016_HIPM-v1/MINIAOD	5.83
/JetHT/Run2016C-21Feb2020_UL2016_HIPM-v1/MINIAOD	2.62
/JetHT/Run2016D-21Feb2020_UL2016_HIPM-v1/MINIAOD	4.29
/JetHT/Run2016E-21Feb2020_UL2016_HIPM-v1/MINIAOD	4.07
/JetHT/Run2016F-21Feb2020_UL2016_HIPM-v1/MINIAOD	2.71
/JetHT/Run2016F-21Feb2020_UL2016-v1/MINIAOD	0.42
/JetHT/Run2016G-21Feb2020_UL2016-v1/MINIAOD	7.65
/JetHT/Run2016H-21Feb2020_UL2016-v1/MINIAOD	8.74
/JetHT/Run2017B-UL2017_MiniAODv2-v1/MINIAOD	4.80
/JetHT/Run2017C-UL2017_MiniAODv2-v1/MINIAOD	9.57
/JetHT/Run2017D-UL2017_MiniAODv2-v1/MINIAOD	4.25
/JetHT/Run2017E-UL2017_MiniAODv2-v1/MINIAOD	9.31
/JetHT/Run2017F-UL2017_MiniAODv2-v1/MINIAOD	13.5
/JetHT/Run2018A-UL2018_MiniAODv2-v1/MINIAOD	14.0
/JetHT/Run2018B-UL2018_MiniAODv2-v1/MINIAOD	7.06
/JetHT/Run2018C-UL2018_MiniAODv2-v1/MINIAOD	6.90
/JetHT/Run2018D-UL2018_MiniAODv2-v1/MINIAOD	31.8
Total integrated luminosity	138

provided in the MiniAOD format, specifically MiniAODv1 for the 2016 data-taking period, and MiniAODv2 for 2017 and 2018. The sample identifiers, as they are labeled in the CMS Data Aggregation Service [177], are listed in Tab. 7.2.

The table lists the samples ordered by the so-called era, which is a subdivision motivated by changes in detector conditions and/or calibration. The eras are denoted by letters B–H in 2016, B–F in 2017, and A–D in 2018. Another distinction is made within the 2016 data-taking period. Early in 2016, the silicon strip tracker was subject to a decrease in signal-to-noise ratio and a loss of hits of tracks caused by saturation effects in the pre-amplifier of the readout chip. In order to mitigate this, the feedback preamplifier bias voltage (VFP) was later adjusted [178]. The data-taking periods before and after this adjustment are thus often referred to as 2016 preVFP and 2016 postVFP, respectively. Table 7.2 labels preVFP samples, which are eras B, C, D, E, and the majority of F, with the tag `HIPM`. This is a reference to the so-called HIP mitigation used in the reconstruction of the affected events.

The centrally produced MiniAOD files were processed privately into the NanoAOD format corresponding to the official production campaign because NanoAOD files are significantly smaller and easier to handle. However, this format by default discards information of individual jet constituents. It is thus extended to include particle flow candidates, using the JetMET PFNano producer [179].

7.2.2 MC Background Simulation Samples

While MC background simulation samples are not directly used in the analysis—the background will be estimated using data-driven methods—they still play a crucial role for optimizing and validating the analysis strategy.

The dominant background process consists of QCD multijet events, which are simulated directly with PYTHIA 8.240 [52]. The more minor SM background processes considered in this analysis are W and Z bosons produced in association with jets (V+jets) and forced to decay into quark pairs, top quark-antiquark ($\mathrm{t\bar{t}}$) pair production, and single top quark production. Other SM processes were found to be negligible with the anticipated selections. The V+jets samples were simulated at leading order (LO) in the matrix element calculations with MADGRAPH5_aMC@NLO v2.6.5 [180]. The $\mathrm{t\bar{t}}$ and single top events were generated at next-to-leading order (NLO) in the matrix element calculations with POWHEG v2 [181–183]. All processes were produced with the NNPDF 3.1 parton distribution functions [184–186] and showered with PYTHIA 8.240 using the CP5 tune [187]. The interactions with the CMS detector were simulated with GEANT 4 [188–190]. The pileup (PU) interactions were simulated by superimposing additional inelastic proton-proton interaction events, simulated in PYTHIA 8.240.

The MC background simulation samples were produced in the UL campaign and provided in the MiniAODv2 format. They were generated individually for each of the four data-taking periods (2016 preVFP, 2016 postVFP, 2017, 2018) and in separate phase space bins, which are summarized in Tab. 7.3 for each process along with their respective cross section. The latter values were obtained for the QCD multijet, V+jets, and $\mathrm{t\bar{t}}$ samples from Ref. [191], which is the collaboration-internal analysis note corresponding to the publication in Ref. [192]. The cross sections for the single top events produced in the tW-channel were taken from Ref. [193], while the cross sections for t-channel single top events were obtained from the CMS cross section database [194]. The detailed list of samples is provided in Tabs. B.1 to B.4 in Appendix B.1.

As with the real data, the MC background simulation MiniAOD samples were privately processed to the NanoAOD format with extended information on individual jet constituents. This was done with the same CMSSW release `CMSSW_10_6_20` and the following global tags per data-taking period[12]:

- 2016, preVFP: `106X_mcRun2_asymptotic_preVFP_v11`
- 2016, postVFP: `106X_mcRun2_asymptotic_v17`
- 2017: `106X_mc2017_realistic_v9`
- 2018: `106X_upgrade2018_realistic_v16_L1v1`.

The QCD multijet samples were produced in bins of jet transverse momentum, the V+jets samples in H_T bins, the $\mathrm{t\bar{t}}$ samples in top quark-antiquark pair mass bins, and the single top process was produced inclusively. The number of available events in each sample is not proportional to the cross section of every process and kinematic bin. Conventional analyses that use MC background simulation primarily for background estimation mitigate this by multiplying each event with a weight determined by cross section and detector acceptance efficiency (in addition to weights resulting from the specific MC generator). However, this does not exactly model the situation of a purely data-driven, ML-based analysis where every event is considered with the same weight and follows Poisson statistics. Moreover, very large or small weights can cause artificial instabilities during training, which is not the case in real data. For this reason, we

[12] The 2016 preVFP period is simulated separately in order to account for the different conditions in the silicon strip tracker.

Table 7.3: Summary table of considered MC background simulation processes, along with their respective phase space binning and the respective cross section. The same composition is generated in each of the four data-taking periods (2016 preVFP, 2016 postVFP, 2017, 2018) individually. A more detailed list of samples is provided in Tabs. B.1 to B.4 in the appendix Sec. B.1.

Process	**Phase Space Region (GeV)**	**σ (pb)**
QCD	$300 < p_\mathrm{T} < 470$	7823
QCD	$470 < p_\mathrm{T} < 600$	648.2
QCD	$600 < p_\mathrm{T} < 800$	186.9
QCD	$800 < p_\mathrm{T} < 1000$	32.293
QCD	$1000 < p_\mathrm{T} < 1400$	9.4183
QCD	$1400 < p_\mathrm{T} < 1800$	0.84265
QCD	$1800 < p_\mathrm{T} < 2400$	0.114943
QCD	$2400 < p_\mathrm{T} < 3200$	0.00682981
QCD	$3200 < p_\mathrm{T}$	0.000165445
W+jets	$400 < H_\mathrm{T} < 600$	315.6
W+jets	$600 < H_\mathrm{T} < 800$	68.57
W+jets	$800 < H_\mathrm{T}$	34.9
Z+jets	$400 < H_\mathrm{T} < 600$	145.4
Z+jets	$600 < H_\mathrm{T} < 800$	34.0
Z+jets	$800 < H_\mathrm{T}$	18.67
$\mathrm{t\bar{t}}$	$700 < m_{\mathrm{t\bar{t}}} < 1000$	76.605
$\mathrm{t\bar{t}}$	$1000 < m_{\mathrm{t\bar{t}}}$	20.578
single t, t-channel	full	115.3
single $\mathrm{\bar{t}}$, t-channel	full	69.09
single t, tW production	full	35.85
single $\mathrm{\bar{t}}$, tW production	full	35.85

chose to eliminate the relative event weights by sampling a realistic number of events N_sampled from each MC simulation sample:

$$N_\mathrm{sampled} = \mathcal{L}_\mathrm{eff} \times \varepsilon_\mathrm{presel} \times \sigma, \tag{7.1}$$

where $\mathcal{L}_\mathrm{eff}$ is the effective target luminosity, $\varepsilon_\mathrm{presel}$ is the fraction of events in the process and bin passing the analysis preselections, and σ is the cross section of the process and bin. The effective luminosity corresponds to the maximum number of unique events that can be obtained from these samples while respecting the relative cross sections of the processes and bins. The lowest transverse momentum bin of the QCD multijet samples was the limiting factor, with an effective luminosity of $6.7\,\mathrm{fb}^{-1}$ in each data-taking period (2016 preVFP, 2016 postVFP, 2017, 2018), according to Eq. 7.1, resulting in a total integrated luminosity of $26.8\,\mathrm{fb}^{-1}$. This is approximately five times less than the total integrated luminosity of the real data. The equal contribution to the total sample from each of the four data-taking periods does not accurately reflect the actual composition of the Run 2 data, but this was neglected in favor of obtaining a larger sample size.

A comparison of the data and the MC background simulation contributions is shown in Sec. 7.2.8 after the corresponding analysis preselections are introduced.

7.2.3 MC Signal Simulation Samples

The MC signal simulation samples in this analysis are primarily used for setting explicit exclusion limits on a few representative types of new physics signals, and to measure the expected experimental sensitivity of the search for the types of signal in question. The focus lies on the boosted topology, i.e., the decay products of the particles B and C are merged into single large-radius jets.

Table 7.4 lists the considered signal models, along with the sample identifiers within the CMS Collaboration, ordered by the pronginess of the decays of the two particles B and C, ranging up to six. Each process was produced at resonance particle (A) masses of 2, 3 and 5 TeV. While the daughter particles (B and C) were initially simulated with all combinations of masses from 25, 80, 170, and 400 GeV, only those that result in a boosted topology and on-shell decays were considered. If the daughter particles are present in the SM, their masses are kept at their respective SM values, unless stated otherwise. All signal samples were generated with a narrow decay width, i.e. it was chosen so narrow that the observed width is entirely due to the detector resolution.

Table 7.4: CMS sample identifiers of MC signal simulation samples used in this analysis for estimating the experimental sensitivity and setting exclusion limits. The three signal processes involving at least one two-prong decay were generated with 50,000 events and the remaining two processes with 100,000 events in each data-taking period. All samples were simulated with resonance masses of 2, 3, and 5 TeV, and the daughter masses were scanned ensuring on-shell decays. Unless otherwise specified, the daughter particles present in the Standard Model are kept at their respective Standard Model masses. The asterisk (*) indicates a scan over masses in GeV.

Process	**Prongs**	**Sample name**	**B and C masses (GeV)**
$\mathrm{Q}^* \rightarrow \mathrm{qW}'$	1+2	QstarToQW_M_*_mW_* _TuneCP2_13TeV-pythia8	$m_{\mathrm{W}} = 25, 80, 170, 400$
$\mathrm{X} \rightarrow \mathrm{YY}' \rightarrow 4\mathrm{q}$	2+2	XToYYprimeTo4Q_MX*_MY*_MYprime* _narrow_TuneCP5_13TeV-madgraph-pythia8	$m_{\mathrm{Y/Y'}} = 25, 80, 170, 400$
$\mathrm{W}' \rightarrow \mathrm{B}'\mathrm{t} \rightarrow \mathrm{bZt}$	3+3	WpToBpT_Wp*_Bp*_Top170_Zbt _TuneCP5_13TeV-madgraphMLM-pythia8	$m_{\mathrm{B'}} = 25, 80, 170, 400$
$\mathrm{W}_{\mathrm{KK}} \rightarrow \mathrm{RW} \rightarrow 3\mathrm{W}$	2+4	WkkToWRadionToWWW_M*_Mr* _TuneCP5_13TeV-madgraph-pythia8	$m_{\mathrm{R}} = 170, 400$
$\mathrm{Y} \rightarrow \mathrm{HH} \rightarrow 4\mathrm{t}$	6+6	YtoHH_Htott_Y*_H* _TuneCP5_13TeV-madgraph-pythia8	$m_{\mathrm{H}} = 400$

The first model listed in Tab. 7.4 postulates an excited quark resonance, Q^*, decaying into a quark and a W' boson [195, 196]. This model follows the class of BSM scenarios discussed in Sec. 2.2.2. The W' subsequently decays into two quarks with the same branching fraction as the SM W boson, which results in a two-prong jet, in addition to the single-prong jet from the quark that was produced in association.

The second model involves a generic heavy resonance, X, that decays into two new particles Y and Y', which subsequently decay into two light quarks each. This results in a double two-prong jet topology. This signal resembles most closely the one used in the LHCO R&D benchmark dataset of Sec. 6.1.

The third considered model consists of a W' boson decaying into a vector-like quark, B', and a top quark [197]. It is inspired by the composite Higgs models introduced in Sec. 2.2.3. The B' then decays into a Z boson and a bottom quark, resulting in a three-prong jet from both the B' and t. A dedicated search for this model was performed in Ref. [198], but not in the parameter space considered in this analysis, where the B$'$ is fully contained within a single large-radius jet because of the large mass difference with respect to the W$'$.

The fourth signal model contains a Kaluza-Klein excitation of the W boson, W_{KK}, decaying into a W boson and a radion, R [35, 199]. The radion decays into two W bosons, resulting in a two-prong and a four-prong jet. This model has been targeted within dedicated searches by the CMS Collaboration [200, 201], and arises from the theoretical framework of warped extra dimensions discussed in Sec. 2.2.4.

In the last considered model, also based on warped extra dimensions, a heavy graviton-like spin-2 particle, Y, decays into two lighter Higgs boson–like scalars, H, which each decay into a top quark-antiquark pair [202]. This results in two six-prong jets.

The samples were produced within the official CMS MC production framework in the UL campaign. The $Q^* \rightarrow qW'$ sample was fully produced in PYTHIA 8.240, using the CP2 tune. The other samples were generated using MadGraph5_aMC@NLO v2.6.5 at LO in QCD, with the NNPDF 3.1 parton distribution functions, and their parton showering was done with PYTHIA 8.240 using the CP5 tune. The $W' \rightarrow B't \rightarrow bZt$ sample additionally used the MLM matching scheme. The samples were provided in the MiniAODv1 format for 2016 and MiniAODv2 for 2017 and 2018.

The signal samples were also processed privately into the NanoAOD format with extended information on individual jet constituents. This was done with the same CMSSW release `CMSSW_10_6_20` and the following global tags per data-taking period:

- 2016 preVFP: `106X_mcRun2_asymptotic_preVFP_v9`
- 2016 postVFP: `106X_mcRun2_asymptotic_v15`
- 2017: `106X_mc2017_realistic_v8`
- 2018: `106X_upgrade2018_realistic_v15_L1v1`.

7.2.4 Quality Filters and Vertex Selection

Several filters were applied in order to remove events that were affected by misreconstructed missing transverse energy due to instrumental effects. These can, for example, appear as anomalous HCAL signals or spurious energy deposits in the ECAL. The filters used in this analysis are the ones recommended by the MET Physics Object Group [203]:

- `goodVertices`
- `globalSuperTightHalo2016Filter`
- `HBHENoiseFilter`
- `HBHENoiseIsoFilter`
- `EcalDeadCellTriggerPrimitiveFilter`
- `BadPFMuonFilter`
- `BadPFMuonDzFilter`
- `eeBadScFilter`
- `CSCTightHaloFilter` (2016 only)
- `ecalBadCalibFilter` (not in 2016).

Additional requirements were imposed on the proton-proton interaction vertex. At least one such primary vertex had to be reconstructed within a 24 cm window along the beam axis, with a transverse distance from the nominal proton-proton interaction region smaller than 2 cm. If multiple vertices were reconstructed, the one with the highest sum of the squared transverse momenta of the associated tracks was chosen.

A failure of several HCAL modules during the 2018 data-taking period caused a negative impact on the jet energy measurements in the affected region. To avoid these instrumental effects being misinterpreted as signal-like anomalies, events in which one of the two highest transverse momentum jets was found to be in the affected region, $-1.57 < \phi < -0.87$ and $-2.5 < \eta < -1.3$, were vetoed. This resulted in a signal efficiency loss of approximately 1%.

7.2.5 Event Reconstruction

A particle-flow (PF) algorithm, as discussed in Sec. 2.4.4, was used to reconstruct individual particles. The resulting PF candidates were then clustered using the anti-k_T algorithm (Sec. 2.5.2) with a radius parameter of $R = 0.8$, referred to as in AK8 jets. The jet clustering was implemented in the FastJet package [153]. The PUPPI algorithm (Sec. 2.5.3) was used to mitigate the effects of pileup on the jets.

As an additional quality criterion on the AK8 jets, they are required to pass the *tight jet ID* defined and recommended by the JetMET Physics Object Group for UL samples and an eta range of $|\eta| < 2.5$ [204]:

- neutral hadron fraction < 0.9,
- neutral electromagnetic fraction < 0.9,
- number of constituents > 1,
- number of charged hadron constituents > 0,
- charged hadron fraction > 0.

Moreover, standard CMS jet energy corrections are applied to the AK8 jets to correct for non-linearities in the jet transverse momentum and rapidity dependent energy response, as described in Ref. [205]. The latest jet energy corrections were automatically applied in the NanoAOD processing, and are described in Ref. [206].

The AK8 jet mass, which will be used as input to the ML algorithms, was calculated using the soft drop algorithm, as described in Sec. 2.5.2. Subjettiness features (Sec. 2.5.5) were reconstructed up to τ_4.

Another reconstructed jet feature is the b-tagging score, i.e., the probability of the jet arising from a B-hadron decay according to a classification algorithm. In this case, the DeepCSV [64] algorithm (briefly introduced in Sec. 2.5.6) was employed, which classifies AK4 jets. The score per AK8 jet in this analysis was obtained as the maximum score per reconstructed AK4 subjet within the AK8 jet. In the case where either no subjet was found or some necessary inputs for DeepCSV were missing in the subjet, the score was set to -1. The resulting score is labeled as *DeepB* in the following.

7.2.6 Analysis Preselections

In addition to the generic quality requirements, further analysis preselections were applied in order to focus on the relevant phase space corresponding to the target topology of a heavy resonance decaying into two highly Lorentz-boosted particles, and thus reduce the overwhelming SM background.

Events are required to contain at least two AK8 jets with $p_{\mathrm{T}} > 300\,\mathrm{GeV}$, $|\eta| < 2.5$, and passing the tight jet ID requirement. The two jets with highest p_{T} are selected as the candidate dijet system. The analysis region further requires these two *leading jets* to have a pseudorapidity difference of $|\Delta\eta| < 1.3$, which significantly reduces the number of QCD multijet events because these are dominated by t-channel interactions and thus typically have large pseudorapidity differences. The complementary region will be used for validation studies, as discussed in Sec. 7.7. The invariant mass of the dijet system, m_{jj}, is required to be larger than 1455 GeV, which is chosen to ensure full trigger efficiency, as discussed in Sec. 7.2.7. In the following, the two leading jets will be sorted by descending mass, i.e., the leading jet (j1) is the one with the higher mass than the subleading jet (j2).

7.2.7 Triggers

The high-level triggers used in the analysis are based on the presence of at least one high transverse momentum jet or the scalar sum of the transverse momenta of all jets in the event, H_{T}. At least one of the listed trigger paths in Tab. 7.5 needs to have been activated during data taking in order to be considered in the analysis. The trigger paths are separated into data-taking periods as the triggers were updated between the years.

Table 7.5: High-level triggers of each data-taking period used to collect data for this analysis. They are all based on either jet transverse momentum p_{T} or the scalar sum of jet transverse momenta in the event H_{T}.

Period	**H_{T} trigger**	**jet p_{T} trigger**
2016 runs B-G	HLT_PFHT800	HLT_PFJet450
2016 run H	HLT_PFHT900	HLT_PFJet450 HLT_AK8PFJet450
2017 runs B-F	HLT_PFHT1050	HLT_AK8PFJet500
2018 runs A-D	HLT_PFHT1050	HLT_AK8PFJet500

The background estimation, discussed in Sec. 7.4, will be based on fitting a monotonically falling function through the m_{jj} distribution. Thus, it is crucial that the trigger requirements do not change the background shape. To ensure this, the phase space is restricted to a region where the triggers are considered fully efficient. This means that the fraction of events passing the triggers is at least 99%, when the analysis preselections are in place. This was studied in data by using a disjoint dataset, the `SingleMuon` dataset, with the requirement of passing either of the reference triggers `HLT_IsoMu27` or `HLT_Mu50`. The resulting trigger efficiency on this dataset is shown as a function of m_{jj} in Fig. 7.1. The 99% efficiency is reached at 1181 GeV in 2016 and at 1455 GeV in 2017 and 2018. The latter value is thus chosen as the threshold for the analysis region.

The peak in trigger efficiency on the left-hand side of Fig. 7.1 is due to the tight $p_{\mathrm{T}} > 300\,\mathrm{GeV}$ selection on the leading jets. For dijet invariant masses below approximately 600 GeV, this peak only arises when both leading jets point to a similar direction. Because of momentum conservation, this is usually balanced by multiple additional jets in the opposite direction. The resulting high value in H_{T} activates the respective triggers, resulting in a high efficiency.

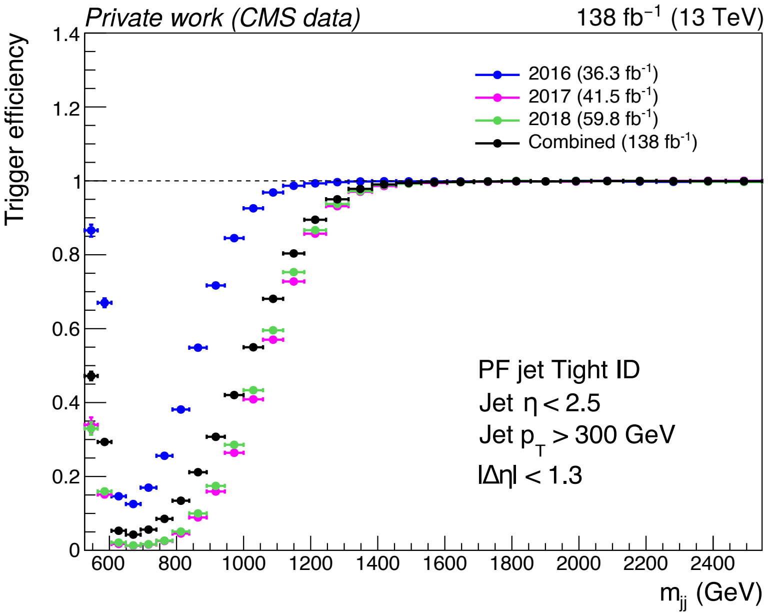

Figure 7.1: Trigger efficiency, i.e., the number of events passing the trigger divided by the total number of events, as a function of the dijet invariant mass in each data-taking year as well as their combination.

7.2.8 Data-to-Simulation Comparison

Figures 7.2 to 7.11 show the comparison of the observed collision data and the MC background simulation samples, discussed in Sec. 7.2.2, after all filters, triggers and analysis preselections are applied. Shown are two per-dijet features m_{jj} and $|\Delta\eta|$ (Fig. 7.2), and four per-jet kinematic observables, m, p_{T}, η, ϕ (Figs. 7.3 to 7.6). Figures 7.7 to 7.11 illustrate jet substructure features τ_1, τ_2, τ_3, τ_4, and the b-tagging score DeepB. The per-bin ratio in the lower panel reveals that there is visible disagreement between data and the MC background simulation. This is expected, as the description of the SM QCD background is known to be limited. This motivates the use of a data-driven background estimation method and is a key reason why the MC simulation is only used for optimizing and validating the analysis strategy.

In addition to the SM background processes, Figs. 7.2 to 7.11 also display an example BSM signal. Specifically, an $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal sample with masses of $m_{\mathrm{X}} = 3000\,\mathrm{GeV}$, $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$, and $m_{\mathrm{Y}'} = 80\,\mathrm{GeV}$ is shown, with a normalization corresponding to a cross section of 30 fb.

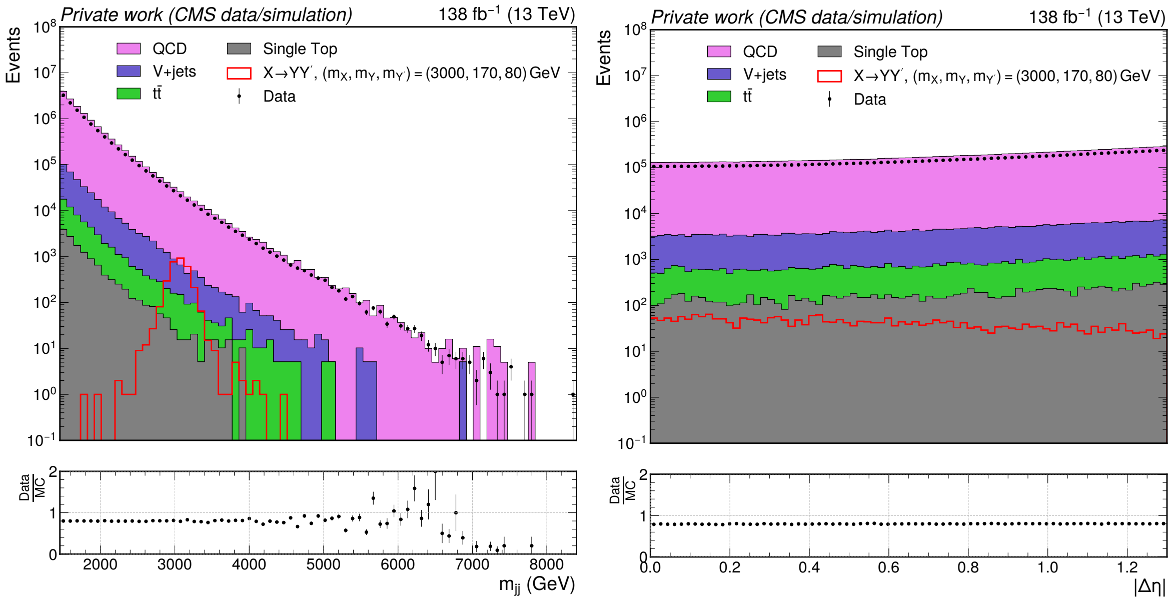

Figure 7.2: Reconstructed dijet invariant mass m_{jj} (left) and absolute jet pseudorapidity difference $|\Delta\eta|$ (right) distribution, with all analysis preselections applied, for collision data, MC background simulation and an example 30 fb X $\to$ YY$'$ $\to$ 4q signal sample with masses of $m_X = 3000\,\text{GeV}$, $m_Y = 170\,\text{GeV}$, and $m_{Y'} = 80\,\text{GeV}$.

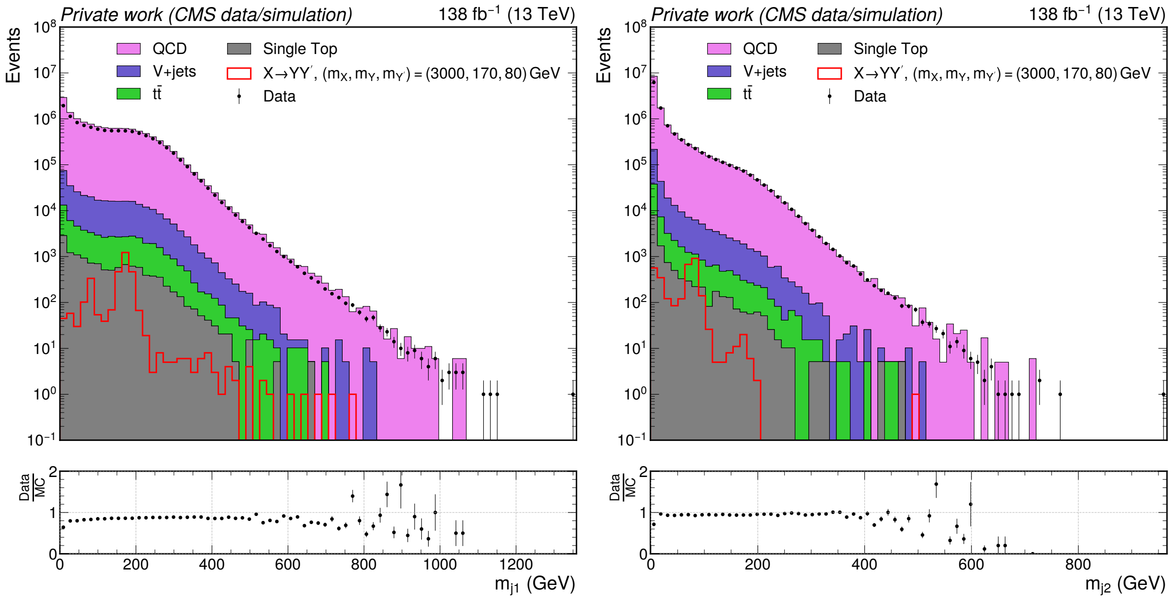

Figure 7.3: Analogously to Fig. 7.2, the soft drop mass distribution of the leading jet (left) and subleading jet (right).

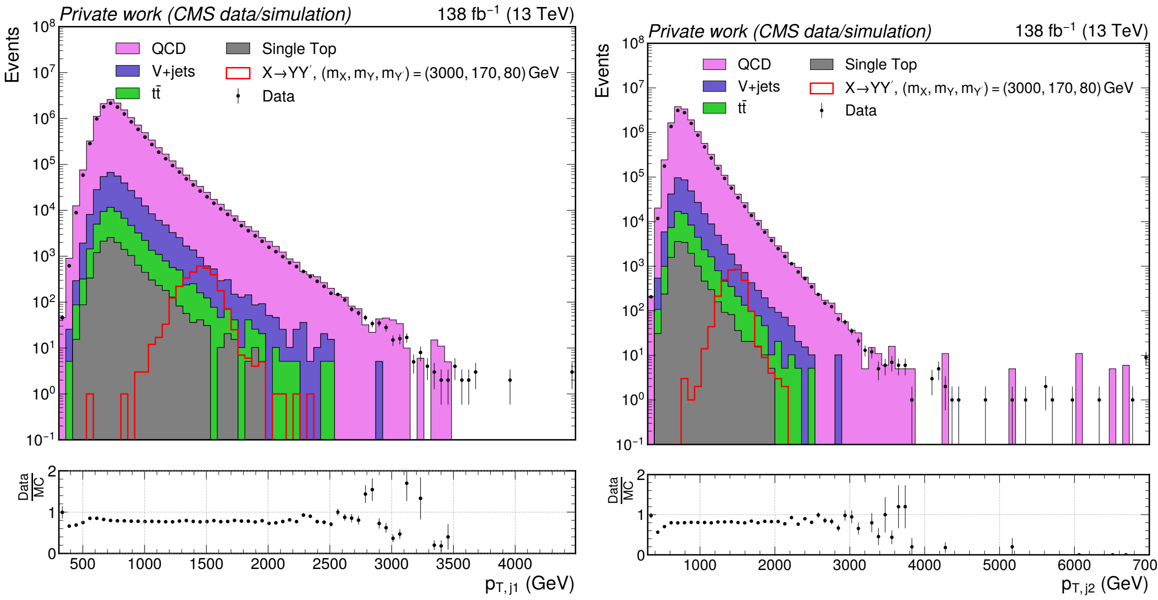

Figure 7.4: Analogously to Fig. 7.2, the transverse momentum distribution of the leading jet (left) and subleading jet (right).

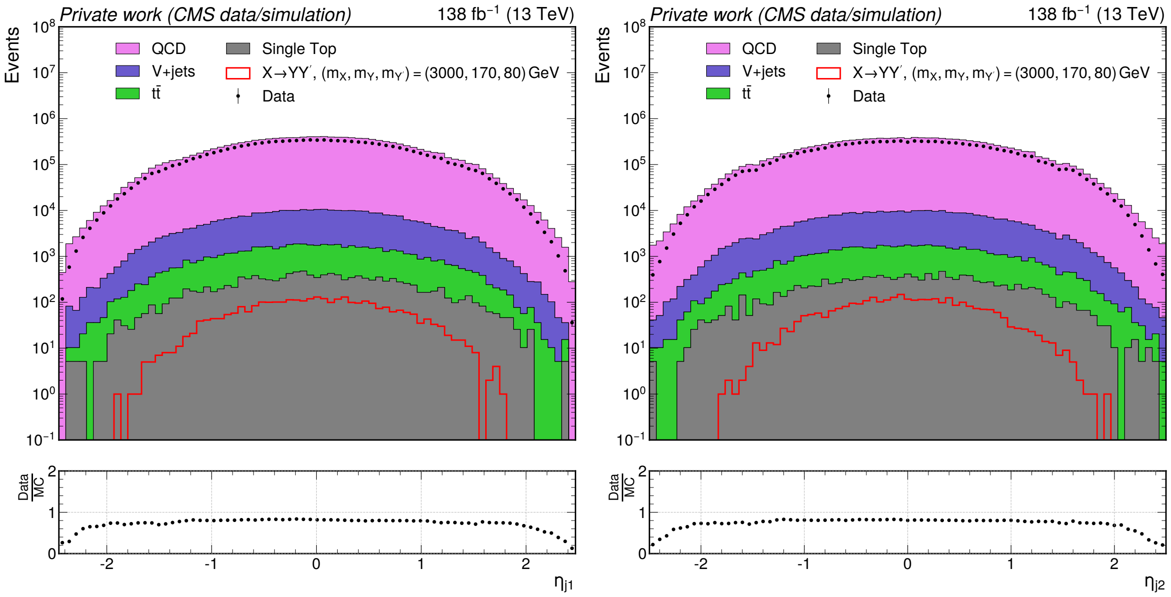

Figure 7.5: Analogously to Fig. 7.2, the pseudorapidity distribution of the leading jet (left) and subleading jet (right).

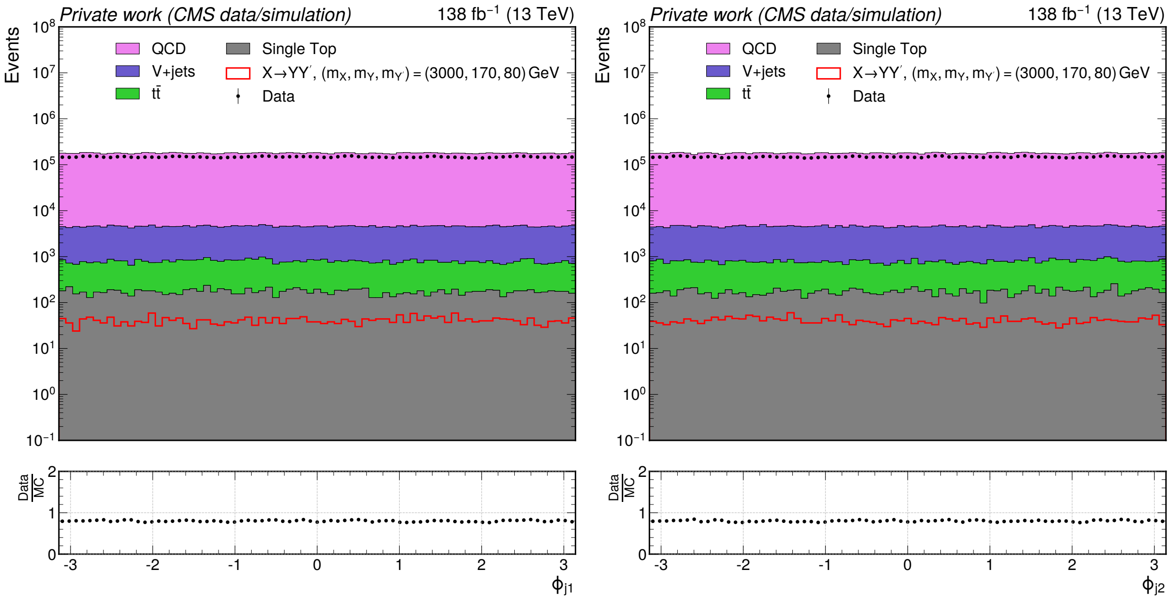

Figure 7.6: Analogously to Fig. 7.2, the azimuthal angle distribution of the leading jet (left) and subleading jet (right).

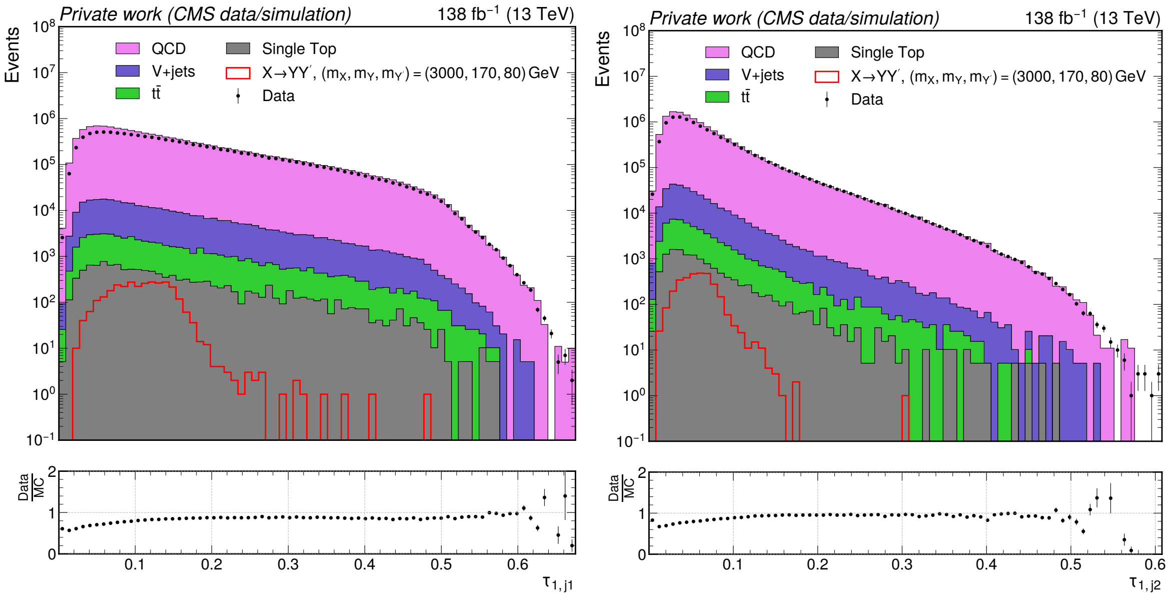

Figure 7.7: Analogously to Fig. 7.2, the 1-subjettiness, τ_1, distribution of the leading jet (left) and subleading jet (right).

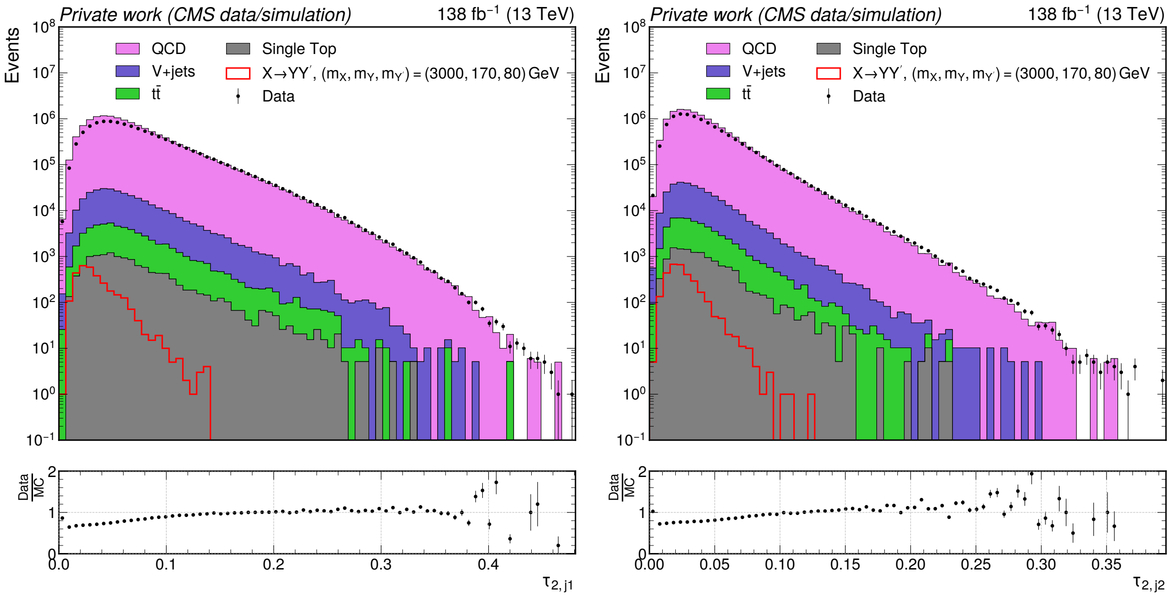

Figure 7.8: Analogously to Fig. 7.2, the 2-subjettiness, τ_2, distribution of the leading jet (left) and subleading jet (right).

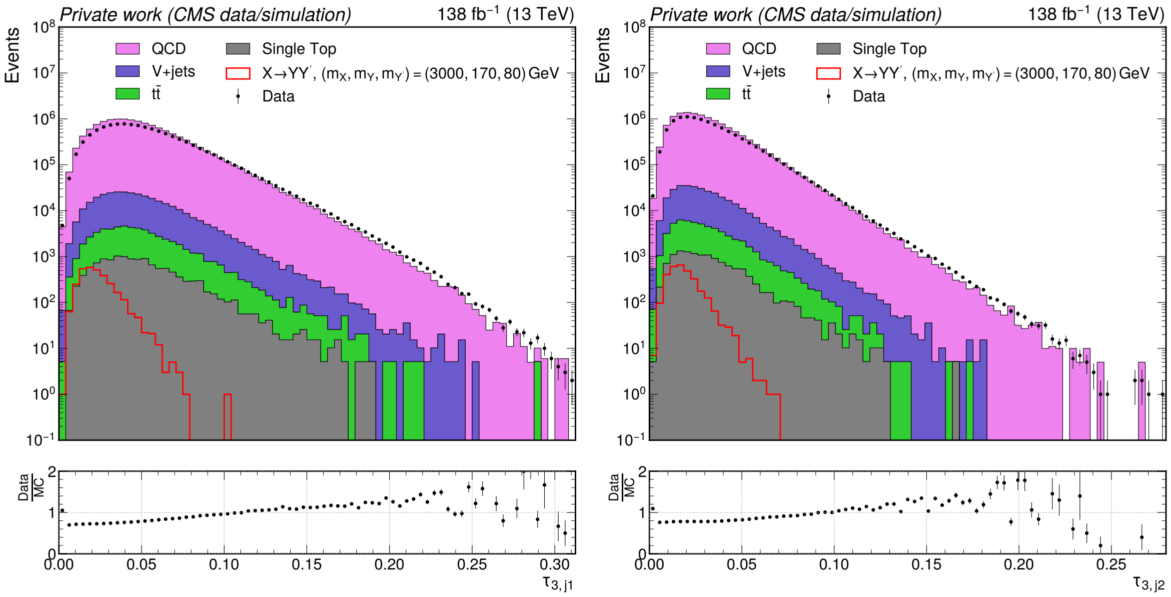

Figure 7.9: Analogously to Fig. 7.2, the 3-subjettiness, τ_3, distribution of the leading jet (left) and subleading jet (right).

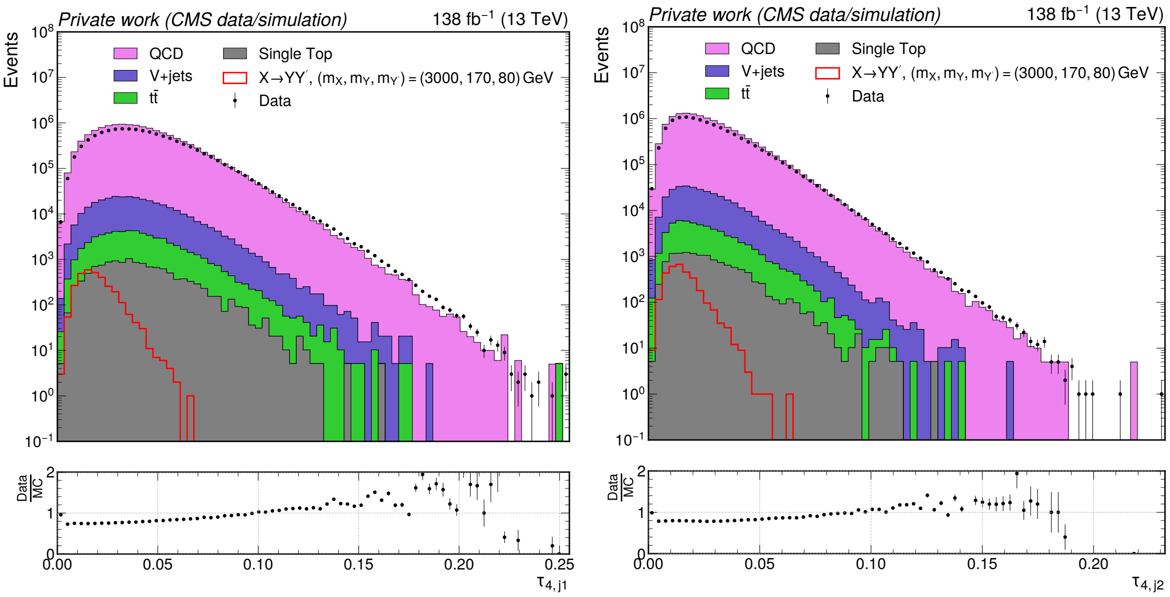

Figure 7.10: Analogously to Fig. 7.2, the 4-subjettiness, τ_4, distribution of the leading jet (left) and subleading jet (right).

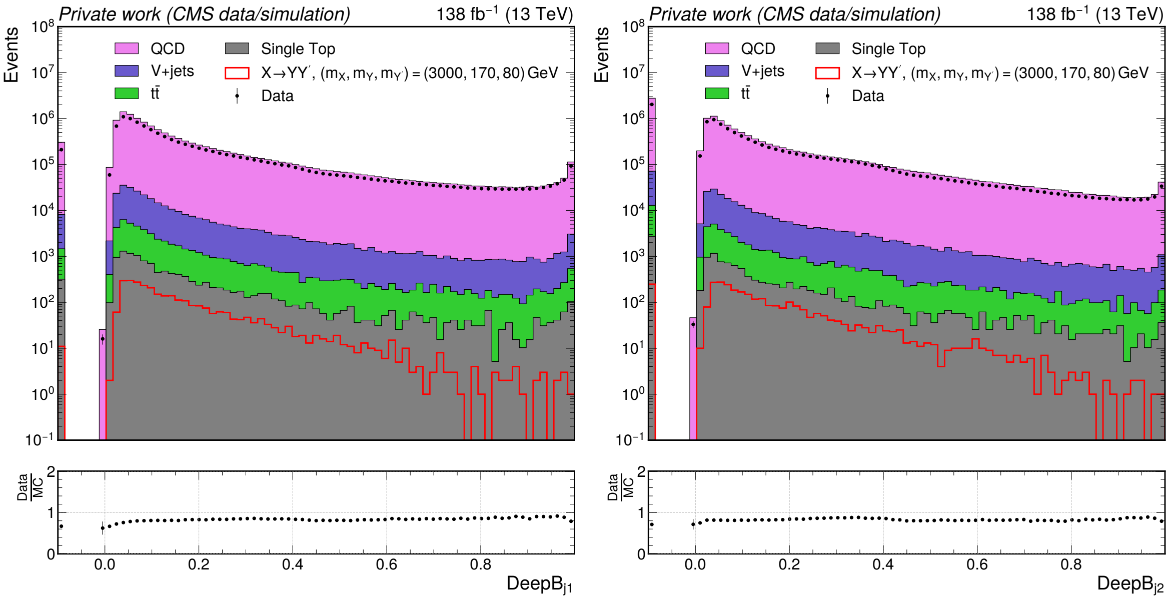

Figure 7.11: Analogously to Fig. 7.2, the b-tagging score, DeepB, distribution of the leading jet (left) and subleading jet (right). Jets with invalid classifier input, e.g., missing AK4 subjets, are displayed at a value of -0.1.

7.3 CATHODE Implementation

The overall implementation of CATHODE for this analysis is based on the proof-of-concept studies discussed in Sec. 6.2. Since the studies shown in Sec. 6.3 and 6.4 only finalized in a later stage, the improvements suggested in these sections could not be incorporated into the final analysis without causing a significant delay. Explicitly decorrelating the auxiliary features from the resonant one as well as the use of boosted decision trees for classification are an interesting avenue for future work.

As the signature of a dijet system with anomalous substructure in this analysis is equivalent to the proof-of-concept studies discussed in Sec. 6.2, the initial choice of input features was made analogously, i.e., the dijet invariant mass m_{jj} serves as a conditional feature m of the algorithm, and the auxiliary input was originally chosen as $x = \big(m_{j1}, \Delta m_j, \tau_{21,j1}, \tau_{21,j1}\big)$, with the only difference that the jets are labeled by descending mass: $m_{j1} \geq m_{j2}$. However, the choice of τ_{21} was found to be suboptimal for a generic search targeting a wide range of signal processes resulting in various numbers of subjets because τ_{21} is only sensitive to the presence of two subjets. In order to gain discrimination power for signal processes with more subjets, one would need to consider higher ratios, such as τ_{32} and τ_{43}. Including these to the feature vector x, on the other hand, would involve the presence of more uninformative feature for a given signal with a specific number of subjets, reducing the sensitivity as discussed in Sec. 6.4. An alternative for increasing sensitivity towards higher pronginess is to replace τ_{21} by τ_{41}. This ratio is expected to yield lower values for jets with two to four prongs. This choice was indeed found to result in higher sensitivity towards a broad range of different signal processes in early simulation-based studies, compared to using τ_{21} or including multiple ratios. The final auxiliary feature vector for the CATHODE implementation in this analysis is thus $x = \big(m_{j1}, \Delta m_j, \tau_{41,j1}, \tau_{41,j2}\big)$. The distributions of these features in the MC background simulation with an example signal region and sidebands are shown in Fig. 7.12.

Many hypothesized new physics models predict particles decaying to final states with b quarks, thus resulting in b jets in the detector. One example is the $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ model considered in the limit setting of this analysis. In order to gain sensitivity towards such b jet signal processes, the b-tagging score DeepB serves as a crucial handle. For other signal processes, without b jets, the DeepB score is expected to be fully uninformative and result in suboptimal sensitivity. In order to gain this b-tagging sensitivity, while retaining high sensitivity for non-b final states, an alternative feature set was introduced for performing an additional CATHODE implementation alongside the one based on x discussed above: $x_b = \big(m_{j1}, \Delta m_j, \tau_{41,j1}, \tau_{41,j2}, \mathrm{DeepB}_{j1}, \mathrm{DeepB}_{j2}\big)$. CATHODE with this feature set is referred to as CATHODE-b in the following, and all results are obtained with both strategies.

The DeepB distributions, shown in Fig. 7.11, consist of two parts: a continuous spectrum ranging from zero to one, and a large peak of events with values at exactly -1 (moved to a value of -0.1 for better visibility). This type of discontinuity is challenging to model with a normalizing flow, which is based on diffeomorphic transformations to a latent distribution, and thus the generated samples from a realistic normalizing flow will likely smoothen out the peak. This type of difference between data and background template would be detected by the classifier and ultimately yield false positives and negatives. In order to mitigate this, events in the single-value peak are associated with a new value, which is randomly drawn from a Gaussian distribution with mean value at the beginning of the continuous spectrum. The exact location is determined by measuring the difference between the first and the second unique value in the continuum d_{12}, and subtracting this from the first one. The standard deviation of the Gaussian distribution is set to d_{12}. This results in a very thin peak, but the subsequent logit preprocessing before training the normalizing flow will further smoothen it out.

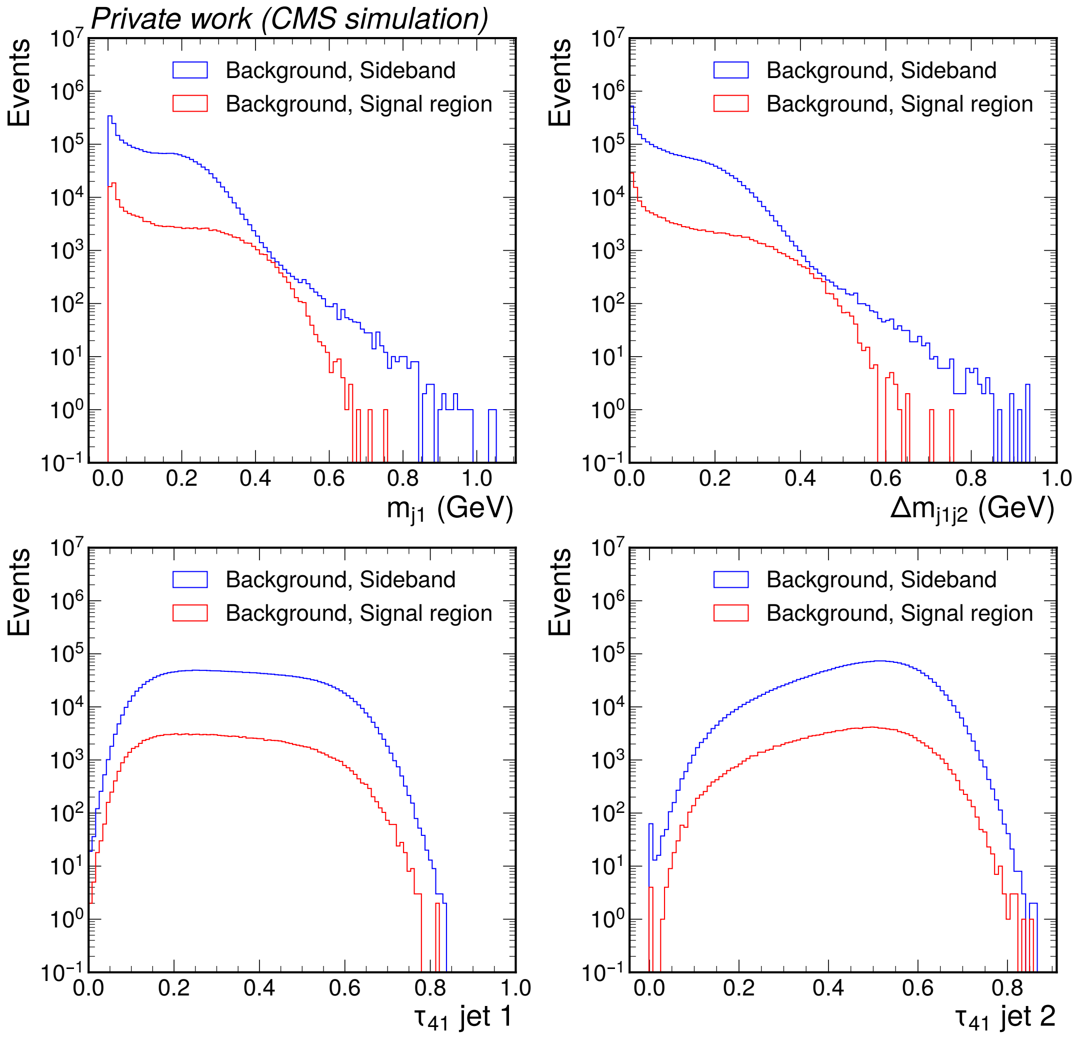

Figure 7.12: MC background simulation distribution of CATHODE auxiliary input features. The red line corresponds to an example signal region of 2230 to 2725 GeV and the blue line is the complementary sideband region.

The conditional normalizing flow model was chosen to be similar to the one used in the proof-of-concept studies in Sec. 6.2: the architecture consists of 15 MADE blocks with a single layer of 128 nodes and ReLu activation function, and each is followed by a feature permutation layer. They map the input features to a standard Gaussian latent space distribution with the same dimensionality. In contrast to the studies in previous sections, the flow was implemented with the Pyro library [207, 208], as it provides a numerically more stable training procedure. Another subtlety is that the transformation parameters in the MADE blocks depend on the latent space direction, which was simpler to implement in the Pyro framework. Thus, the flow is strictly speaking an inverse autoregressive flow. As in Sec. 6.2, the auxiliary input features are preprocessed with a logit transformation, followed by centering the distribution at zero with standard deviation of one. The model is trained with the Adam optimizer using a learning rate of 10^{-4}, and a batch size of 256. Three additional measures are applied to reduce overfitting and increase training stability: L_2 regularization (as introduced in Eq. 3.11) with $\lambda = 10^{-6}$, batch normalization [209] between MADE blocks with momentum of 0.1, and clipping the scaling transformation parameters in each MADE block ($\alpha_{\theta t}$ in Eq. 3.27) to lie within a range of $[-3, 3]$. Each training is performed for 100 epochs, which has been seen to be sufficient to cover the validation loss minimum.

A SR background template is created by sampling four times as many events as are present in the SR data, and distributed proportionally into training and validation sets for the classifier. In order to increase the stability with respect to per-epoch fluctuations, these background events are sampled in equal parts from the model states at the five epochs with the lowest validation loss. As conditional input to the SR background sampling, m_{jj} values are sampled from a KDE model that has been fitted to the training data. The KDE model is implemented with the `scikit-learn` library [160], using a Gaussian kernel and a bandwidth determined as $n^{-0.4}$, where n is the number of events in the training data. The KDE model is trained on the logit-transformed m_{jj} values, and the sampled values are transformed back to the original space.

The classifier architecture is fully analogous to Sec. 6.2, i.e., a fully connected neural network with three hidden layers with 64 nodes each, and ReLu activation function, whereas the output layer has a single node with sigmoid activation. It is implemented with the PyTorch library [159]. The model is trained with the Adam optimizer, a learning rate of 10^{-3}, a batch size of 128, and reweighting of the data and the background template to have equal total contributions in the training and validation sets. The training is performed for 100 epochs, and the model states at the ten epochs with the lowest validation loss are ensembled via the mean prediction per event for inference. It was found that, in contrast to MC simulation studies, the classifier training did not always fully converge on the collision data in the low end of the m_{jj} region because of the large number of events due to the exponential shape of the distribution. The training duration was nevertheless limited to 100 epochs to avoid overfitting on small mismodelings of the background template, which become more evident with large sample numbers. This can be better understood from the discussion in Appendix B.5.

A double k-fold cross-validation scheme is employed, as discussed in Sec. 3.2.2, in order to obtain separate training, validation, and test (inference) sets while using every data point for each of these tasks. The data are randomly partitioned into $k = 5$ subsets, and the training, validation, and test sets are rotated through these subsets, as illustrated in Fig. 7.13. For each of the k choices of test partition, the remaining $l = 4$ partitions are separated into a validation set (one partition) and a training set (three partitions), and this is performed for all l choices of validation partition. This results in $k \times l = 20$ different data splits $D^{k,l} = \{D^{k,l}_{\mathrm{train}}, D^{k,l}_{\mathrm{val}}, D^{k,l}_{\mathrm{test}}\}$, each corresponding to one individually trained normalizing flow and classifier model pair. Each $D^{k,l}$ is split into SR and SB contributions. The conditional normalizing flow for one $D^{k,l}$ is trained on the SB, and its epoch-ensembled background template is sampled in the SR to form training and validation sets for the classifier model, together with the respective $D^{k,l}_{\mathrm{train}}$ and $D^{k,l}_{\mathrm{val}}$. The l (epoch-ensembled) classifiers per test partition are ensembled via their mean prediction on each test set event. For each test partition, the threshold on the classifier prediction for selecting the most anomalous 1% of events in the SR is determined, and all events from SR and SB passing this threshold are merged into a single output dataset for statistical analysis.

In general, SIC curves in MC signal simulation studies suggest that the more restrictive the selection on the anomaly score, the higher is the resulting significance improvement. The motivation for choosing a selection efficiency of 1% is that it results in a sufficient number of events to perform a statistical analysis at every considered m_{jj} region. Finding a more optimal selection efficiency while retaining a model-agnostic approach is left for future work.

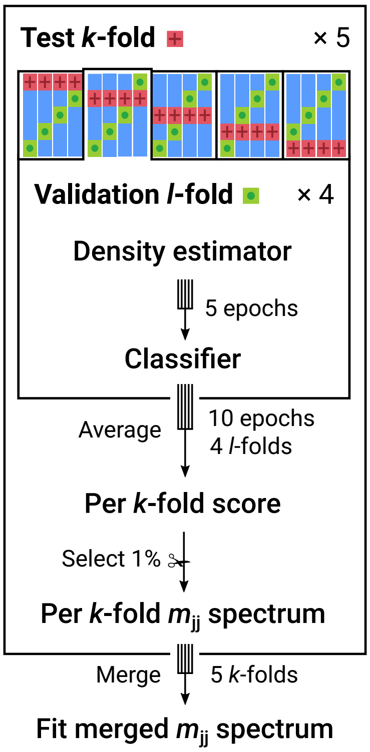

Figure 7.13: Illustration of the k-l-fold cross-validation procedure and epoch ensembling strategy. Figure courtesy of Louis Moureaux.

7.4 Statistical Analysis

The search for dijet resonances with CATHODE is performed over a range of m_{jj} from 1,650 to 5,500 GeV in steps of 100 GeV. This step size is well below the observable decay width, which is dominated by the detector resolution in the case of narrow-width signal processes, as shown later in Sec. 7.4.2. Since the CATHODE protocol relies on an a-priori choice of SR and SB, the m_{jj} range is divided into bins, where each bin constitutes the SR for one double k-folded CATHODE training, as discussed in Sec. 7.3, with the complementary m_{jj} space as SB. Two overlapping sets of bins are chosen as shown in Tab. 7.6. Odd bin numbers are adjacent to each other, covering the full m_{jj} range, while even bin numbers are chosen to overlap with odd ones, each covering half of the m_{jj} range of each of their odd neighbors. This second set of bins aims to capture potential signals that peak on the boundary between two odd bins and might thus be missed. The increasing bin widths have been designed to yield approximately even spacing in $\log m_{jj}$, and such that all SRs, as well as the corresponding upper and lower SBs, contain enough events ($\mathcal{O}(10^3)$) to allow a robust ML training and statistical analysis. The region from the m_{jj} analysis threshold of $1,455$ GeV to the first lower bin boundary, as well as the region beyond the last upper bin boundary of $5,500$ GeV, are only used as SB, in order to avoid training CATHODE with one-sided SBs. Table 7.6 indicates which signal mass hypothesis in the later statistical analysis corresponds to which SR bin. The underlying CATHODE trainings, and the resulting most anomalous event selection, are the same between multiple signal masses in the same SR bin. In order to avoid statistical analyses of signal hypotheses that are too close to the bin boundaries, the signal masses are tested within a more restricted range of 1800 GeV to 5000 GeV.

Table 7.6: Binning choice for scanning CATHODE and the subsequent statistical analysis over the full dijet invariant mass range.

Bin number	Range (GeV)	Signal masses (GeV)	Number of data events
1	1650–2017	1800, 1900	4,500,000
2	1824–2230	2000, 2100	2,100,000
3	2017–2465	2200, 2300	1,400,000
4	2230–2725	2400, 2500	630,000
5	2465–3013	2600, 2700, 2800	400,000
6	2725–3331	2900, 3000, 3100	170,000
7	3013–3682	3200, 3300, 3400, 3500	100,000
8	3331–4071	3600, 3700, 3800	42,000
9	3682–4500	3900, 4100, 4200, 4300	22,000
10	4071–4975	4400, 4500, 4600, 4700	8500
11	4500–5500	4800, 4900, 5000	3900

7.4.1 Background Estimation

The background estimation is performed via a bump hunt, i.e., the procedure is based on fitting the selected m_{jj} distribution with a parametric function, as described in Sec. 2.6.3. As the optimal choice of functional form is not known a priori, the fit is performed using multiple choices with increasing flexibility, and a criterion is constructed to decide which one to use as background model in the hypothesis tests. Since the selected background shape might change between CATHODE trainings at various SR positions, the data-driven background estimation procedure is performed separately for each choice of SR.

The overwhelming majority of background events are expected to arise from QCD multijet interactions, whose distribution is known to have a smooth and monotonically decreasing behavior with respect to the dijet invariant mass. These properties are assumed in the choice of family of fit functions, which are as follows:

$$\text{3 parameters:} \quad \left(\frac{dN}{dm_{jj}}\right)_3 (x) = p_0 \frac{(1-x)^{p_1}}{x^{p_2}} \tag{7.2}$$

$$\text{4 parameters:} \quad \left(\frac{dN}{dm_{jj}}\right)_4 (x) = p_0 \frac{(1-x)^{p_1}}{x^{p_2+p_3\log(x)}} \tag{7.3}$$

$$\text{5 parameters:} \quad \left(\frac{dN}{dm_{jj}}\right)_5 (x) = p_0 \frac{(1-x)^{p_1}}{x^{p_2+p_3\log(x)+p_4\log^2(x)}}. \tag{7.4}$$

The variable $x \equiv m_{jj}/\sqrt{s}$ is a rescaling of the dijet invariant mass into units of the center-of-mass energy, $\sqrt{s} = 13\,\text{TeV}$. The parameter p_0 describes the normalization of the probability density function, while p_1, p_2, p_3, and p_4 describe its shape. All p_i are treated as freely floating parameters during the fit. The choice of this specific functional form is an extension of the fit function used in the dijet resonance search of Ref. [142]. It is motivated by the mass dependence of lowest order QCD cross section calculations. The full selected m_{jj} distribution is fitted with each parameter number choice, i.e., the signal region is not explicitly excluded from the background shape fit. This tends to result in a more stable background prediction, but might bias the selection of the background shape towards false negatives, which is the more conservative direction. The fit is performed via a binned maximum likelihood optimization using the RooFit [210] software package with the MINUIT2 [211] minimization library, setting the optimizer to MIGRAD and estimating errors with MINOS.

For each choice of parameterization, the goodness of fit is quantified by the χ^2 metric:

$$\chi^2 = \sum_{\text{bins } i} \frac{(N_i - N_i^{\text{fit}})^2}{\sigma_i^2}, \tag{7.5}$$

where N_i is the number of observed events in the bin number i, N_i^{fit} is the number of events predicted by the fit, and σ_i is the uncertainty in the observed number of events. More parameters in the fit function will generally result in lower χ^2 (better goodness of fit). The Fisher F-test [212] is used as a criterion to decide whether the improvement in goodness of fit is significant enough to warrant the inclusion of additional parameters. The F-test statistic comparing two function choices, a and b, is defined as:

$$F_{ab} = \left(\frac{\chi_a^2 - \chi_b^2}{n_{\text{dof},a} - n_{\text{dof},b}} \right) / \left(\frac{\chi_b^2}{n_{\text{dof},b}} \right), \tag{7.6}$$

where the number of degrees of freedom n_{dof} is computed as the number of bins minus the number of function parameters. Starting from the 3-parameter function choice, Eq. 7.2, the F-test statistic is computed with respect to the function with one more parameter. Under the null hypothesis that the simpler function is sufficient, F_{ab} follows an F-distribution. If the resulting observed p-value is found lower than a chosen confidence level of 10%, the gain in goodness of fit is considered significant and the more complex function is chosen. The procedure is repeated until the null hypothesis is accepted, i.e., the simpler function is chosen. In some instances, the inclusion of more parameters can lead to correlations that result in unreasonably large fit uncertainties. In order to avoid these, fit functions are excluded if their relative error at the tested resonance mass is larger than 50%.

Two different choices of m_{jj} binning are used for the procedure discussed above. The background fit itself is performed with a *fine binning* that has a constant width of 4 GeV, which approximates the limit of an unbinned likelihood fit without the computational burden of modeling every data point individually. The range is chosen to start at 1,460 GeV and ranging either until the maximum observed m_{jj} in the data, or 1.2 times the hypothesized signal mass, whichever is larger. The F-test relies on the χ^2 value, which is only well described with a sufficient number of entries in each bin. To this end, the second binning choice, also referred to as the *dijet binning*, was introduced. The dijet bin boundaries are as follows:

$$\begin{aligned} \text{dijet binning} = [&1460, 1530, 1607, 1687, 1770, 1856, 1945, 2037, 2132, 2231, 2332, 2438, \\ &2546, 2659, 2775, 2895, 3019, 3147, 3279, 3416, 3558, 3704, 3854, 4010, \\ &4171, 4337, 4509, 4700, 4900, 5100, 5300, 5500, 5800, 6100, 6400, 6800]\ \text{GeV}. \end{aligned} \tag{7.7}$$

They have been chosen in previous dijet searches, such as Ref. [213], to roughly accommodate the detector resolution, and they increase in width to account for the exponentially small number of events at larger m_{jj}. Bins with fewer than five events are iteratively merged with the next higher bin, starting from the least populated one. This choice of binning is also used for visualization in the figures of this thesis.

Different choices of background fit functions might lead to different results in the statistical analysis. This potential bias by sampling toy experiments from one fit function choice and fitting with another is tested in Sec. 7.7.3.

7.4.2 Signal Modeling

In order to perform binned likelihood–based hypothesis tests, as described in Sec. 2.6.1, the signal shape needs to be modeled across multiple bins[13]. Just as for the background, this is handled in this analysis by fitting a parametric function. The significance computation as a function of the resonance mass is aimed to be as model agnostic as possible. Thus, the corresponding signal hypothesis is obtained by selecting a representative shape, motivated primarily by generic detector resolution features. For the limit setting, the test is model specific in nature and thus the signal shape is chosen to be obtained directly from the MC simulation of each tested signal.

In either case, the signal shape is modeled as a Double Crystal Ball (DCB) function [214], which is a combination of a Gaussian core and two power law tails on each side. It is defined as:

$$f_{\mathrm{DCB}}(x) = \mathcal{N} \begin{cases} A_1(B_1 - \frac{x-\mu}{\sigma})^{n_1} & \frac{x-\mu}{\sigma} \leq \alpha_1 \\ \exp(\frac{(x-\mu)^2}{\sigma^2}) & \alpha_1 < \frac{x-\mu}{\sigma} < \alpha_2 \\ A_2(B_2 - \frac{x-\mu}{\sigma})^{n_2} & \frac{x-\mu}{\sigma} \geq \alpha_2, \end{cases} \tag{7.8}$$

where the parameters A_i and B_i are defined as

$$A_i = \left(\frac{n_i}{|\alpha_i|}\right)^{n_i} \exp\left(-\frac{\alpha_i^2}{2}\right), \tag{7.9}$$

$$B_i = \frac{n_i}{|\alpha|} - |\alpha|. \tag{7.10}$$

The parameters μ and σ describe the mean and width of the Gaussian core, respectively. The power laws are controlled by the location of the transition points, α_1 and α_2, and the power law slopes, n_1 and n_2. These shape parameters will be held fixed in the later profile likelihood fit. The normalization parameter $\mathcal{N}$, on the other hand, will remain a freely floating parameter, related to the signal cross section. In Eq. 7.8, x corresponds to the independent variable, which is m_{jj} in this case. The Gaussian core is motivated by the finite detector resolution, and the power law tails capture the effects of the parton distribution functions. Similar to the background function, the DCB parameter values are obtained by fitting the signal MC simulation m_{jj} values via a binned maximum likelihood optimization in bins of 4 GeV and using the RooFit software package with the MINUIT2 optimizer and MIGRAD.

The generic signal template for the significance scan is obtained by fitting the $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal, with intermediate particle masses of $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$ and $m_{\mathrm{Y}'} = 80\,\mathrm{GeV}$. This signal is chosen because the decay products are seen to be fully contained within the AK8 jets, resulting in a generic shape that is representative of the detector resolution. In other MC signal simulation samples, where the decay products are partially not captured by the AK8 jets, the shape is distorted towards longer tails. The $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ MC simulation samples have been produced at masses of 2, 3, and 5 TeV, and the corresponding fits are shown in Fig. 7.14. Since the signal MC simulation sample has been produced with a negligible decay width, the resulting shapes confirm that the significance scan step size of 100 GeV is smaller than the detector resolution and would thus capture any physical peak.

The intermediate resonance masses are tested by interpolating the parameters of the three fitted DCB shapes. The interpolation is performed with a third-degree spline for μ and σ, and linear interpolation for the power law parameters A_1, B_1, A_2, and B_2. This interpolation

[13] A multi-bin signal shape model could be avoided by performing a “counting analysis”, i.e., computing the likelihood function from Eq. 2.22 in a single bin. However, this would greatly reduce the statistical power with respect to the “shape analysis” that exploits well-motivated information on bin-to-bin correlations.

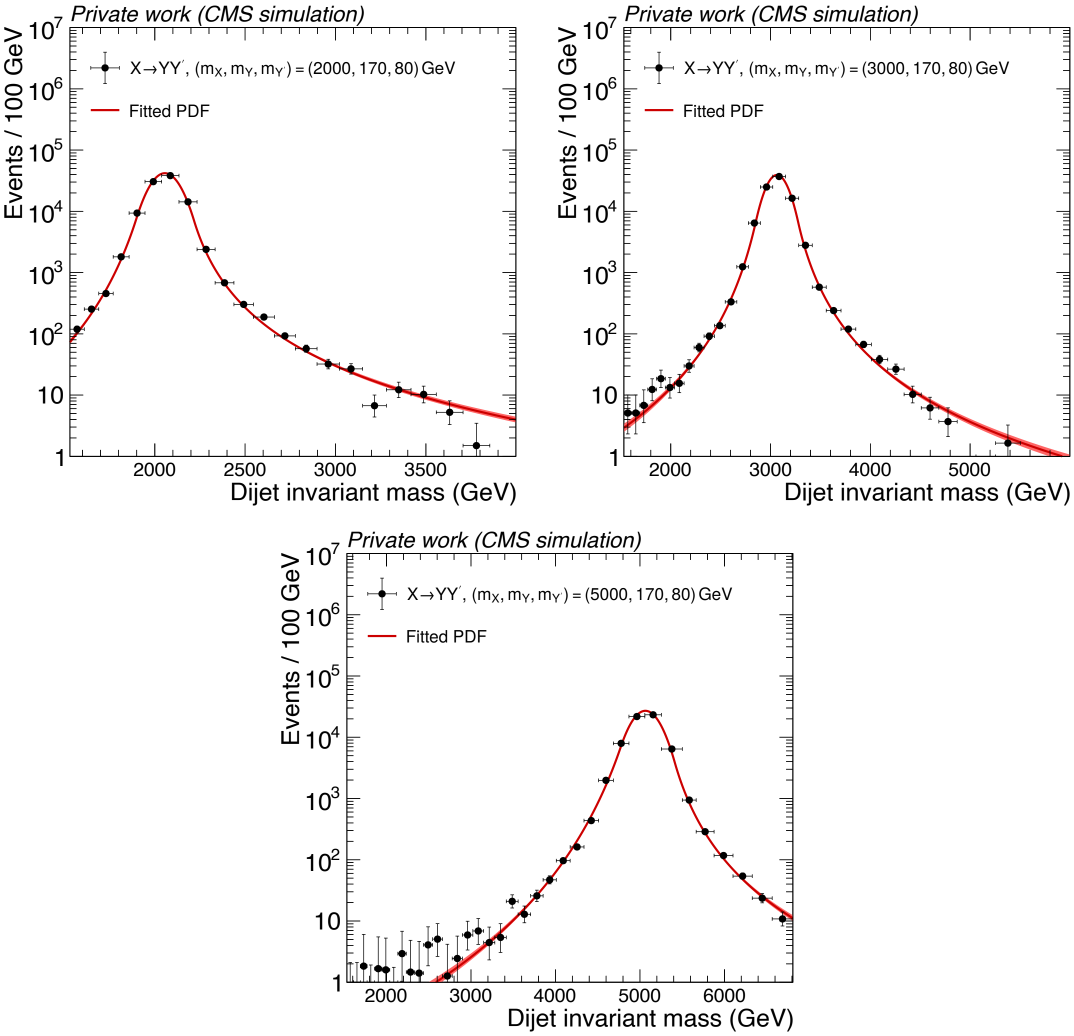

Figure 7.14: Signal shape template obtained from fitting the X $\rightarrow$ YY$'$ $\rightarrow$ 4q signal MC simulation samples with intermediate resonance masses of $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$ and $m_{\mathrm{Y'}} = 80\,\mathrm{GeV}$ at resonance masses of 2 TeV (top left), 3 TeV (top right), and 5 TeV (bottom), using a Double Crystal Ball function.

scheme is similar to the procedure used in Ref. [215]. The interpolation function is also slightly extrapolated to the lower signal masses of 1.8 and 1.9 TeV. The resulting signal shapes for all tested resonance masses are shown in Fig. 7.15. The validity of the interpolation scheme was confirmed by performing it without the 3 TeV mass point and comparing the interpolated shape to the explicit 3 TeV signal shape, which was subject to only minor differences.

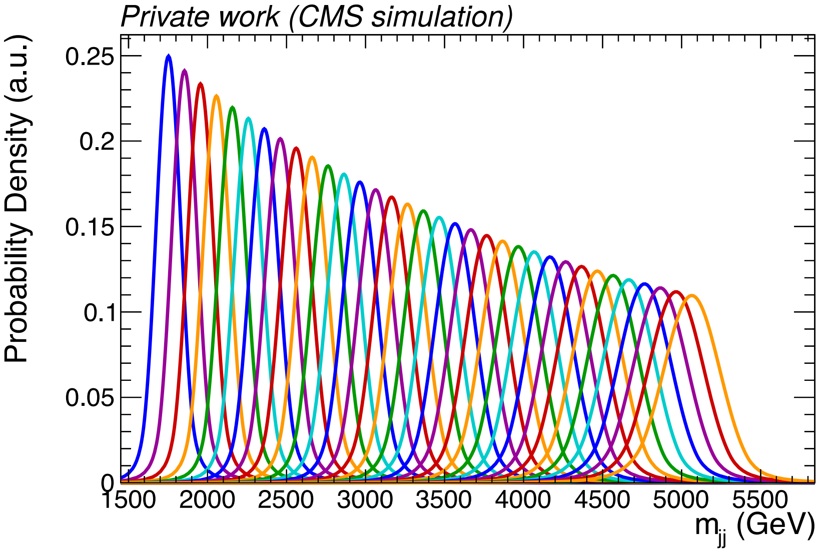

Figure 7.15: Interpolated Double Crystal Ball signal shapes for all considered resonance mass hypotheses in the significance scan.

Since the signal shape templates for the limit setting are obtained directly from the MC simulation of the respective signal hypothesis, no interpolation is necessary in this case. The signal shape template fits are shown for a representative set of signal processes in the Appendix B.3.

The statistical analysis is performed after selecting the most anomalous events via CATHODE. It needs to be shown that the signal shape, obtained as described above on the signal MC simulation samples without CATHODE-based selections, is a valid proxy for the shape after selections. Figure 7.16 demonstrates this check on the $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ ($m_{\mathrm{Y}} = 170\,\mathrm{GeV}$, $m_{\mathrm{Y'}} = 80\,\mathrm{GeV}$) signal for resonance masses of 2, 3, and 5 TeV. The signal is injected into the MC background simulation with cross sections of 61.75 fb, 14.52 fb, and 1.95 fb, respectively, and the full double k-folded CATHODE procedure is trained on these pseudo data. The figure shows the resulting signal shape after selecting the most anomalous events at various selection efficiencies and then performing the parametric fit. Only minor differences are observed, which confirms the validity of modeling the signal shape via the pre-CATHODE distributions. This greatly simplifies the analysis procedure, e.g., by ensuring a sufficient number of signal events in every shape fit. Moreover, a non-trivial shape distortion via the CATHODE selection would likely have resulted in a more complicated mass dependence, which is more difficult to model in the interpolation scheme.

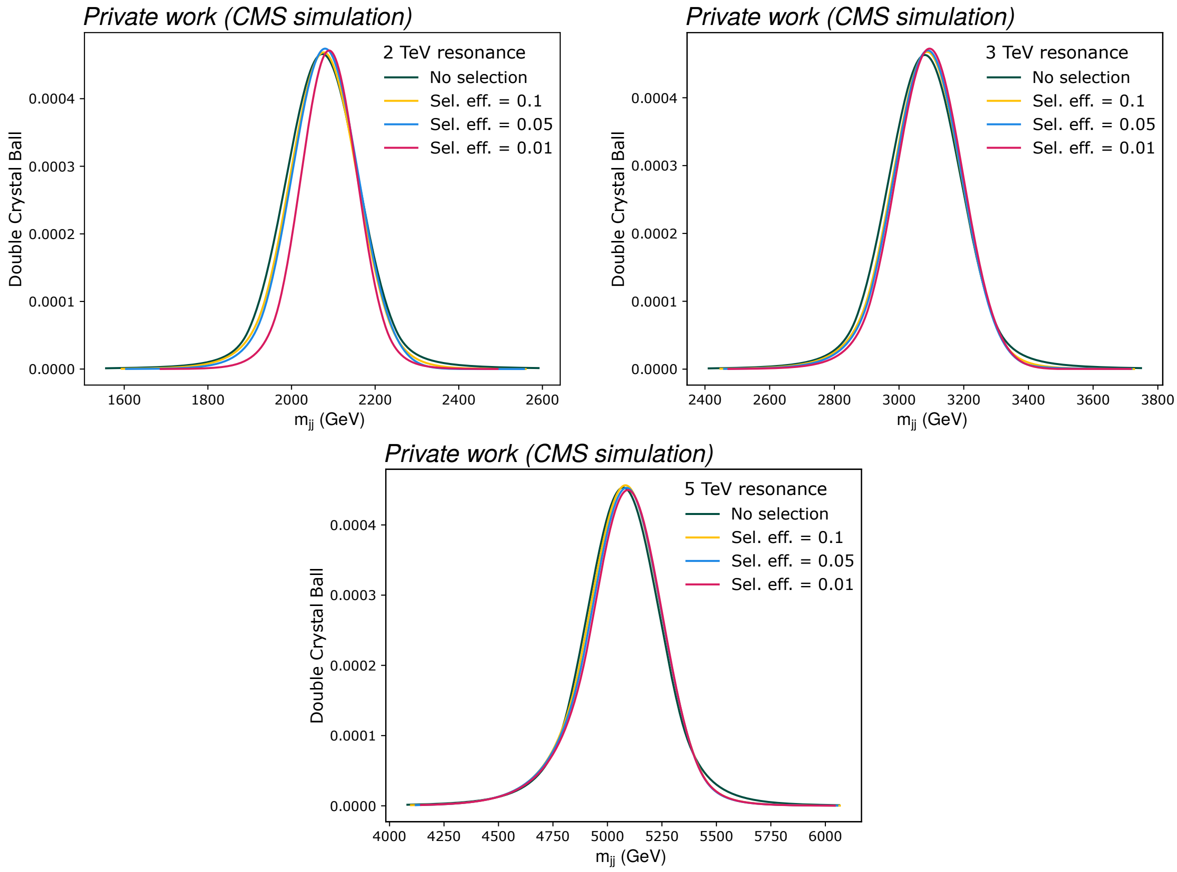

Figure 7.16: The Double Crystal Ball signal shape obtained from fitting the data before and after applying a CATHODE-based anomaly selection, with different choices of selection efficiency. The results are obtained by injecting 61.75 fb of the $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ ($m_{\mathrm{Y}} = 170\,\mathrm{GeV}$, $m_{\mathrm{Y}'} = 80\,\mathrm{GeV}$) signal at 2 TeV (top left), 14.52 fb of the otherwise same signal at 3 TeV (top right), and 1.95 fb at 5 TeV (bottom) into the MC background simulation.

7.4.3 Hypothesis Tests

The hypothesis tests for computing the significance of excesses and setting exclusion limits follow the frequentist paradigm, using profile likelihood ratios, as described in Sec. 2.6.1. The test statistic for quoting significances is the one defined in Eq. 2.31, while exclusion limits are based on Eq. 2.34. The test statistic distributions are obtained via the asymptotic formulae from Ref. [66], involving Asimov datasets. The treatment of nuisance parameters in the statistical model is discussed in Sec. 7.5.

The profile likelihood fits are implemented with the Combine tool [216] using the options `-M Significance --usePLC` for significance computation and `-M AsymptoticLimits` for upper limits. The same range and binning are used as for selecting the background functional form in Sec. 7.4.1.

The resulting exclusion limits are determined via the CL_s method, as discussed in Sec. 2.6.2, with a confidence level of 95%. The output of the Combine tool is a set of upper limits, observed and expected with bands, on the signal strength parameter μ (Eq. 2.36). By multiplying these values with the input signal shape normalization, they can be interpreted as upper limits on the number of signal events. The input signal normalization is a hyperparameter, which has been chosen to result in numerically stable fits, as well as resulting in no significant numerical changes to the output when varied up and down by an order of magnitude. The conversion of the limit on the number of events to a limit on the model cross section will be discussed in Sec. 7.6, as the weak supervision paradigm requires additional considerations.

7.5 Systematic Uncertainties

Systematic uncertainties are treated as nuisance parameters in the profile likelihood fits of the hypothesis tests in Sec. 7.4.3, meaning that they are floating parameters with additional constraints. Equations 2.25 and 2.26 showed how a nuisance parameter following a Gaussian distribution can be included in the likelihood function.

7.5.1 Calibration and Scale Factors

Many calibration procedures for correcting differences between MC simulation and real data involve event weights. When counting events, e.g., when computing a histogram of an observable, these weights are multiplied with each event count of the MC simulation. These weights primarily arise from *scale factors* (SFs), which are the ratio between the efficiency of some selection in data and simulation:

$$\mathrm{SF} = \frac{\varepsilon_{\mathrm{data}}}{\varepsilon_{\mathrm{MC}}}, \tag{7.11}$$

where the efficiency ε is the fraction of events passing the selection in question. A common example is the tagging efficiency of b-tagging algorithms. The SFs are usually measured as a function of kinematic variables, such as p_{T} and η. An often-used example for translating the measured b-tagging SFs to an event weight is by computing the probabilities of a given configuration of jets in data (P_{data}) and simulation (P_{MC}), then using the ratio of these probabilities as the MC event weight w_{MC} [217]:

$$P_{\mathrm{MC}} = \prod_{i:\ \text{selected jets}} \varepsilon_{\mathrm{MC},i} \prod_{j:\ \text{unselected jets}} \left(1 - \varepsilon_{\mathrm{MC},j}\right), \tag{7.12}$$

$$P_{\mathrm{data}} = \prod_{i:\ \text{selected jets}} \mathrm{SF}_i \varepsilon_{\mathrm{MC},i} \prod_{j:\ \text{unselected jets}} \left(1 - \mathrm{SF}_j \varepsilon_{\mathrm{MC},j}\right), \tag{7.13}$$

$$w_{\mathrm{MC}} = \frac{P_{\mathrm{data}}}{P_{\mathrm{MC}}}. \tag{7.14}$$

Every SF measurement is subject to statistical and systematic uncertainties, which are commonly propagated to the event weights by varying the SFs within their uncertainties and recomputing the event weights.

Another type of calibration is the direct change in the observable itself, e.g., by applying jet energy corrections. Uncertainties on these calibrations are typically propagated by varying the source of uncertainty within its bounds and observing the resulting change in the observable.

Both types of calibrations are applied to the signal MC simulation samples, and the resulting uncertainties are propagated to the signal shape and normalization in the likelihood function, as described in the following sections. The full list of considered systematic uncertainty sources is provided in Sec. 7.5.5.

7.5.2 Background Treatment

The background estimation is fully data driven, and thus there are no model-specific uncertainties to consider. The parameters of the background fit function (Sec. 7.4.1), both shape and normalization, are treated as nuisance parameters with uniform priors, which effectively translates to omitting a constraint term $\pi(m, \theta)$ in Eq. 2.26, considering that the parameter boundaries are chosen to be wide enough to not affect the fit result.

7.5.3 Signal Shape Uncertainties

Since the signal shape (Sec. 7.4.2) is modeled according to the detector response to the presence of a narrow-width resonance, it is subject to uncertainties affecting said response. In particular, the *jet energy scale* (JES) and *jet energy resolution* (JER) are determined in the jet energy calibration procedure, including their systematic uncertainties [218, 219]. While the JES affects the peak position of the resonance shape, the JER impacts the width of the detector response.

These effects are propagated by varying the jet energy calibration within expected bounds and observing the resulting effect in the fitted DCB signal shape in m_{jj}. Performing this procedure with the representative $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal at 2, 3, and 5 TeV, the observed variance in the DCB parameters μ and σ are $\delta_\mu = 1\%$ and $\delta_\sigma = 3.5\%$, respectively. For other signals, the resulting δ_μ and δ_σ are found to be comparable, and thus the same values are used for both the significance scan and the limit setting. They are treated as shifts to the nominal parameter values in the DCB shape, Eq. 7.8: $\mu \to \mu + \theta_\mu$ and $\sigma \to \sigma + \theta_\sigma$, where θ_μ and θ_σ are Gaussian-distributed nuisance parameters with mean of zero and standard deviations of δ_μ and δ_σ, respectively. This is handled via Gaussian constraint terms as in Eq. 2.25.

7.5.4 Signal Normalization Uncertainties

Since the exclusion limits are set based on the signal strength in the profile likelihood fit, it is crucial to consider systematic uncertainties in the signal normalization. Multiple uncertainty sources are considered and combined into a single signal normalization uncertainty δ_{sig} and treated by multiplying the signal strength by $\delta_{\mathrm{sig}}^{\nu}$, where the exponent ν is a Gaussian-distributed nuisance parameter with mean of zero and standard deviation of one. This constraint is implemented again via a Gaussian constraint term in the likelihood function. This treatment of rate uncertainties via a log-normal distribution is a common practice in HEP analyses, as it ensures that the signal strength remains positive definite.

The propagation of systematic uncertainty sources towards the combined signal normalization uncertainty for limit setting involves retraining and/or reevaluating CATHODE classifiers with signal simulation samples whose distributions are varied within uncertainties. This will be discussed in Sec. 7.6. The respective uncertainty sources are listed in the following.

7.5.5 List of Systematic Uncertainty Sources

- **Jet substructure modeling**: The AK8 jet substructure variables used for training CATHODE are not perfectly modeled in the MC simulation. The standard method of calibrating the substructure of large-radius signal jets would be to compare them to SM jets with the same number of prongs. However, we are setting limits on signal models with more prongs than are available in the SM (with a sufficient number of events), i.e., more than the three subjets from boosted top decays. Therefore, a novel method [220][14] was developed to compute a per-prong substructure correction weight via the Lund jet plane [222] ratio between data and simulation. In essence, the Lund plane captures the showering history

[14] The Lund plane ratio reweighting method was first publically described in Ref. [220]. It has been updated in the meanwhile, in particular concerning the treatment of uncertainties, and is currently under review within the CMS Collaboration. The current handling of uncertainties reflects the state described in the internal analysis note [221].

of a jet. In this case, it is derived for reclustered subjets within the AK8 jet that should be representative of single quarks (one prong). The resulting weights are then applied to generic multi-prong resonances. These weights and their uncertainty can thus be translated to a scale factor for any substructure-based selection. The uncertainty in the Lund plane ratio reweighting method is decomposed into several individual contributions:

  - The combination of systematic uncertainties during the measurement of the Lund jet plane ratio.
  - The Lund plane ratio is derived in a flavor-agnostic manner via the two subjets within boosted W jets, but the showering pattern of b quarks is known to be different from lighter quarks. An additional correction is derived via a b-to-$\mathrm{q}_{\mathrm{light}}$ Lund jet plane ratio, which is then used as uncertainty estimate for signal b subjets based on generator-level information. It is propagated to the substructure weights by scaling the uncertainties up (multiplication with b-to-q_{light} ratio) and down (multiplication with q_{light}-to-b ratio).
  - The statistical uncertainty in the Lund plane ratio measurement, implemented by sampling 100 new values per Lund jet plane bin from a Gaussian distribution centered at the nominal value and with the Poisson uncertainty as width.
  - The Lund plane ratio is computed as a function of the subjet p_{T}, and it is extrapolated to values outside the measured range. Similar to the statistical uncertainty, 100 new values are sampled per Lund jet plane bin from a Gaussian smearing of the nominal value with the functional fit uncertainty as width.
  - While the number of prongs is a signal process–dependent input to the reclustering scheme of the AK8 jet, there is some ambiguity associated with this choice if the quarks cannot be correctly captured with reclustered subjets. This is relevant if quarks fall close to the boundary of the AK8 jet as parts of their radiation will leak in or out of the jet cone, or if two generator-level quarks are associated to the same subjet. In these cases, a variation of the number of prongs up and/or down by one is considered for the weights derivation and these are used as up and down variations in the signal weights, respectively.
  - If the reclustering algorithm fails to match a close subjet to a generator-level quark, this quark cannot be calibrated with the Lund plane ratio weights. In this case, a conservative uncertainty is applied by varying the event weight up and down by a factor of 5, which is the maximum correction factor.
  - The Lund plane ratio method is based on the assumption that the showering pattern of SM quarks is independent of their source. However, this assumption can be violated by multiple factors, such as color reconnection effects. To account for this, the Lund plane ratio is measured for each signal model and compared to the MC simulation of boosted W jets, for which the correction was derived. The ratio of these two measurements is used as an up and down variation of the signal weights.

- **ML training variance**: The neural network classifier predictions are subject to variance from one training to another. Two sources of variance are considered. First, there is the *intrinsic variance* from retraining the same model on the same data with different random initializations of the network weights. Second, the limit setting procedure, outlined in Sec. 7.6, involves retraining the classifier model on small injections of signal MC simulation events into the data. Because the signal presence in the training data is very small, the resulting classifier predictions will depend on which specific signal events are randomly included in the training set. In order to account for both effects simultaneously, we perform five retrainings of the neural network model with different signal event samples and random initializations of the network weights. When computing the signal efficiency of the anomaly selection, the mean is used as the nominal value and the standard deviation as the uncertainty.

- **Jet energy/mass scale and resolution**: Uncertainties on the jet energy scale and resolution from the jet energy calibration scheme are treated by varying the 4-momentum vectors of both leading AK8 jets within uncertainties and recording the observed change in analysis observables. This results in a different shape and normalization of the signal model. Similarly, we apply a jet mass scale uncertainty of 5% and jet mass resolution uncertainty of 8%, following the recommendations of the JetMET Physics Object Group for W jet tagging [223].

- **Pileup reweighting**: The simulation is corrected for the number of pileup interactions via event weights, based on the predicted cross section of inelastic collisions. This cross section is varied up and down by its uncertainty of 4.6% to compute the event weights for the systematic uncertainty.

- **Parton distribution functions**: The MC event generators use a specific set of parton distribution functions to model parton-parton interactions. These are based on fits to experimental data. Changes of the PDFs within fit uncertainties are provided via 100 event weights. We compute the standard deviation of these weights for each event, and vary the nominal weight up and down by this standard deviation in order to obtain a reduced set of event weights capturing the PDF uncertainty.

- **Renormalization and factorization scale**: The theoretical computation of the signal cross section is performed at a fixed order in perturbation theory, which introduces a dependence on the renormalization and factorization scales (μ_R and μ_F in Eq. 2.9). The uncertainty in the signal cross section is estimated by the conventional method of varying these scales up and down by a factor of two, and obtaining event weights accordingly. The varying is performed for each variable separately, as well as in unison to capture correlations between the scales. The unphysical case of anti-correlations is not considered.

- **Parton shower modeling**: Similar to the fixed-order cross section computation, also the parton shower modeling of MC event generators depends on the renormalization scale through the strong coupling constant α_s [52]. An uncertainty is obtained via weights that arise from varying the renormalization scale up and down by a factor of two. The impact factorizes into initial state radiation and final state radiation effects. Only the former is considered in this analysis, as the latter is redundant with the Lund jet plane ratio reweighting.

- **Top quark transverse momentum reweighting**: The top quark p_T spectrum in the simulation is known to be mismodeled. Correction weights, as well as up- and down-scaled weights, are derived from the difference between unfolded data and parton-level information in the MC event generator. The procedure is outlined in Ref. [224].

- **L1 trigger prefiring**: In 2016 and 2017, the L1 trigger system was subject to a slowly increasing shift of reconstructed cluster time in the ECAL, in particular at high p_T and η. This resulted in some clusters being assigned to the wrong LHC bunch crossing and thus the trigger firing on the wrong event [225]. The resulting loss of trigger efficiency is not present in the simulation, so an event weight is computed from the probability of each AK4 jet or photon to cause this prefiring, according to the recommended procedure by the L1 Detector Performance Group [226]. An uncertainty is estimated via weights from varying the probabilities up and down by the uncertainty of 20%.

- **B-tagging corrections**: CATHODE-b relies on the DeepCSV b-tagging algorithm for increasing the sensitivity for signals involving b quarks. The difference between data and MC simulation efficiencies is corrected by scale factors, which are used to derive event weights that correct the full shape of the distribution in MC simulation, according to the treatment recommended by the BTV Physics Object Group for shape recalibration [227]. The scale factors are subject to statistical and systematic uncertainties. We add these in quadrature to obtain a single total uncertainty that is propagated to a set of up- and down-scaled weights. These weights are normalized such that only the shape changes before training CATHODE-b, but not the normalization.
- **Luminosity**: The total integrated luminosity of the Run 2 data-taking period is known with a precision of 1.6% [228–230]. This can be directly applied as a normalization uncertainty in the signal yield. However, compared to the other uncertainties in the combination, it is found to be negligible for every considered signal model.

7.6 Weakly Supervised Limit Setting

The statistical treatment, described in Sec. 7.4.3, results in an upper limit on the number of events N_{exc} arising from the signal process in question. The number of signal events usually translates to a signal cross section through a rescaling with the integrated luminosity L_{int} and the signal efficiency. The latter factorizes into two parts. The first one is the signal acceptance $\varepsilon_{\mathrm{pre}}$, which is the fraction of signal events that pass all the analysis preselections. The second part is the tagging efficiency $\varepsilon_{\mathrm{tag}}$, which corresponds to the fraction of pre-selected signal events that pass further signal-enhancing selections, in this case a selection based on the CATHODE classifier output:

$$\sigma_{\mathrm{exc}} = \frac{N_{\mathrm{exc}}}{L_{\mathrm{int}} \cdot \varepsilon_{\mathrm{pre}} \cdot \varepsilon_{\mathrm{tag}}}. \tag{7.15}$$

In the case where the signal-enhancing selections are based on a weakly supervised classifier, $\varepsilon_{\mathrm{tag}}$ becomes implicitly dependent on the actual signal cross section. If no signal is present in the data, a well-constructed background template sample would be identical to the SR data in the classifier training, and the classifier would only learn random noise. The resulting $\varepsilon_{\mathrm{tag}}$ would then approximate the selection efficiency of 1%. On the other hand, with an increasing number of signal events in the training data, the difference between SR data and background template becomes larger, and the classifier can learn to distinguish signal from background more effectively, leading to a higher signal efficiency. This behavior was demonstrated, for example, in Fig. 6.8.

In the more likely case that a specific signal process is not present in the data and thus $\varepsilon_{\mathrm{tag}}$ is small, the σ_{exc} obtained from Eq. 7.15 would become relatively large. However, if this amount of signal were present in the data, the classifier would have learned to become more sensitive to it, resulting in a larger $\varepsilon_{\mathrm{tag}}$ and smaller σ_{exc}. On the other hand, if signal events were abundant in the data, the classifier would achieve high $\varepsilon_{\mathrm{tag}}$ and thus low σ_{exc}, to an extent that one might set an upper cross section limit lower than what is actually present in the data. Both of these two options are not valid limits, as they are either overly conservative or too tight, respectively.

In order to obtain a well-defined limit, we inject signal MC simulation events into the data and retrain the classifier. By scanning over different values of an injected signal cross section

σ_{inj}, one can estimate the function $\sigma_{\text{exc}}(\sigma_{\text{inj}})$, through Eq. 7.15 and find the value of σ_{inj} such that $\sigma_{\text{exc}} = \sigma_{\text{inj}}$. This value then serves as the desired limit on the signal cross section. This procedure is illustrated in Fig. 7.17 and is analogous to what was adopted in Ref. [18]. The detailed technical implementation of this procedure is described in the Sec. 7.6.1. It relies on the key assumption that the presence of actual signal in the data, on top of the injected MC simulation events, would not bias the classifier as much as to result in ε_{tag} so large that the excluded cross section becomes smaller than the present one. This assumption is tested more quantitatively in Sec. 7.6.2.

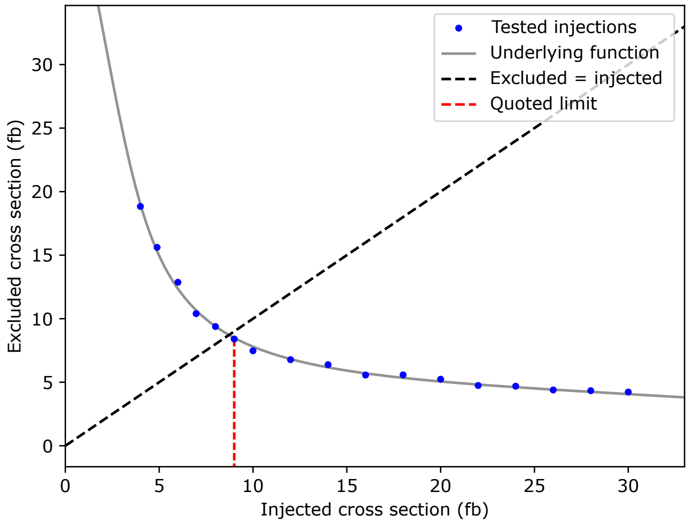

Figure 7.17: Illustrative example of a scan over various cross sections of injected MC simulation signal events, resulting in different excluded cross sections. Each injection (blue dots) can be seen as a sample of the hidden underlying function (gray line), with small noise due to the ML training. The aim is to find the value on the diagonal, where the excluded cross section equals the injected cross section. The quoted limit (red dashed line) is the maximum of the σ_{inj} and σ_{exc} at the injection point closest to the diagonal.

The inclusion of systematic uncertainties into the limit setting procedure factorizes into three subsequent steps. First, the cross section scan procedure outlined above, and detailed in Sec. 7.6.1, is performed *without* signal normalization uncertainties in the hypothesis test that yields N_{exc}. This results in an optimal injection point $\sigma_{\text{inj}}^{\text{opt}}$ such that $\sigma_{\text{exc}}(\sigma_{\text{inj}}^{\text{opt}}) \approx \sigma_{\text{inj}}^{\text{opt}}$. As a second step, the uncertainty in the signal efficiency at $\sigma_{\text{inj}}^{\text{opt}}$ is evaluated with a procedure discussed in Sec. 7.6.3, which translates to a systematic uncertainty in the signal normalization, δ_{sig}. Finally, a new hypothesis test is performed with δ_{sig} included as a nuisance parameter.

7.6.1 Cross Section Scan Implementation

When scanning over different σ_{inj} values, Eq. 7.15 is used to compute the corresponding σ_{exc}, where the retrained CATHODE classifiers will result in a dedicated $\varepsilon_{\text{tag}}(\sigma_{\text{inj}})$ in the denominator. The N_{exc} in the numerator is chosen to remain at the value obtained without any injection, rather than re-obtaining it from another bump hunt on top of the retrained classifiers. The reason for choosing the data-only N_{exc} is that the alternative would have multiple disadvantages. The classifiers and thus the event selection would differ from the model-independent p-value scan and instead correspond more to a suboptimal supervised search. Moreover, the event selection itself would depend on the confidence level, and the natural statistical fluctuations of the observed limit within the expected bands would translate to oscillations of $\sigma_{\mathrm{exc}}(\sigma_{\mathrm{inj}})$ and thus obstruct finding an unambiguous $\sigma_{\mathrm{exc}}(\sigma_{\mathrm{inj}}) = \sigma_{\mathrm{inj}}$ point.

Another simplification of the scan procedure is the choice of only retraining the classifier models of CATHODE at each value of σ_{inj}, and not the normalizing flows. This is justified by the narrow width of the considered signals, resulting in only relatively small event fractions outside the SR and a negligible impact of these few signal events in the flow training on the SB data. As training a single normalizing flow model involves multiple GPU hours, this choice significantly reduces the computational cost of the scan, which would likely not have been feasible otherwise.

The treatment discussed in Sec. 7.5.1 results in a weight associated with each event. Rather than performing a weighted training, which is not realistically mimicking the training on real detector data, the weights are used as the sampling probability when drawing a random subset of events with size σ_{inj} from the full MC signal simulation dataset. The remaining available signal events are used for computing the efficiency $\varepsilon_{\mathrm{tag}}(\sigma_{\mathrm{inj}})$, which minimizes the effect of statistical fluctuations from a limited sample size compared to a σ_{inj}-sized evaluation set. Weights are taken into account by computing $\varepsilon_{\mathrm{tag}}$ as the sum of weights of the events passing the selection threshold divided by the full sum of weights.

Despite the double k-folded CATHODE classifier training, the signal efficiency at the different σ_{inj} values can still fluctuate significantly due to the small number of signal events in the training data. This is mitigated by performing the full training procedure five times, each with a different random neural network weight initialization, and sampling a new subset of signal events from the full MC simulation for each training.

In order to find the value of σ_{inj} such that $\sigma_{\mathrm{exc}} = \sigma_{\mathrm{inj}}$ most efficiently, a binary search algorithm is employed, which assumes a monotonically decreasing $\sigma_{\mathrm{exc}}(\sigma_{\mathrm{inj}})$ function. An initial value of $\sigma_{\mathrm{inj}}^{t=0}$ returns a value of $\sigma_{\mathrm{exc}}^{t=0}$. The maximum of the two serves as upper boundary of a search window, x_{upper}, and the minimum as the lower boundary, x_{lower}. At each time step t, the logarithmic center, $\sigma_{\mathrm{inj}}^{t} = \exp(0.5(\log(x_{\mathrm{lower}}) + \log(x_{\mathrm{upper}})))$, is used to compute the next value of $\sigma_{\mathrm{inj}}^{t}$. The lower value of $\left\{\sigma_{\mathrm{inj}}^{t}, \sigma_{\mathrm{exc}}^{t}\right\}$ is then used to update x_{lower}, if it is larger than the previous value, and the higher value updates x_{upper}, if applicable. At each step, all cross section values are rounded to two digits precision (in units of fb), and the search is terminated when the fractional difference between the two values is below 10%. The final limit (without considering signal normalization uncertainties) is the maximum of the $\{\sigma_{\mathrm{inj}}^{t}, \sigma_{\mathrm{exc}}^{t}\}$ pair with the smallest difference.

The scanning procedure outlined above is performed four times: for the observed limit, for the expected limit, and the expected $\pm 1\sigma$ bands. Each limit definition results in a distinct N_{exc} value, which results in a different $\sigma_{\mathrm{exc}} = \sigma_{\mathrm{inj}}$ point. Since the trained classifier models at each σ_{inj} are independent of the limit definition, they are reused between subsequent binary searches on a given signal model.

7.6.2 Behavior With Signal in Data

The limit setting procedure based on injecting signal events assumes that the potential presence of an actual signal in the data does not impact the classifier training enough to yield a lower limit than what is actually present. We tested this assumption explicitly in the MC background simulation dataset for two signal processes, $\mathrm{X} \to \mathrm{Y}\mathrm{Y}' \to 4\mathrm{q}$ with $m_{\mathrm{X}} = 3\,\mathrm{TeV}$ and $m_{\mathrm{Y}} = m_{\mathrm{Y}'} = 170\,\mathrm{GeV}$, and $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ with $m_{\mathrm{W}'} = 3\,\mathrm{TeV}$ and $m_{\mathrm{B}'} = 400\,\mathrm{GeV}$. Figure 7.18 shows the observed and expected limits as a function of a signal cross section that has been injected into the dataset before performing the cross section scan.

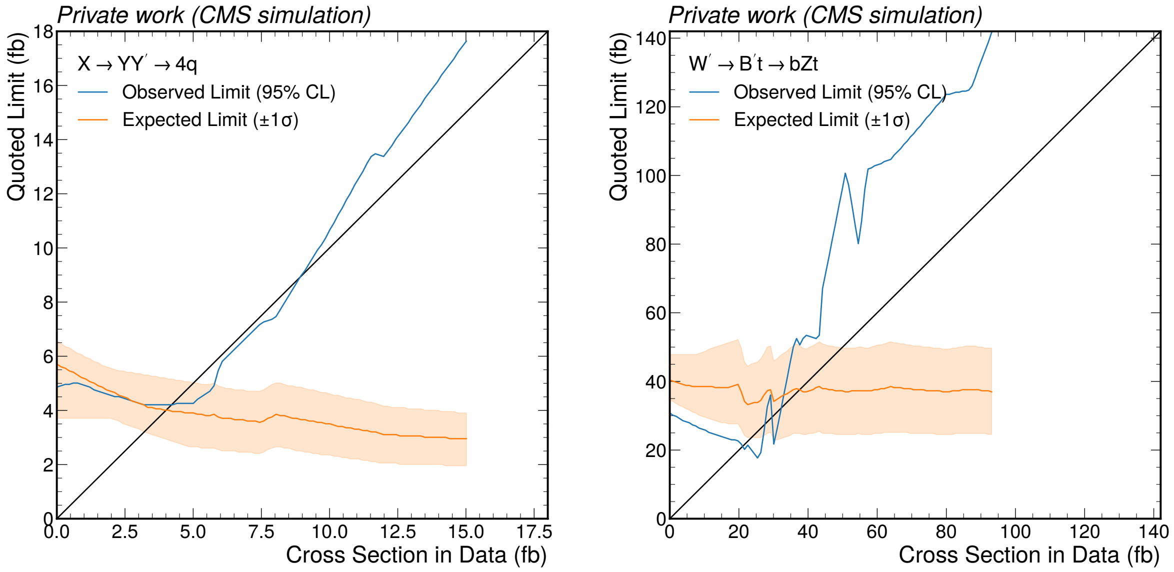

Figure 7.18: Observed and expected limits when the tested signal is present in the data with various cross sections. The two plots show results obtained with CATHODE on the MC background simulation dataset for 3 TeV resonances: X $\rightarrow$ YY$'$ $\rightarrow$ 4q with $m_{\mathrm{Y}} = m_{\mathrm{Y'}} = 170\,\mathrm{GeV}$ (left) and W$'$ $\rightarrow$ B$'$t $\rightarrow$ bZt with $m_{\mathrm{B'}} = 400\,\mathrm{GeV}$ (right).

The expected limits in Fig. 7.18 decrease slightly as a function of the signal cross section in data. This is due to the classifiers becoming more discriminating with more signal events in the training data, resulting in an increase of $\varepsilon_{\mathrm{tag}}$, while the expected event limit for N_{exc} is not reacting to larger signal presence. Such constant N_{exc} is the desired behavior as the expected limit corresponds to a hypothetical background-only result.

The observed limit, on the other hand, is subject to fluctuations of the order of magnitude of the expected limit bands, because every point corresponds to a new CATHODE classifier training and thus a different N_{exc}. Within fluctuations, the observed limit does not significantly enter the region below the diagonal where it is lower than the present amount of signal. Two regions can be identified: at low injected cross sections, the observed limit decreases as a function of the cross section because the classifiers become more discriminant and thus $\varepsilon_{\mathrm{tag}}$ increases. At a point relatively close to the diagonal, the increase in N_{exc} becomes more dominant as it reacts to the presence of signal during the bump hunt. In the latter region, one can see the observed limit significantly exceeding the upper band of the expected limit due to the noticeable presence of the signal in the data.

7.6.3 Incorporating Systematic Uncertainties

If the shape of the input feature distributions changes within the expected variance due to systematic uncertainty sources, the performance of a weakly supervised classifier trained on these features will also change. In particular, if the transformed signal distribution becomes more similar to the background, the task of identifying overdensities becomes more difficult and the classifier efficiency, $\varepsilon_{\mathrm{tag}}$, decreases. Therefore, an uncertainty in the signal efficiency translates to one in the signal normalization, as described in Sec. 7.5.4. The following paragraphs describe how the efficiency uncertainty is systematically evaluated individually for each source of uncertainty and then combined into a total uncertainty in the signal normalization.

Once the optimal injection point $\sigma_{\mathrm{inj}}^{\mathrm{opt}}$ with $\sigma_{\mathrm{exc}}(\sigma_{\mathrm{inj}}^{\mathrm{opt}}) \approx \sigma_{\mathrm{inj}}^{\mathrm{opt}}$ is identified without considering signal normalization uncertainties, the uncertainty in $\varepsilon_{\text{tag}}(\sigma_{\text{inj}}^{\text{opt}})$ is evaluated by considering its variance under systematic shifts of the signal input feature distributions at $\sigma_{\text{inj}}^{\text{opt}}$.

Ideally, we would retrain the CATHODE classifier models for every considered uncertainty source shifted up and down, each resulting in a varied signal efficiency. However, the computational burden of this becomes large considering $\mathcal{O}(20)$ sources of systematic uncertainty, each varied up and down, and trained with five random samplings of the signal. This is avoided by limiting the procedure to uncertainty sources that induce statistically significant deviations on the input features. To test this, histograms of the nominal signal event distribution are created from the full MC simulation sample, using ten equal-width bins that contain the central 98% of the events, i.e., removing the upper and lower 1% tails to reduce sensitivity to outliers. For each uncertainty source, the feature distribution for the signal in question is then obtained by varying the uncertainty up and down, and the per-bin deviation is measured. This difference is then scaled down to a signal injection of $\sigma_{\text{inj}}^{\text{opt}}$ in order to compare it to the statistical (Poisson) uncertainty at the expected signal injection in this bin. If the deviation is larger than 20% of this statistical uncertainty in any bin of any feature, the systematic uncertainty source is considered significant for the signal in question and the classifier will be fully retrained. The 20% threshold is to some extent arbitrary, but has been used as a conventional choice for a conservative estimate of the significance of a deviation.

In the following, this procedure is illustrated for the case of the observed upper limit of a $\text{W}' \to \text{B}'\text{t} \to \text{bZt}$ signal with $m_{\text{W}'} = 3\,\text{TeV}$ and $m_{\text{B}'} = 400\,\text{GeV}$ in the MC simulation dataset. The optimal injection point was found to be $\sigma_{\text{inj}}^{\text{opt}} = 48.4\,\text{fb}$, corresponding to an excluded cross section of $\sigma_{\text{exc}}(\sigma_{\text{inj}}^{\text{opt}}) = 47.7\,\text{fb}$. The criterion for uncertainty sources resulting in significant enough shape changes to warrant retraining is shown for two examples: the parton distribution function uncertainty in Fig. 7.19 and the final state radiation parton shower uncertainty in Fig. 7.20. The latter is a predecessor of the Lund plane–based set of jet substructure modeling uncertainties, used in an earlier iteration of the analysis. The shape change from the PDF uncertainties is only minor in this case and does not qualify for retraining, while the parton shower uncertainty does, as the shape change is larger than 20% of the statistical uncertainty in at least one bin for at least one feature (in this case all features).

For those uncertainties that are considered significant, the full double k-folded classifier setup is retrained with the uncertainty shifted up and down. In each direction, the retraining is performed five times and then averaged, just as for the nominal training, in order to factor out random training variance. For weight-based corrections, the up- and down-shifted weights are each used as per-event probabilities for sampling the $\sigma_{\text{inj}}^{\text{opt}}$ events from the full MC simulation sample, as well as for computing the weighted signal efficiency. For corrections resulting in shifted feature values, the same five signal realizations from the nominal training are reused, just with the feature values shifted accordingly. The five values of ε_{tag} per varied direction are then averaged to obtain an up- and down-shifted signal efficiency per uncertainty source. The mean of the absolute fractional difference between the shifted and nominal efficiencies in either direction is used as the signal efficiency uncertainty value of each source.

An estimate of the variance of ε_{tag} is also obtained for all uncertainty sources, regardless of whether they qualify for retraining via the criterion above, by simply evaluating the nominally trained classifier models with shifted systematic uncertainties, i.e., only the efficiency computation itself is affected from the change in either event weights or feature values. This is done for each of the five signal realizations, and they are averaged for each direction of shifting the uncertainty. Again, the mean of the absolute fractional difference between the shifted and nominal efficiencies is used as uncertainty estimate. If an uncertainty source qualifies for retraining, the maximum of the retrained and reevaluated efficiency is used as the final uncertainty value for

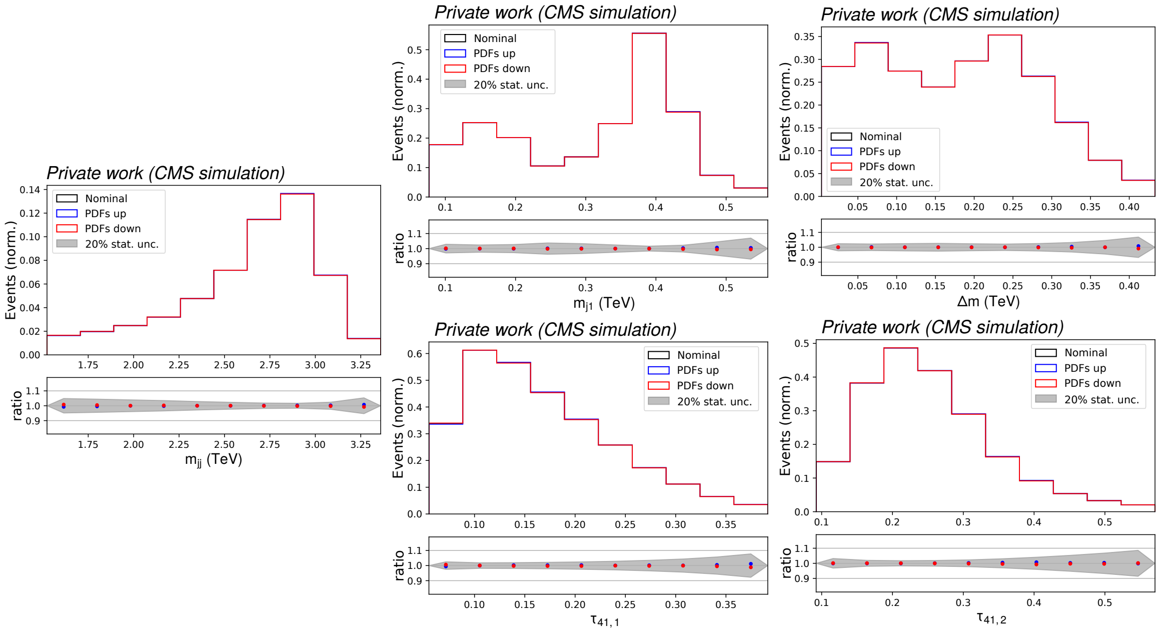

Figure 7.19: Example of the test in the MC background simulation dataset to decide which sources of systematic uncertainties are significant enough to retrain the classifier models. The feature distribution for the $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ signal with $m_{\mathrm{W}'} = 3\,\mathrm{TeV}$ and $m_{\mathrm{B}'} = 400\,\mathrm{GeV}$ is shown for the nominal case (black), as well as the distribution resulting from the parton distribution function uncertainty shifted up (blue) and down (red). The bottom panel shows the bin-wise ratio, where the gray band denotes 20% of the statistical uncertainty at an injection of $\sigma_{\mathrm{inj}}^{\mathrm{opt}} = 48.4\,\mathrm{fb}$. The shape change from the PDF uncertainties is only minor in this case and does not qualify for retraining.

this source.

A special case is the treatment of random variance in classifier training from both random initialization and the signal event realization through randomly sampling only $\sigma_{\mathrm{inj}}^{\mathrm{opt}}$ events. This uncertainty is measured as the maximum deviation away from the nominal efficiency, which is the mean of the same five efficiency values.

Another subtlety to consider is the change of $\varepsilon_{\mathrm{pre}}$ due to a shift in JES and JER. The preselections, discussed in Sec 7.2.6 are based on the reconstructed p_{T} of the jets, and a shift in the jet energy scale or resolution generally results in a different fraction of events passing the selections. The $\varepsilon_{\mathrm{pre}}$ values for each up- and down-shifted JES and JER are computed during the sample preprocessing and multiplied to the respective varied $\varepsilon_{\mathrm{tag}}$ values before measuring the fractional difference to the nominal efficiency.

The resulting effects on the signal efficiency for the $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ signal example are summarized in Fig. 7.21. The most dominant uncertainty is the final state radiation parton shower modeling uncertainty shown in Fig. 7.20. The uncertainty from retraining classifiers is larger than evaluating modified shapes with the nominal classifiers, so the retrained efficiency is used as the final uncertainty value for this source. The next most impactful uncertainty originates from retraining the classifiers with random resamplings of the signal events, which is by definition only derived via retraining classifiers. The first five uncertainty sources have qualified for retraining, while the remaining ones were evaluated with the nominal classifiers, all with visibly negligible impact on the signal efficiency.

For every signal model, all the efficiency uncertainties are combined by adding them in

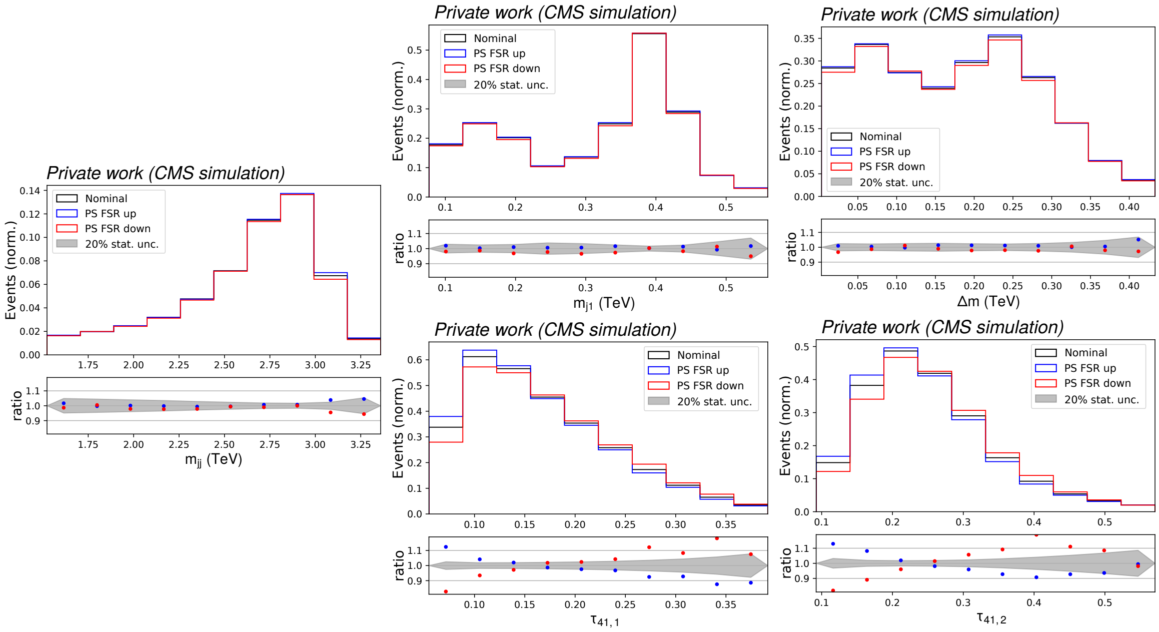

Figure 7.20: The same example as in Fig. 7.19, but for the final state radiation parton shower uncertainties. In this case, the shape change from shifting this uncertainty up and down exceeds the 20% statistical uncertainty threshold and thus a classifier retraining is deemed necessary for this uncertainty source.

quadrature, which serves as the total signal normalization uncertainty. Another profile likelihood fit is performed with this normalization uncertainty included as a log-normally constrained nuisance parameter, yielding a new exclusion limit on the number of events $N_{\text{exc}}^{\text{syst}}$. With the same nominal ε_{tag} as used for the initial $\sigma_{\text{inj}}^{\text{opt}}$ determination, the new $\sigma_{\text{exc}}^{\text{syst}}$ is computed via Eq. 7.15. The maximum of $\sigma_{\text{inj}}^{\text{opt}}$ and $\sigma_{\text{exc}}^{\text{syst}}$ serves as the final limit on the signal cross section.

In the $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ example above, the combination of uncertainties shown in Fig. 7.21 results in a total signal normalization uncertainty of 19%. The exclusion limit after including this uncertainty as a nuisance parameter in the profile likelihood fit is $\sigma_{\text{exc}}^{\text{syst}} = 50.4\,\text{fb}$. As this is larger than the optimal injection point, this value serves as the final (observed) upper limit on the signal cross section. In rare cases, the excluded limit is lower than the optimal injection point, both before and after systematic uncertainties are included. In these cases, the optimal injection point remains the choice of limit, as it is the more conservative estimate of the true limit.

As mentioned in Sec. 7.6.1, not one but four upper limits are computed via a scan over the excluded vs injected cross section: observed, expected and the expected $\pm 1\sigma$ bands. These limits generally differ, and relate to four different values of σ_{inj}. Since the uncertainty, just as the signal efficiency itself is dependent on σ_{inj}, the procedure above would ideally be repeated for each of the four limits. However, the computational cost of this is prohibitive, and we instead use the approximation that the efficiency uncertainty derived for the observed limit can also be applied to the expected limit and its bands. The observed limit is chosen as it corresponds the most to a physics result, derived on the real manifestation of the collider data.

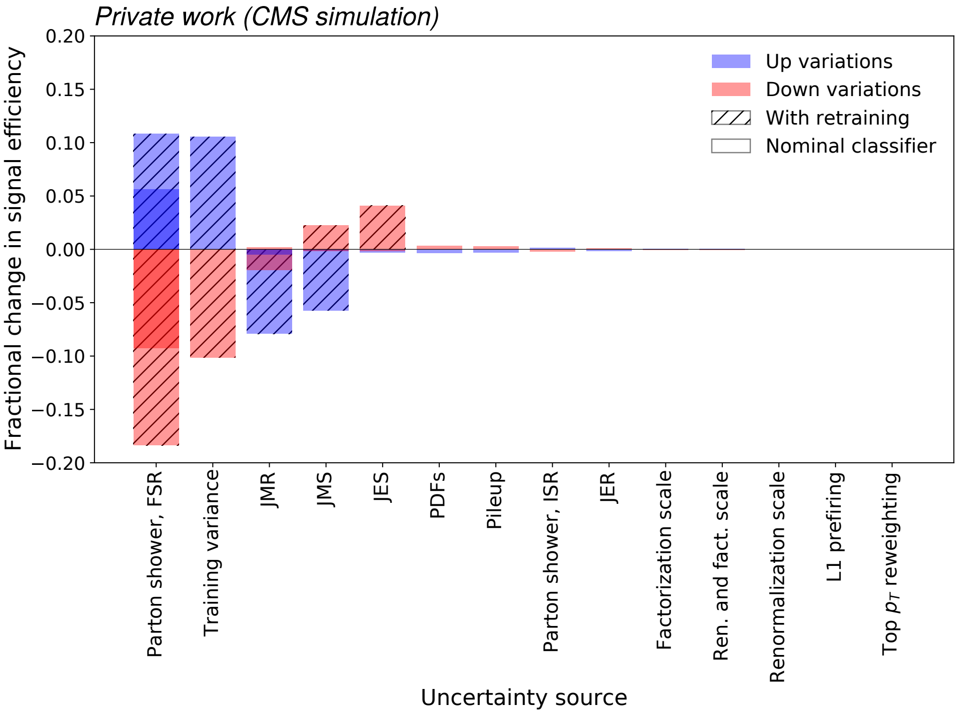

Figure 7.21: Summary of the effects of all systematic uncertainties in the signal efficiency for the MC background simulation dataset example of a $W' \to B't \to bZt$ signal with $m_{W'} = 3\,\text{TeV}$ and $m_{B'} = 400\,\text{GeV}$. The uncertainties are ordered by their impact on the signal efficiency. If a classifier retraining was performed, the resulting fractional efficiency change is denoted in hatched bars, while the unhatched bars correspond to an evaluation of the shifted distribution with the nominally trained classifier models. If both retraining and reevaluation were performed, the maximum of the two is taken as the uncertainty value for the respective source.

7.7 Validation Studies

Before performing the outlined analysis strategy on real collision data, it is crucial to validate the procedure in a controlled environment. To this end, validation studies were performed in two stages. First, a background-only MC simulation dataset is used to ensure that the analysis chain does not sculpt the background in a way that would lead to false excesses. Moreover, by injecting various amounts of simulated signal events into the MC background simulation dataset, the sensitivity of the analysis procedure towards the presence of a signal is tested. Second, the analysis chain is validated in a real data control region where no signal presence is expected. This is done to ensure that the analysis chain does not induce false excesses when exposed to unforeseen features intrinsic to real data. These validation studies are complemented by a set of bias studies, which are used to quantify the robustness of the analysis strategy against potential systematic effects.

7.7.1 Monte Carlo Simulation

The MC background simulation dataset, described in Sec. 7.2.2, is composed of a realistic mixture of SM processes, namely QCD multijet events, W and Z bosons produced in association with jets, top quark-antiquark pair production, and single top quark production. Each of the four data-taking periods (2016 preVFP, 2016 postVFP, 2017, 2018) contributes equally, each with a size equivalent to an integrated luminosity of $6.7\,\mathrm{fb}^{-1}$, resulting in a total dataset equivalent to $26.8\,\mathrm{fb}^{-1}$. This dataset is used to validate the analysis chain in a controlled background-only environment, where no significant excess is expected in a well-calibrated analysis. By additionally injecting MC signal simulation events into this dataset, one can study the sensitivity of the analysis procedure towards the presence of the respective signal.

The primary focus of the MC simulation validation studies is to verify that no significant excess is found in the absence of any signal. Such an excess might arise due to the type of background sculpting discussed in Sec. 2.6.3 and the Appendix B.5. This was tested by performing the full analysis procedure on the MC background simulation sample, i.e., training CATHODE(-b), selecting the most anomalous percentile of events for every choice of SR, fit the background models via a smoothly falling functional form, and then compute p-values under the background-only hypothesis as a function of the dijet invariant mass.

The distributions of events that were selected as most anomalous in every SR bin are shown in Fig. 7.22 for CATHODE and in Fig. 7.23 for CATHODE-b. A direct comparison with the shapes before selection, as studied in Figs. 6.18, would be misleading as the lack of a change in background shape is a sufficient but not strictly necessary requirement for the absence of excesses. Rather, the figures demonstrate that the smoothly falling background shape is preserved, and the chosen background fit functions describe the passing m_{jj} distributions well within statistical fluctuations. The latter can be seen from the pull plots in the lower panels under every histogram, quantifying the difference between the data and the fits in units of the statistical uncertainty. Moreover, it is the quadratic sum of these pulls that determines the χ^2 value, which is shown divided by the number of degrees of freedom n_{dof} in each panel and used as fit quality metric in the F-test. The resulting χ^2/n_{dof} values do not deviate excessively from unity.

The p-values obtained from the subsequent tests of the background-only hypothesis are shown in Fig. 7.24 as a function of m_{jj} for both CATHODE and CATHODE-b. As expected, most values in both cases correspond to a significance lower than 1σ. The p-value of 0.5 serves as a hard upper boundary because the test statistic from Eq. 2.31 has a lower bound of zero in the case of signal underfluctuations, i.e., if the unconditionally fitted signal strength is negative. As can be seen, for a few of the tested mass points the p-values exceed 1σ, which is expected due to statistical fluctuations. No excesses beyond 2σ were observed in the background-only MC simulation study.

While this study establishes the lack of significant sculpting in the absence of signal, it is also crucial that the presence of signal would result in an excess. This was tested by injecting increasing amounts of signal into the MC simulation sample and repeating the full analysis procedure for every such dataset. Specifically, the test was performed with the $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal with $m_{\mathrm{X}} = 3000\,\mathrm{GeV}$, $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$, and $m_{\mathrm{Y}'} = 80\,\mathrm{GeV}$, and different amounts corresponding to cross sections from 5 and 30 fb in steps of 5 fb were injected. In order to save computational resources, the classifier was retrained exclusively within the SR bin centered around the nominal signal mass, i.e., between 2.725 and 3.331 TeV. Moreover, the normalizing flows were not retrained with injected signal. Instead, the background-only trained models were used for generating a background template. These two simplifications follow the assumption that the signal contamination outside the central SR is negligible.

Figure 7.25 shows the resulting background fits to the selected m_{jj} distributions, after performing CATHODE in the central SR around the nominal signal mass. Starting from 15 fb, a signal bump becomes increasingly visible, with more significant deviations between the data and the smooth background fit. The resulting significance is assessed more quantitatively as a function of m_{jj} in Fig. 7.26. There, the excess at 3 TeV indeed starts exceeding 5σ at 15 fb. For reference, each panel shows the significance from an “inclusive search” at the respective signal cross section as a green dashed line, which corresponds to the statistical analysis performed

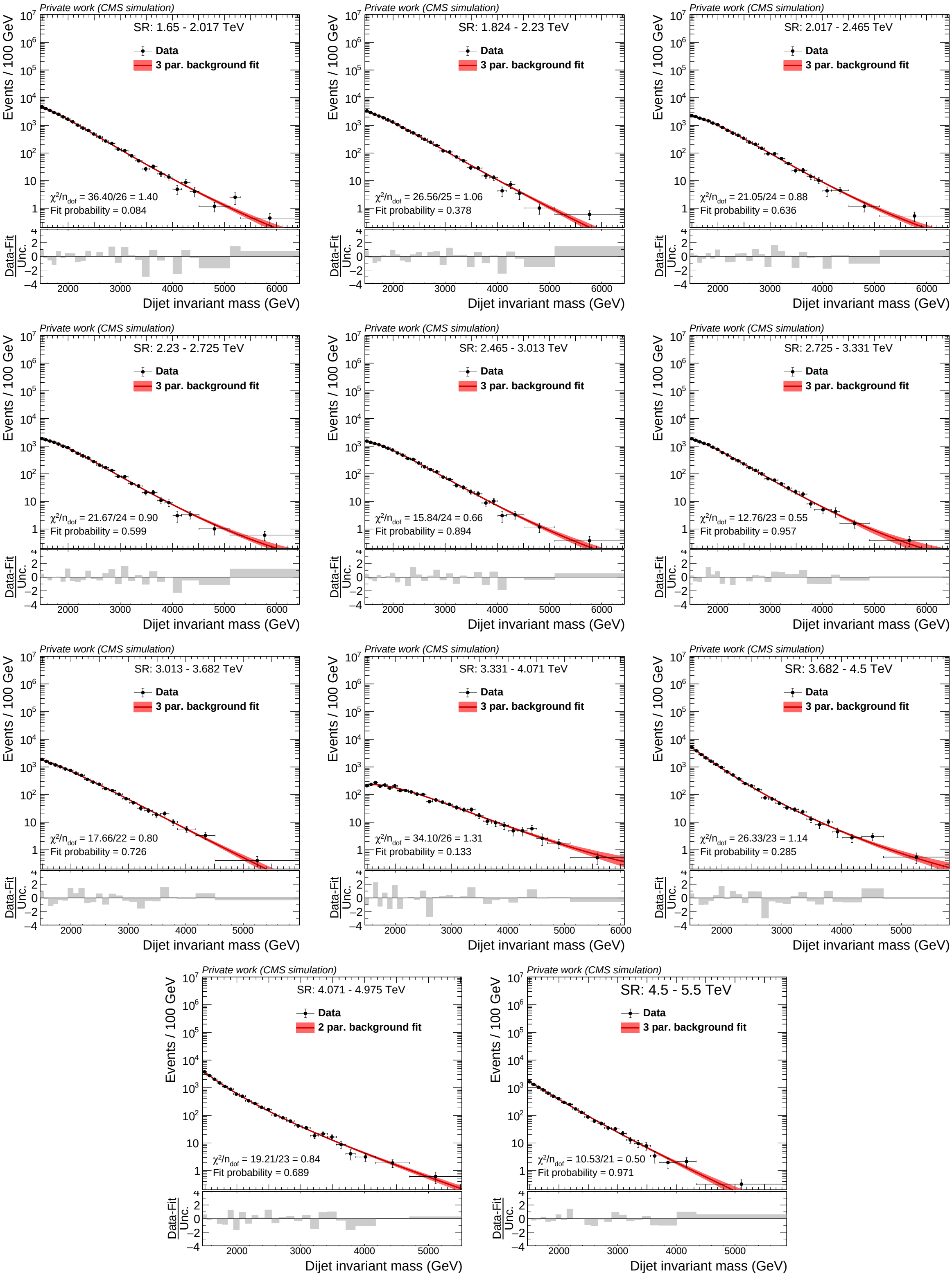

Figure 7.22: MC background simulation m_{jj} spectra after selecting the most anomalous 1% of events using CATHODE for every choice of signal region, and the background fit function chosen by the F-test.

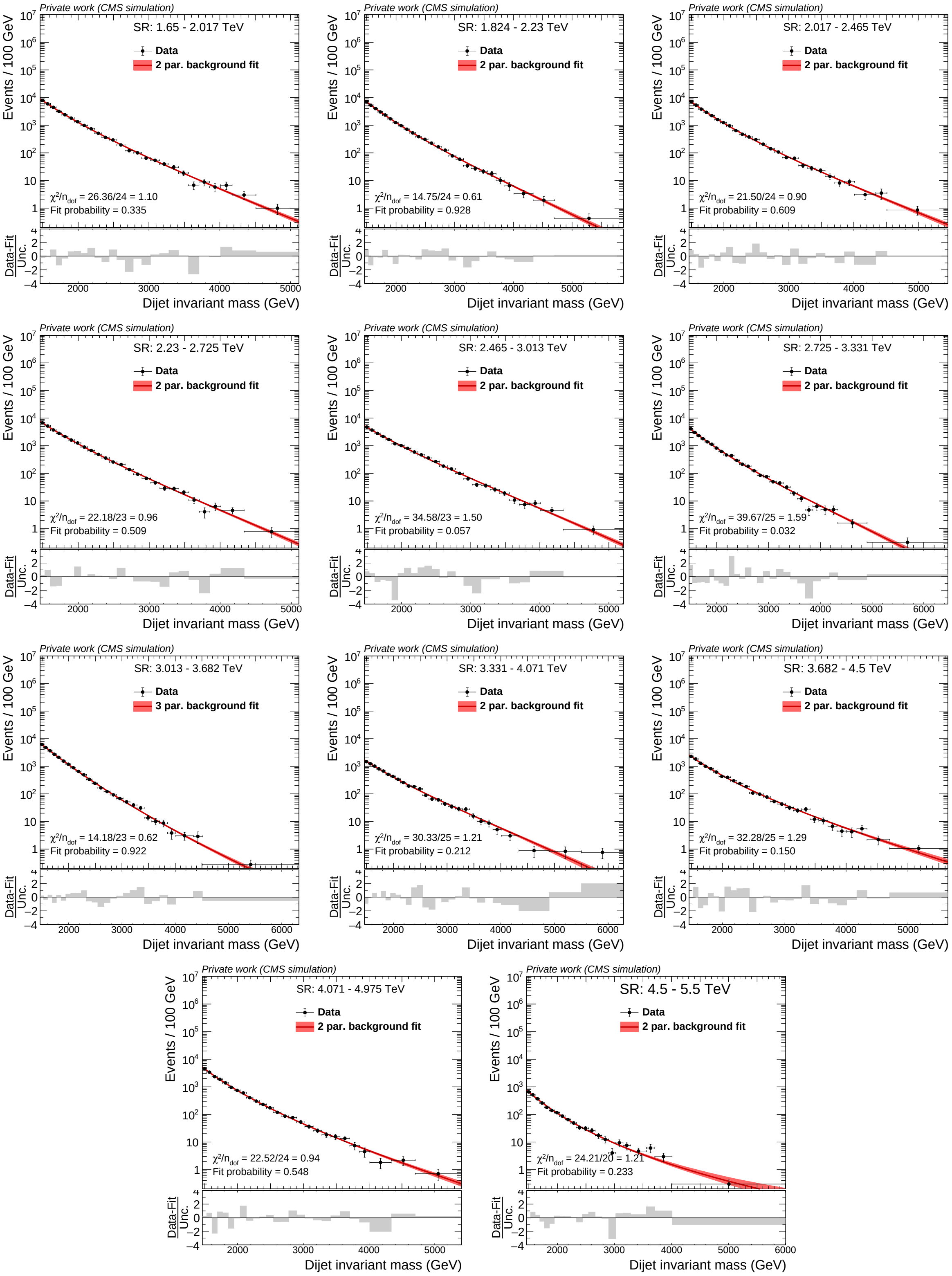

Figure 7.23: MC background simulation m_{jj} spectra after selecting the most anomalous 1% of events using CATHODE-b for every choice of signal region, and the background fit function chosen by the F-test.

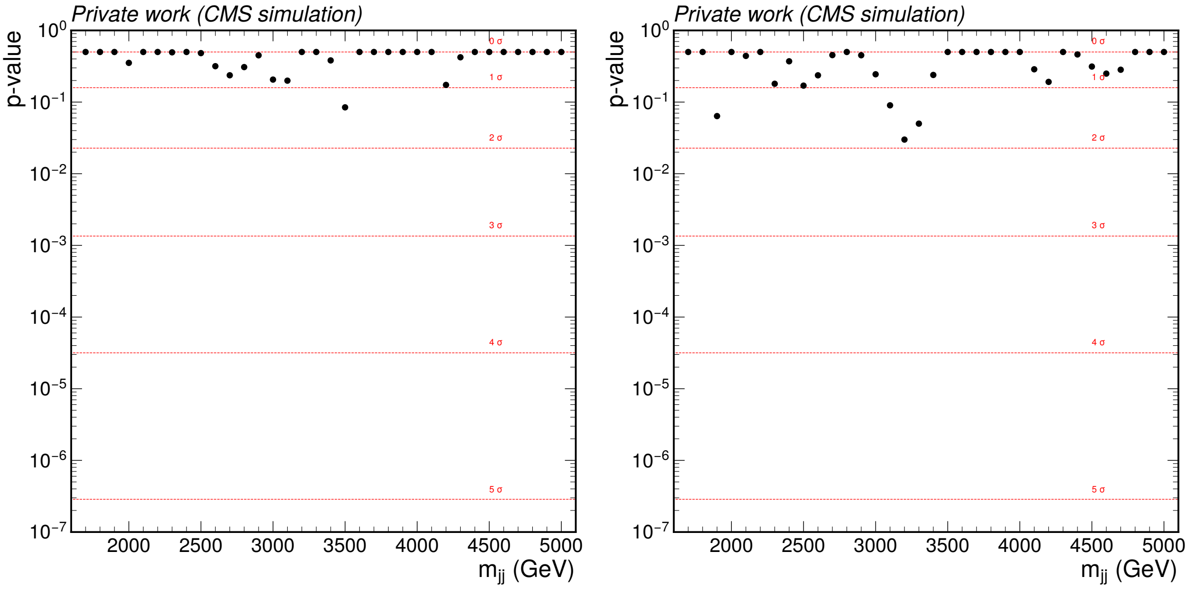

Figure 7.24: MC background simulation p-values as a function of m_{jj} after selecting the most anomalous 1% of events using CATHODE (left) and CATHODE-b (right).

without any anomaly score–based selections and will be introduced in Sec. 7.9.5. Even at the maximum signal injection of 30 fb, this CATHODE-free analysis strategy fails to obtain a significance beyond 2.5σ, which demonstrates the non-trivial signal extraction performance of our CATHODE-based analysis strategy.

A similar behavior is observed for CATHODE-b. Figure 7.27 shows the respective selected m_{jj} distributions and the fit functions, where a visible bump is appearing between 15 and 20 fb. Analogously, Fig. 7.28 shows the significance scan results, yielding a $> 3\sigma$ evidence at an injection of 15 fb and a $> 5\sigma$ discovery with a cross section of 20 fb. CATHODE-b thus seems similarly able to extract an X $\rightarrow$ YY$'$ $\rightarrow$ 4q signal if it is present, however with somewhat reduced sensitivity compared to CATHODE due to the lack of discrimination power in the additional input features for this specific signal model.

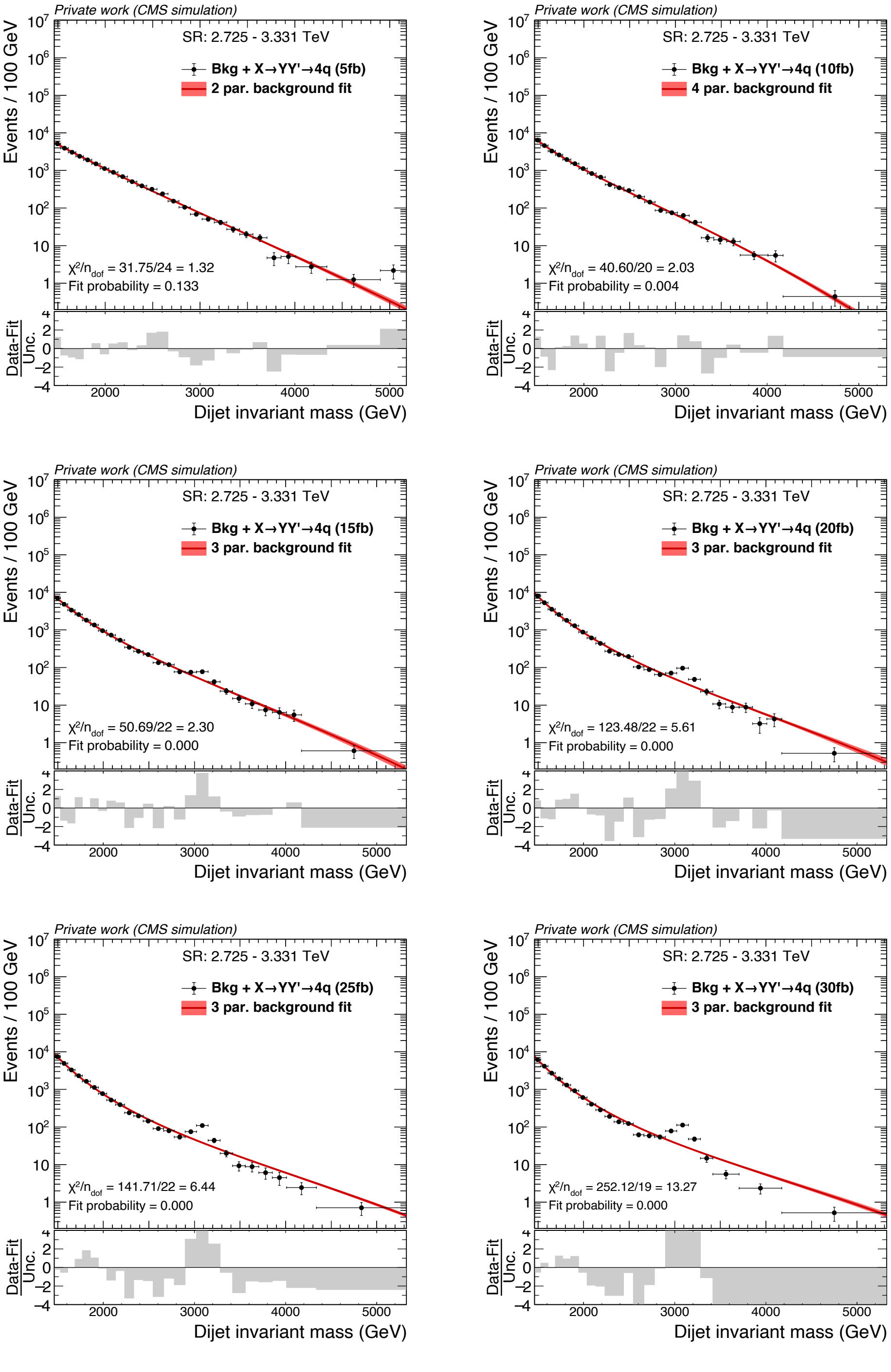

Figure 7.25: MC simulation dataset m_{jj} spectra and smooth background fits after selecting the most anomalous 1% of events using CATHODE when injecting an increasing amount of the $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal with $m_{\mathrm{X}} = 3000\,\mathrm{GeV}$, $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$, and $m_{\mathrm{Y}'} = 80\,\mathrm{GeV}$. The injected cross sections are 5 fb (top left), 10 fb (top right), 15 fb (center left), 20 fb (center right), 25 fb (bottom left), and 30 fb (bottom right).

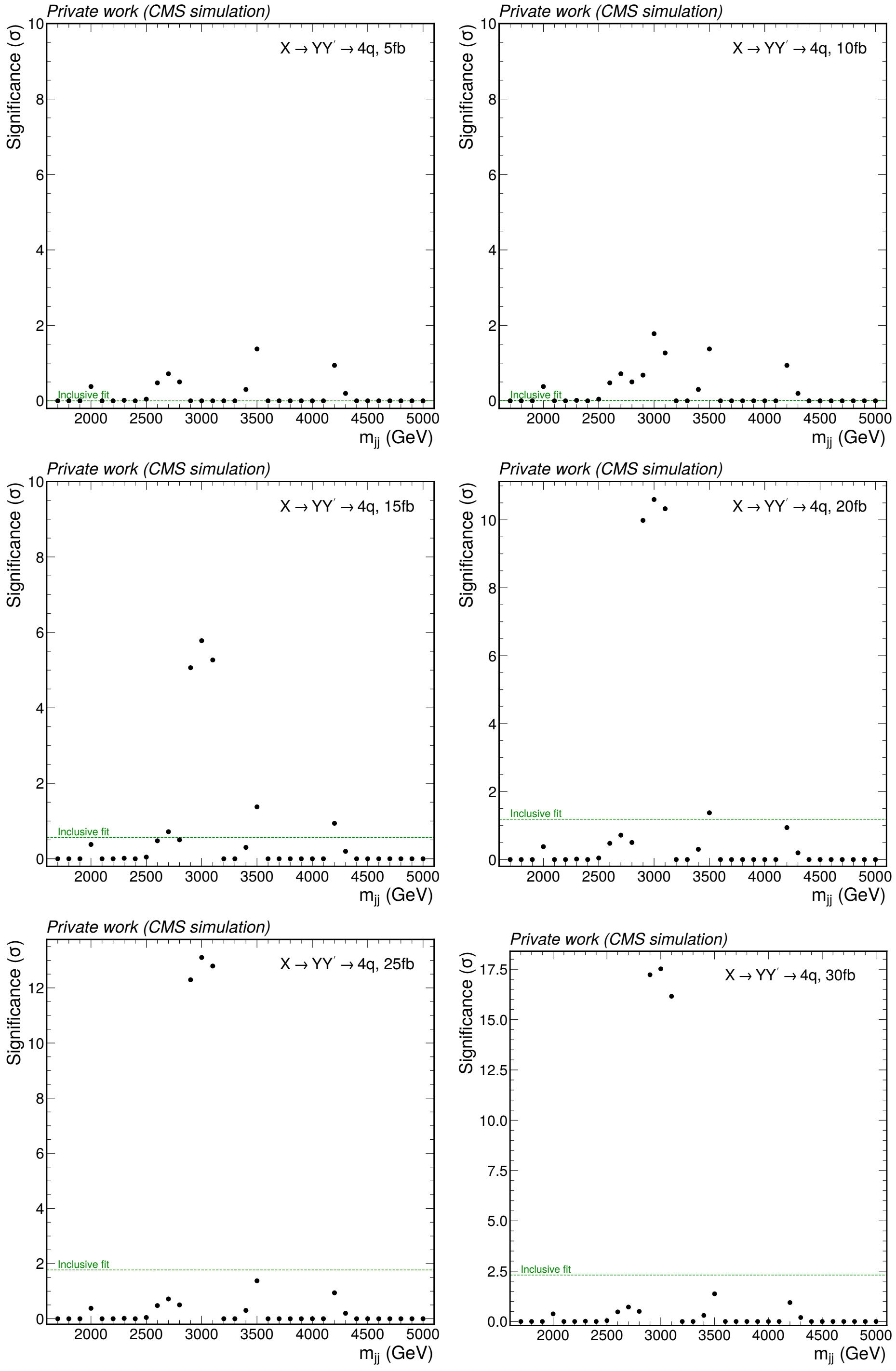

Figure 7.26: MC simulation dataset significance as a function of the injected signal cross section after selecting the most anomalous 1% of events using CATHODE. The injected signal is X → YY′ → 4q with $m_{\mathrm{X}} = 3000\,\mathrm{GeV}$, $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$, and $m_{\mathrm{Y'}} = 80\,\mathrm{GeV}$. The injected cross sections are 5 fb (top left), 10 fb (top right), 15 fb (center left), 20 fb (center right), 25 fb (bottom left), and 30 fb (bottom right). The green line indicates the significance obtained from an inclusive fit, i.e., performing the same statistical procedure but without CATHODE-based event selection. The significance values outside the central SR of 2.725 to 3.331 TeV are adopted from the background-only trainings and fits.

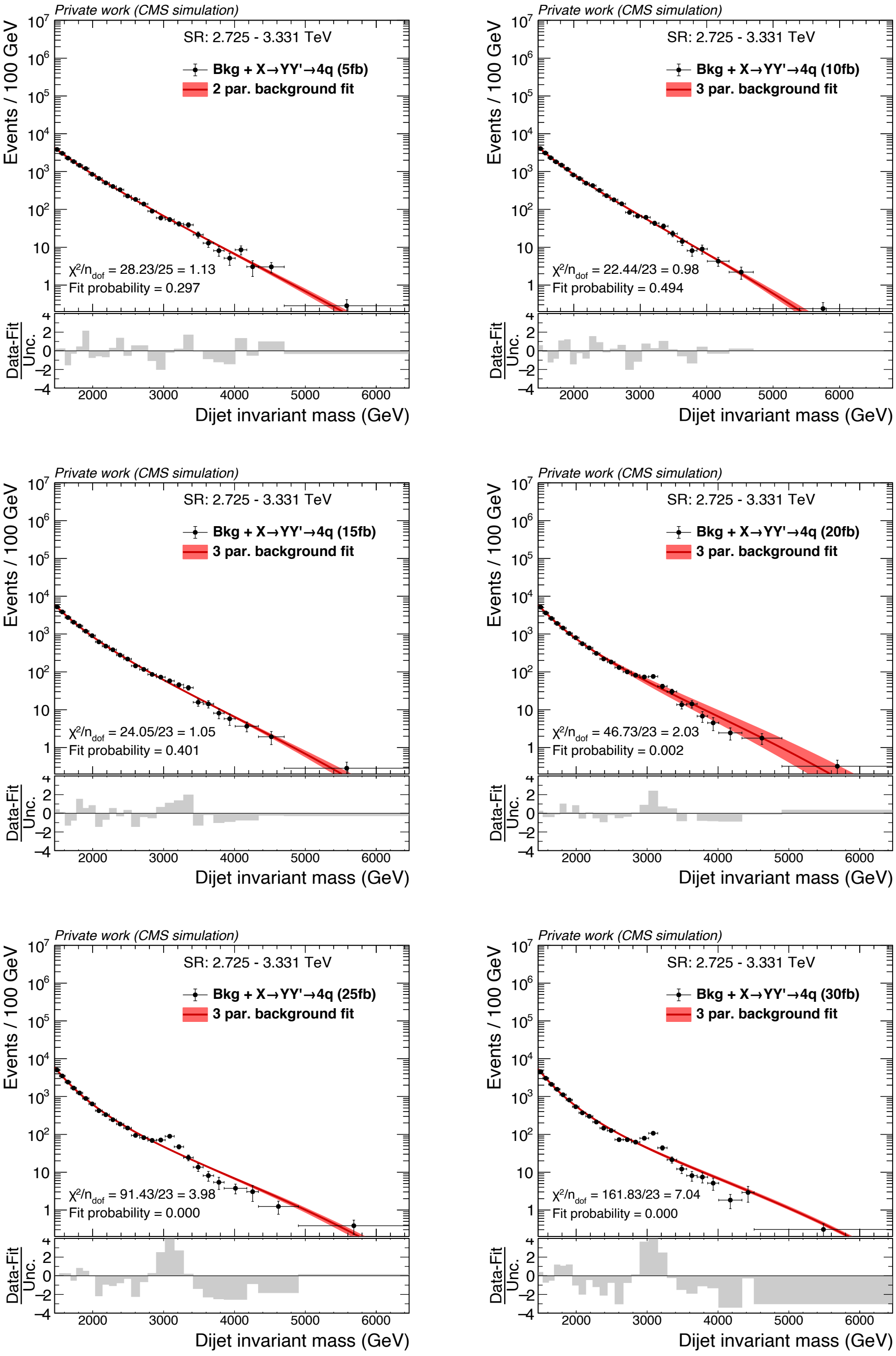

Figure 7.27: MC simulation dataset m_{jj} spectra and smooth background fits after selecting the most anomalous 1% of events using CATHODE-b when injecting an increasing amount of the $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal with $m_{\mathrm{X}} = 3000\,\mathrm{GeV}$, $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$, and $m_{\mathrm{Y}'} = 80\,\mathrm{GeV}$. The injected cross sections are 5 fb (top left), 10 fb (top right), 15 fb (center left), 20 fb (center right), 25 fb (bottom left), and 30 fb (bottom right).

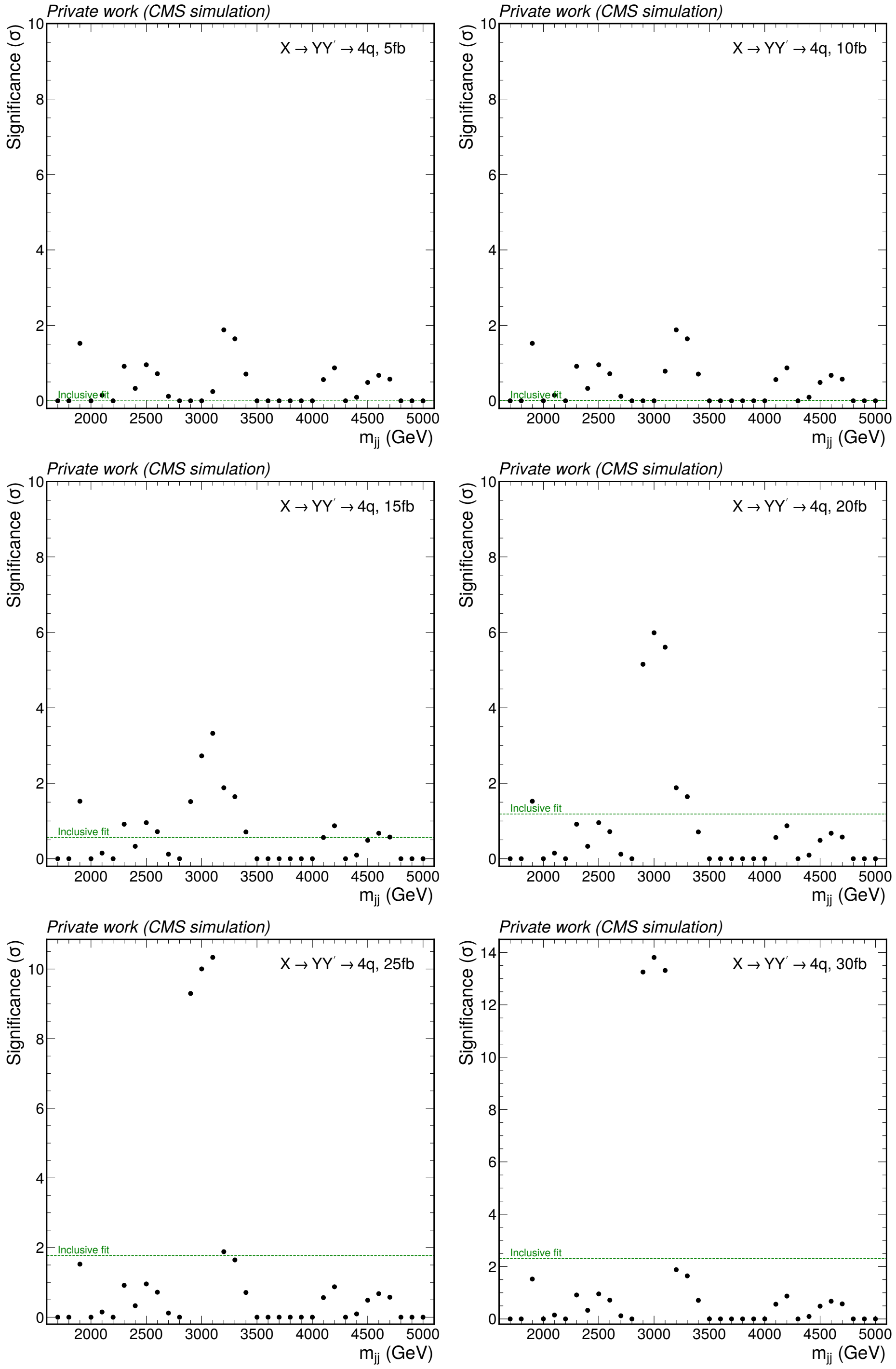

Figure 7.28: MC simulation dataset significance as a function of the injected signal cross section after selecting the most anomalous 1% of events using CATHODE-b. The injected signal is X → YY′ → 4q with $m_{\mathrm{X}} = 3000\,\mathrm{GeV}$, $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$, and $m_{\mathrm{Y'}} = 80\,\mathrm{GeV}$. The injected cross sections are 5 fb (top left), 10 fb (top right), 15 fb (center left), 20 fb (center right), 25 fb (bottom left), and 30 fb (bottom right). The green line indicates the significance obtained from an inclusive fit, i.e., performing the same statistical procedure but without CATHODE-based event selection. The significance values outside the central SR of 2.725 to 3.331 TeV are adopted from the background-only trainings and fits.

7.7.2 Data Control Region

The next crucial step of the validation studies is to ensure that there are no unforeseen features intrinsic to real detector data that induce false excesses unseen in simulation. The standard approach for validation studies within real data is to define a control region (CR) in phase space that is both disjoint from the analysis region, and in which the presence of any considered signal is suppressed. The latter constraint is relevant in the case of an observed excess in the CR, as it could otherwise not unambiguously be attributed to methodological issues.

Since the present analysis is designed to be sensitive to a large variety of signal models, the signal suppression potential of different CR definitions was studied. The considered signal processes are summarized in the first column of Tab. 7.7 and listed with the full MC simulation sample names in Tab. B.5 in Appendix B.2. This set comprises models featuring a graviton (G) decaying to a pair of either W or Z bosons, a W$'$ boson decaying to W and Z, a Z$'$ boson decaying to a W pair, an excited b quark (b*) decaying to a top quark and a W, and a W$_{\mathrm{KK}}$ decaying to a W and a radion (R). The last signal has a varying radion mass, quantified by the ratio of the radion mass to the W$_{\mathrm{KK}}$ mass, $R_0 = m_{\mathrm{R}}/m_{\mathrm{W_{KK}}}$. The weakness of using explicit signal processes is that there could hypothetically be infinitely many other signal models that peak exactly in any CR definition. However, this choice of signals corresponds to a relatively well-motivated set of processes and is representative of the type of signals that is expected to be probed by the analysis.

The metric for signal suppression is chosen based on the approximate theoretical significance $N_{\mathrm{sig}}/\sqrt{N_{\mathrm{bkg}}}$, where N_{sig} and N_{bkg} are the number of signal and background events in a region, respectively. The target is to find a CR which suppresses this significance by roughly a factor of ten compared to the analysis region. While N_{bkg} is approximated by the number of data events in a chosen phase space region, N_{sig} is chosen to correspond to the 95% CL quark-antiquark final state cross section limit that was found in the inclusive CMS dijet search [17], assuming a signal acceptance of 50% to convert the cross section into a number of events. In order to reduce the effect of statistical fluctuation, the full MC simulation samples were used to determine the efficiency of a phase space selection and scaled down to the respective target N_{sig}. The full summary of the significance fraction with respect to the analysis region is shown in Tab. 7.7 for all signal models and different CR definitions discussed in the following.

The criterion of complementarity to the analysis region is met trivially by inverting the $|\Delta\eta| < 1.3$ selection defining the analysis region. Figure 7.29 shows the passing fraction of data events (left) and example signal MC simulation events (right) as a function of a varying lower $|\Delta\eta|$ boundary. The signal suppression of a mere $|\Delta\eta| > 1.3$ selection is visibly insufficient, since the fraction of W$_{\mathrm{KK}}$ and b* events in this CR is approximately the same as in the analysis region $|\Delta\eta| < 1.3$. We thus restrict the CR lower $|\Delta\eta|$ boundary to 2.0 instead.

From Fig. 7.30 (left), it can be seen that the $|\Delta\eta| > 2.0$ selection results in a significant change in m_{jj} distribution with respect to the analysis region. This can be understood from the invariant mass of a dijet system in the highly relativistic limit. i.e., with negligible jet masses compared to their total energies:

$$m_{jj}^2 = 2p_{\mathrm{T},1}p_{\mathrm{T},2}\left(\cosh\Delta\eta - \cos\Delta\phi\right), \tag{7.16}$$

where $p_{\mathrm{T},1}$ and $p_{\mathrm{T},2}$ are the transverse momenta of the two jets. Higher values of $|\Delta\eta|$ correlate with high m_{jj}. For a more realistic proxy of the m_{jj} distribution in the analysis region, we additionally restrict the CR by an upper $|\Delta\eta|$ boundary of 2.5. The resulting m_{jj} distribution is shown in Fig. 7.30 (right), which follows closely the distribution in the analysis region apart from an initial increase in the number of events at low m_{jj}.

Table 7.7: Summary of the significance fraction of different CR definitions with respect to the analysis region, shown for all considered signal models. The columns each correspond to a CR definition, and the rows denote the ratio of the significance with respect to the significance of an inclusive dijet resonance search, except the first and second rows, which show the number of data events and the fraction with respect to a pure $2.0 < |\Delta\eta| < 2.5$ sideband, respectively. Blue values indicate a suppression by a factor of ten or more, while orange values do not meet this criterion. Bold orange values indicate missing the target suppression by a factor of two or more. The CR definitions are defined in the text.

Sample	$\Delta\eta$ only	$\Delta\eta$&jet veto	$\Delta\eta$&jet veto&p_T asymmetry	$\Delta\eta$&jet veto&p_T asymmetry$\|A$
Data events in control region	14 251 002	13 311 775	4 830 300	7 599 923
Data efficiency $\Delta\eta$ only	100%	93%	34%	53%
G → WW, $M = 2.0\,\text{TeV}$	0.03	0.03	0.01	0.02
G → WW, $M = 2.5\,\text{TeV}$	0.04	0.03	0.02	0.01
G → WW, $M = 3.0\,\text{TeV}$	0.03	0.03	0.01	0.01
G → WW, $M = 3.5\,\text{TeV}$	0.04	0.03	0.01	0.01
G → WW, $M = 4.0\,\text{TeV}$	0.03	0.03	0.01	0.01
G → ZZ, $M = 2.0\,\text{TeV}$	0.02	0.02	0.02	0.02
G → ZZ, $M = 2.5\,\text{TeV}$	0.03	0.03	0.01	0.01
G → ZZ, $M = 3.0\,\text{TeV}$	0.03	0.03	0.01	0.01
G → ZZ, $M = 3.5\,\text{TeV}$	0.04	0.04	0.01	0.01
G → ZZ, $M = 4.0\,\text{TeV}$	0.04	0.03	0.01	0.01
G → ZZ, $M = 4.5\,\text{TeV}$	0.04	0.03	0.01	0.01
G → ZZ, $M = 5.0\,\text{TeV}$	0.04	0.03	0.01	0.01
G → ZZ, $M = 5.5\,\text{TeV}$	0.04	0.03	0.01	0.01
G → ZZ, $M = 6.0\,\text{TeV}$	0.04	0.03	0.01	0.01
W′ → WZ $M = 2.0\,\text{TeV}$	0.09	0.09	0.03	0.03
W′ → WZ $M = 2.5\,\text{TeV}$	0.09	0.09	0.03	0.03
W′ → WZ $M = 3.0\,\text{TeV}$	0.09	0.09	0.03	0.03
W′ → WZ $M = 3.5\,\text{TeV}$	0.10	0.09	0.02	0.03
W′ → WZ $M = 4.0\,\text{TeV}$	0.09	0.09	0.03	0.03
W′ → WZ $M = 4.5\,\text{TeV}$	0.09	0.09	0.02	0.02
Z′ → WW $M = 2.0\,\text{TeV}$	0.08	0.08	0.02	0.02
Z′ → WW $M = 2.5\,\text{TeV}$	0.09	0.09	0.03	0.03
Z′ → WW $M = 3.0\,\text{TeV}$	0.09	0.09	0.02	0.02
Z′ → WW $M = 3.5\,\text{TeV}$	0.09	0.09	0.02	0.02
Z′ → WW $M = 4.0\,\text{TeV}$	0.09	0.09	0.02	0.02
Z′ → WW $M = 4.5\,\text{TeV}$	0.08	0.08	0.02	0.02
b^* → tW $M = 2.0\,\text{TeV}$	**0.22**	**0.21**	0.12	0.10
b^* → tW $M = 2.6\,\text{TeV}$	**0.20**	0.18	0.09	0.08
b^* → tW $M = 3.0\,\text{TeV}$	**0.21**	0.19	0.09	0.08
b^* → tW $M = 3.6\,\text{TeV}$	**0.22**	0.18	0.08	0.07
b^* → tW $M = 4.0\,\text{TeV}$	**0.23**	0.19	0.07	0.07
W_{KK} → RW → 3W $M = 2.0\,\text{TeV}$, $R_0 = 0.1$	**0.28**	**0.28**	0.10	0.09
W_{KK} → RW → 3W $M = 2.0\,\text{TeV}$, $R_0 = 0.2$	**0.22**	0.20	0.14	0.14
W_{KK} → RW → 3W $M = 2.0\,\text{TeV}$, $R_0 = 0.3$	**0.21**	0.17	0.12	0.16
W_{KK} → RW → 3W $M = 2.5\,\text{TeV}$, $R_0 = 0.08$	**0.33**	**0.32**	0.09	0.08
W_{KK} → RW → 3W $M = 2.5\,\text{TeV}$, $R_0 = 0.1$	**0.32**	**0.32**	0.10	0.09
W_{KK} → RW → 3W $M = 2.5\,\text{TeV}$, $R_0 = 0.2$	**0.30**	**0.23**	0.15	0.18
W_{KK} → RW → 3W $M = 2.5\,\text{TeV}$, $R_0 = 0.3$	**0.25**	0.16	0.12	0.15
W_{KK} → RW → 3W $M = 3.0\,\text{TeV}$, $R_0 = 0.06$	**0.34**	**0.34**	0.09	0.08
W_{KK} → RW → 3W $M = 3.0\,\text{TeV}$, $R_0 = 0.08$	**0.33**	**0.32**	0.07	0.06
W_{KK} → RW → 3W $M = 3.0\,\text{TeV}$, $R_0 = 0.1$	**0.33**	**0.32**	0.09	0.09
W_{KK} → RW → 3W $M = 3.5\,\text{TeV}$, $R_0 = 0.06$	**0.34**	**0.33**	0.08	0.07
W_{KK} → RW → 3W $M = 3.5\,\text{TeV}$, $R_0 = 0.08$	**0.33**	**0.32**	0.07	0.06
W_{KK} → RW → 3W $M = 3.5\,\text{TeV}$, $R_0 = 0.1$	**0.34**	**0.33**	0.08	0.07
W_{KK} → RW → 3W $M = 3.5\,\text{TeV}$, $R_0 = 0.2$	**0.33**	0.19	0.09	0.15
W_{KK} → RW → 3W $M = 4.0\,\text{TeV}$, $R_0 = 0.08$	**0.36**	**0.35**	0.09	0.08
W_{KK} → RW → 3W $M = 4.0\,\text{TeV}$, $R_0 = 0.1$	**0.35**	**0.34**	0.09	0.09
W_{KK} → RW → 3W $M = 4.5\,\text{TeV}$, $R_0 = 0.06$	**0.35**	**0.33**	0.06	0.05
W_{KK} → RW → 3W $M = 4.5\,\text{TeV}$, $R_0 = 0.08$	**0.33**	**0.33**	0.06	0.06
W_{KK} → RW → 3W $M = 4.5\,\text{TeV}$, $R_0 = 0.1$	**0.33**	**0.32**	0.09	0.08

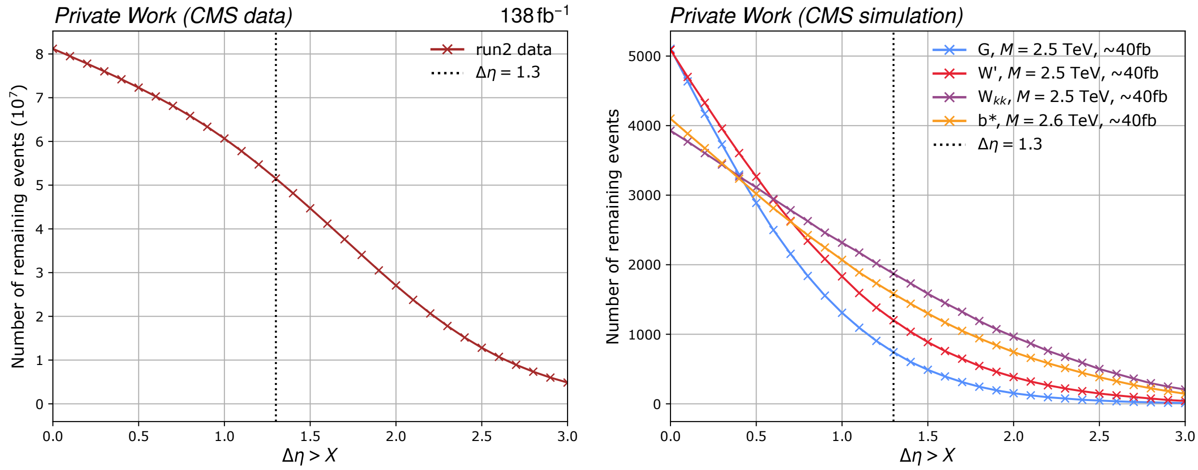

Figure 7.29: Number of data events (left) and different signal MC simulation events (right) passing a varying lower $|\Delta\eta|$ boundary. The other selections are the same as in the analysis region.

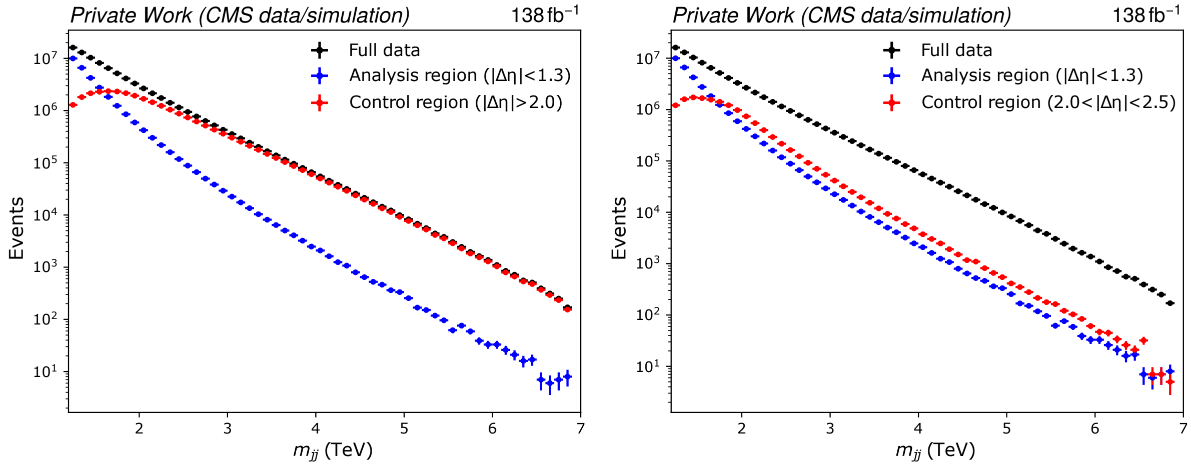

Figure 7.30: Distribution of the dijet invariant mass in the full data (black), the analysis region (blue), and a $|\Delta\eta|$ CR definition (red). The left plot shows a one-sided $|\Delta\eta| > 2.0$ selection, while the right plot shows a $2.0 < |\Delta\eta| < 2.5$ selection.

Table 7.7 shows that this $2.0 < |\Delta\eta| < 2.5$ selection already suppresses a significant fraction of events for all G, W$'$ and Z$'$ signal samples. In the b* and W$_{\mathrm{KK}}$ samples, however, the suppression is not yet sufficient by our criterion. These signals are characterized by an intermediate heavy particle (top quark or radion) whose decay products are resolved into two separate jets. We reduce their presence by vetoing on a third jet with a transverse momentum above 300 GeV. This type of third-jet veto is 90% efficient on the data, but further reduces the presence of the b* (except the lightest) and W$_{\mathrm{KK}}$ (except the $R_0 \leq 0.1$) signals, as shown in the second column of Tab. 7.7.

For the W$_{\mathrm{KK}}$ signals with low radion mass, the decay products are sufficiently boosted to be merged into a balanced dijet system. We can thus further suppress these signals by requiring a minimum jet p_T asymmetry:

$$\frac{|p_{\mathrm{T},1} - p_{\mathrm{T},2}|}{p_{\mathrm{T},1} + p_{\mathrm{T},2}} > 0.1. \tag{7.17}$$

The distribution of the jet p_T asymmetry can be seen in Fig. 7.31 (left), and the effect of this selection in the third column of Tab. 7.7. While this selection does result in the desired significance suppression of approximately a factor of 10, it has a low data efficiency of 34%. It should be noted that it is possible to define an additional variable A based on how well the relativistic approximation of Eq. 7.16 holds (setting $\cos\Delta\phi = -1$):

$$A \equiv \frac{2p_{\mathrm{T},1}p_{\mathrm{T},2}}{m_{jj}^2}\left(\cosh\Delta\eta + 1\right). \tag{7.18}$$

This quantity is approximately unity if the approximation were fully accurate. Observing the distribution of A in Fig. 7.31 (right), we can see that the $\mathrm{W_{KK}}$ signal samples are also suppressed by requiring $A \notin [0.95, 1]$. We therefore require data events to pass either the jet p_T asymmetry or the A selection, which results in a 53% efficiency on top of the $|\Delta\eta|$ window.

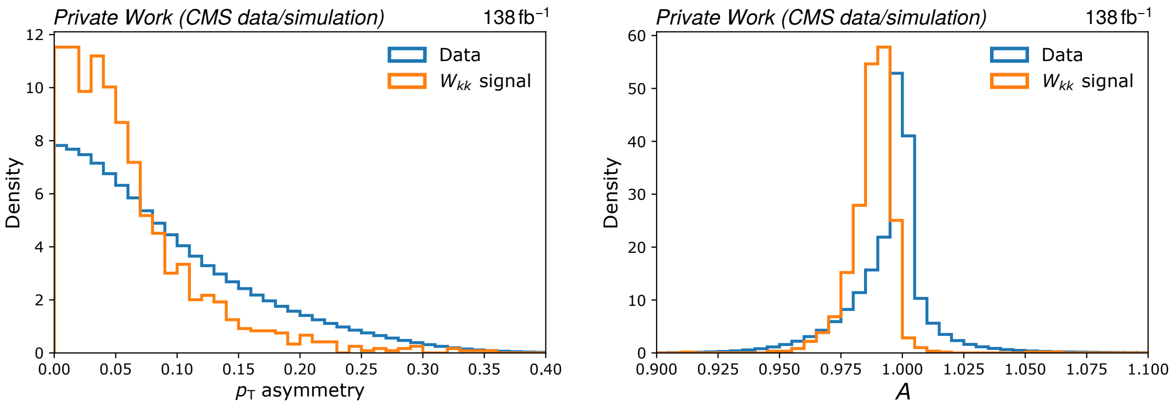

Figure 7.31: Jet p_T asymmetry (left) and dijet A distributions, as defined in Eq. 7.17 and 7.18, respectively. Shown are the real collision data (blue) and the $\mathrm{W_{KK}}$ signal sample with $m_{\mathrm{W_{KK}}} = 2500\,\mathrm{GeV}$ and $R_0 = 0.08$ (orange).

The final definition of the CR can be summarized as:

$$\text{AND}\begin{cases} 2.0 < |\Delta\eta| < 2.5 \\ \text{No third jet with } p_\mathrm{T} > 300\,\mathrm{GeV} \\ \text{OR}\begin{cases} \frac{|p_\mathrm{T}1 - p_\mathrm{T}2|}{p_\mathrm{T}1 + p_\mathrm{T}2} > 0.1 \\ A = p_{\mathrm{T},1}p_{\mathrm{T},2}(2\cosh\Delta\eta + 2)/m_{jj}^2 \notin [0.95, 1]. \end{cases} \end{cases} \tag{7.19}$$

There are a few $\mathrm{W_{KK}}$ mass combinations in Tab. 7.7 where the significance suppression is not exactly a factor of ten, but the CR definition is close to achieving this target while accepting a reasonable number of data events.

Analogous to the MC simulation studies in Sec. 7.7.1, the full CATHODE(-b) analysis strategy was performed in the data control region of Eq. 7.19. The CR selections result in an increasing m_{jj} count at the lower end of the considered region, which only changes into the desired negative slope after approximately 2 TeV. For this reason, the likelihood fits are restricted to a region above 2 TeV and the p-value scan only starts from 2.3 TeV. The selected m_{jj} spectra from CATHODE are shown in Fig. 7.32, and the corresponding CATHODE-b plots are collected in Fig. 7.33. As in the MC simulation case, the m_{jj} shapes remain smoothly falling with a good description via the background fit functions without the sudden appearance of artificial bumps. This is confirmed in the p-value scan results in Fig. 7.34, where no excess beyond 2σ is observed.

In the data CR, no signal sensitivity studies are performed, as the CR is defined to suppress any signal presence.

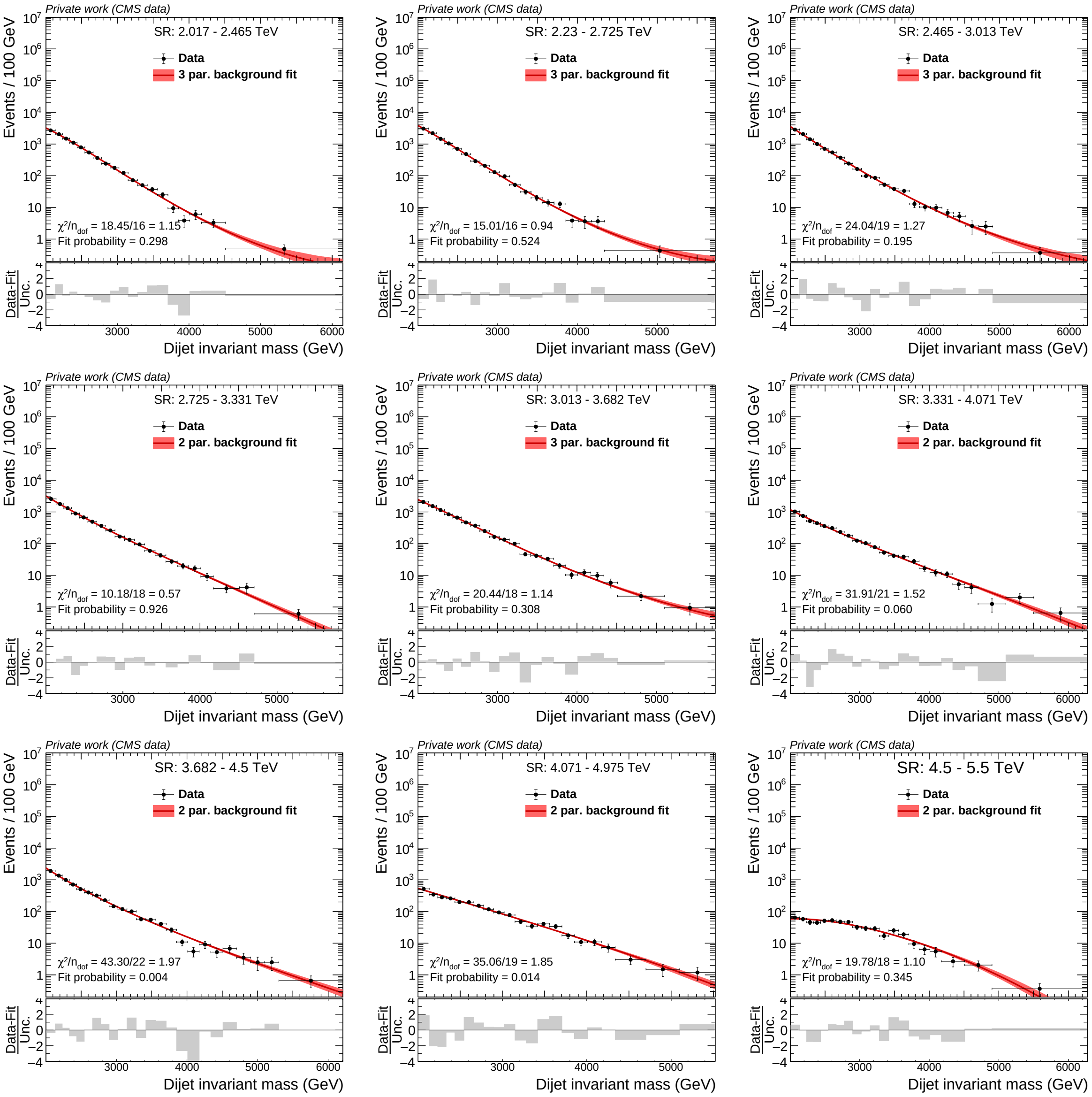

Figure 7.32: Data control region m_{jj} spectra after selecting the most anomalous 1% of events using CATHODE, as well as the best background fit, for every choice of signal region.

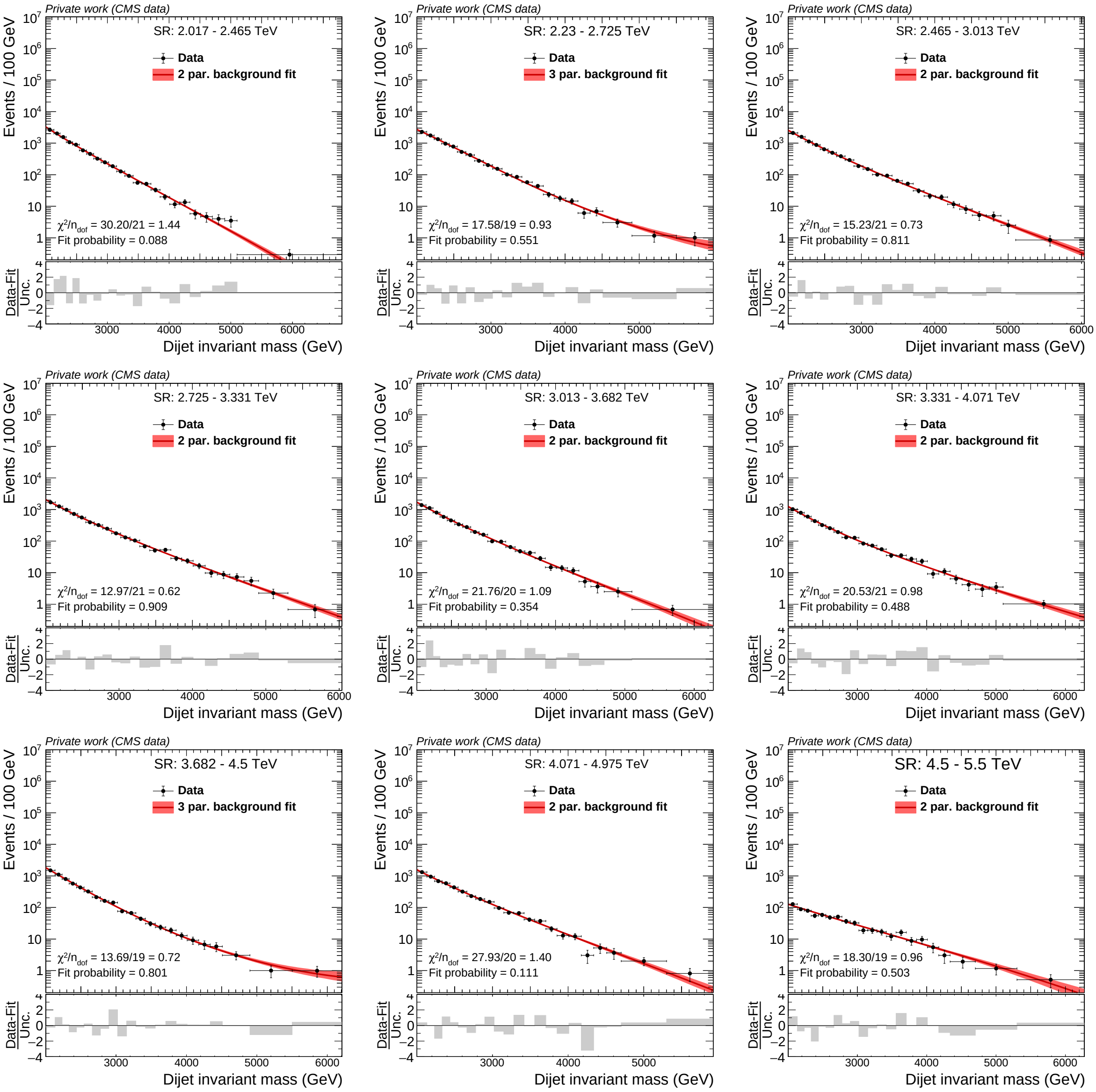

Figure 7.33: Data control region m_{jj} spectra after selecting the most anomalous 1% of events using CATHODE-b, as well as the best background fit, for every choice of signal region.

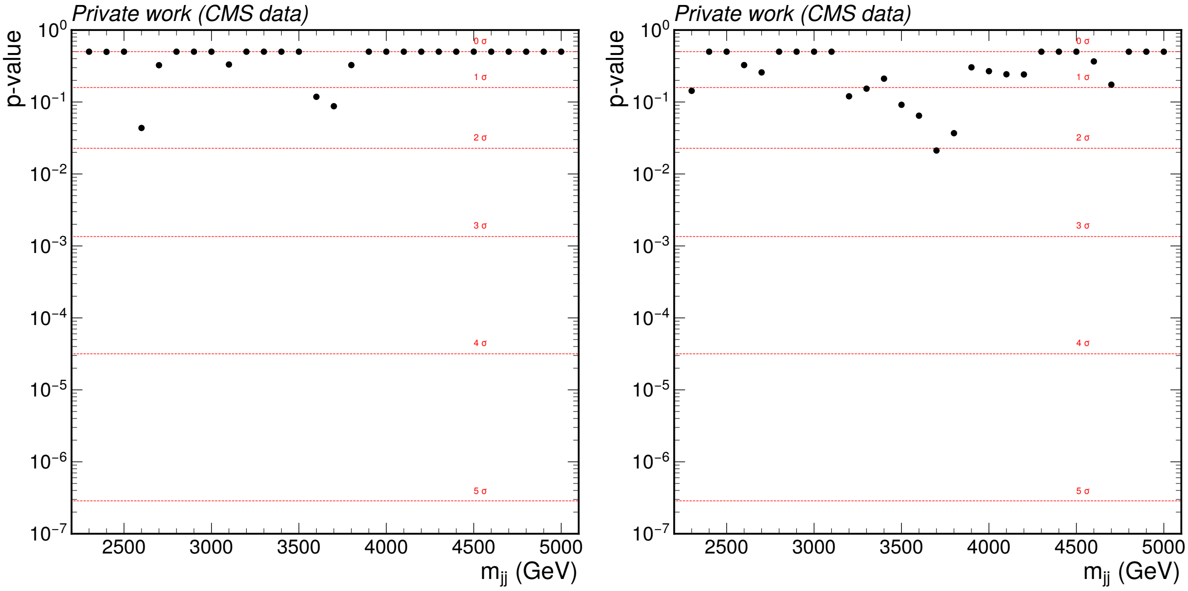

Figure 7.34: Data control region p-values as a function of m_{jj} after selecting the most anomalous 1% of events using CATHODE (left) and CATHODE-b (right).

7.7.3 Bias From Background Function Choice

A key assumption in the bump hunt is that the chosen fit function represents an unbiased estimate of the background. Specifically, the choice of fit function family (Eqs. 7.2 to 7.4) should ideally not have a significant effect on the result of the statistical analysis, but rather be sufficiently generic. We test this explicitly by sampling an m_{jj} toy dataset from another conventional fit function choice and then perform the profile likelihood fit ratio procedure with ours. This is repeated with different signal strengths in the toy sample and the difference between the true and the measured signal strength is quantified as a function of m_{jj}.

In order to quantify this fit function bias at some nominal value in m_{jj}, we first fit the 5-parameter function of an alternative function family choice to the most anomalous 1% of events in the MC background simulation dataset according to CATHODE trained with the most central SR around the nominal m_{jj}:

$$f_{\text{alt}}(x) \equiv \left(\frac{dN}{dm_{jj}}\right)_{\text{5, alt}}(x) = \frac{p_0}{x^{p_1}} \exp\left(-p_2 x - p_3 x^2 - p_4 x^3\right). \tag{7.20}$$

Here, again $x = m_{jj}/\sqrt{s}$ is the dijet invariant mass in units of the center-of-mass energy. We then combine the background fit function $f_{\text{alt}}(x)$ with the addition of a DCB signal shape at the nominal m_{jj}, which has a normalization corresponding to either zero (no signal contribution), a 2σ excess (where the injection strength is estimated from the 95% CL upper limit in the MC simulation dataset), or a 5σ excess (estimated via 2.5 times the 2σ injection). For each choice of signal strength, we generate 100 toy datasets from the combined alternative function and fit them with the 5-parameter background function choice used in this analysis (Eq. 7.4):

$$f_{\text{nom}}(x) \equiv \left(\frac{dN}{dm_{jj}}\right)_{5}(x) = p_0 \frac{(1-x)^{p_1}}{x^{p_2 + p_3 \log(x) + p_4 \log^2(x)}}, \tag{7.21}$$

plus the DCB signal shape with freely floating normalization.

The difference between the true and the measured signal strength, divided by the signal strength uncertainty, corresponds to the bias from fitting with one background function, f_{nom}, while the data follow the alternative choice, f_{alt}. These 100 individual bias values per signal strength are collected in a histogram, as shown in Fig. 7.35 (left), and fitted with a Gaussian function. The mean value from the fit is interpreted as the mean bias at the respective signal strength and nominal m_{jj} hypothesis. This entire procedure is repeated at every step in m_{jj} that is targeted in the significance scan. The result is shown in Fig. 7.35 (right), where the three colors correspond to the three different signal strengths. The uncertainty at every point is estimated by the fitted Gaussian standard deviation, divided by the square root of the number of toy datasets.

The result in Fig. 7.35 (right) shows a maximum bias of approximately 1.5σ in the absence of signal, i.e., one might at worst observe on average a 1.5σ excess in the data when fitting with the wrong function. This is deemed acceptable as excesses of this order of magnitude are not considered significant. Once a measurable signal strength is present, the maximum bias decreases further to only 0.6σ.

7.7.4 Bias From Full Statistical Procedure

While the study in Sec. 7.7.3 examined the bias introduced by fitting the background with a different functional form than the data actually follow, there might be other bias-inducing components of the full statistical procedure. This includes the choice of parametric background

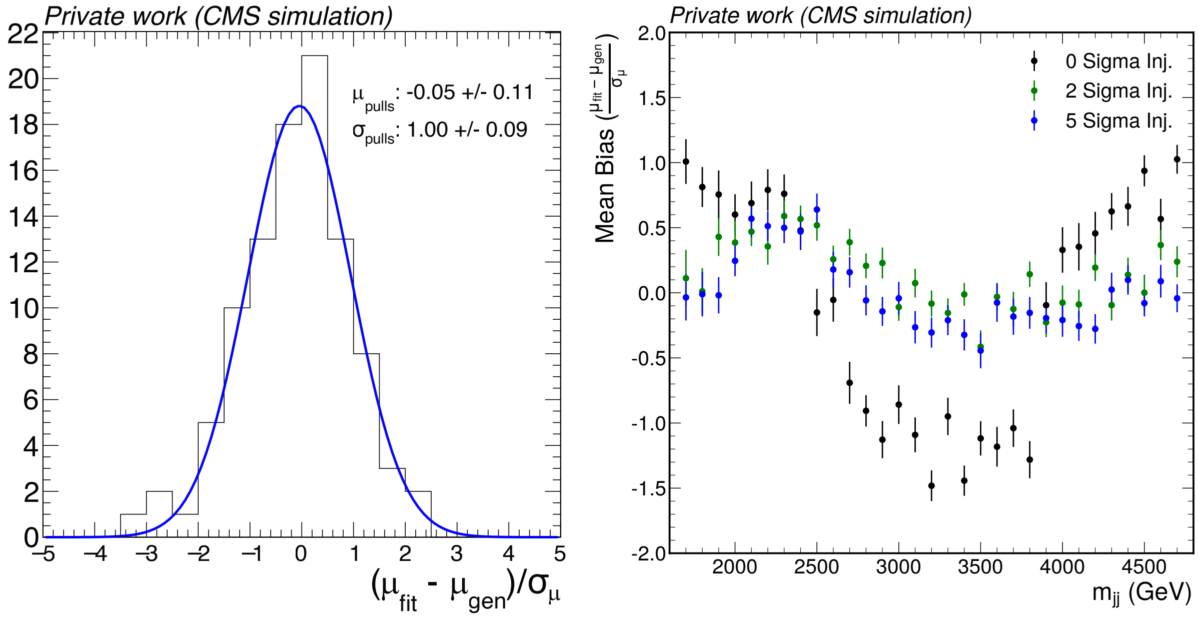

Figure 7.35: Left: histogram of the difference between measured and true signal strength in units of the signal strength uncertainty for 100 toy experiments where the background is sampled from one function and fitted with another. A Gaussian fit is applied to extract the mean and standard deviation. This example corresponds to the 2σ injection at a nominal m_{jj} of 3000 GeV. Right: mean bias and uncertainty extracted from the procedure shown in the left plot as a function of the nominal m_{jj}. The three colors correspond to different injections of signal into the toy dataset: without signal (black), one corresponding to a 2σ significance (green), and one corresponding to 5σ (blue).

fit function family, the selection of background fit parameterization via the F-test as well as the signal shape and the profile likelihood fits to determine the p-values under the background-only hypothesis.

Rather than employing CATHODE, we test the full bias by sampling a *random* 1% subset of the MC background simulation dataset and performing the full statistical procedure for computing background-only p-values. Since no signal is present in the dataset and the selection is randomized rather than based on an anomaly score, which might be subject to background sculpting, the resulting p-value should be uniformly distributed in the case of a well-calibrated statistical procedure. To estimate the p-value distribution, we repeat this sampling 1000 times with replacement for every dijet mass value tested during the p-value scan.

The resulting distribution of p-values, for all masses combined, is shown in Fig. 7.36, where the histogram has been normalized by the width of each bin. The left plot shows the full range of unique p-values, while the right plot zooms in on the region below 0.1, with a binning choice reflecting the denoted significance ranges in units of σ. It should be noted that all 1000 samples per mass value are combined into a single histogram, which assumes p-values at various masses to be approximately uncorrelated. Moreover, because of the choice of test statistic that only interprets upwards fluctuations of the data as an excess, the maximum p-value that can be obtained is 0.5. This is why all p-values of exactly 0.5 are collected in a single bin whose width extends up to 1. This bin is only partially displayed, but its content reflects the normalization with respect to the full width.

Figure 7.36 (left) shows two narrow peaks: one at the low end of the distribution, and one

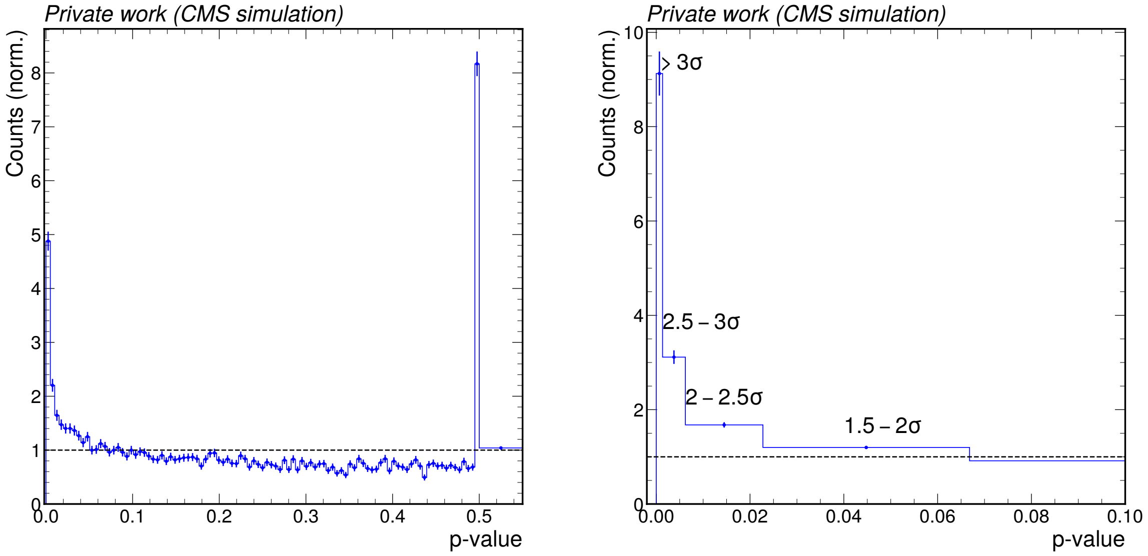

Figure 7.36: Area-normalized distribution of p-values obtained from the full statistical procedure on 1000 randomly sampled 1% subsets of the MC background simulation dataset, combining results for all considered nominal dijet masses. The left plot shows the full range of unique p-values. The chosen definition of test statistic results in a maximum numeric value of 0.5. These values are thus collected in a single bin whose full width is not displayed. The right plot shows a zoomed-in view of the p-values below 0.1 with a binning that corresponds to a significance in steps of 0.5σ.

at p-values just below 0.5. The latter corresponds to too many p-values with continuous values before the maximum of 0.5, likely due to numeric precision issues. It is deemed harmless, since it is not dominant compared to the full area of the neighboring highest bin, and the exact value of p-values around 0.5 are not relevant for the analysis. The peak towards low p-values is more striking, as it indicates an excess of p-values that would be interpreted as a hint for new physics. This is resolved more finely in Fig. 7.36 (right), which shows that a substantial fraction of the p-values correspond to a significance beyond 3σ.

One way of recalibrating the p-value distribution in Fig. 7.36 is to map it a posteriori to the desired uniform distribution. This can be achieved by a *quantile transformation*, which is a function that maps any distribution to a uniform one. This is shown in Fig. 7.37 (left) with a quantile transformation implemented with the Scikit-learn library [160]. There are two benefits of computing the quantile transformation in this context. First, one might consider applying this type of recalibration scheme as part of an analysis strategy if deemed necessary. Second, the resulting function captures the full bias introduced by the statistical procedure and can be used to better understand its impact. In order to achieve the latter, the transformation is visualized in terms of the significance in units of σ (i.e., the Z-score) in Fig. 7.37 (right). The plot reveals a departure from the identity map, which is increasing with higher significance. At a measured significance of 4σ, the transformation indicates that actually a $\sim 3\sigma$ excess is present, i.e., the statistical procedure overestimates the significance by approximately up to one unit of σ. Higher significance values could not be probed due to their low occurrence with only $\mathcal{O}(1000)$ trials.

The test above assumes a constant bias across the full dijet mass spectrum. The same methodology can be applied individually at each mass, i.e., deducing a quantile transformation

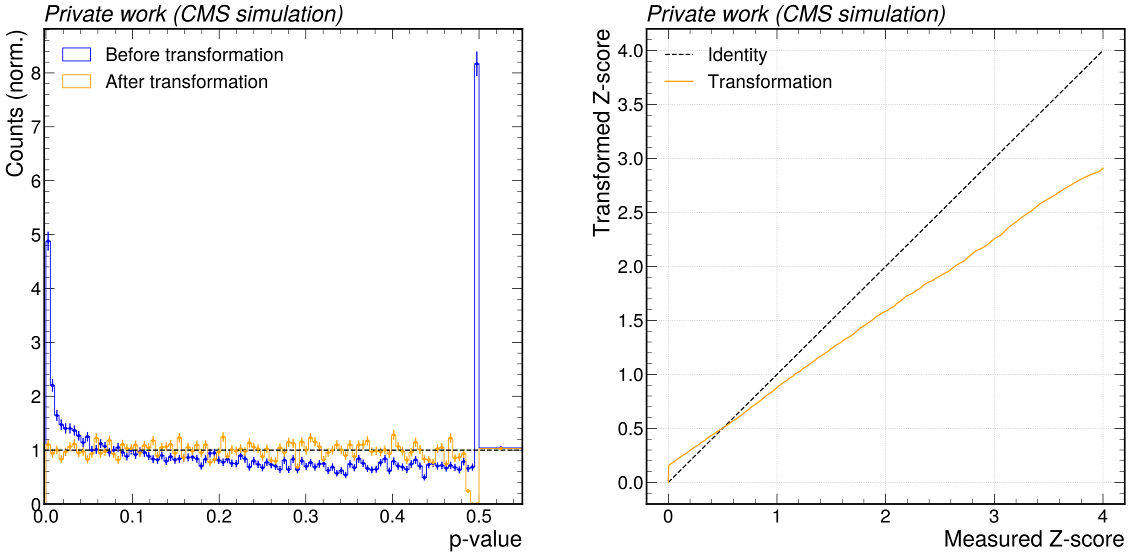

Figure 7.37: Left: quantile transformation defined to map to p-values of the random 1% samples of Fig. 7.36 (blue) to a uniform distribution (orange). Right: the transformation function visualized in terms of the significance in units of σ.

on distributions of p-values at each mass point. These transformations are visualized in Fig. 7.38. The transformed significance values are shown as a function of the dijet invariant mass, for measured significances of 1–4σ. Linear fits are shown for visual reference. The result shows that the bias is largest at low dijet masses and (as seen before) at high measured significance. At 1σ measured significance, the bias is very low and overall it decreases with increasing mass.

In order to understand at which point the statistical procedure introduces the $\mathcal{O}(1\sigma)$ biases, we repeat the procedure above, replacing the 1% random sampling with toy experiments drawn from the best fitting background function to the full MC background simulation dataset. As these m_{jj} values follow exactly one of the parametric functions in the fit function family of the F-test, it should factor out any issues arising from a suboptimal description of the data by the choice of fit function family. The resulting p-value distribution is shown in Fig. 7.39. The distribution is visibly more uniform than the one obtained from the random sampling of the MC background simulation dataset. The zoomed-in view in the right plot shows that the peak at high significance is still present, but substantially less pronounced compared to Fig. 7.36.

The bias can again be visualized in terms of the transformed and measured significance values after applying a quantile transformation. The resulting transformation is shown in Fig. 7.40 (left) when considering a mass-independent bias. Only a minor departure from the identity map is observed, above a measured Z-score of 3σ. The mass-dependent picture is shown in Fig. 7.40 (right), which confirms that the bias remains small across the full dijet mass spectrum. Only few values could be measured at a 4σ significance because at most masses this significance is not reached due to the (expectedly) low occurrence of these p-values, and thus the transformation is not well-defined in this case.

One thus observes that the bias of up to 1.5σ seen in Fig. 7.38 at high significances and low dijet mass does not persist when the dataset is replaced by toy experiments drawn from the best fitting background function. This indicates that the bias is primarily arising from data not being perfectly captured by the chosen family of fit functions. In fact, this relates to the

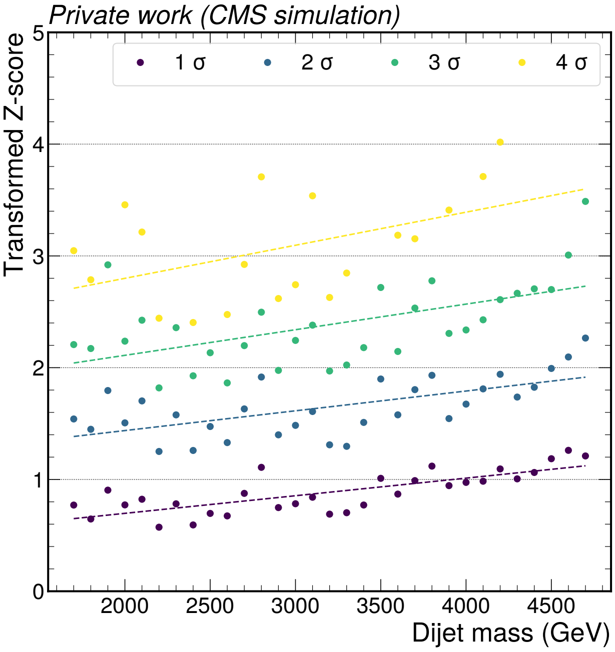

Figure 7.38: Illustration of the quantile transformations that individually recalibrate the p-value distributions at each dijet mass point. For significances of $1-4\sigma$, the resulting Z-scores are shown for each mass point. For each measured significance, a linear fit is shown as a dashed line.

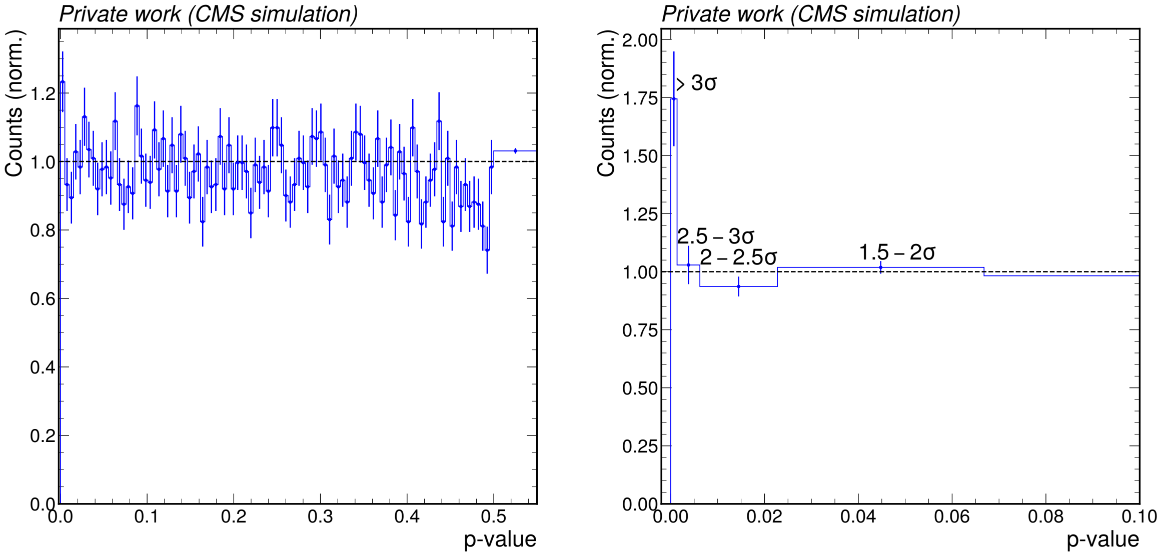

Figure 7.39: Area-normalized p-value distribution as in Fig. 7.36, but here the p-values are obtained from the full statistical procedure on toy datasets drawn from the best fitting background function to the full dataset.

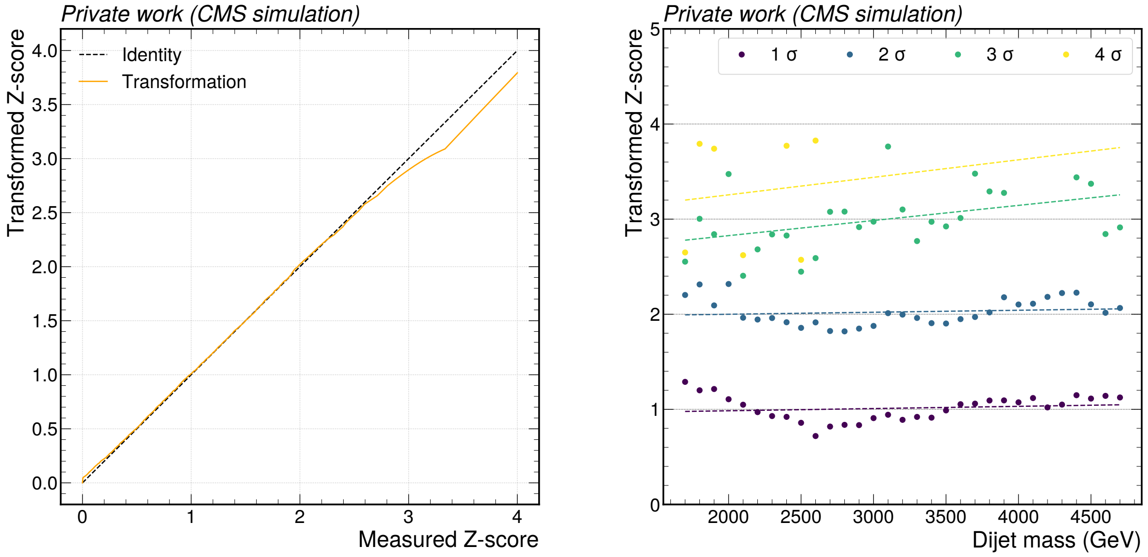

Figure 7.40: Visualization of the quantile transformation recalibration as in Figs. 7.37 (right) and 7.38, but for the p-value distributions obtained from toy datasets drawn from the best fitting background function to the full dataset.

same source of bias that was discussed in Sec. 7.7.3. There, the bias in the absence of signal was found to be mass dependent and reaching a maximum of approximately 1.5σ, as well. The bias in presence of signal has not been tested in the current study, but it is expected that it decreases with increasing signal strength, as seen in Sec. 7.7.3.

Following the discussion of Sec. 7.7.3, the bias shown here was deemed acceptable, as it is not expected to lead to a significant excess in the data. The suggested recalibration procedure based on a quantile transformation is thus not applied in the analysis. However, it could be a consideration for future analyses, especially if the statistical procedure is found to be biased at a level that could lead to a misinterpretation of the data.

7.8 Excess Interpretation

In case a significant excess is found in a model-agnostic search, it is important to characterize the distinct features of the events contributing to the excess. On the one hand, this serves as a sanity check to ensure that the excess is of a physical nature and not due to methodological issues, such as a systematic mismodeling of the background template. On the other hand, if the excess seems to be physical, one can try to interpret what type of signal is present. This understanding can be used to design dedicated studies, potentially with other datasets and experiments, eventually leading to the discovery of specific new physics phenomena.

In order to determine what phase space regions are found to be most signal like according to the CATHODE classifier, we developed two types of tests: observing the most anomalous patterns and evaluating the training feature importance. In this section, we assess their expressiveness by injecting two signal models into the MC simulation dataset with cross sections roughly corresponding to 5σ excesses in the CATHODE-based analysis.

The first test involves comparing the distributions of the most anomalous events in interpretable observables to those found without any anomaly selection. The most straightforward choice of features are the auxiliary features that are directly used as input to the classifier. However, by tracking the necessary event information, this test can be extended to any per-event feature. In particular, we choose to test the distributions of the 100 events with the highest anomaly score according to CATHODE. The features are chosen to be the four CATHODE classifier inputs (m_{j1}, Δm, $\tau_{41,j1}$, and $\tau_{41,j2}$) as well as the b-tagging score and the subjettiness ratios τ_{21}, τ_{32}, and τ_{43}, each with respect to the leading and the subleading jet. This allows us to qualitatively estimate the masses of intermediate particles, the presence of decays into b quarks, and the pronginess of the jets. It should be noted that the CATHODE classifier is expected to associate high output scores to background events as well if they are similar to an identifiable signal. This test thus shows generally what phase space patterns are most signal like, but does not provide a precise signal extraction method, as the background may be highly sculpted in these features.

The second test is inspired by popular interpretability methods in the ML literature. We evaluate the importance of the individual classifier input features by measuring the average change in the output score when the value of that feature is randomly permuted between events. In particular, we select the 100 events with the highest anomaly score, replace the values of one input feature with the values from 100 randomly selected (non-anomalous) events, and measure the average change in the anomaly score. This is repeated for each of the four input features individually. The aim is to rank the importance of each feature by their impact on the anomaly score. This may guide the decision on which features capture anomalous patterns and need to be considered in a dedicated study.

Events from the $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal sample with masses of $m_{\mathrm{X}} = 3000\,\mathrm{GeV}$ and $m_{\mathrm{Y}} = m_{\mathrm{Y}'} = 170\,\mathrm{GeV}$, are injected into the MC background simulation dataset with a cross section of 18.1 fb, which results in a 4.3σ excess with the analysis procedure based on CATHODE. The resulting feature distribution test is shown in the first three rows of Fig. 7.41. The m_{j1} distribution of the most anomalous events peaks at an intermediate particle mass of 170 GeV, whereas the Δm distribution peaks at zero, which accurately reflects the equality of the two daughter particle masses. The τ_{41} distributions of the two leading jets indicate a higher pronginess than one. This is resolved more finely into subsequent subjettiness ratios, where the τ_{21} values of both jets are visibly lower for the most anomalous events, while the τ_{32} and τ_{43} values approximately overlap with the inclusive distribution. This indicates a two-prong structure in both jets. The b-tagging scores are not significantly different between the two distributions, indicating a lack of b-quark content in the signal. While this test provides a correct qualitative description of the known signal properties, the permutation score, shown in the bottom of Fig. 7.41, is less informative, as it is limited to the four training features and does not account for the correlations between them. It primarily confirms that the mass of the heavier jet is a key feature to distinguish the signal from the background.

A $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ signal with $m_{\mathrm{W}'} = 3000\,\mathrm{GeV}$ and $m_{\mathrm{B}'} = 400\,\mathrm{GeV}$ is injected with a cross section of 120 fb, resulting in a 4.8σ excess using CATHODE. The interpretation study results are shown in Fig. 7.42. There, most anomalous events shift the m_{j1} distribution to peak at the B' mass of 400 GeV, and their Δm distribution peaks at $m_{\mathrm{B}'} - m_{\mathrm{t}} \approx 230\,\mathrm{GeV}$. This correctly reflects the expected daughter particle masses. The non-trivial pronginess of the jets is again indicated in the τ_{41} distributions, and resolved more cleanly in the subsequent subjettiness ratios. Here, a shift between full and anomalous samples is visible in τ_{32} of both jets, in addition to τ_{21}, indicating a three-prong structure in each leading jet. We also observe an increase in the average b-tagging score, as both jets are expected to contain b quarks. The permutation score, shown in the bottom of Fig. 7.42, again confirms that the mass of the leading jet is the most important feature to distinguish the signal from the background, but does not provide additional insights into the other features.

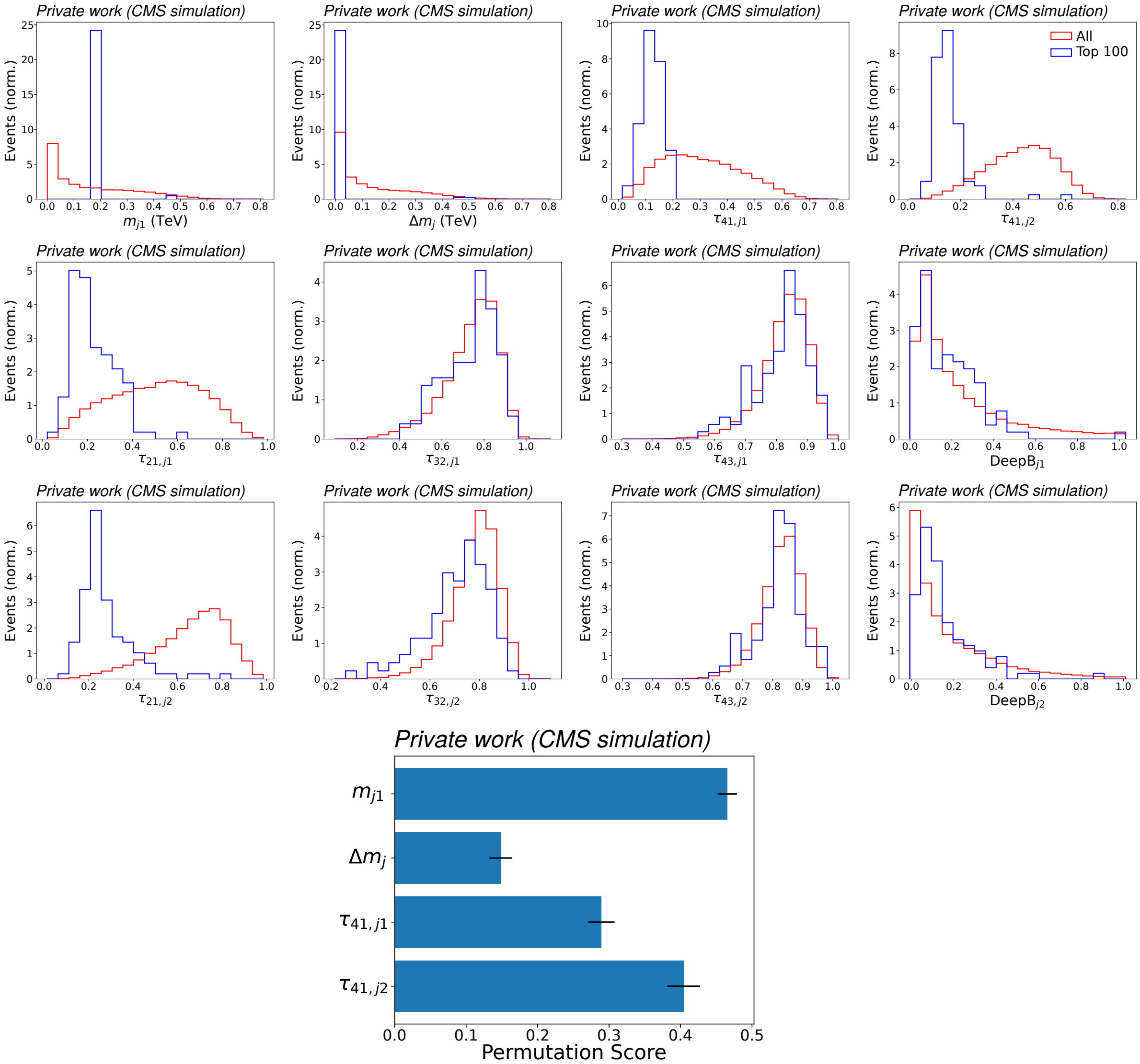

Figure 7.41: Interpretation study of the CATHODE classifier with the presence of an X → YY′ → 4q signal with $m_{\mathrm{X}} = 3000\,\mathrm{GeV}$ and $m_{\mathrm{Y}} = m_{\mathrm{Y'}} = 170\,\mathrm{GeV}$, resulting in a 4.3$\sigma$ excess. The first row shows the distributions of the most anomalous events (blue) compared to all events (red) in the features used for training. The second and third rows show additional substructure features of the leading and subleading jet, respectively. The bottom plot shows the training feature importance evaluation based on the permutation score.

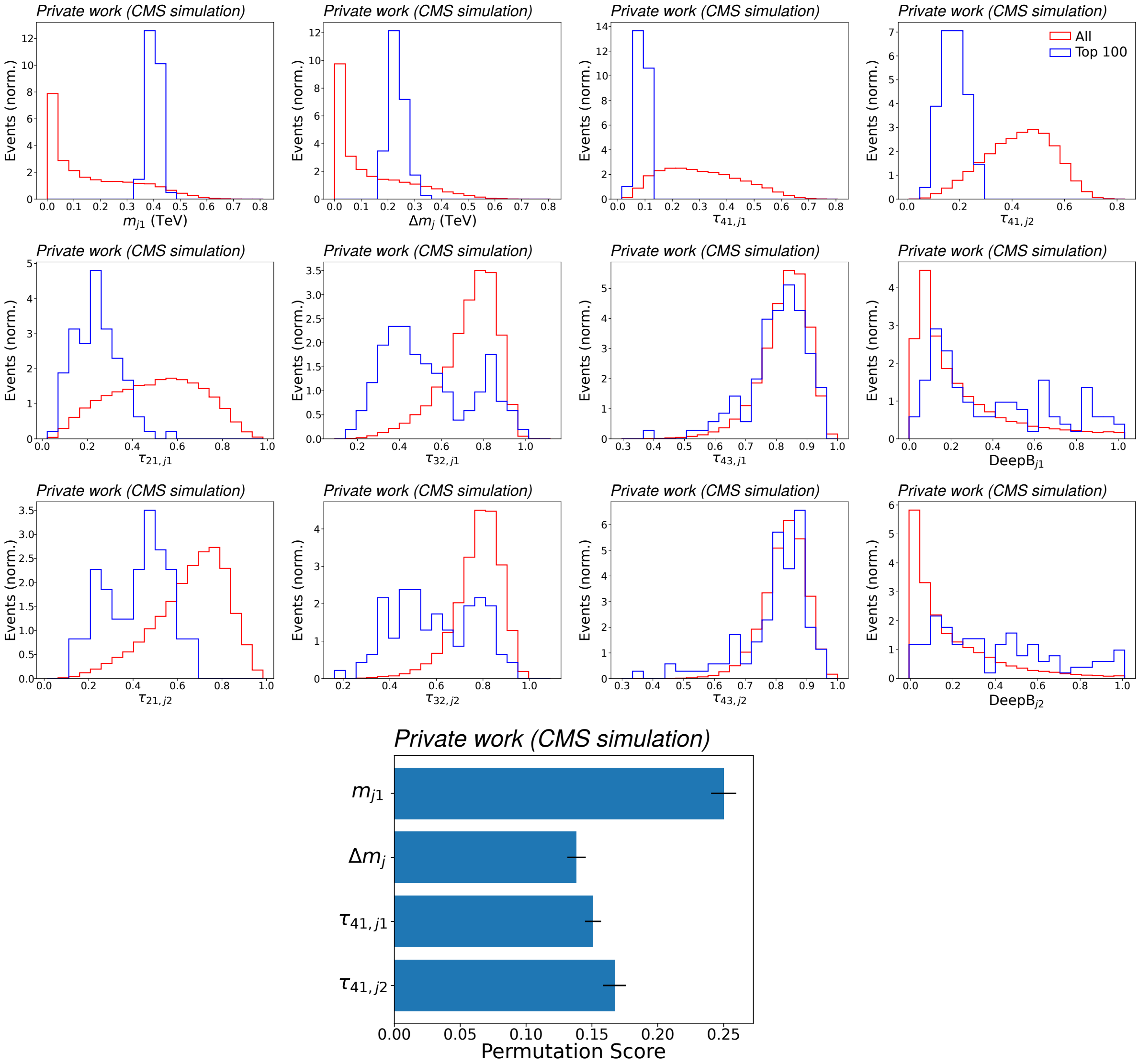

Figure 7.42: Interpretation study of the CATHODE classifier with the presence of a $\mathrm{W}' \rightarrow \mathrm{B}'\mathrm{t} \rightarrow \mathrm{bZt}$ signal with $m_{\mathrm{W}'} = 3000\,\mathrm{GeV}$ and $m_{\mathrm{B}'} = 400\,\mathrm{GeV}$, resulting in a 4.8σ excess. The first row shows the distributions of the most anomalous events (blue) compared to all events (red) in the features used for training. The second and third rows show additional substructure features of the leading and subleading jet, respectively. The bottom plot shows the training feature importance evaluation based on the permutation score.

7.9 Other Anomaly Detection Methods

While this thesis focuses on the CATHODE(-b) application within a CMS dijet search, it should be noted that four more methods were applied alongside within a common analysis strategy and publication [5]. These methods, which are briefly described in Sec. 4, are CWoLa Hunting, Tag N' Train, Variational Autoencoder With Quantile Regression, and Quasi Anomalous Knowledge. The following sections will discuss how these methods are applied to this analysis, and how they compare in MC simulation studies. Their results on real collision data will be shown in Sec. 7.10, in addition to the CATHODE(-b) results.

While these methods significantly differ in how the ML models are trained and applied to select the most anomalous events, they share the underlying data and MC simulation samples as well as reconstruction and selections, outlined in Sec. 7.2, before they are applied. Moreover, once the most anomalous events are selected, the statistical analysis described in Sec. 7.4 is performed analogously for all methods. Since CWoLa Hunting and TNT also follow the weakly supervised paradigm, their limit setting strategy follows the description in Sec. 7.6.

7.9.1 CWoLa Hunting

The application of the CWoLa Hunting method—Sec. 4.3.3 introduced the method conceptually and Sec. 5.3.1 described its application to a previous search for new physics—shares many similarities with the CATHODE search. The SR binning choice is equivalent to the one outlined in Tab. 7.6. However, the SB region for each choice of SR is restricted to only the most adjacent bins, in order to reduce the impact of correlations with the dijet mass. This necessity arises as the (weighted) SB regions are used directly as the background template, instead of interpolating a learned background description.

Another major difference with respect to the CATHODE-based dijet search is that the CWoLa Hunting classifier is implemented per jet instead of per event. This choice further reduces the correlations between input features and m_{jj}. Two separate classifier models are trained, one on the heavier jet per event and another on the lighter jet. The input features are chosen to be $x = \big(m, \tau_{21}, \tau_{32}, \tau_{43}, n_{\mathrm{pf}}, \mathrm{LSF}, \mathrm{DeepB}\big)$ where m is the soft drop mass of the corresponding jet, n_{pf} is the number of jet constituents according to the particle-flow algorithm, and LSF is the lepton subjet fraction [231] (with a subjet number input of $n = 3$ as used in Ref. [231]), which is a measure of enhanced contributions from leptons in the jet. These classifiers are implemented as fully connected neural networks.

A two-stage reweighting procedure is applied in order to produce an effective background template with the same distributions as the SR background. First, the left and right SB regions are reweighted to contribute equally to the total background and together match the effective number of events as the SR. In a second step, the SB jets are reweighted to match the p_{T} distribution of the SR jets. The sample weights resulting from multiplying the two contributions are used exclusively during classifier training.

An equivalent k-l-folding as in the CATHODE search is applied, with the addition that three models are trained per l-fold, each with a different network weights initialization. The model with the best performance is selected for further analysis. The performance metric for this model selection is based on the fraction of SR data events that are selected at a given selection efficiency in the background template. The two per-jet classifiers are combined into a single event-based score by first computing the percentile of each jet classifier output value with respect to the total distribution, and then taking the maximum of these two percentile scores.

The selection efficiency of 1% is chosen for SRs with bin centers below 3.1 TeV, 3% between 3.1 and 4.5 TeV, and 5% above 4.5 TeV. This choice is made to partially account for the exponentially decreasing number of events with increasing dijet mass. For this method, the selection efficiency is defined on the weighted average of the SB regions, rather than the SR. The corresponding anomaly score threshold is then applied to the SR and SB regions alike.

7.9.2 Tag N' Train

The TNT method, introduced in Sec. 4.3.3, is adapted to the dijet search in close analogy to CWoLa Hunting, as it essentially adds another measure to increase the signal faction in the SR data sample, while further purifying the background sample. This is based on the assumption that both jets from a signal event are anomalous, while background jets are uncorrelated from each other.

In order to reduce correlations with the dijet mass, the two leading jets in each event are not sorted by mass but randomly assigned to be "Jet 1" or "Jet 2" of the event. An autoencoder is used as an unsupervised preselection on "Jet 1" to decide whether "Jet 2" is signal like or background like, and vice versa, and these assignments are then used to build purer samples for the CWoLa classification task. While the original TNT proposal [125] trains two separate classifiers, one for "Jet 1" and one for "Jet 2", where each sample is purified by the autoencoder selection on the other jet per event, this search uses a simplified procedure based on a single classifier training per SR. The two sets of background-like and signal-like samples, one from each leading jet, are mixed into a single training set, and the classifier is trained on this combined jet sample.

The input to the autoencoder is a p_T-weighted two-dimensional histogram of jet constituents in the η-ϕ plane, which can be interpreted as a *jet image*. The autoencoder is implemented as a convolutional neural network where the encoder yields a 6-dimensional latent space bottleneck, followed by a decoder mirroring the encoder architecture. The loss function is the MSE between the input and the reconstructed image. It is trained on jets from the SB region (defined as the neighboring SR bins), where both leading jets per event are used as training samples.

The signal-enriched and background-enriched samples of the CWoLa classifier training are then constructed as follows. The signal-enriched sample comprises the "Jet 1" of every SR event where the autoencoder loss of "Jet 2" is in the upper 20% of the distribution, and vice-versa every SR "Jet 2" with the respective "Jet 1" in the upper 20% of autoencoder loss values. The background-enriched sample consists of SB jets (as used in CWoLa Hunting), and it is enhanced by adding the bottom 40% of the SR data jets. The classifier is trained to distinguish these modified samples of jets with otherwise the same procedure as in CWoLa Hunting, including the input features and the reweighting scheme.

7.9.3 Variational Autoencoder With Quantile Regression

This VAE-based analysis strategy follows conceptually the outlier detection paradigm described in Sec. 4.2. A VAE is trained to reconstruct jets from the CR data, defined in Eq. 7.19, which is dominated by QCD jets. A high loss in the reconstruction of the SR jets is interpreted as more anomalous, as these jets are less likely to be QCD jets. Contrary to the weakly supervised approaches, the VAE does not rely on further splitting the analysis region into SR and SB regions, thus a single training is sufficient to cover the entire m_{jj} spectrum.

As there are inherent kinematic differences between the CR and the analysis region, a sampling procedure is first applied to match the CR jets to the expected background distribution. The MC background simulation sample serves as a prior for the kinematic jet distributions. In every event, the two leading jets are each replaced by the CR jet that is most similar in terms of p_T, η and ϕ. Each of the two jets is then represented by the Cartesian 3-momentum vectors of the 100 highest-p_{T} jet constituents according to the particle-flow algorithm, ordered by a reclustering of the jet via the Cambridge-Aachen algorithm. The resulting 100×3 matrix, with zero-padded values in case of fewer jet constituents than 100, serves as input to a 1D convolutional neural network encoder with 12 output nodes that model the mean and variance of the 6-dimensional Gaussian latent space, followed by a decoder network. The loss function is the sum of the Chamfer distance [232] between the input and the reconstructed jet constituent matrix, and the Kullback-Leibler divergence between the latent space distribution and the unit Gaussian prior. The same loss serves as per-jet anomaly score on analysis region events. A per-event score is computed as the minimum of the scores of the two leading jets.

Substantial correlations are expected between the anomaly score and the dijet invariant mass, thus the method relies on explicit measures to avoid background sculpting. In this case, the highly correlated anomaly score is decorrelated after training via a quantile regression (QR) procedure. Three quantile working points are chosen: 90%, 95%, and 99%, corresponding to the 10%, 5%, and 1% selection efficiency in the SR, respectively. For each of the three quantiles, a threshold function is learned depending on the dijet mass values, such that the same selection efficiency persists across the entire analysis region m_{jj} spectrum along this threshold. Consequently, the shape of the m_{jj} distribution remains the same before and after this selection on the VAE score. The QR models are implemented as fully connected neural networks, and a k-fold cross-validation procedure with $k = 4$ ensures that the QR training samples are orthogonal to inference events, and the decision threshold function is smoothened across m_{jj} with a third-order polynomial.

The significance scan procedure is performed on the most anomalous 10% of the events. For the limit setting, these are separated more finely into three distinct categories: the most anomalous 1%, the events from the 95th to the 99th percentile, and the events from the 90th to the 95th percentile. A single profile likelihood fit is performed in these three categories simultaneously, using the relative signal yields that are known from simulation of the particular signal model that is tested.

7.9.4 Quasi Anomalous Knowledge

The QUAK method was introduced in Sec. 4.4 and can be regarded as a bridge between supervised and unsupervised anomaly detection methods. It is the only method in the CMS dijet anomaly search where MC simulation samples are directly used to train the ML model.

Prior signal and background knowledge is incorporated by normalizing flows that learn the density of simulated samples. This is a departure from the original QUAK implementation, which was based on VAEs. One flow model is trained on the MC background simulation sample, where the m_{jj} distribution is flattened through a resampling scheme in order to avoid sculpting. For the signal, six normalizing flows are trained, each with a mixture of multiple available MC signal simulation samples with the same intermediate particle masses: (80, 80), (80, 170), (80, 400), (170, 170), (170, 400), and (400, 400) GeV, where the parentheses denote the masses of the two daughter particles resulting from the heavy resonance decay. The input features are properties of the two leading jets, which are ordered by decreasing mass and each described by $\rho = \frac{m}{p_{\mathrm{T}}}$, τ_{21}, τ_{32}, τ_{43}, $\tau_s = \frac{\sqrt{\tau_{21}}}{\tau_1}$, the DeepB score, and the number of jet constituents n_{pf}.

The losses of the flows correspond to the negative log likelihood on each of the respective datasets. After training, they are decorrelated from m_{jj} with a principal component analysis projection, followed by a Box-Cox transformation [233] and standardization that transforms the loss distributions into an approximate standard normal distribution. In order to reduce the QUAK space to two dimensions, the losses of the six signal flows are combined into a single value via the signed L5 norm: $L_5^{\mathrm{sgn}}(x) = \mathrm{sgn}(\bar{\ell})|\bar{\ell}|^{1/5}$ where $\bar{\ell} = \sum_{i=1}^{6} \mathrm{sgn}(\ell_i)|\ell_i|^5$ and ℓ_i are the loss values of the six signal flows.

QUAK differs from the other methods insofar that it returns a two-dimensional anomaly score, rather than a scalar one. In order to select events from this plane, a data-driven contour building algorithm is employed, using background templates constructed from the MC background simulation in m_{jj} sidebands. For each tested signal mass hypothesis m_{H}, the SR is defined as $m_{jj} \in [m_{\mathrm{H}} - 400\,\mathrm{GeV}, m_{\mathrm{H}} + 200\,\mathrm{GeV}]$, and the SB is defined as the two 500 GeV-wide intervals above and below the SR. First, the 50% least background-like events are selected, and a 2D QUAK space histogram is constructed in the SB. The same binning is applied to SR events. Starting from the least populated SB bin, more bins are iteratively included in the contour region until at least 1500 events or $500 \times (m_{\mathrm{H}}/\mathrm{TeV} - 3)^3$ events have been selected within the corresponding SR histogram. The SR events falling into the contour region are selected for the common downstream analysis via a bump hunt.

While the procedure above mitigates the higher dimensionality of the QUAK anomaly score, it is still useful for comparison with other methods to reduce it to a scalar value. In particular, Sec. 7.9.6 quantifies the correlation between the anomaly scores of each method, which would be difficult with one method having a two-dimensional score. For this purpose, the QUAK score is compressed into a one-dimensional score for each event by locating it in the 2D histograms discussed above and scoring it by the inverse of the MC background simulation sideband event occupancy in the corresponding bin, as well as its alignment with the signal-like axis.

7.9.5 Classical Benchmarks

In addition to CATHODE(-b) and the four other anomaly detection methods, we consider four benchmarks that serve as proxy for classical search strategies with different degrees of model assumptions.

Firstly, the *inclusive* search refers to performing the full analysis chain without any signal-enhancing anomaly detection method, i.e., computing p-values and upper limits directly from the preselected events. This is both the simplest and the most model-agnostic approach, and serves as a lower bound on what performance can be seen as non-trivial.

Another interesting point of comparison is a small set of loose selections that are expected to enhance signal events involving jets with more complex substructure than the QCD background. For this purpose, two selections are applied: a *2-prong* selection with $\tau_{21} < 0.4$ and a soft drop mass above 50 GeV, and a *3-prong* selection with $\tau_{32} < 0.65$ and the same criterion on the soft drop mass. These two benchmarks can be seen as a semi-agnostic approach, making assumptions about the number of subjets but otherwise retaining a degree of generality.

Lastly, the *model-specific QUAK* approach relies on the same working principle as QUAK but using the true signal MC simulation sample for training only a single signal flow for the QUAK space. This mimics a fully supervised search with the advantage that the existing framework for QUAK could be reused, in particular the background sculpting mitigation, which in the case of an independent supervised search strategy would need to be developed from scratch.

7.9.6 Complementarity

A key question that arises from deploying multiple anomaly detection methods in parallel is whether they will provide complementary information. If that is not the case, and the methods all find the same events to be anomalous based on the same patterns, it would be sufficient to use a single method and thus heavily reduce the computational cost.

One way to test the complementarity of methods is to study the correlations between their respective anomaly scores. For this purpose, we compute the anomaly score resulting from training each method on some MC simulation sample, and then compare their values per event in pairs. Figure 7.43 shows an example where 2D scatter plots are created from the anomaly scores of CATHODE in comparison to TNT, VAE-QR, and QUAK. The training was performed on the MC simulation dataset containing 20 fb of $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal events with masses of $m_{\mathrm{X}} = 3\,\mathrm{TeV}$ and $m_{\mathrm{Y}} = m_{\mathrm{Y}'} = 170\,\mathrm{GeV}$, and the plot is restricted to signal events. Before plotting, each anomaly score distribution is shaped to follow a unit Gaussian distribution, which accounts for the varying scales of anomaly scores between methods. The correlation is further quantified in terms of the Pearson correlation coefficient, which measures the linear correlation between variables, and the distance correlation (DisCo) score [234], which is a more general measure of dependence between variables. In all three cases shown, there is some correlation between methods, in particular between CATHODE and QUAK, but not to a degree that would suggest that the methods are redundant.

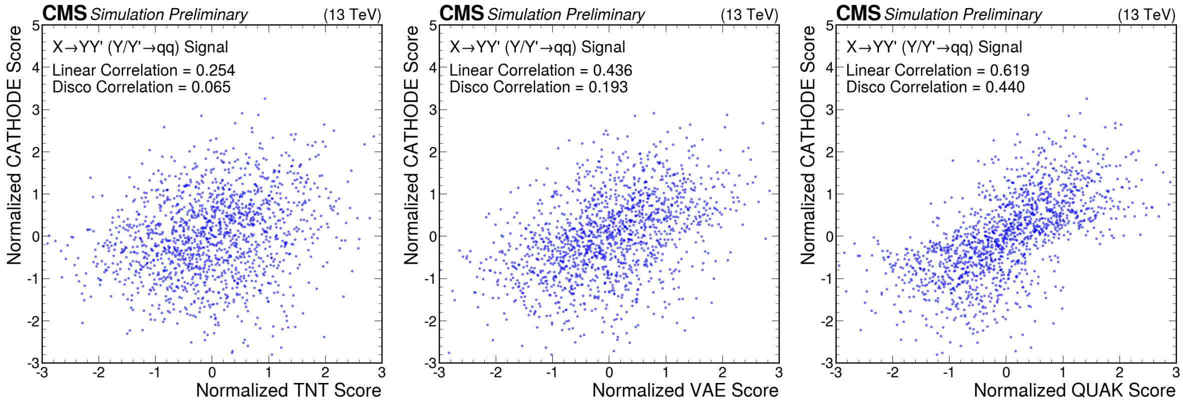

Figure 7.43: Scatter plots of the normalized anomaly score of CATHODE on $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal events, compared to the anomaly scores from TNT (left), VAE-QR (center), and QUAK (right). The correlation is further quantified by the Pearson (linear) correlation coefficients and the distance correlation (DisCo) score. Figure taken from Ref. [4].

The linear correlation coefficients of all pairs of methods are summarized in Fig. 7.44 for the same $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ signal described above (left), an injection of 60 fb of $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ signal events with masses of $m_{\mathrm{W}'} = 3\,\mathrm{TeV}$ and $m_{\mathrm{B}'} = 400\,\mathrm{GeV}$ (center), and the QCD background sample (right). The largest correlations in all three cases are observed between CWoLa Hunting and TNT, which is expected because the final classifier is trained on the same input features. CATHODE correlates most with QUAK on the signal events. In the background, minor correlations are only observed with respect to the VAE-QR method. CATHODE-b was not included in this study. The relationship between CATHODE and CATHODE-b is more straightforward, as the latter only adds two additional features to the former.

7.9.7 Performance Comparison

Since the MC simulation dataset facilitates full control over the exact signal and background contents, it can serve as benchmark dataset to compare the analysis strategies based on the various anomaly detection methods. The main metric for signal extraction performance during the proof-of-concept studies in Sec. 6 was the SIC curve. However, there exist multiple shortcomings when comparing conceptually very different methods with SIC curves. For instance, the weakly supervised approaches depend on how much signal is present in the dataset during training, whereas VAE-QR and QUAK are expected to perform similarly regardless of the signal

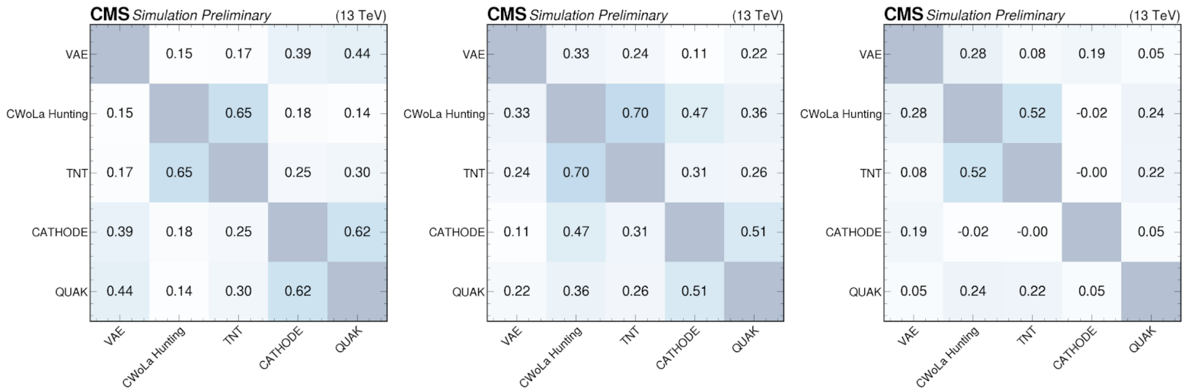

Figure 7.44: Linear correlation matrices of the normalized anomaly scores of VAE-QR, CWoLa Hunting, TNT, CATHODE and QUAK, on the $X \to YY' \to 4q$ signal events (left), the $W' \to B't \to bZt$ signal events (center), and the QCD background events (right). Figure taken from Ref. [4].

content. It also does not account for the later steps in the analysis chain, such as the contour construction in the QUAK space, or how well the signal bump can be extracted with the bump hunt. For this reason, this section does not focus on SIC curves as a performance metric.

A more complete performance indicator is the expected significance as a function of how much of a particular signal is present in the MC simulation sample. This involves the full analysis chain, including common preselections, the chosen selection efficiency, the background estimation and statistical analysis. This metric is compared in Fig. 7.45 for the same $X \to YY' \to 4q$ signal (left) and the $W' \to B't \to bZt$ process (right) as tested in Sec. 7.9.6.

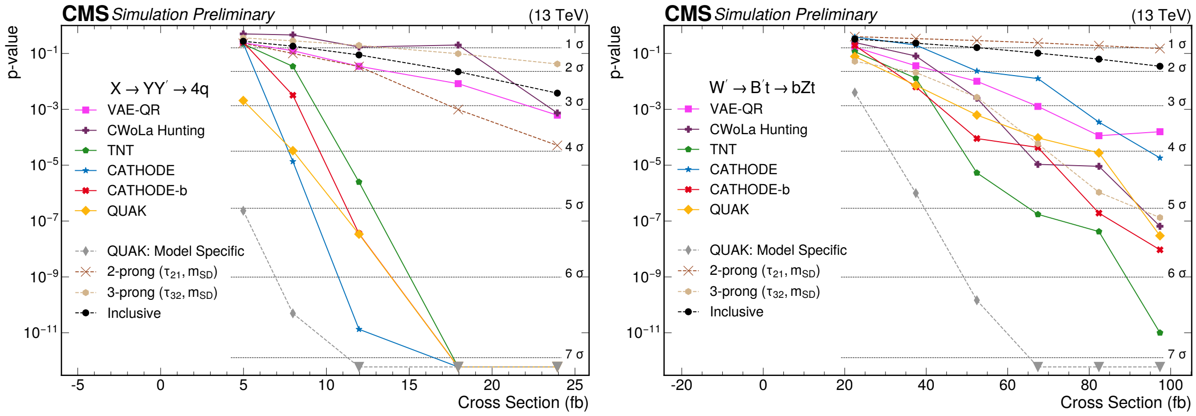

Figure 7.45: Expected p-values under the background-only hypothesis as a function of the signal cross section in the MC simulation, obtained from all anomaly detection methods in this analysis, as well as four benchmark methods that are either more supervised or more conventionally unsupervised. The left plot shows the results for the $X \to YY' \to 4q$ signal with masses of $m_X = 3\,\text{TeV}$ and $m_Y = m_{Y'} = 170\,\text{GeV}$, while the right plot shows the results for the $W' \to B't \to bZt$ signal with masses of $m_{W'} = 3\,\text{TeV}$ and $m_{B'} = 400\,\text{GeV}$. Figure taken from Ref. [4].

The performance comparison for the $X \to YY' \to 4q$ signal in Fig. 7.45 (left) shows injections up to an amount that results in inclusive p-values corresponding to less than 3σ. As expected from this signal with both jets containing two subjets, the 2-prong selection increases the sensitivity by up to 1σ, whereas the 3-prong selection further reduces it by a similar amount. A more striking effect is seen for most anomaly detection methods, notably CATHODE(-b), QUAK and TNT. At the lowest tested injection of 5 fb, QUAK has the highest sensitivity with almost 3σ. CATHODE, which depends more substantially on the amount of signal in the dataset surpasses it already at an injection of 8 fb with more than 4σ and exceeds 6σ with 12 fb. In this middle signal strength range, TNT and CATHODE-b also suddenly increase in sensitivity, crossing the 5σ threshold around 12 fb. The lower performance of CATHODE-b with respect to CATHODE is expected, given the absence of b decays in the signal. On the other hand, the VAE-QR and CWoLa Hunting methods struggle at outperforming the inclusive search for this signal and injection strengths. The model-specific QUAK benchmark remains significantly more sensitive compared to the model-agnostic anomaly detection methods, already reaching the 5σ threshold at the lowest injection of 5 fb. This is expected and confirms that while model-agnostic searches are powerful for their generality, they will always be outperformed by methods that are tailored to the specific signal model.

The $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ signal, on the other hand, provides an example of a signal with three subjets within both leading jets, with expected decays into b quarks. Figure 7.45 (right) shows an injection range where the inclusive paradigm reaches less than 2σ at the highest point, which can be significantly improved by the 3-prong selection that enables a 5σ discovery at the largest injection. The sensitivity of CATHODE is similar to the VAE-QR and roughly in the middle ground between inclusive and a more informed 3-prong selection. The latter is similarly sensitive on this signal as QUAK, CWoLa Hunting and CATHODE-b. The 3σ and 5σ thresholds are achieved roughly 15 fb earlier with CATHODE-b than with the simple 3-prong selection. The superiority of CATHODE-b over CATHODE can be understood by the non-trivial discrimination power in this signal that stems from the b-tagging information. This information is also used by TNT, which visibly performs best among all model-agnostic methods on this signal and achieves a 5σ discovery at 67 fb. Again, the model-specific benchmark shows that substantial improvements can still be obtained within the context of a dedicated search.

7.10 Results

Following the methodology described in the previous sections and after the extensive validation studies in Sec. 7.7, CATHODE(-b) was applied to the Run 2 collision data of the CMS Experiment. The results are presented separately for the significance scan (Sec. 7.10.1) and exclusion limits (Sec. 7.10.2) with specific focus on CATHODE and CATHODE-b, and then compared to the other anomaly detection methods.

7.10.1 Significance Scan

Figure 7.46 shows the dijet mass spectra of the most anomalous 1% of events in every SR when training and evaluating CATHODE on the collision data. The background fit function chosen by the F-test is also displayed in each case along with the corresponding χ^2/n_{dof} value, which quantifies the data-to-fit agreement. The background fit functions describe the selected m_{jj} distribution well in most SRs. The lowest agreement is seen in the SR from 2.465 to 3.013 TeV, where the χ^2/n_{dof} value is 1.75, translating to a fit probability of 1%. The pull plots of this fit reveal the largest upward fluctuation within the SR itself, between 2.6 and 2.8 TeV. The largest downward fluctuation is located at approximately 3.5 TeV.

The selected m_{jj} distributions of CATHODE-b and the corresponding background fit functions are shown in Fig. 7.47. In this case, the agreement between data and fit is generally good, with a fit probability of at least 8% in all SRs.

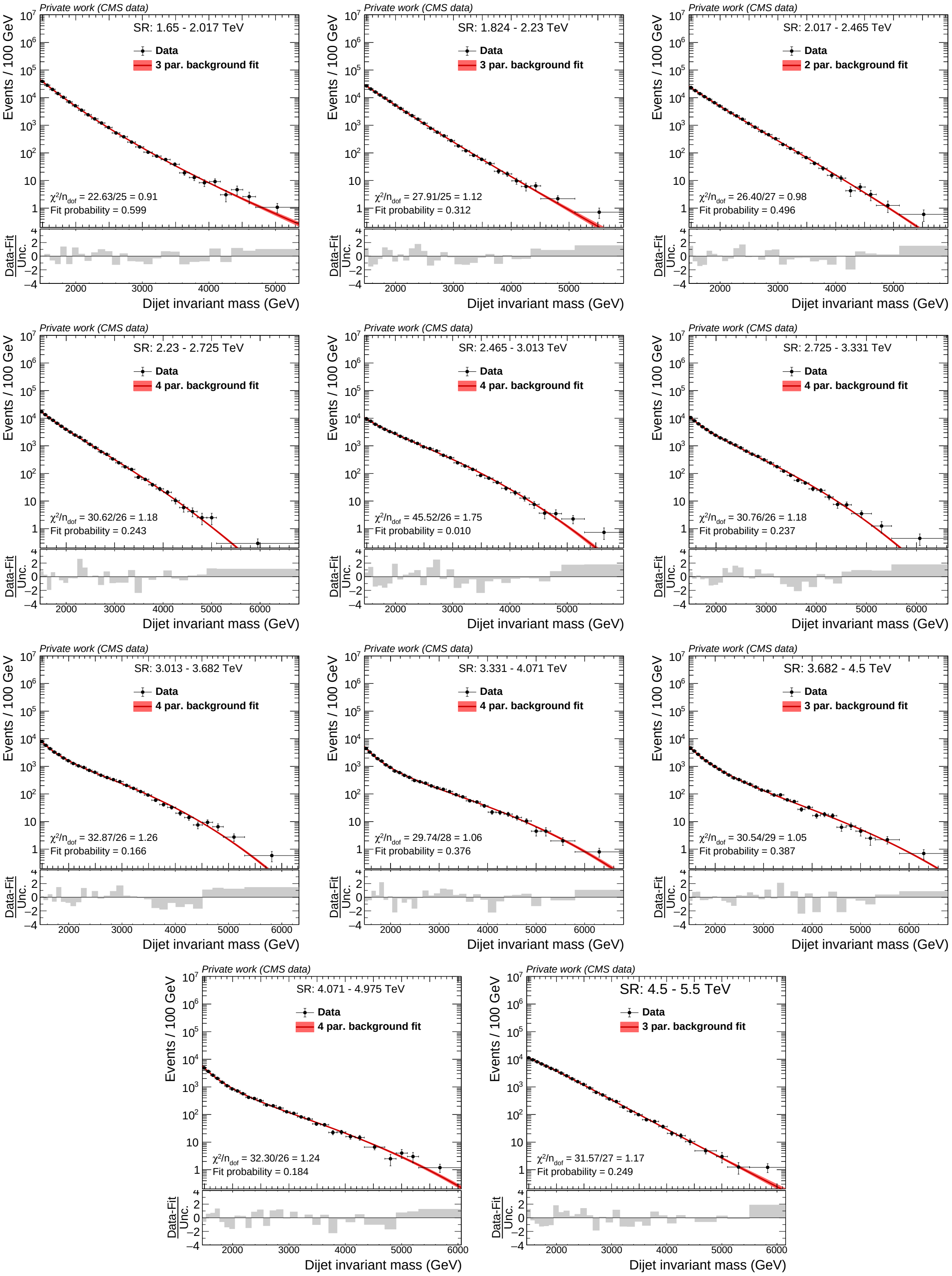

Figure 7.46: Data m_{jj} spectra after selecting the most anomalous 1% of events using CATHODE, for every choice of signal region, and the background fit function chosen by the F-test.

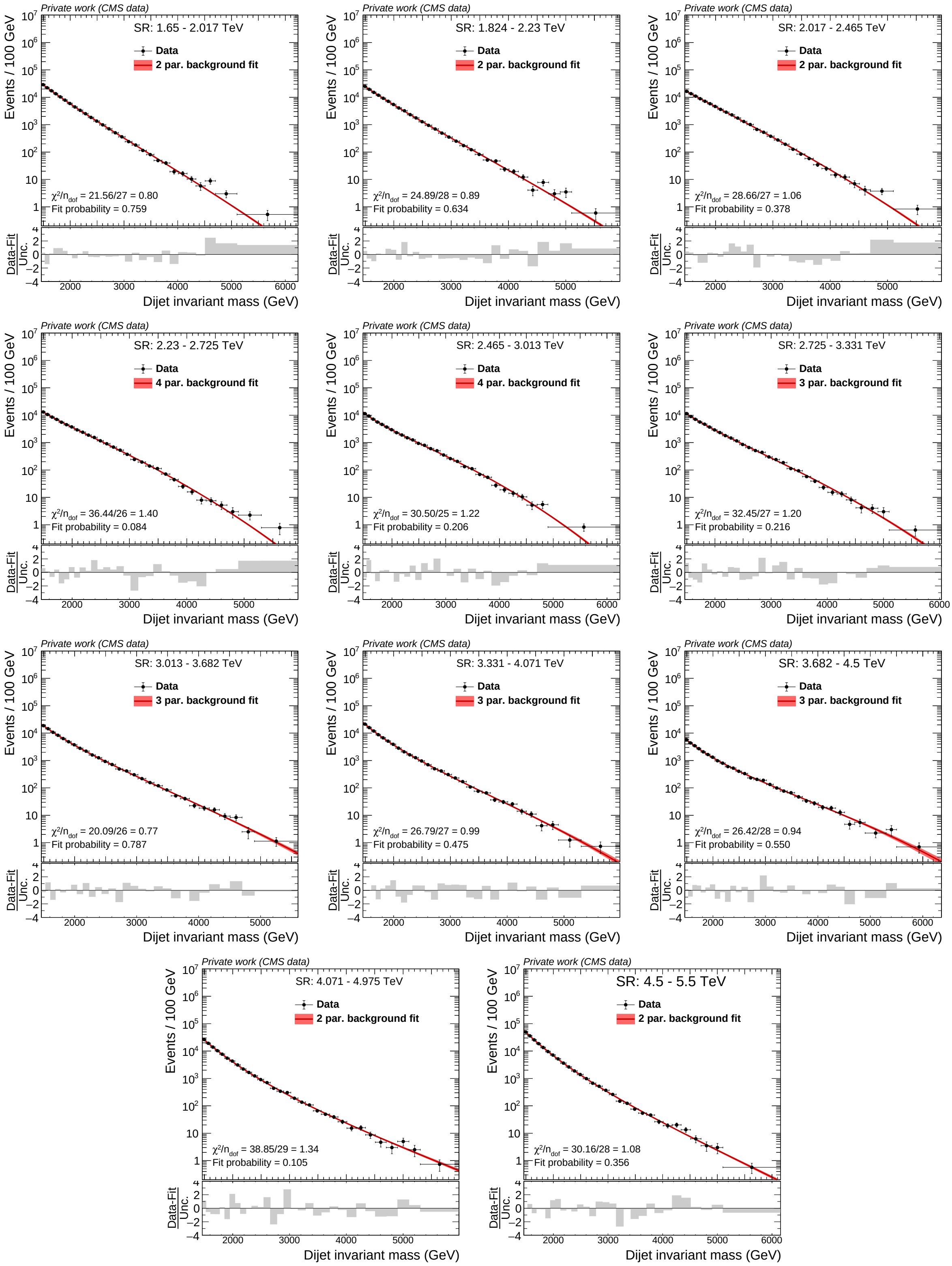

Figure 7.47: Data m_{jj} spectra after selecting the most anomalous 1% of events using CATHODE-b, for every choice of signal region, and the background fit function chosen by the F-test.

The resulting p-values under the background-only hypothesis are shown in Fig. 7.48 for CATHODE (left) and CATHODE-b (right). Neither of the two methods resulted in a local excess with a significance beyond 3σ. Both plots have in common that the region above 3.2 TeV yields relatively large p-values, resulting in a significance well below 1σ. In the lower mass region, the p-values partially become lower, resulting in a significance of up to 2.7σ. In the case of CATHODE, only three points are above 2σ, and the values overall change continuously without sudden jumps, except at 2.5 TeV. CATHODE-b results in eight points above 2σ, with more pronounced short-range fluctuations.

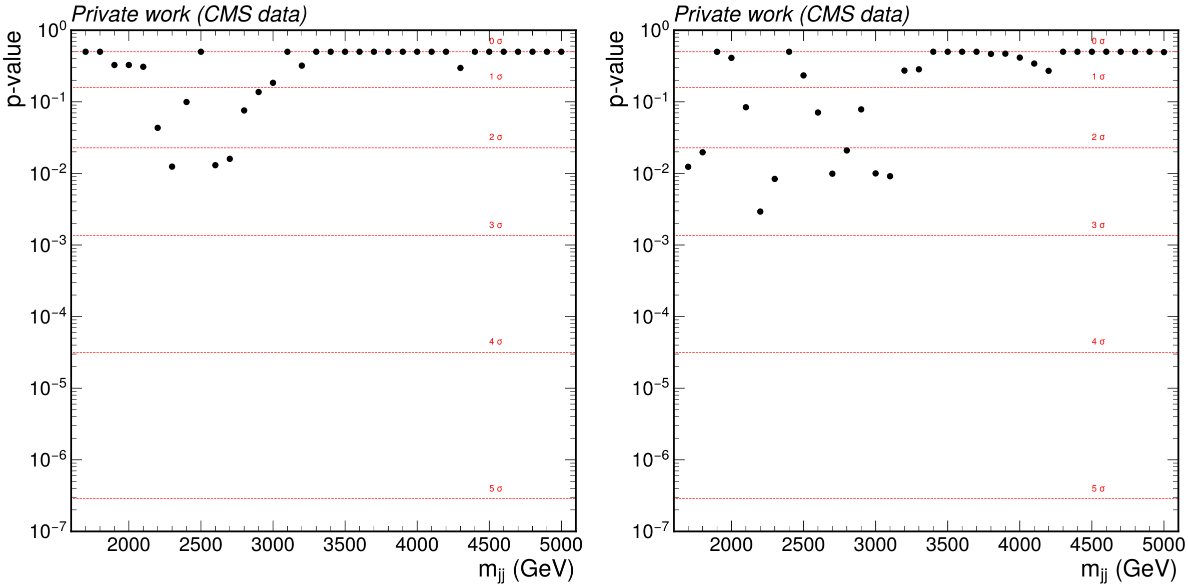

Figure 7.48: Data p-values as a function of m_{jj}, after selecting the most anomalous 1% of events using CATHODE (left) and CATHODE-b (right).

Similarly, none of the other anomaly detection methods in this CMS dijet resonance search reported any local excesses of more than 3σ. Thus, no hints of new physics were found with this analysis.

Treating the full data as background, one can repeat the same study as previously shown in Fig. 7.45, where the expected significance of various signals is tested as a function of an injected cross section. The result is shown in Fig. 7.49 for all anomaly detection methods, as well as the classical benchmarks defined in Sec. 7.9.5. For each of the example signals, the cross section corresponding to a 3σ and 5σ excess is shown. The signals are sorted by the pronginess of the two jets: the $\mathrm{X} \to \mathrm{YY'} \to 4\mathrm{q}$ decay with daughter masses of $m_{\mathrm{Y}} = m_{\mathrm{Y'}} = 170\,\mathrm{GeV}$ leads to two subjets in each large-radius jet, the $\mathrm{W'} \to \mathrm{B't} \to \mathrm{bZt}$ decay with $m_{\mathrm{B'}} = 400\,\mathrm{GeV}$ results in three subjets in each jet, the $\mathrm{W_{KK}} \to \mathrm{RW} \to 3\mathrm{W}$ signal with $m_{\mathrm{R}} = 400\,\mathrm{GeV}$ is associated with jets containing two and four subjets, and the $\mathrm{Y} \to \mathrm{HH} \to 4\mathrm{t}$ decay with $m_{\mathrm{H}} = 400\,\mathrm{GeV}$ leads to six subjets per jet.

On the $\mathrm{X} \to \mathrm{YY'} \to 4\mathrm{q}$ signal, which is the first column in Fig. 7.49, CATHODE is the most sensitive method in terms of achieving a 5σ excess and the second-most sensitive approach for finding a 3σ excess, there only outperformed by QUAK, which is the least model-agnostic approach and uses the signal process in question for training. CATHODE-b has lower sensitivity than CATHODE and performs similar to either QUAK or TNT in terms of 5σ- and 3σ-relating cross sections, respectively. This sensitivity loss is in line with the lack of b-quark decays in the signal. Both CATHODE and CATHODE-b improve upon the simple 2-prong selection, and all

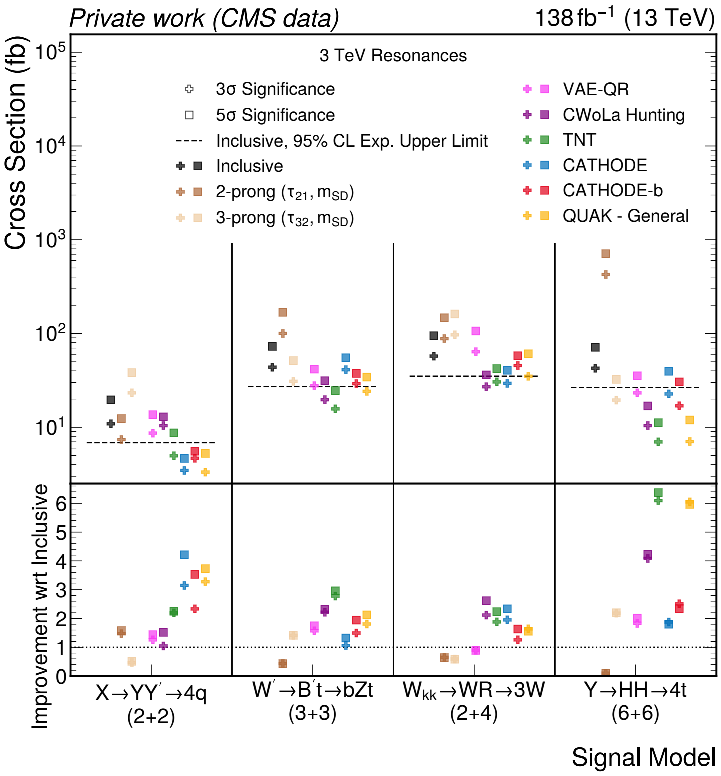

Figure 7.49: Signal cross sections (top panel) at which the expected significance of each method would result in a 3σ and 5σ excess, for four signal models with a resonance mass of 3 TeV: a $\mathrm{X} \to \mathrm{YY}' \to 4\mathrm{q}$ decay with masses of $m_{\mathrm{Y}} = m_{\mathrm{Y}'} = 170\,\mathrm{GeV}$, a $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ decay with $m_{\mathrm{B}'} = 400\,\mathrm{GeV}$, a $\mathrm{W}_{\mathrm{KK}} \to \mathrm{RW} \to 3\mathrm{W}$ decay with $m_{\mathrm{R}} = 400\,\mathrm{GeV}$, and a $\mathrm{Y} \to \mathrm{HH} \to 4\mathrm{t}$ decay with $m_{\mathrm{H}} = 400\,\mathrm{GeV}$. They are sorted by the pronginess of the two jets, denoted in parentheses. The 95% CL upper limit from an inclusive search is displayed for reference as a dashed line. The ratio plot (bottom panel) shows the improvement with respect to the inclusive search strategy.

anomaly detection methods are more sensitive than the inclusive search. The dashed line marks the 95% CL upper limit from the inclusive search as a reference, which would correspond to a 5σ excess in both CATHODE and CATHODE-b.

In the case of the $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ signal, CATHODE-b becomes more sensitive than CATHODE, which is expected from the b decays. The visibly most sensitive method is TNT, followed by CWoLa Hunting, QUAK, and then CATHODE-b and VAE on an equal footing. While CATHODE does not improve over the simple 3-prong selection, all anomaly detectors yield greater sensitivity than an inclusive search. The cross section upper limit from the inclusive search is expected to have appeared as a 5σ excess with TNT and at least a 3σ excess with CWoLa Hunting, QUAK, the VAE and CATHODE-b.

The third signal in Fig. 7.49, a $\mathrm{W}_{\mathrm{KK}} \to \mathrm{RW} \to 3\mathrm{W}$ decay, is most sensitively detected by CWoLa Hunting, followed by CATHODE and TNT, which have similar sensitivity. All anomaly detection methods except the VAE improve notably upon an inclusive search. Due to the lack of b-quark coupling, CATHODE-b is less sensitive than the other weakly supervised methods. The cross section upper limit from the inclusive search would have resulted in a more than a 3σ excess with CATHODE and TNT, and a 5σ excess with CWoLa Hunting.

For the $\mathrm{Y} \to \mathrm{HH} \to 4\mathrm{t}$ signal, TNT is the most sensitive method, followed closely by QUAK and with some distance CWoLa Hunting. CATHODE-b is in the middle ground, benefiting from the b-quark decay sensitivity over CATHODE. All the anomaly detection methods have higher sensitivity than the inclusive search, and the inclusive cross section upper limit is expected to have resulted in an excess of almost 5σ with CATHODE-b, and more than 3σ with CATHODE.

7.10.2 Exclusion Limits

The 95% CL upper limits on the cross section using CATHODE are summarized for all considered signal models with a resonance mass of 3 TeV and varying daughter masses in Fig. 7.50. The results with a resonance mass of 5 TeV are shown in Fig. 7.51. The same is shown with CATHODE-b in Fig. 7.52 and 7.53, respectively. While the MC signal simulation samples necessary to set limits on resonances with a mass of 2 TeV would have been available, this mass was ultimately not included in the analysis due to the extremely high computational cost of training so many CATHODE classifiers at different signal cross sections and varied systematic uncertainties. Moreover, the expected sensitivity in this range of m_{jj} was deemed low because of the high background rate.

A dedicated search has already been performed for a W_{KK} resonance in Ref. [201], and the corresponding observed limits have been added to the figures. The search in Ref. [215] is sensitive to the $X \to YY' \to 4q$ signal, but only with daughter masses of $m_Y = m_{Y'} = 80\,\text{GeV}$. As expected, the dedicated search limits are more stringent than the ones obtained with CATHODE(-b), if they exist. The advantage of the model-agnostic search paradigm is that the same analysis can be used to search for a broad range of signal models, in this case all other shown signal models.

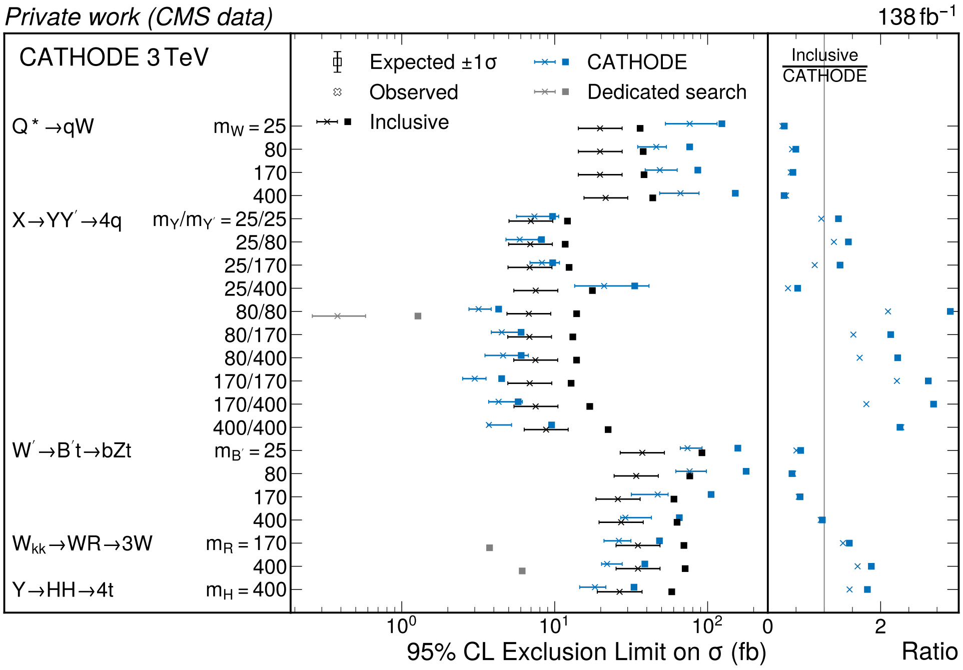

Figure 7.50: Summary of the CATHODE observed and expected upper 95% CL limits on the cross section of signals with a resonance mass of 3 TeV with varying daughter masses, in comparison with the limits that are obtained in the same analysis without any anomaly detection method applied (inclusive). The right panel shows the ratio between the CATHODE and inclusive limit, thus quantifying the improvement in sensitivity. The dedicated search limit reference for the $W_{KK} \to RW \to 3W$ signal is taken from Ref. [201], and for $X \to YY' \to 4q$ from Ref. [215].

The 3 TeV CATHODE limits, shown in Fig. 7.50, improve significantly over a pure inclusive search in most limits, especially in the X → YY′ → 4q decays when both daughter masses are at least 80 GeV. If the daughter masses are lower, the final decay products are more boosted and the discrimination power in the substructure variables is smaller due to the higher similarity to QCD jets. On the Q* → qW′ signal, the CATHODE limits are consistently worse than an inclusive search for similar reasons. Since one of the two jets has only one prong, it cannot be discriminated by the substructure variables, and CATHODE does not sufficiently learn to find it. For the W′ → B′t → bZt signal, CATHODE only finds equivalent sensitivity to the inclusive search at the highest B′ mass, and otherwise worsens the limits.

The improvement of CATHODE with respect to the inclusive search paradigm becomes less pronounced for 5 TeV limits, as seen in Fig. 7.51. Only for the X → YY′ → 4q decays with the largest daughter masses and the Y → HH → 4t signal the application of CATHODE results in lower limits. In this upper end of the m_{jj} spectrum, the number of events available for the CATHODE classifier training is relatively small, and thus a larger signal presence with respect to the background is necessary to achieve a good separation.

The results for CATHODE-b are similar to CATHODE. On the 3 TeV resonance mass signals, Fig. 7.52, there is an increase in sensitivity for the signals where a b quark is present in the decay chain, namely the W′ → B′t → bZt and Y → HH → 4t signals. The resulting limits for the W′ → B′t → bZt signal thus become comparable or partially better than the inclusive search. In the remaining signals, the lack of b activity results in a more noise-dominated training and consequently worse limits than CATHODE with only four features.

The 5 TeV limits for CATHODE-b, shown in Fig. 7.53, are not as sensitive as the inclusive search in almost all cases. Compared to CATHODE, they become either worse (for signals without b decays) or are roughly the same (for signals with b decays). The small number of events available for training the classifier in the high-mass region is likely not enough to

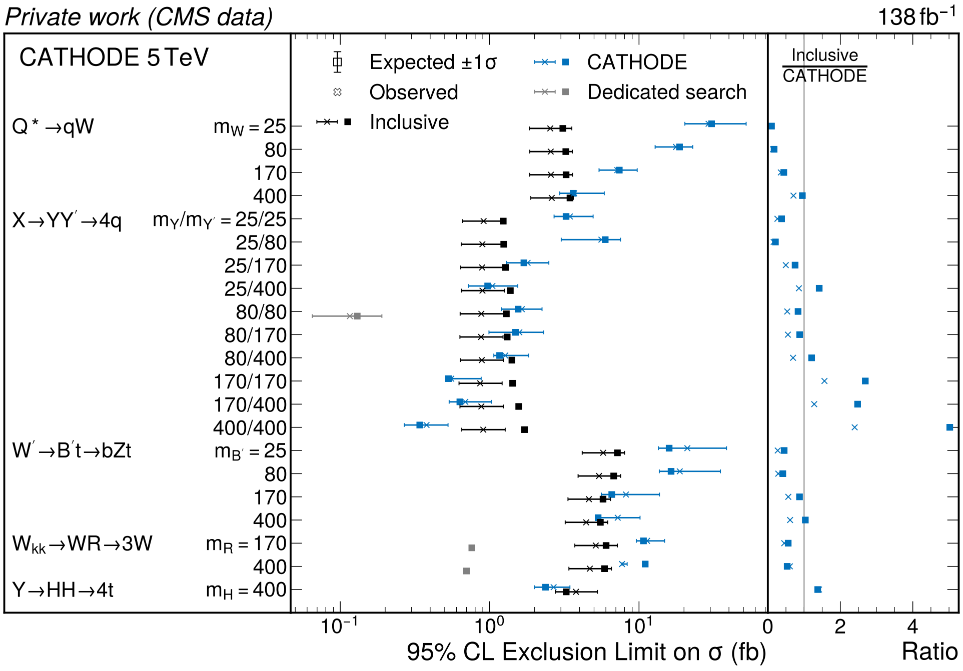

Figure 7.51: Summary of the CATHODE observed and expected upper 95% CL limits on the cross section of signals with a resonance mass of 5 TeV with varying daughter masses, analogous to Fig. 7.50.

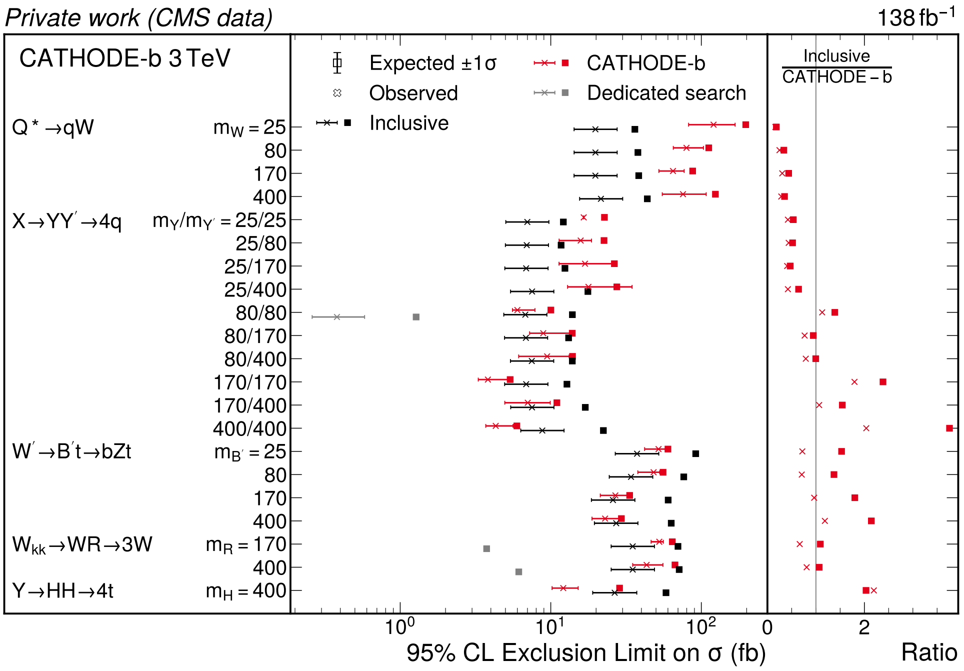

Figure 7.52: Summary of the CATHODE-b observed and expected upper 95% CL limits on the cross section of signals with a resonance mass of 3 TeV with varying daughter masses, analogous to Fig. 7.50.

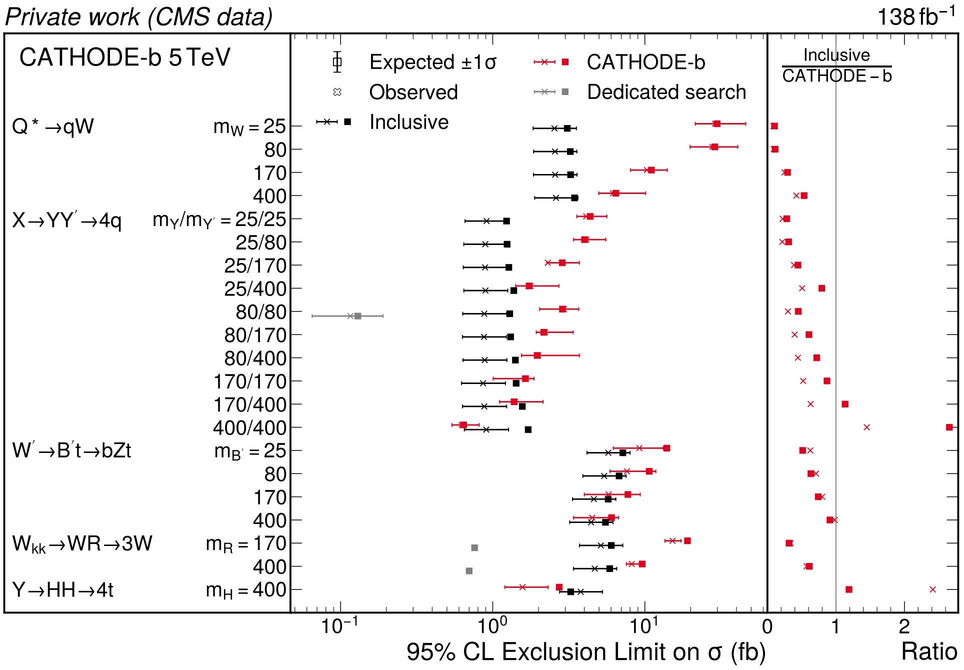

Figure 7.53: Summary of the CATHODE-b observed and expected upper 95% CL limits on the cross section of signals with a resonance mass of 5 TeV with varying daughter masses, analogous to Fig. 7.50.

counteract the larger number of features in CATHODE-b.

A direct comparison between the exclusion limits achieved with different anomaly detection methods and additional benchmarks is shown in Fig. 7.54 for the 3 TeV (left) and 5 TeV (right) resonance masses. Four representative signal examples are selected, corresponding to the same choice as in Fig. 7.49. As expected, for the 3 TeV resonance mass, the limits found by the different methods follow a similar ordering as the expected significances, which can be seen by comparing Fig. 7.49 and 7.54 (left). The 5 TeV comparison reveals that the lack of a substantial sensitivity improvement in most signal processes with respect to the inclusive search, which was observed in CATHODE(-b), is a common feature among all methods.

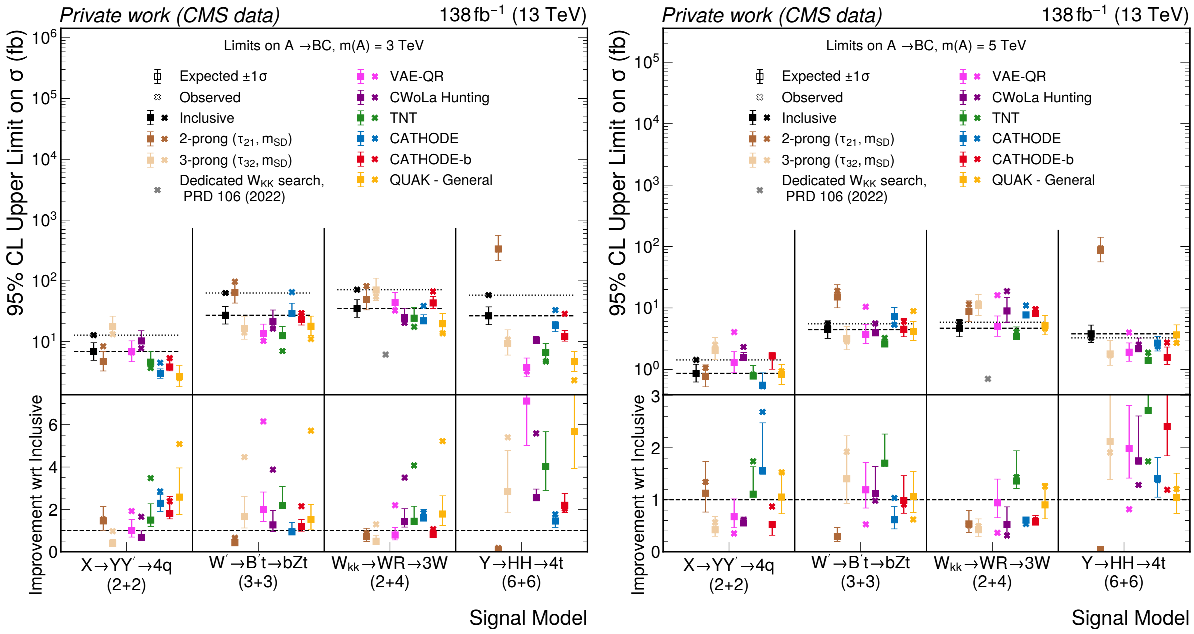

Figure 7.54: Upper 95% CL limits of all anomaly detection and benchmark methods on the cross section of four signal models (top panels) with a resonance mass of 3 TeV (left) and 5 TeV (right): a $\mathrm{X} \to \mathrm{YY'} \to 4\mathrm{q}$ decay with masses of $m_{\mathrm{Y}} = m_{\mathrm{Y'}} = 170\,\mathrm{GeV}$, a $\mathrm{W'} \to \mathrm{B't} \to \mathrm{bZt}$ decay with $m_{\mathrm{B'}} = 400\,\mathrm{GeV}$, a $\mathrm{W_{KK}} \to \mathrm{RW} \to 3\mathrm{W}$ decay with $m_{\mathrm{R}} = 400\,\mathrm{GeV}$, and a $\mathrm{Y} \to \mathrm{HH} \to 4\mathrm{t}$ decay with $m_{\mathrm{H}} = 400\,\mathrm{GeV}$. They are sorted by the pronginess of the two jets, denoted in parentheses. The ratio plots (bottom panels) show the improvement with respect to the inclusive search strategy.

A more quantitative overview of the upper 95% CL cross section limits is provided in Tab. 7.8 for 3 TeV resonances and in Tab. 7.9 for 5 TeV resonances. They show the numeric values obtained from the most sensitive anomaly detection method per signal. The last column computes the improvement with respect to the inclusive search strategy.

In line with the findings in Sec. 7.9.6, the tables show that the different anomaly detection methods are sensitive to different anomalous patterns. No one method performs best on all signals, but instead each method excels in different signal models. CATHODE works particularly well on the double two-prong $\mathrm{X} \to \mathrm{YY'} \to 4\mathrm{q}$ signal and TNT on the $\mathrm{W'} \to \mathrm{B't} \to \mathrm{bZt}$ signal. CWoLa Hunting, QUAK and the VAE improve on individual mass combinations. CATHODE-b is not the most sensitive method for any of the signals, because in the cases where a b quark is present, TNT is more sensitive. Moreover, the results show that the sensitivity improvement from employing these data-driven anomaly detection methods is present in the cases where anomalous substructure patterns could be exploited. For the $\mathrm{Q}^* \to \mathrm{qW'}$ signal, the lack of higher pronginess in both jets results in a worse limit than the inclusive search. The improvement depends to a large degree on a sufficient number of events, which is not given anymore at 5 TeV.

The inclusion of systematic uncertainties is one of the most substantial differences between the results of this analysis and previous proof-of-principle studies. Appendix B.4 shows a comparison of the upper 95% CL cross section limits before and after the inclusion of systematic uncertainties in the signal strength. The resulting 20–50% increase in the limits need to be taken into consideration when making estimations based on proof-of-concept studies without systematic uncertainties.

Table 7.8: Observed and expected upper 95% CL limits on the cross section of signals with a resonance mass of 3 TeV with varying daughter particle masses. Only the values of the most sensitive anomaly detection method (denoted in the third column) per signal are shown.

Model A → BC	**Daughter**	**Method**	**Exp. (Obs.)**	**Improvement**
(m_A = 3 TeV)	**Masses**		**Limit**	**w.r.t.**
	(GeV)		**(fb)**	**Inclusive**
$Q^* \rightarrow qW'$	25	CWoLa Hunting	61.1 (30.1)	0.3
	80	CATHODE	46.2 (76.3)	0.4
	170	CATHODE	48.7 (86.3)	0.4
	400	CWoLa Hunting	45.8 (24.3)	0.5
$X \rightarrow YY' \rightarrow 4q$	25/25	CATHODE	7.4 (9.7)	0.9
	25/80	CATHODE	5.9 (8.2)	1.2
	25/170	CATHODE	8.3 (9.7)	0.8
	25/400	VAE-QR	13.6 (12.5)	0.6
	80/80	CATHODE	3.2 (4.3)	2.1
	80/170	CATHODE	4.5 (6.0)	1.5
	80/400	CATHODE	4.6 (6.0)	1.6
	170/170	QUAK	2.7 (2.5)	2.6
	170/400	CATHODE	4.3 (5.8)	1.7
	400/400	VAE-QR	2.1 (1.9)	4.2
$W' \rightarrow B't \rightarrow bZt$	25	TNT	22.6 (13.9)	1.7
	80	TNT	18.2 (11.3)	1.9
	170	TNT	12.2 (7.3)	2.1
	400	TNT	12.5 (7.0)	2.2
$W_{KK} \rightarrow RW \rightarrow 3W$	170	TNT	22.1 (15.2)	1.6
	400	QUAK	19.7 (13.7)	1.8
$Y \rightarrow HH \rightarrow 4t$	400	VAE-QR	3.7 (3.2)	7.1

Table 7.9: Observed and expected upper 95% CL limits on the cross section of signals with a resonance mass of 5 TeV with varying daughter particle masses. Only the values of the most sensitive anomaly detection method (denoted in the third column) per signal are shown.

Model A → BC	**Daughter**	**Method**	**Exp. (Obs.)**	**Improvement**
(m_A = 5 TeV)	**Masses**		**Limit**	**w.r.t.**
	(GeV)		**(fb)**	**Inclusive**
$Q^* \to qW'$	25	QUAK	3.5 (3.1)	0.7
	80	QUAK	3.2 (2.8)	0.8
	170	QUAK	3.3 (3.6)	0.8
	400	CATHODE	3.7 (3.6)	0.7
$X \to YY' \to 4q$	25/25	QUAK	1.7 (1.6)	0.5
	25/80	QUAK	1.3 (1.3)	0.7
	25/170	QUAK	1.1 (1.1)	0.8
	25/400	VAE-QR	1.0 (3.4)	0.9
	80/80	TNT	1.1 (1.2)	0.8
	80/170	QUAK	0.9 (1.0)	0.9
	80/400	VAE-QR	0.9 (3.0)	0.9
	170/170	CATHODE	0.6 (0.5)	1.6
	170/400	CATHODE	0.7 (0.6)	1.3
	400/400	CATHODE	0.4 (0.3)	2.4
$W' \to B't \to bZt$	25	TNT	3.6 (5.2)	1.6
	80	TNT	3.5 (5.0)	1.6
	170	TNT	2.5 (3.4)	1.9
	400	TNT	2.6 (3.2)	1.7
$W_{KK} \to RW \to 3W$	170	TNT	4.4 (5.9)	1.2
	400	TNT	3.4 (4.1)	1.4
$Y \to HH \to 4t$	400	TNT	1.4 (1.9)	2.7

7.11 Conclusion

This analysis presents a novel application of CATHODE, in parallel to four other state-of-the-art machine learning anomaly detection methods, to search for physics beyond the Standard Model in the CMS Experiment with only minimal model assumptions. It is performed in the context of a search for heavy narrow resonances produced in proton-proton collisions and decaying to two large-radius jets with anomalous substructure, using data collected by the CMS detector from 2016 to 2018, with a center-of-mass energy of 13 TeV

CATHODE is a weakly supervised anomaly detection technique, developed in the context of this thesis, where a background template is sampled from a generative machine learning model that has been trained on a sideband region, and then a classifier model is trained to find difference between the actual data and the interpolated background template, thus detecting overdensities with respect to the expected background distribution. It is applied to this analysis in two configurations: one with four input features that generally capture the substructure of the two highest-p_T jets, and another with two additional features that encode the presence of b-tagged jets in the event.

In order to ensure that the method is well calibrated, i.e., it only results in significant excesses when a new physics signal is present, the method has been validated in both simulation and a data control regions. Ultimately, no significant excess was observed in the collision data of interest, neither by CATHODE nor by any of the other methods.

The sensitivity of the various anomaly detection approaches is quantified and compared via the expected significance and 95% CL upper limits on the cross section. The limit setting procedure for CATHODE and other weakly supervised methods is more involved than in most searches, as it requires accounting for the signal strength dependence of the classifier efficiency. For many types of signal models, CATHODE is found to be substantially more sensitive than traditional model-agnostic searches, which rely either on simple generic selections or none at all. The exclusion limits presented in this analysis are in many cases the most stringent of their kind, in particular because many of these signals have not been targeted before. This simultaneous coverage of a broad range of signal models is a unique strength of the model-agnostic search paradigm. Existing dedicated searches are seen to be more sensitive than the model-agnostic approach, which demonstrates the complementarity of the two types of searches.

Not only is this the first application of CATHODE within a high-energy physics experiment, it is also the first time that so many state-of-the-art anomaly detection methods could be compared in a real analysis setting. This provides a more comprehensive view of the strengths and weaknesses of each method than the usual proof-of-concept studies, as it involved a fully realistic event content, background estimation, and statistical analysis taking into account systematic uncertainties. Nevertheless, a direct comparison between the methods is complicated by the plethora of design choices that need to be made, such as the input features, model hyperparameters, and selection efficiency. More studies are being performed to further understand the specifics of each method and to compare them on a more equal footing, e.g., by using the same set of input features across methods.

As this is the first application of anomaly detection for model-agnostic searches for new physics within the CMS experiment, it opens the door to many more analyses of this type in the future. On the one hand, these methods can be applied to other types of final states. On the other hand, new data are being collected in the current Run 3 of the LHC. The larger dataset with higher energy and updated trigger and reconstruction algorithms will provide a promising discovery potential for data-driven methods, such as CATHODE.

8 Summary & Conclusions

There are many open questions in high-energy physics (HEP) that cannot be answered with the Standard Model (SM) alone. Many searches have been performed at the Large Hadron Collider (LHC) that target specific signals of physics beyond the SM (BSM), however no such processes have been observed so far. This motivates the development of new search strategies that are more model agnostic and can target a wider range of signals. A promising approach is the use of anomaly detection—a subfield of machine learning (ML) that aims to identify unusual patterns in data.

This thesis has presented the full development of a new anomaly detection method, CATHODE, for BSM searches. CATHODE is a novel approach for weakly supervised anomaly detection. The core of the method is a conditional generative model that is trained on data from a sideband region. The conditioning facilitates the interpolation into the signal region where a background template is sampled from the model. A classifier is then trained to distinguish between the real data and the sampled background, thus gaining sensitivity to signal-induced data overdensities. I demonstrated that the method has state-of-the-art performance on the LHCO R&D dataset, also in the more challenging case of correlated input features. The original study only marginally investigated the compatibility with a bump hunt–based background estimation. It also provided hints that the sensitivity for a specific signal deteriorates due to the inclusion of uninformative input features.

A later and more thorough investigation of the background shape after applying CATHODE-based selections revealed that the presence of significant correlations between the classifier input features and the resonant observable can lead to a highly obscured background shape in the latter, which might severely impair an estimation of the background. In fact, this issue is intrinsic to weakly supervised anomaly detection methods. To mitigate it, a modification to CATHODE was proposed, called LaCATHODE: the conditional generative model is chosen to be a normalizing flow with an unconditional prior latent space. Then, the invertible transformation of the flow is used to map input features to the decorrelated latent space before a classifier is trained on these transformed features. While much of the signal sensitivity of CATHODE is retained, the background shape after selecting on the LaCATHODE classifier output remains unchanged, thus providing a more reliable basis for background estimation based on a bump hunt.

The impact of uninformative input features on the performance of weakly supervised learning tasks was investigated more systematically, both in terms of uncorrelated artificial noise and the inclusion of realistic physical features where information is non-trivially spread out. It was found that the resilience with respect to many features with no or limited information content can be substantially improved by using a classifier based on an ensemble of boosted decision trees rather than the neural network used in the CATHODE prototype. This improvement can be understood as a consequence of the generally more noise-resistant nature of tree-based methods, in particular for tabular data structures. The inclusion of more input features means that a broader range of signal-like signatures can be captured. Thus, this improvement is expected to lead to more model-agnostic results as they are less dependent on the specific choice of input features.

The ultimate test for CATHODE was to perform an actual search for BSM physics at an LHC experiment, including all the necessary steps that come with it, e.g., a well-calibrated background estimation, the statistical interpretation of the results, a strategy of interpreting a potential excess, and setting upper limits on cross sections that take into account systematic uncertainties. The search for heavy resonances decaying into dijet final states arising from proton-proton collisions at a center-of-mass energy of 13 TeV in the CMS Experiment was an ideal showcase for this purpose. On the one hand, its final state topology coincides with the LHCO R&D dataset used for prototyping CATHODE, so it can be seen as a direct reality check. On the other hand, the simultaneous application of multiple other state-of-the-art anomaly detection methods in the same analysis enabled a cross-comparison of methods that has never before been done within a single anomaly detection search in HEP. While none of the methods found a significant excess over the SM background expectation, they could be used to set upper limits on the cross sections of multiple explicit signals covering a wide range of different jet substructure signatures. It was observed that the methods varied in sensitivity depending on the signal process, with some methods performing better on certain signals and worse on others. In many cases, the use of data-driven ML techniques visibly improved the sensitivity compared to a more traditional generic search strategy, which was performed in addition to serve as a benchmark. The resulting exclusion limits are the most stringent to date for almost all considered signal models, mainly because hardly any of them had been targeted by a dedicated search before. This underlines the main advantage of the anomaly detection approach: it is capable of targeting a wide range of signal processes in a single analysis.

Because of time constraints, several of the methodological improvements developed in the course of this thesis could not be applied to the experimental analysis, such as LaCATHODE or the use of tree-based classifiers. Moreover, multiple further studies on CATHODE and other anomaly detection methods have been sparked by the ones shown in this thesis. For example, Ref. [7] compared the most state-of-the-art weakly supervised anomaly detection methods on an equal footing and discussed ways of combining them. Reference [235] confirmed the promising discovery potential of CATHODE on a different simulation dataset and discussed its feasibility with input features where anomalies are located in the tails of the distribution. A different route to incorporating more input features than tree-based classifiers is presented in Ref. [236], where the authors propose to perform weakly supervised anomaly detection directly on lower-level detector information. They achieved this by using more sophisticated architectures for both generation and classification, which showed non-trivial performance in a more abundant signal presence. An evolution of this low-level approach, which yielded promising sensitivity at a realistic order of signal cross sections, was presented in Ref. [237] using a foundation model that had been pre-trained on related tasks.

The anomaly detection–based search for new physics presented in this thesis is the first of its kind in the CMS Experiment and the most complex in terms of the number of methods used at the LHC. The lessons learned from this analysis and the steady progress on the methodological side will open the door to many more anomaly detection searches in the future. To date, a plethora of final state topologies and datasets are yet to be explored by this model-agnostic search paradigm. From this point of view, it seems that we are gradually entering a new era of more data-driven (and simultaneously heavily innovation-driven) searches for new physics. New particles might be hiding in already-recorded data, possibly in unexpected corners of phase space.

Acknowledgements

I would like to express my deepest gratitude to everyone who has supported me throughout my PhD journey. Listing all names would unfortunately be impossible as it would make this section longer than the thesis itself, considering that my research was funded by the German Federal Ministry of Education and Research, and thus by all the millions of inhabitants of Germany. Even more so, my research would not have been possible without the facilities, tools and data provided by the CMS Collaboration at CERN, publically funded by many countries. Thus, while I am highly grateful for being part of a society that values scientific progress and education, I will focus the remainder of this section on highlighting a few individuals that have been particularly important for reaching the point of submitting this thesis.

First and foremost, I would like to thank my supervisor Gregor Kasieczka for his unwavering support and guidance throughout my PhD. I always valued his honest and respectful approach, which fostered an open and constructive working environment. I would also like to thank my second supervisor Peter Schleper for his valuable inputs at our annual meetings.

A very special thanks goes to my colleagues Louis Moureaux and Karim El Morabit for their thorough feedback on the thesis draft.

All the projects in this thesis were a team effort, and I appreciated the close collaboration within our local anomaly detection subgroup. It began as a duo with Tobias Quadfasel, whose daily cheerful “good morning” messages and virtual coffee chats I truly valued. Later, we were joined by Louis Moureaux, whose expertise and enthusiasm were invaluable to the project. The recent additions of Tore von Schwartz and Chitrakshee Yede brought a great boost of energy and contributed to a lively and fun office atmosphere. I also drew great inspiration from our external collaborators, such as David Shih and Oz Amram.

I am grateful to have been part of the larger Kasieczka group, where I always felt welcomed, supported, and entertained by my colleagues, many of whom became good friends. My time in Hamburg would not have been the same without them.

Another very special thanks goes to my wonderful wife, Rosa, whose care, support and patience made this journey possible. I would also like to thank my lovely daughter, Flora, whose birth last year brought immense joy into my life. Another invaluable source of happiness and support has been our dog and best boy, Elvis.

Finally, I dedicate this work to my mother, Ursula Sommerhalder Riner, who unfortunately passed away during my PhD. Her strength and determination, even in the face of the most challenging times, have been a profound source of inspiration to me. I would not have reached this point without her unconditional support on every path I chose to take—even when it meant doing something as silly as becoming a particle physicist.

References

[1] Anna Hallin, Joshua Isaacson, Gregor Kasieczka, Claudius Krause, Benjamin Nachman, Tobias Quadfasel, Matthias Schlaffer, David Shih, and Manuel Sommerhalder. "Classifying anomalies through outer density estimation". In: *Phys. Rev. D* 106 (Sept. 2022), p. 055006. DOI: 10.1103/physrevd.106.055006.

[2] Anna Hallin, Gregor Kasieczka, Tobias Quadfasel, David Shih, and Manuel Sommerhalder. "Resonant anomaly detection without background sculpting". In: *Phys. Rev. D* 107 (June 2023), p. 114012. DOI: 10.1103/physrevd.107.114012.

[3] Thorben Finke, Marie Hein, Gregor Kasieczka, Michael Krämer, Alexander Mück, Parada Prangchaikul, Tobias Quadfasel, David Shih, and Manuel Sommerhalder. "Tree-based algorithms for weakly supervised anomaly detection". In: *Phys. Rev. D* 109 (Feb. 2024), p. 034033. DOI: 10.1103/PhysRevD.109.034033.

[4] Philip Harris, William Patrick McCormack, Sang Eon Park, Tobias Quadfasel, Manuel Sommerhalder, Louis Jean Moureaux, Gregor Kasieczka, Oz Amram, Petar Maksimovic, Benedikt Maier, Maurizio Pierini, Kinga Anna Wozniak, Thea Klaeboe Åarrestad, Jennifer Ngadiuba, Irene Zoi, Samuel Kai Bright-Thonney, David Shih, and Aritra Bal. "Machine learning techniques for model-independent searches in dijet final states". CMS Note CMS-NOTE-2023-013. 2023.

[5] CMS Collaboration. "Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV". CMS Physics Analysis Summary CMS-PAS-EXO-22-026. 2024.

[6] CMS Collaboration. "Search for narrow resonances in the b-tagged dijet mass spectrum in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. D* 108 (2023), p. 012009. DOI: 10.1103/PhysRevD.108.012009.

[7] Tobias Golling, Gregor Kasieczka, Claudius Krause, Radha Mastandrea, Benjamin Nachman, John Andrew Raine, Debajyoti Sengupta, David Shih, and Manuel Sommerhalder. "The interplay of machine learning-based resonant anomaly detection methods". In: *Eur. Phys. J. C* 84 (Mar. 2024), p. 241. DOI: 10.1140/epjc/s10052-024-12607-x.

[8] Tore von Schwartz. "Semi-supervised learning as a new tool for particle physics analysis". Bachelor Thesis. University of Hamburg, 2021.

[9] Parada Prangchaikul. "Mitigating uninformative features in weak supervision". Master Thesis. University of Hamburg, 2023.

[10] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett. B* 716 (Sept. 2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020.

[11] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Phys. Lett. B* 716 (Sept. 2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021.

[12] ATLAS Collaboration. "A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment". In: *Eur. Phys. J. C* 79 (Feb. 2019), p. 120. DOI: 10.1140/epjc/s10052-019-6540-y.

[13] CMS Collaboration. "MUSiC: a model-unspecific search for new physics in proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$". In: *Eur. Phys. J. C* 81 (July 2021), p. 629. DOI: 10.1140/epjc/s10052-021-09236-z.

[14] ATLAS Collaboration. “Search for New Particles in Two-Jet Final States in 7 TeV Proton-Proton Collisions with the ATLAS Detector at the LHC”. In: *Phys. Rev. Lett.* 105 (2010), p. 161801. DOI: 10.1103/PhysRevLett.105.161801.

[15] ATLAS Collaboration. “Search for new phenomena in dijet mass and angular distributions from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”. In: *Phys. Lett. B* 754 (2016), pp. 302–322. DOI: 10.1016/j.physletb.2016.01.032.

[16] CMS Collaboration. “Search for Dijet Resonances in 7 TeV *pp* Collisions at CMS”. In: *Phys. Rev. Lett.* 105 (2010), p. 211801. DOI: 10.1103/PhysRevLett.105.211801.

[17] CMS Collaboration. “Search for high mass dijet resonances with a new background prediction method in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$”. In: *J. High Energ. Phys.* 2020 (May 2020), p. 033. DOI: 10.1007/jhep05(2020)033.

[18] ATLAS Collaboration. “Dijet Resonance Search with Weak Supervision Using $\sqrt{s} =$ 13 TeV *pp* Collisions in the ATLAS Detector”. In: *Phys. Rev. Lett.* 125 (Sept. 2020), p. 131801. DOI: 10.1103/physrevlett.125.131801.

[19] ATLAS Collaboration. “Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV *pp* collisions with the ATLAS detector”. In: *Phys. Rev. D* 108 (Sept. 2023), p. 052009. DOI: 10.1103/physrevd.108.052009.

[20] ATLAS Collaboration. “Search for New Phenomena in Two-Body Invariant Mass Distributions Using Unsupervised Machine Learning for Anomaly Detection at $\sqrt{s} = 13$ with the ATLAS Detector”. In: *Phys. Rev. Lett.* 132 (Feb. 2024), p. 081801. DOI: 10.1103/physrevlett.132.081801.

[21] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to quantum field theory.* Reading, USA: Addison-Wesley, 1995. DOI: 10.1201/9780429503559.

[22] Chris Ferrie. *Quantum Physics for Babies.* Sourcebooks Explore, 2017.

[23] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model.* Cambridge University Press, 2013.

[24] Gordon Kane. *Modern Elementary Particle Physics: Explaining and Extending the Standard Model.* 2nd ed. Cambridge University Press, 2017.

[25] Mark Thomson. *Modern Particle Physics.* Cambridge University Press, 2013.

[26] Particle Data Group. “Review of Particle Physics”. In: *Phys. Rev. D* 110 (Aug. 2024), p. 030001. DOI: 10.1103/PhysRevD.110.030001.

[27] François Englert and Robert Brout. “Broken Symmetry and the Mass of Gauge Vector Mesons”. In: *Phys. Rev. Lett.* 13 (Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[28] Peter W. Higgs. “Broken Symmetries and the Masses of Gauge Bosons”. In: *Phys. Rev. Lett.* 13 (Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[29] Gerald S. Guralnik, Carl R. Hagen, and Tom W. B. Kibble. “Global Conservation Laws and Massless Particles”. In: *Phys. Rev. Lett.* 13 (Nov. 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585.

[30] Pierre Fayet and Sergio Ferrara. “Supersymmetry”. In: *Phys. Rep.* 32 (1977), pp. 249–334. DOI: 10.1016/0370-1573(77)90066-7.

[31] Wilfried Buchmüller. "Composite Quarks and Leptons". In: *Nucleon-Nucleon and Nucleon-Antinucleon Interactions*. Ed. by H. Mitter and W. Plessas. Vienna: Springer Vienna, 1985, pp. 517–595. DOI: 10.1007/978-3-7091-8830-9_8.

[32] Kaustubh Agashe, Roberto Contino, and Alex Pomarol. "The minimal composite Higgs model". In: *Nucl. Phys. B* 719 (July 2005), pp. 165–187. DOI: 10.1016/j.nuclphysb.2005.04.035.

[33] Stefania De Curtis, Michele Redi, and Andrea Tesi. "The 4D composite Higgs". In: *J. High Energ. Phys.* 2012 (Apr. 2012), p. 042. DOI: 10.1007/jhep04(2012)042.

[34] Lisa Randall and Raman Sundrum. "Large Mass Hierarchy from a Small Extra Dimension". In: *Phys. Rev. Lett.* 83 (Oct. 1999), pp. 3370–3373. DOI: 10.1103/physrevlett.83.3370.

[35] Kaustubh Agashe, Peizhi Du, Sungwoo Hong, and Raman Sundrum. "Flavor universal resonances and warped gravity". In: *J. High Energ. Phys.* 2017 (Jan. 2017), p. 016. DOI: 10.1007/jhep01(2017)016.

[36] Vladimir N. Gribov and Lev N. Lipatov. "Deep inelastic e p scattering in perturbation theory". In: *Sov. J. Nucl. Phys.* 15 (1972), pp. 438–450.

[37] Yuri L. Dokshitzer. "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics." In: *Sov. Phys. JETP* 46 (1977), pp. 641–653.

[38] Guido Altarelli and Giorgio Parisi. "Asymptotic Freedom in Parton Language". In: *Nucl. Phys. B* 126 (1977), pp. 298–318. DOI: 10.1016/0550-3213(77)90384-4.

[39] Oliver Brüning, Helmut Burkhardt, and Stephen Myers. "The large hadron collider". In: *Prog. Part. Nucl. Phys.* 67 (2012), pp. 705–734. DOI: 10.1016/j.ppnp.2012.03.001.

[40] High Luminosity LHC Project. *LS3 schedule change.* URL: https://hilumilhc.web.cern.ch/article/ls3-schedule-change (visited on 09/13/2024).

[41] CMS Collaboration. "The CMS experiment at the CERN LHC". In: *J. Instrum.* 3 (Aug. 2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[42] Tai Sakuma and Thomas McCauley. "Detector and Event Visualization with SketchUp at the CMS Experiment". In: *J. Phys. Conf. Ser.* 513 (June 2014), p. 022032. DOI: 10.1088/1742-6596/513/2/022032.

[43] V. Veszpremi. "Performance verification of the CMS Phase-1 Upgrade Pixel detector". In: *J. Instrum.* 12 (Dec. 2017), p. C12010. DOI: 10.1088/1748-0221/12/12/c12010.

[44] Christian W. Fabjan and Fabiola Gianotti. "Calorimetry for particle physics". In: *Rev. Mod. Phys.* 75 (Oct. 2003), pp. 1243–1286. DOI: 10.1103/RevModPhys.75.1243.

[45] CMS Collaboration. "The CMS trigger system". In: *J. Instrum.* 12 (Jan. 2017), P01020. DOI: 10.1088/1748-0221/12/01/p01020.

[46] Mia Tosi. "The CMS trigger in Run 2". In: *PoS* EPS-HEP2017 (2017), p. 523. DOI: 10.22323/1.314.0523.

[47] Hale Sert. "CMS Run 2 High Level Trigger Performance". In: *PoS* EPS-HEP2019 (2020), p. 165. DOI: 10.22323/1.364.0165.

[48] CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *J. Instrum.* 12 (Oct. 2017), P10003. DOI: 10.1088/1748-0221/12/10/p10003.

[49] Wolfgang. Adam, Boris Mangano, Thomas Speer, and Teddy Todorov. “Track reconstruction in the CMS tracker”. CMS Note CMS-NOTE-2006-041. Dec. 2005.

[50] R. Keith Ellis, W. James Stirling, and Bryan R. Webber. *QCD and collider physics.* Vol. 8. Cambridge University Press, Feb. 2011. DOI: 10.1017/CBO9780511628788.

[51] Bo Andersson, Gosta Gustafson, Gunnar Ingelman, and Torbjörn Sjöstrand. “Parton fragmentation and string dynamics”. In: *Phys. Rep.* 97 (1983), pp. 31–145. DOI: 10.1016/0370-1573(83)90080-7.

[52] 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”. In: *Comput. Phys. Commun.* 191 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024.

[53] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. “The anti-kt jet clustering algorithm”. In: *J. High Energ. Phys.* 2008 (Apr. 2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063.

[54] Stefano Catani, Yuri L. Dokshitzer, Michael H. Seymour, and Bryan R. Webber. “Longitudinally invariant K_t clustering algorithms for hadron hadron collisions”. In: *Nucl. Phys. B* 406 (1993), pp. 187–224. DOI: 10.1016/0550-3213(93)90166-M.

[55] Yuri L. Dokshitzer, Garth D. Leder, Stefano Moretti, and Bryan R. Webber. “Better jet clustering algorithms”. In: *J. High Energ. Phys.* 1997 (Aug. 1997), p. 001. DOI: 10.1088/1126-6708/1997/08/001.

[56] CMS Collaboration. *Public CMS Luminosity Information.* URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults (visited on 09/08/2024).

[57] CMS Collaboration. “Pileup mitigation at CMS in 13 TeV data”. In: *J. Instrum.* 15 (Sept. 2020), P090180. DOI: 10.1088/1748-0221/15/09/p09018.

[58] Daniele Bertolini, Philip Harris, Matthew Low, and Nhan Tran. “Pileup per particle identification”. In: *J. High Energ. Phys.* 2014 (Oct. 2014), p. 059. DOI: 10.1007/jhep10(2014)059.

[59] David Krohn, Jesse Thaler, and Lian-Tao Wang. “Jet trimming”. In: *J. High Energ. Phys.* 2010 (Feb. 2010), p. 084. DOI: 10.1007/jhep02(2010)084.

[60] Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh. “Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches”. In: *Phys. Rev. D* 81 (May 2010), p. 094023. DOI: 10.1103/physrevd.81.094023.

[61] Jonathan M. Butterworth, Adam R. Davison, Mathieu Rubin, and Gavin P. Salam. “Jet Substructure as a New Higgs-Search Channel at the Large Hadron Collider”. In: *Phys. Rev. Lett.* 100 (June 2008), p. 242001. DOI: 10.1103/physrevlett.100.242001.

[62] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, and Jesse Thaler. “Soft drop”. In: *J. High Energ. Phys.* 2014 (May 2014), p. 146. DOI: 10.1007/jhep05(2014)146.

[63] Jesse Thaler and Ken Van Tilburg. “Identifying boosted objects with N-subjettiness”. In: *J. High Energ. Phys.* 2011 (Mar. 2011), p. 015. DOI: 10.1007/jhep03(2011)015.

[64] CMS Collaboration. “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”. In: *J. Instrum.* 13 (May 2018), P05011. DOI: 10.1088/1748-0221/13/05/p05011.

[65] CMS Collaboration. “Performance of heavy-flavour jet identification in boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV”. CMS Physics Analysis Summary CMS-PAS-BTV-22-001. 2023.

[66] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. “Asymptotic formulae for likelihood-based tests of new physics”. In: *Eur. Phys. J. C* 71 (Feb. 2011), p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0.

[67] Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. “IX. On the problem of the most efficient tests of statistical hypotheses”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (1933), pp. 289–337. DOI: 10.1098/rsta.1933.0009.

[68] Thomas Junk. “Confidence level computation for combining searches with small statistics”. In: *Nucl. Instrum. Methods Phys. Res. A* 434 (Sept. 1999), pp. 435–443. DOI: 10.1016/s0168-9002(99)00498-2.

[69] Alexander L. Read. “Presentation of search results: the CLs technique”. In: *J. Phys. G* 28 (Sept. 2002), p. 2693. DOI: 10.1088/0954-3899/28/10/313.

[70] Janice Button, George R. Kalbfleisch, Gerald R. Lynch, Bogdan C. Maglić, Arthur H. Rosenfeld, and M. Lynn Stevenson. “Pion-Pion Interaction in the Reaction $\bar{p} + p \rightarrow 2\pi^+ + 2\pi^- + n\pi^0$”. In: *Phys. Rev.* 126 (June 1962), pp. 1858–1863. DOI: 10.1103/PhysRev.126.1858.

[71] Georgios Choudalakis. “On hypothesis testing, trials factor, hypertests and the BumpHunter”. 2011. arXiv: 1101.0390 [physics.data-an].

[72] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* http://www.deeplearningbook.org. MIT Press, 2016.

[73] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. 2014. arXiv: 1412.6980 [cs.LG].

[74] Warren S. McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *Bull. Math. Biol.* 5 (1943), pp. 115–133.

[75] D.O. Hebb. *The Organization of Behavior: A Neuropsychological Theory.* 1949.

[76] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychol. Rev.* 65 (1958), pp. 386–408.

[77] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators”. In: *Neural Netw.* 2 (1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

[78] Seppo Linnainmaa. “The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors”. (in Finnish). Master Thesis. University of Helsinki, 1970.

[79] Paul J. Werbos. “Applications of advances in nonlinear sensitivity analysis”. In: *System Modeling and Optimization.* Ed. by R. F. Drenick and F. Kozin. Berlin, Heidelberg: Springer Berlin Heidelberg, 1982, pp. 762–770. DOI: 10.1007/BFb0006203.

[80] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323 (1986), pp. 533–536. DOI: 10.1038/323533a0.

[81] Leo Breiman. “Bagging predictors”. In: *Mach. Learn.* 24 (1996), pp. 123–140. DOI: 10.1007/BF00058655.

[82] Leo Breiman. "Random forests". In: *Mach. Learn.* 45 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[83] Yoav Freund and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *J. Comput. Syst. Sci.* 55 (1997), pp. 119–139. DOI: https://doi.org/10.1006/jcss.1997.1504.

[84] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Ann. Stat.* 29 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.

[85] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '16. ACM, Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[86] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Adv. Neural Inf. Process Syst.* Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. NIPS'17. Curran Associates, Inc., 2017, pp. 3149–3157.

[87] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". 2013. arXiv: 1312.6114 [stat.ML].

[88] Irina Higgins, Loic Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. "beta-VAE: Learning basic visual concepts with a constrained variational framework." In: *ICLR (Poster)* 3 (2017).

[89] Emanuel Parzen. "On estimation of a probability density function and mode". In: *Ann. Math. Stat.* 33 (1962), pp. 1065–1076.

[90] Esteban G. Tabak and Eric Vanden-Eijnden. "Density estimation by dual ascent of the log-likelihood". In: *Commun. Math. Sci.* 8 (2010), pp. 217–233.

[91] Esteban G. Tabak and Cristina V. Turner. "A family of nonparametric density estimation algorithms". In: *Commun. Pure Appl. Math.* 66 (2013), pp. 145–164. DOI: 10.1002/cpa.21423.

[92] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". 2016. arXiv: 1605.08803 [cs.LG].

[93] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. "Improving Variational Inference with Inverse Autoregressive Flow". 2016. arXiv: 1606.04934 [cs.LG].

[94] George Papamakarios, Theo Pavlakou, and Iain Murray. "Masked Autoregressive Flow for Density Estimation". 2017. arXiv: 1705.07057 [stat.ML].

[95] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. "MADE: Masked Autoencoder for Distribution Estimation". 2015. arXiv: 1502.03509 [cs.LG].

[96] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. "Learning Likelihoods with Conditional Normalizing Flows". 2023. arXiv: 1912.00042 [cs.LG].

[97] Jason Gallicchio, John Huth, Michael Kagan, Matthew D. Schwartz, Kevin Black, and Brock Tweedie. "Multivariate discrimination and the Higgs+W/Z search". In: *J. High Energ. Phys.* 2011 (Apr. 2011), p. 069. DOI: 10.1007/jhep04(2011)069.

[98] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". In: *ACM Comput. Surv.* 41 (July 2009). DOI: 10.1145/1541880.1541882.

[99] Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih. "Machine learning in the search for new fundamental physics". In: *Nature Rev. Phys.* 4 (2022), pp. 399–412. DOI: 10.1038/s42254-022-00455-1.

[100] Vasilis Belis, Patrick Odagiu, and Thea Klaeboe Åarrestad. "Machine learning for anomaly detection in particle physics". In: *Rev. Phys.* 12 (Dec. 2024), p. 100091. DOI: 10.1016/j.revip.2024.100091.

[101] Charline Le Lan and Laurent Dinh. "Perfect Density Models Cannot Guarantee Anomaly Detection". In: *Entropy* 23 (2021). DOI: 10.3390/e23121690.

[102] Gregor Kasieczka, Radha Mastandrea, Vinicius Mikuni, Benjamin Nachman, Mariel Pettee, and David Shih. "Anomaly detection under coordinate transformations". In: *Phys. Rev. D* 107 (Jan. 2023), p. 015009. DOI: 10.1103/physrevd.107.015009.

[103] Marco Farina, Yuichiro Nakai, and David Shih. "Searching for new physics with deep autoencoders". In: *Phys. Rev. D* 101 (Apr. 2020), p. 075021. DOI: 10.1103/physrevd.101.075021.

[104] Theo Heimel, Gregor Kasieczka, Tilman Plehn, and Jennifer Thompson. "QCD or what?" In: *SciPost Phys.* 6 (Mar. 2019), p. 030. DOI: 10.21468/scipostphys.6.3.030.

[105] Jan Hajer, Ying-Ying Li, Tao Liu, and He Wang. "Novelty detection meets collider physics". In: *Phys. Rev. D* 101 (Apr. 2020), p. 076015. DOI: 10.1103/physrevd.101.076015.

[106] Olmo Cerri, Thong Q. Nguyen, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant. "Variational autoencoders for new physics mining at the Large Hadron Collider". In: *J. High Energ. Phys.* 05 (May 2019), p. 036. DOI: 10.1007/jhep05(2019)036.

[107] Taoli Cheng, Jean-François Arguin, Julien Leissner-Martin, Jacinthe Pilette, and Tobias Golling. "Variational autoencoders for anomalous jet tagging". In: *Phys. Rev. D* 107 (Jan. 2023), p. 016002. DOI: 10.1103/physrevd.107.016002.

[108] Barry Dillon, Tilman Plehn, Christof Sauer, and Peter Sorrenson. "Better latent spaces for better autoencoders". In: *SciPost Phys.* 11 (Sept. 2021). DOI: 10.21468/scipostphys.11.3.061.

[109] Barry M. Dillon, Luigi Favaro, Tilman Plehn, Peter Sorrenson, and Michael Krämer. "A normalized autoencoder for LHC triggers". In: *SciPost Phys. Core* 6 (2023), p. 074. DOI: 10.21468/SciPostPhysCore.6.4.074.

[110] Rob Verheyen. "Event Generation and Density Estimation with Surjective Normalizing Flows". In: *SciPost Phys.* 13 (Sept. 2022), p. 047. DOI: 10.21468/scipostphys.13.3.047.

[111] Claudius Krause, Benjamin Nachman, Ian Pang, David Shih, and Yunhao Zhu. "Anomaly detection with flow-based fast calorimeter simulators". In: *Phys. Rev. D* 110 (Aug. 2024), p. 035036. DOI: 10.1103/PhysRevD.110.035036.

[112] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. "Do Deep Generative Models Know What They Don't Know?" 2018. arXiv: 1810.09136 [stat.ML].

[113] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. "Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality". 2019. arXiv: 1906.02994 [stat.ML].

[114] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. "Why normalizing flows fail to detect out-of-distribution data". In: *Adv. Neural Inf. Process Syst.* 33 (2020), pp. 20578–20589.

[115] Jerónimo Hernández-González, Iñaki Inza, and Jose A. Lozano. "Weak supervision and other non-standard classification problems: A taxonomy". In: *Pattern Recognit. Lett.* 69 (2016), pp. 49–55. DOI: 10.1016/j.patrec.2015.10.008.

[116] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles". In: *Artif. Intell.* 89 (1997), pp. 31–71. DOI: 10.1016/S0004-3702(96)00034-3.

[117] Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. "Classification with Asymmetric Label Noise: Consistency and Maximal Denoising". 2013. arXiv: 1303.1208 [stat.ML].

[118] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. "Classification without labels: learning from mixed samples in high energy physics". In: *J. High Energ. Phys.* 10 (Oct. 2017), p. 174. DOI: 10.1007/jhep10(2017)174.

[119] Kees Benkendorfer, Luc Le Pottier, and Benjamin Nachman. "Simulation-assisted decorrelation for resonant anomaly detection". In: *Phys. Rev. D* 104 (Aug. 2021), p. 035003. DOI: 10.1103/physrevd.104.035003.

[120] Jack Collins, Kiel Howe, and Benjamin Nachman. "Anomaly Detection for Resonant New Physics with Machine Learning". In: *Phys. Rev. Lett.* 121 (Dec. 2018), p. 241803. DOI: 10.1103/physrevlett.121.241803.

[121] Jack H. Collins, Kiel Howe, and Benjamin Nachman. "Extending the search for new resonances with machine learning". In: *Phys. Rev. D* 99 (Jan. 2019), p. 014038. DOI: 10.1103/physrevd.99.014038.

[122] Anders Andreassen, Benjamin Nachman, and David Shih. "Simulation assisted likelihood-free anomaly detection". In: *Phys. Rev. D* 101 (May 2020), p. 095004. DOI: 10.1103/physrevd.101.095004.

[123] Anders Andreassen and Benjamin Nachman. "Neural networks for full phase-space reweighting and parameter tuning". In: *Phys. Rev. D* 101 (May 2020), p. 091901. DOI: 10.1103/physrevd.101.091901.

[124] Benjamin Nachman and David Shih. "Anomaly detection with density estimation". In: *Phys. Rev. D* 101 (Apr. 2020), p. 075042. DOI: 10.1103/physrevd.101.075042.

[125] Oz Amram and Cristina Mantilla Suarez. "Tag N' Train: a technique to train improved classifiers on unlabeled data". In: *J. High Energ. Phys.* 01 (Jan. 2021), p. 153. DOI: 10.1007/jhep01(2021)153.

[126] John Andrew Raine, Samuel Klein, Debajyoti Sengupta, and Tobias Golling. "CURTAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals". In: *Front. Big Data* 6 (Mar. 2023), p. 899345. DOI: 10.3389/fdata.2023.899345.

[127] Debajyoti Sengupta, Sam Klein, John Andrew Raine, and Tobias Golling. "CURTAINs flows for flows: Constructing unobserved regions with maximum likelihood estimation". In: *SciPost Phys.* 17 (2024), p. 046. DOI: 10.21468/SciPostPhys.17.2.046.

[128] Tobias Golling, Samuel Klein, Radha Mastandrea, and Benjamin Nachman. "Flow-enhanced transportation for anomaly detection". In: *Phys. Rev. D* 107 (May 2023), p. 096025. DOI: 10.1103/physrevd.107.096025.

[129] Sang Eon Park, Dylan Rankin, Silviu-Marian Udrescu, Mikaeel Yunus, and Philip Harris. "Quasi anomalous knowledge: searching for new physics with embedded knowledge". In: *J. High Energ. Phys.* 06 (June 2021), p. 030. DOI: 10.1007/jhep06(2021)030.

[130] DØ Collaboration. "Search for new physics in $e\mu X$ data at DØ using SLEUTH: A quasi-model-independent search strategy for new physics". In: *Phys. Rev. D* 62 (Oct. 2000), p. 092004. DOI: 10.1103/physrevd.62.092004.

[131] DØ Collaboration. "Quasi-model-independent search for new physics at large transverse momentum". In: *Phys. Rev. D* 64 (June 2001), p. 012004. DOI: 10.1103/physrevd.64.012004.

[132] DØ Collaboration. "Quasi-Model-Independent Search for New High p_T Physics at DØ". In: *Phys. Rev. Lett.* 86 (Apr. 2001), pp. 3712–3717. DOI: 10.1103/physrevlett.86.3712.

[133] DØ Collaboration. "Model independent search for new phenomena in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV". In: *Phys. Rev. D* 85 (May 2012), p. 092015. DOI: 10.1103/physrevd.85.092015.

[134] CDF Collaboration. "Model-independent and quasi-model-independent search for new physics at CDF". In: *Phys. Rev. D* 78 (July 2008), p. 012002. DOI: 10.1103/physrevd.78.012002.

[135] CDF Collaboration. "Global search for new physics with 2.0 fb^{-1} at CDF". In: *Phys. Rev. D* 79 (Jan. 2009), p. 011101. DOI: 10.1103/physrevd.79.011101.

[136] H1 Collaboration. "A general search for new phenomena in *ep* scattering at HERA". In: *Phys. Lett. B* 602 (Nov. 2004), pp. 14–30. DOI: 10.1016/s0370-2693(04)01396-6.

[137] H1 Collaboration. "A general search for new phenomena at HERA". In: *Phys. Lett. B* 674 (Apr. 2009), pp. 257–268. DOI: 10.1016/j.physletb.2009.03.034.

[138] UA1 Collaboration. "Two-jet mass distributions at the CERN proton-antiproton collider". In: *Phys. Lett. B* 209 (1988), pp. 127–134. DOI: 10.1016/0370-2693(88)91843-6.

[139] UA2 Collaboration. "A measurement of two-jet decays of the W and Z bosons at the CERN $p\bar{p}$ collider". In: *Z. Phys. C* 49 (1991), pp. 17–28. DOI: 10.1007/BF01570793.

[140] CDF Collaboration. "Two-jet invariant-mass distribution at $\sqrt{s} = 1.8$ TeV". In: *Phys. Rev. D* 41 (1990), pp. 1722–1725. DOI: 10.1103/PhysRevD.41.1722.

[141] DØ Collaboration. "Search for new particles in the two-jet decay channel with the DØ detector". In: *Phys. Rev. D* 69 (2004), p. 111101. DOI: 10.1103/PhysRevD.69.111101.

[142] CDF Collaboration. "Search for new particles decaying into dijets in proton-antiproton collisions at $\sqrt{s} = 1.96$ TeV". In: *Phys. Rev. D* 79 (June 2009), p. 112002. DOI: 10.1103/physrevd.79.112002.

[143] ATLAS Collaboration. "Search for resonances in the mass distribution of jet pairs with one or two jets identified as *b*-jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev. D* 98 (2018), p. 032016. DOI: 10.1103/PhysRevD.98.032016.

[144] Duccio Pappadopulo, Andrea Thamm, Riccardo Torre, and Andrea Wulzer. "Heavy vector triplets: bridging theory and data". In: *J. High Energ. Phys.* 2014 (Sept. 2014), p. 60. DOI: 10.1007/jhep09(2014)060.

[145] Alan Kahn, Julia Gonski, Inês Ochoa, Daniel Williams, and Gustaaf Brooijmans. "Anomalous jet identification via sequence modeling". In: *J. Instrum.* 16 (Aug. 2021), P08012. DOI: 10.1088/1748-0221/16/08/p08012.

[146] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. "A Recurrent Latent Variable Model for Sequential Data". 2016. arXiv: 1506.02216 [cs.LG].

[147] Andrew J. Larkoski, Ian Moult, and Duff Neill. "Analytic Boosted Boson Discrimination". In: *J. High Energ. Phys.* 05 (2016), p. 117. DOI: 10.1007/JHEP05(2016)117. eprint: 1507.03018.

[148] Sergei Chekanov. "Imaging particle collision data for event classification using machine learning". In: *Nucl. Instrum. Methods Phys. Res. A* 931 (July 2019), pp. 92–99. DOI: 10.1016/j.nima.2019.04.031.

[149] Sergei Chekanov and Walter Hopkins. "Event-Based Anomaly Detection for Searches for New Physics". In: *Universe* 8 (Sept. 2022), p. 494. DOI: 10.3390/universe8100494.

[150] Gregor Kasieczka, Ben Nachman, and David Shih. "R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge". Version v5. Apr. 2019. DOI: 10.5281/zenodo.6466204.

[151] Gregor Kasieczka, Benjamin Nachman, David Shih, Oz Amram, Anders Andreassen, Kees Benkendorfer, Blaz Bortolato, Gustaaf Brooijmans, Florencia Canelli, Jack H Collins, Biwei Dai, Felipe F De Freitas, Barry M Dillon, Ioan-Mihail Dinu, Zhongtian Dong, Julien Donini, Javier Duarte, D A Faroughy, Julia Gonski, Philip Harris, Alan Kahn, Jernej F Kamenik, Charanjit K Khosa, Patrick Komiske, Luc Le Pottier, Pablo Martín-Ramiro, Andrej Matevc, Eric Metodiev, Vinicius Mikuni, Christopher W Murphy, Inês Ochoa, Sang Eon Park, Maurizio Pierini, Dylan Rankin, Veronica Sanz, Nilai Sarda, Urŏ Seljak, Aleks Smolkovic, George Stein, Cristina Mantilla Suarez, Manuel Szewc, Jesse Thaler, Steven Tsan, Silviu-Marian Udrescu, Louis Vaslin, Jean-Roch Vlimant, Daniel Williams, and Mikaeel Yunus. "The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics". In: *Rep. Prog. Phys.* 84 (Dec. 2021), p. 124201. DOI: 10.1088/1361-6633/ac36b9.

[152] 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". In: *J. High Energ. Phys.* 2014 (Feb. 2014), p. 057. DOI: 10.1007/jhep02(2014)057.

[153] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet user manual: (for version 3.0.2)". In: *Eur. Phys. J. C* 72 (Mar. 2012), p. 1896. DOI: 10.1140/epjc/s10052-012-1896-2.

[154] David Shih. "Additional QCD Background Events for LHCO2020 R&D (signal region only)". Version v1. Dec. 2021. DOI: 10.5281/zenodo.5759087.

[155] Manuel Sommerhalder, Anna Hallin, Joshua Isaacson, Claudius Krause, and Tobias Quadfasel. "CATHODE". https://github.com/HEPML-AnomalyDetection/CATHODE. Commit: e03dc9f2ef9ea73848ed096b8b60e822493fea10. 2021.

[156] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Adv. Neural Inf. Process Syst.* Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014.

[157] Christopher M Bishop. “Mixture density networks”. Technical Report. Birmingham, UK: Aston University, 1994.

[158] Ilya Kostrikov. “pytorch-flows”. `https://github.com/ikostrikov/pytorch-flows`. Commit: `bf12ed91b86867b38d74982f5e2d44c248604df9`. 2018.

[159] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Adv. Neural Inf. Process Syst.* 32 (2019).

[160] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *J. Mach. Learn. Res.* 12 (2011), pp. 2825–2830.

[161] Anja Butter, Sascha Diefenbacher, Gregor Kasieczka, Benjamin Nachman, and Tilman Plehn. “GANplifying event samples”. In: *SciPost Phys.* 10 (June 2021), p. 139. DOI: `10.21468/scipostphys.10.6.139`.

[162] Manuel Sommerhalder, Anna Hallin, Joshua Isaacson, Claudius Krause, and Tobias Quadfasel. “LaCATHODE”. `https://github.com/HEPML-AnomalyDetection/CATHODE/tree/LaCATHODE`. Commit: `01433508a4dd50f2cefe5d6c9a7ca0c41a76baef`. 2023.

[163] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. “Array programming with NumPy”. In: *Nature* 585 (Sept. 2020), pp. 357–362. DOI: `10.1038/s41586-020-2649-2`.

[164] François Chollet et al. “Keras”. `https://keras.io`. 2015.

[165] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems”. Software available from tensorflow.org. 2015.

[166] Marie Hein, Tobias Quadfasel, and Manuel Sommerhalder. “Tree-Based Algorithms for Weakly Supervised Anomaly Detection”. `https://github.com/uhh-pd-ml/treebased_anomaly_detection`. Commit: `f6b7017cec4d4c80df6ecf4fb9d8f6710a97c91b`. 2023.

[167] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. “Searching for exotic particles in high-energy physics with deep learning”. In: *Nat. Commun.* 5 (2014), p. 4308.

[168] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. “Why do tree-based models still outperform deep learning on tabular data?” 2022. arXiv: `2207.08815 [cs.LG]`.

[169] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. “Deep Neural Networks and Tabular Data: A Survey”. In: *IEEE Trans. Neural Netw. Learn. Syst.* 35 (2022), pp. 7499–7519. DOI: 10.1109/TNNLS.2022.3229161.

[170] Gregor Kasieczka, Tilman Plehn, Anja Butter, Kyle Cranmer, Dipsikha Debnath, Barry M. Dillon, Malcolm Fairbairn, Darius A. Faroughy, Wojtek Fedorko, Christophe Gay, Loukas Gouskos, Jernej F. Kamenik, Patrick T. Komiske, Simon Leiss, Alison Lister, Sebastian Macaluso, Eric M. Metodiev, Liam Moore, Ben Nachman, Karl Nordström, Jannicke Pearkes, Huilin Qu, Yannik Rath, Marcel Rieger, David Shih, Jennifer M. Thompson, and Sreedevi Varma. “The Machine Learning landscape of top taggers”. In: *SciPost Phys.* 7 (2019), p. 014. DOI: 10.21468/SciPostPhys.7.1.014.

[171] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, A. Christov, D. Dannheim, K. Danielowski, S. Henrot-Versille, M. Jachowski, K. Kraszewski, A. Krasznahorkay Jr. au2, M. Kruk, Y. Mahalalel, R. Ospanov, X. Prudent, A. Robert, D. Schouten, F. Tegenfeldt, A. Voigt, K. Voss, M. Wolter, and A. Zemla. “TMVA - Toolkit for Multivariate Data Analysis”. 2007. arXiv: physics/0703039 [physics.data-an].

[172] Rene Brun and Fons Rademakers. “ROOT — An object oriented data analysis framework”. In: *Nucl. Instrum. Methods Phys. Res. A* 389 (1997). New Computing Techniques in Physics Research V, pp. 81–86. DOI: 10.1016/S0168-9002(97)00048-X.

[173] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. “Optuna: A Next-generation Hyperparameter Optimization Framework”. 2019. arXiv: 1907.10902 [cs.LG].

[174] Marat Freytsis, Maxim Perelstein, and Yik Chuen San. “Anomaly detection in the presence of irrelevant features”. In: *J. High Energ. Phys.* 2024 (Feb. 2024), p. 220. DOI: 10.1007/JHEP02(2024)220.

[175] CMS Collaboration. “Searching for new physics detecting anomalies in jets”. CMS Analysis Note AN-2020/051. (Collaboration internal). 2024.

[176] Katarzyna Maria Dziedziniewicz, Domenico Giordano, Vincenzo Innocente, Anne-Catherine Le Bihan, Antonio Pierro, and Zhen Xie. “CMS conditions database web application service”. In: *J. Phys. Conf. Ser.* 219 (2010). Ed. by Jan Gruntorad and Milos Lokajicek, p. 072048. DOI: 10.1088/1742-6596/219/7/072048.

[177] CMS Collaboration. *Data Aggregation Service*. URL: https://cmsweb.cern.ch/das/ (visited on 08/21/2024).

[178] CMS Collaboration. “Simulation of the Silicon Strip Tracker pre-amplifier in early 2016 data”. CMS Detector Performance Summary CMS-DP-2020-045. 2020.

[179] CMS Collaboration. “PFNano”. https://github.com/cms-jet/PFNano. Commit: 37a791cd3dbc88f09d402e0d9e1777a0941c9a51. 2020.

[180] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. In: *J. High Energ. Phys.* 2014 (July 2014), p. 79. DOI: 10.1007/jhep07(2014)079.

[181] Paolo Nason. “A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms”. In: *J. High Energ. Phys.* 2004 (Nov. 2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040.

[182] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". In: *J. High Energ. Phys.* 2007 (Nov. 2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070.

[183] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *J. High Energ. Phys.* 06 (June 2010), p. 043. DOI: 10.1007/jhep06(2010)043.

[184] Richard D. Ball, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, José I. Latorre, Juan Rojo, and Maria Ubiali. "A first unbiased global NLO determination of parton distributions and their uncertainties". In: *Nucl. Phys. B* 838 (Oct. 2010), pp. 136–206. DOI: 10.1016/j.nuclphysb.2010.05.008.

[185] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Christopher S. Deans, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, José I. Latorre, Juan Rojo, and Maria Ubiali. "Parton distributions for the LHC run II". In: *J. High Energ. Phys.* 04 (Apr. 2015), p. 40. DOI: 10.1007/jhep04(2015)040.

[186] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Patrick Groth-Merrild, Alberto Guffanti, Nathan P. Hartland, Zahari Kassabov, José I. Latorre, Emanuele R. Nocera, Juan Rojo, Luca Rottoli, Emma Slade, and Maria Ubiali. "Parton distributions from high-precision collider data: NNPDF Collaboration". In: *Eur. Phys. J. C* 77 (Oct. 2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5.

[187] CMS Collaboration. "Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements". In: *Eur. Phys. J. C* 80 (Jan. 2020), p. 4. DOI: 10.1140/epjc/s10052-019-7499-4.

[188] S. Agostinelli et al. "Geant4—a simulation toolkit". In: *Nucl. Instrum. Methods Phys. Res. A* 506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[189] J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce Dubois, M. Asai, G. Barrand, R. Capra, S. Chauvie, R. Chytracek, G.A.P. Cirrone, G. Cooperman, G. Cosmo, G. Cuttone, G.G. Daquino, M. Donszelmann, M. Dressel, G. Folger, F. Foppiano, J. Generowicz, V. Grichine, S. Guatelli, P. Gumplinger, A. Heikkinen, I. Hrivnacova, A. Howard, S. Incerti, V. Ivanchenko, T. Johnson, F. Jones, T. Koi, R. Kokoulin, M. Kossov, H. Kurashige, V. Lara, S. Larsson, F. Lei, O. Link, F. Longo, M. Maire, A. Mantero, B. Mascialino, I. McLaren, P. Mendez Lorenzo, K. Minamimoto, K. Murakami, P. Nieminen, L. Pandola, S. Parlati, L. Peralta, J. Perl, A. Pfeiffer, M.G. Pia, A. Ribon, P. Rodrigues, G. Russo, S. Sadilov, G. Santin, T. Sasaki, D. Smith, N. Starkov, S. Tanaka, E. Tcherniaev, B. Tome, A. Trindade, P. Truscott, L. Urban, M. Verderi, A. Walkden, J.P. Wellisch, D.C. Williams, D. Wright, and H. Yoshida. "Geant4 developments and applications". In: *IEEE Trans. Nucl. Sci.* 53 (2006), pp. 270–278. DOI: 10.1109/TNS.2006.869826.

[190] J. Allison, K. Amako, J. Apostolakis, P. Arce, M. Asai, T. Aso, E. Bagli, A. Bagulya, S. Banerjee, G. Barrand, B.R. Beck, A.G. Bogdanov, D. Brandt, J.M.C. Brown, H. Burkhardt, Ph. Canal, D. Cano-Ott, S. Chauvie, K. Cho, G.A.P. Cirrone, G. Cooperman, M.A. Cortés-Giraldo, G. Cosmo, G. Cuttone, G. Depaola, L. Desorgher, X. Dong, A. Dotti, V.D. Elvira, G. Folger, Z. Francis, A. Galoyan, L. Garnier, M. Gayer, K.L. Genser, V.M. Grichine, S. Guatelli, P. Guèye, P. Gumplinger, A.S. Howard, I. Hřivnáčová, S. Hwang, S. Incerti, A. Ivanchenko, V.N. Ivanchenko, F.W. Jones, S.Y. Jun, P. Kaitaniemi, N. Karakatsanis, M. Karamitros, M. Kelsey, A. Kimura, T. Koi, H. Kurashige, A. Lechner, S.B. Lee, F. Longo, M. Maire, D. Mancusi, A. Mantero, E. Mendoza, B. Morgan, K. Murakami, T. Nikitina, L. Pandola, P. Paprocki, J. Perl, I. Petrović, M.G. Pia, W.

Pokorski, J.M. Quesada, M. Raine, M.A. Reis, A. Ribon, A. Ristić Fira, F. Romano, G. Russo, G. Santin, T. Sasaki, D. Sawkey, J.I. Shin, I.I. Strakovsky, A. Taborda, S. Tanaka, B. Tomé, T. Toshito, H.N. Tran, P.R. Truscott, L. Urban, V. Uzhinsky, J.M. Verbeke, M. Verderi, B.L. Wendt, H. Wenzel, D.H. Wright, D.M. Wright, T. Yamashita, J. Yarba, and H. Yoshida. "Recent developments in Geant4". In: *Nucl. Instrum. Methods Phys. Res. A* 835 (2016), pp. 186–225. DOI: `10.1016/j.nima.2016.06.125`.

[191] CMS Collaboration. "A multi-dimensional search for new heavy resonances decaying to boosted WW, WZ, ZZ, WH or ZH boson pairs in the all jets final state at 13 TeV". CMS Analysis Note AN-2019/131. (Collaboration internal). 2022.

[192] CMS Collabotation. "A multi-dimensional search for new heavy resonances decaying to boosted WW, WZ, or ZZ boson pairs in the dijet final state at 13 TeV". In: *Eur. Phys. J. C* 80 (Mar. 2020), p. 237. DOI: `10.1140/epjc/s10052-020-7773-5`.

[193] CMS Collaboration. *Summary table of samples produced for the 1 Billion campaign, with 25ns bunch-crossing.* (Collaboration internal). URL: `https://twiki.cern.ch/twiki/bin/viewauth/CMS/SummaryTable1G25ns` (visited on 06/19/2024).

[194] CMS Collaboration. *XSDB.* URL: `https://xsdb-temp.app.cern.ch/` (visited on 06/19/2024).

[195] Ulrich Baur, Ian Hinchliffe, and Dieter Zeppenfeld. "Excited quark production at hadron colliders". In: *Int. J. Mod. Phys. A* 02 (1987), pp. 1285–1297. DOI: `10.1142/S0217751X87000661`.

[196] Ulrich Baur, Michael Spira, and Peter Matthias Zerwas. "Excited-quark and -lepton production at hadron colliders". In: *Phys. Rev. D* 42 (Aug. 1990), pp. 815–824. DOI: `10.1103/PhysRevD.42.815`.

[197] Daniele Barducci, Alexander Belyaev, Stefania De Curtis, Stefano Moretti, and Giovanni Marco Pruna. "Exploring Drell-Yan signals from the 4D Composite Higgs Model at the LHC". In: *J. High Energ. Phys.* 04 (Apr. 2013), p. 152. DOI: `10.1007/jhep04(2013)152`.

[198] CMS Collaboration. "Search for a W' boson decaying to a vector-like quark and a top or bottom quark in the all-jets final state at $\sqrt{s} = 13$ TeV". In: *J. High Energ. Phys.* 09 (Sept. 2022), p. 088. DOI: `10.1007/jhep09(2022)088`.

[199] Kaustubh Agashe, Jack H. Collins, Peizhi Du, Sungwoo Hong, Doojin Kim, and Rashmish K. Mishra. "Dedicated strategies for triboson signals from cascade decays of vector resonances". In: *Phys. Rev. D* 99 (Apr. 2019), p. 075016. DOI: `10.1103/physrevd.99.075016`.

[200] CMS Collaboration. "Search for Resonances Decaying to Three W Bosons in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. Lett.* 129 (July 2022), p. 021802. DOI: `10.1103/physrevlett.129.021802`.

[201] CMS Collaboration. "Search for resonances decaying to three W bosons in the hadronic final state in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. D* 106 (2022), p. 012002. DOI: `10.1103/PhysRevD.106.012002`.

[202] Alexandra Carvalho. "Gravity particles from Warped Extra Dimensions, predictions for LHC". 2018. arXiv: `1404.0102 [hep-ph]`.

[203] CMS JetMET Physics Object Group. *Noise Filter Recommendations for Run II & Run III.* (Collaboration internal). URL: `https://twiki.cern.ch/twiki/bin/view/CMS/MissingETOptionalFiltersRun2` (visited on 06/19/2024).

[204] CMS JetMET Physics Object Group. *Jet Identification for the 13 TeV UL data.* (Collaboration internal). URL: https://twiki.cern.ch/twiki/bin/view/CMS/JetID13TeVUL (visited on 06/19/2024).

[205] CMS Collaboration. "Determination of jet energy calibration and transverse momentum resolution in CMS". In: *J. Instrum.* 6 (Nov. 2011), P11002. DOI: 10.1088/1748-0221/6/11/p11002.

[206] CMS Collaboration. *Recommended Jet Energy Corrections and Uncertainties For Data and MC.* (Collaboration internal). URL: https://twiki.cern.ch/twiki/bin/view/CMS/JECDataMC (visited on 06/19/2024).

[207] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. "Pyro: Deep Universal Probabilistic Programming". In: *J. Mach. Learn. Res.* (2018).

[208] Du Phan, Neeraj Pradhan, and Martin Jankowiak. "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro". 2019. arXiv: 1912.11554 [stat.ML].

[209] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". 2015. arXiv: 1502.03167 [cs.LG].

[210] Wouter Verkerke and David Kirkby. "The RooFit toolkit for data modeling". 2003. arXiv: physics/0306116 [physics.data-an].

[211] Fred James and Matthias Winkler. *Minuit 2.* URL: https://root.cern/root/htmldoc/guides/minuit2/Minuit2.html (visited on 07/08/2024).

[212] Richard G Lomax and Debbie L Hahs-Vaughn. *Statistical Concepts-A Second Course.* Routledge, 2012.

[213] CMS Collaboration. "Search for heavy resonances in the W/Z-tagged dijet mass spectrum in pp collisions at 7 TeV". In: *Phys. Lett. B* 723 (June 2013), pp. 280–301. DOI: 10.1016/j.physletb.2013.05.040.

[214] Mark Joseph Oreglia. "A study of the reactions $\psi' \to \gamma\gamma\psi$". SLAC Report SLAC-R-236. PhD thesis. Stanford University, 1980.

[215] CMS Collaboration. "Search for new heavy resonances decaying to WW, WZ, ZZ, WH, or ZH boson pairs in the all-jets final state in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Lett. B* 844 (Sept. 2023), p. 137813. DOI: 10.1016/j.physletb.2023.137813.

[216] CMS Collaboration. "The CMS statistical analysis and combination tool: COMBINE". 2024. arXiv: 2404.06614 [physics.data-an].

[217] CMS B Tag & Vertexing Physics Object Group. *Methods to apply b-tagging efficiency scale factors.* (Collaboration internal). URL: https://twiki.cern.ch/twiki/bin/view/CMS/BTagSFMethods (visited on 09/09/2024).

[218] CMS Collaboration. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *J. Instrum.* 12 (Feb. 2017), P02014. DOI: 10.1088/1748-0221/12/02/p02014.

[219] CMS Collaboration. "Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV". CMS Detector Performance Summary CMS-DP-2021-033. 2021.

[220] CMS Collaboration. "Lund Plane Reweighting for Jet Substructure Correction". CMS Detector Performance Summary CMS-DP-2023-046. 2023.

[221] CMS Collaboration. "Jet Substructure Corrections with Sub-Jet Lund Plane Reweighting". CMS Analysis Note AN-2022/180. (Collaboration internal). 2024.

[222] Frédéric A. Dreyer, Gavin P. Salam, and Grégory Soyez. "The Lund jet plane". In: *J. High Energ. Phys.* 12 (Dec. 2018), p. 064. DOI: `10.1007/jhep12(2018)064`.

[223] CMS JetMET Physics Object Group. *W/Z-tagging of Jets.* (Collaboration internal). URL: `https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetWtagging` (visited on 07/12/2024).

[224] CMS Collaboration. "Measurement of the top quark mass using charged particles in $\sqrt{s} = 8$ TeV". In: *Phys. Rev. D* 93 (May 2016), p. 092006. DOI: `10.1103/physrevd.93.092006`.

[225] CMS Collaboration. "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$TeV". In: *J. Instrum.* 15 (Oct. 2020), P10017. DOI: `10.1088/1748-0221/15/10/p10017`.

[226] CMS L1 Detector Performance Group. *Reweighting recipe to emulate Level 1 ECAL prefiring.* (Collaboration internal). URL: `https://twiki.cern.ch/twiki/bin/viewauth/CMS/L1ECALPrefiringWeightRecipe` (visited on 07/12/2024).

[227] CMS B Tag & Vertexing Physics Object Group. *B-tagging discriminant shape calibration using event weights with a tag-and-probe method.* (Collaboration internal). URL: `https://twiki.cern.ch/twiki/bin/view/CMS/BTagShapeCalibration` (visited on 07/12/2024).

[228] CMS Collaboration. "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS". In: *Eur. Phys. J. C* 81 (Sept. 2021), p. 800. DOI: `10.1140/epjc/s10052-021-09538-2`.

[229] CMS Collaboration. "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV". CMS Physics Analysis Summary CMS-PAS-LUM-17-004. 2018.

[230] CMS Collaboration. "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV". CMS Physics Analysis Summary CMS-PAS-LUM-18-002. 2019.

[231] Christopher Brust, Petar Maksimovic, Alice Sady, Prashant Saraswat, Matthew T. Walters, and Yongjie Xin. "Identifying boosted new physics with non-isolated leptons". In: *J. High Energ. Phys.* 04 (Apr. 2015), p. 079. DOI: `10.1007/jhep04(2015)079`.

[232] Haoqiang Fan, Hao Su, and Leonidas Guibas. "A Point Set Generation Network for 3D Object Reconstruction from a Single Image". 2016. arXiv: `1612.00603 [cs.CV]`.

[233] George E. P. Box and David R. Cox. "An Analysis of Transformations". In: *J. R. Stat. Soc. Ser. B Methodol.* 26 (July 1964), pp. 211–243. DOI: `10.1111/j.2517-6161.1964.tb00553.x`.

[234] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. "Measuring and testing dependence by correlation of distances". In: *Ann. Stat.* 35 (2007), pp. 2769–2794. DOI: `10.1214/009053607000000505`.

[235] Gerrit Bickendorf, Manuel Drees, Gregor Kasieczka, Claudius Krause, and David Shih. "Combining resonant and tail-based anomaly detection". In: *Phys. Rev. D* 109 (May 2024), p. 096031. DOI: `10.1103/physrevd.109.096031`.

[236] Erik Buhmann, Cedric Ewen, Gregor Kasieczka, Vinicius Mikuni, Benjamin Nachman, and David Shih. "Full phase space resonant anomaly detection". In: *Phys. Rev. D* 109 (Mar. 2024), p. 055015. DOI: `10.1103/PhysRevD.109.055015`.

[237] Vinicius Mikuni and Benjamin Nachman. “OmniLearn: A Method to Simultaneously Facilitate All Jet Physics Tasks”. 2024. arXiv: 2404.16091 [hep-ph].

A General Appendix

A.1 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence between two continuous probability density functions $p(x)$ and $q(x)$ with the same support $x \in \chi$ is defined as:

$$D_{\mathrm{KL}}(p(x)||q(x)) = \int_{\chi} p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right]. \tag{A.1}$$

It is always non-negative and equals zero if and only if $p(x) = q(x)$ almost everywhere.

A.2 Negative Log-Likelihood minimization

The aim is to approximate the unknown true likelihood $p(x)$ via a parametric model $q_\theta(x)$ with parameters θ. From the properties of the Kullback-Leibler divergence, this is achieved by finding an optimal parameter set θ^* that minimizes the KL divergence between the true likelihood and the model:

$$\theta^* = \arg\min_{\theta} D_{\mathrm{KL}}(p(x)||q_\theta(x)). \tag{A.2}$$

The KL divergence can be rewritten as:

$$D_{\mathrm{KL}}(p(x)||q_\theta(x)) = \mathbb{E}_p \log p(x) - \mathbb{E}_p \log q_\theta(x). \tag{A.3}$$

The first term in Eq. A.3 is a constant and does not depend on θ. The second term can be estimated via the empirical average over a dataset $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$ following the true distribution $p(x)$:

$$-\hat{\mathbb{E}}_p \log q_\theta(x) = -\frac{1}{N} \sum_{i=1}^{N} \log q_\theta(x_i), \tag{A.4}$$

which is (up to a constant factor) the negative log-likelihood (NLL) of the dataset $\mathcal{D}$ under the model $q_\theta(x)$:

$$\mathrm{NLL} = -\log \left(\prod_{i=1}^{N} q_\theta(x_i) \right) = -\sum_{i=1}^{N} \log q_\theta(x_i) \tag{A.5}$$

A.3 Properties of the cross-entropy Loss

The general definition of the cross-entropy between two probability density functions $p(x)$ and $q(x)$ with the same support $x \in \chi$ is:

$$\mathcal{H}(p, q) = \int_{\chi} p(x) \log \frac{1}{q(x)} dx = -\mathbb{E}_p \left[\log q(x) \right], \tag{A.6}$$

where in the discrete case the integral is replaced by a sum. It can also be written in terms of the KL divergence (Sec. A.1) as:

$$\mathcal{H}(p, q) = D_{\mathrm{KL}}(p(x)||q(x)) + H(p), \tag{A.7}$$

where $H(p) = H(p, p)$ is the entropy of p. The cross-entropy measures the average number of bits needed to encode the true distribution $p(x)$ using a different distribution $q(x)$, and it is minimized when $q(x) = p(x)$ almost everywhere, in which case it equals the entropy of $p(x)$. The expectation value in the cross-entropy definition of Eq. A.6 can be estimated with the empirical average over a dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ following the distribution $p(x)$:

$$-\hat{\mathbb{E}}_p \log q_\theta(x) = -\frac{1}{N}\sum_{i=1}^{N} \log q_\theta(x_i), \tag{A.8}$$

The cross-entropy is a common choice of loss function to measure the discrepancy between the true distribution $p(x)$ and the parametric model $q_\theta(x)$. In a classification task, one may represent the true distribution as a so-called one-hot encoded vector $y \in \{0,1\}^K$, where K is the number of classes, i.e., only one element of y is one and the others are zero. Correspondingly, the parametric model outputs a probability distribution over the classes, $q_\theta(x) \in [0,1]^K$, such that the element $q_\theta(x)_k$ represents the probability of the sample x belonging to class k. Using the estimation in Eq. A.8, the cross-entropy loss for a dataset $\mathcal{D}$ becomes:

$$\mathcal{L}(q_\theta, p, \mathcal{D}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log q_\theta(x_i)_k. \tag{A.9}$$

The correspondence between minimizing the cross-entropy loss and the minimization of the negative log-likelihood can be illustrated by rearranging the NLL minimization:

$$\begin{aligned}\theta^* = \arg\min_\theta \left(- \sum_{x\sim p(x)} \log q_\theta(x) \right) &= \arg\min_\theta \left(-N\frac{1}{N} \sum_{x\sim p(x)} \log q_\theta(x) \right) \\ &= \arg\min_\theta N \left(-\hat{\mathbb{E}}_p \log q_\theta(x) \right) = \arg\min_\theta \hat{\mathcal{H}}(p, q_\theta). \end{aligned} \tag{A.10}$$

In the case of binary classification ($K = 2$), the label simplifies to a scalar number $y \in \{0,1\}$ and the probabilities $q_\theta(x|y=1)$ and $q_\theta(x|y=0)$ are related by $q_\theta(x|y=0) = 1 - q_\theta(x|y=1)$. The discrete binary cross-entropy thus becomes

$$\mathcal{H}_{k=2}(p, q_\theta) = -\sum_{x\in\chi} \left[p(x|y=1)\log q_\theta(x|y=1) + p(x_i|y=0)\log(1 - q_\theta(x|y=1)) \right], \tag{A.11}$$

where the sum is evaluated on the entire support χ of the random variable x. In terms of a loss function, estimated on dataset $\mathcal{D}$ and identifying $q_\theta(x) \equiv q_\theta(x|y=1)$ for simplicity, this becomes:

$$\mathcal{L}(q_\theta, p, \mathcal{D}) = -\frac{1}{N}\sum_{i=1}^{N} \left[y_i \log q_\theta(x_i) + (1 - y_i)\log(1 - q_\theta(x_i)) \right]. \tag{A.12}$$

An important property of this loss function is that it approximates the likelihood ratio estimator in the limit of a perfect classifier. Starting from Eq. A.11 (with the same shorthand notation of q_θ), one can find a minimum by taking the functional derivative with respect to $q_\theta(x)$ and setting it to zero:

$$\frac{\delta\mathcal{H}_{k=2}(p, q_\theta)}{\delta q_\theta(x)} = -\sum_{x\in\chi} \left[p(x|y=1)\frac{1}{q_\theta(x)} - p(x_i|y=0)\frac{1}{1 - q_\theta(x)} \right] \overset{!}{=} 0. \tag{A.13}$$

This can be solved by

$$\frac{p(x|y=1)}{p(x|y=0)} = \frac{q_\theta(x)}{1 - q_\theta(x)}, \tag{A.14}$$

or equivalently, solving for $q_\theta(x)$:

$$q_\theta(x) = \frac{p(x|y=1)}{p(x|y=1) + p(x|y=0)}. \tag{A.15}$$

Consequently, the function $q_\theta(x)$ approximates a monotonic rescaling of the likelihood ratio $p(x|y=1)/p(x|y=0)$ in the limit of an ideally converged training. In practice, this approximation is limited by the expressiveness of the model, the number of training data, and the convergence of the numeric optimization algorithm towards a global optimum. The same argument applies to some other loss functions, such as the MSE loss.

B CMS Anomalous Dijet Search Appendix

B.1 MC Background Simulation Samples

The following tables list all the MC background simulation samples used in the CMS dijet search, Sec. 7. They are separated into Tab. B.1 for 2016 preVFP datasets, Tab. B.2 for the 2016 postVFP datasets, Tab. B.3 for the 2017 datasets, and Tab. B.4 for the 2018 datasets.

Table B.1: CMS sample identifiers of 2016 preVFP MC background simulation datasets, used for optimizing and validating the analysis strategy, and their corresponding cross sections σ. The QCD multijet samples are binned in jet transverse momentum, the V+jets samples in H_{T}, and the $\mathrm{t\bar{t}}$ samples in top-antitop pair mass. For brevity, the following replacement is used: [**CAMPAIGN**]≡RunIISummer20UL16MiniAODAPVv2-106X_mcRun2_asymptotic_preVFP_v11

Sample name	**Events**
/QCD_Pt_300to470_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	54,096,000
/QCD_Pt_470to600_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	50,782,000
/QCD_Pt_600to800_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	61,972,000
/QCD_Pt_800to1000_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	35,527,000
/QCD_Pt_1000to1400_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	19,077,000
/QCD_Pt_1400to1800_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	11,000,000
/QCD_Pt_1800to2400_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	5,262,000
/QCD_Pt_2400to3200_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	2,999,000
/QCD_Pt_3200toInf_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	1,000,000
/WJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	5,144,427
/WJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	7,668,058
/WJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	7,740,501
/ZJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	3,454,056
/ZJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	1,623,377
/ZJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	3,726,992
/TT_Mtt-700to1000_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	18,020,206
/TT_Mtt-1000toInf_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	23,219,884
/ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/[**CAMPAIGN**]-v3/MINIAODSIM	55,961,000
/ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/[**CAMPAIGN**]-v3/MINIAODSIM	31,024,000
/ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	2,300,000
/ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	2,300,000

Table B.2: CMS sample identifiers of 2016 postVFP MC background simulation datasets, analogous to Tab. B.1 with the replacement [**CAMPAIGN**]≡RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17

Sample name	**Events**
/QCD_Pt_300to470_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	55,264,000
/QCD_Pt_470to600_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	52,408,000
/QCD_Pt_600to800_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	65,088,000
/QCD_Pt_800to1000_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	37,782,000
/QCD_Pt_1000To1400_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	19,892,000
/QCD_Pt_1400to1800_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	10,722,000
/QCD_Pt_1800to2400_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	5,236,000
/QCD_Pt_2400to3200_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	2,848,000
/QCD_Pt_3200toInf_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	996,000
/WJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	4,455,853
/WJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	6,793,578
/WJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	6,769,101
/ZJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	6,942,718
/ZJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	5,500,386
/ZJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	3,740,901
/TT_Mtt-700to1000_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	33,586,554
/TT_Mtt-1000toInf_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	24,177,380
/ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/[**CAMPAIGN**]-v3/MINIAODSIM	63,073,000
/ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/[**CAMPAIGN**]-v3/MINIAODSIM	30,609,000
/ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	2,491,000
/ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	2,554,000

Table B.3: CMS sample identifiers of 2017 MC background simulation datasets, analogous to Tab. B.1 with the replacement [**CAMPAIGN**]≡RunIISummer20UL17MiniAODv2-106X_mc2017_realistic_v9

Sample name	**Events**
/QCD_Pt_300to470_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	55,690,000
/QCD_Pt_470to600_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	50,885,000
/QCD_Pt_600to800_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	67,379,000
/QCD_Pt_800to1000_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	36,890,000
/QCD_Pt_1000to1400_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	19,781,000
/QCD_Pt_1400to1800_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	10,994,000
/QCD_Pt_1800to2400_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	5,488,000
/QCD_Pt_2400to3200_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	2,997,000
/QCD_Pt_3200toInf_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	1,000,000
/WJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	9,927,793
/WJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	14,667,933
/WJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	14,722,417
/ZJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	14,884,962
/ZJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	11,702,567
/ZJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	9,384,525
/TT_Mtt-700to1000_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	35,862,238
/TT_Mtt-1000toInf_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	22,724,532
/ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	127790000
/ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	69509000
/ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	5,649,000
/ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	5,674,000

Table B.4: CMS sample identifiers of 2018 MC background simulation datasets, analogous to Tab. B.1 with the replacement [**CAMPAIGN**]≡RunIISummer20UL18MiniAODv2-106X_upgrade2018_realistic_v16_L1v1

Sample name	**Events**
/QCD_Pt_300to470_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	57,910,000
/QCD_Pt_470to600_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	52,448,000
/QCD_Pt_600to800_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	67,508,000
/QCD_Pt_800to1000_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	37,160,000
/QCD_Pt_1000to1400_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	19,730,000
/QCD_Pt_1400to1800_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	10,982,000
/QCD_Pt_1800to2400_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	5,491,000
/QCD_Pt_2400to3200_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	2,997,000
/QCD_Pt_3200toInf_TuneCP5_13TeV_pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	1,000,000
/WJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	9,335,298
/WJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	13,633,226
/WJetsToQQ_HT-800toInf/[**CAMPAIGN**]-v2/MINIAODSIM	13,581,343
/ZJetsToQQ_HT-400to600/[**CAMPAIGN**]-v2/MINIAODSIM	13,930,474
/ZJetsToQQ_HT-600to800/[**CAMPAIGN**]-v2/MINIAODSIM	12,029,507
/ZJetsToQQ_HT-800toInf/[**CAMPAIGN**]-v2/MINIAODSIM	9,681,521
/TT_Mtt-700to1000_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	30,720,345
/TT_Mtt-1000toInf_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	23,758,200
/ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	178,756,000
/ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/[**CAMPAIGN**]-v1/MINIAODSIM	95,833,000
/ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	5,649,000
/ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[**CAMPAIGN**]-v2/MINIAODSIM	5,674,000

B.2 Signal Samples for Control Region Studies

Table B.5 lists all the signal MC simulation samples that were used for testing various data validation CR definitions in Sec. 7.7.2. They are grouped into the same signal processes with varying masses.

Table B.5: Signal MC simulation samples used for testing various data validation CR definitions, grouped into the same signal processes with varying masses.

Sample name
/BulkGravToWW_narrow_M-2000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToWW_narrow_M-2500_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToWW_narrow_M-3000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToWW_narrow_M-3500_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToWW_narrow_M-4000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToWW_narrow_M-4500_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToWW_narrow_M-5000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToWW_narrow_M-5500_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToWW_narrow_M-6000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-2000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-2500_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-3000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-3500_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-4000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-4500_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-5000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-5500_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BulkGravToZZToZhadZhad_narrow_M-6000_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WprimeToWZToWhadZhad_narrow_M-2000_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/WprimeToWZToWhadZhad_narrow_M-2500_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/WprimeToWZToWhadZhad_narrow_M-3000_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/WprimeToWZToWhadZhad_narrow_M-3500_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/WprimeToWZToWhadZhad_narrow_M-4000_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/WprimeToWZToWhadZhad_narrow_M-4500_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/ZprimeToWW_narrow_M-2000_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/ZprimeToWW_narrow_M-2500_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/ZprimeToWW_narrow_M-3000_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/ZprimeToWW_narrow_M-3500_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/ZprimeToWW_narrow_M-4000_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/ZprimeToWW_narrow_M-4500_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BstarToTW_M-2000_LH_TuneCP5_13TeV-madgraph-pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BstarToTW_M-2600_LH_TuneCP5_13TeV-madgraph-pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BstarToTW_M-3000_LH_TuneCP5_13TeV-madgraph-pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BstarToTW_M-3600_LH_TuneCP5_13TeV-madgraph-pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/BstarToTW_M-4000_LH_TuneCP5_13TeV-madgraph-pythia8/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM
/WkkToWRadionToWWW_M2000-R0-1_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M2000-R0-2_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M2000-R0-3_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M2500-R0-08_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M2500-R0-1_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M2500-R0-2_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M2500-R0-3_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M3000-R0-06_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M3000-R0-08_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M3000-R0-1_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M3500-R0-06_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M3500-R0-08_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M3500-R0-1_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M3500-R0-2_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M4000-R0-08_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M4000-R0-1_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M4500-R0-06_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M4500-R0-08_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM
/WkkToWRadionToWWW_M4500-R0-1_TuneCP5_13TeV-madgraph/RunIIFall17MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v2/MINIAODSIM

B.3 Signal Shape Templates

This section collects the explicit signal template fits for the signal processes shown in Figs. 7.49 and 7.54, namely a $\mathrm{X} \to \mathrm{Y}\mathrm{Y}' \to 4\mathrm{q}$ decay with masses of $m_{\mathrm{Y}} = m_{\mathrm{Y}'} = 170\,\mathrm{GeV}$ (Fig. B.1), a $\mathrm{W}' \to \mathrm{B}'\mathrm{t} \to \mathrm{bZt}$ decay with $m_{\mathrm{B}'} = 400\,\mathrm{GeV}$ (Fig. B.2), a $\mathrm{W}_{\mathrm{KK}} \to \mathrm{RW} \to 3\mathrm{W}$ decay with $m_{\mathrm{R}} = 400\,\mathrm{GeV}$ (Fig. B.3), and a $\mathrm{Y} \to \mathrm{HH} \to 4\mathrm{t}$ decay with $m_{\mathrm{H}} = 400\,\mathrm{GeV}$ (Fig. B.4). These signal templates are used for obtaining the upper cross section limits on the respective signals, and are accordingly shown with resonance masses of 3 and 5 TeV, used for limit setting.

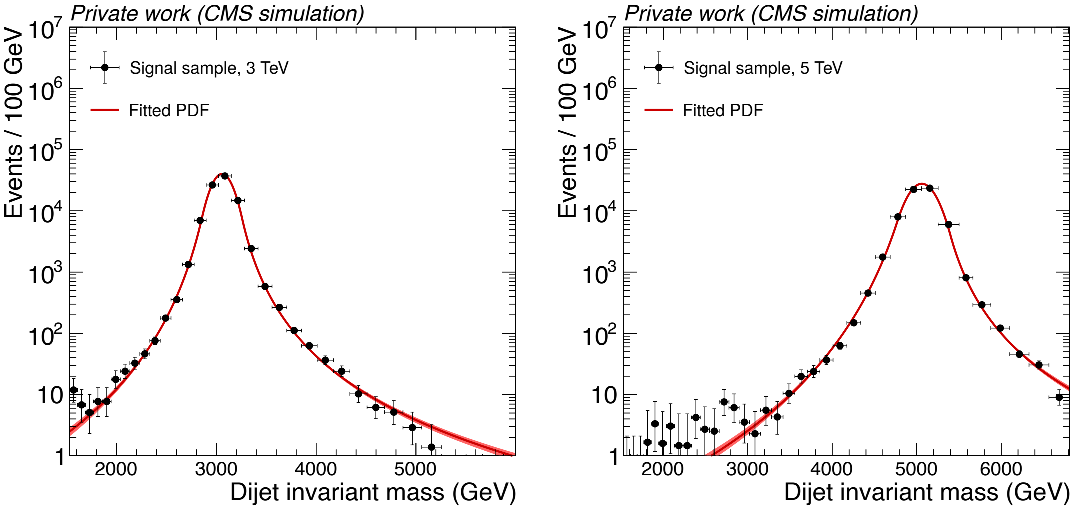

Figure B.1: Signal shape template obtained from fitting the X → YY′ → 4q signal MC simulation samples with intermediate resonance masses of $m_{\mathrm{Y}} = 170\,\mathrm{GeV}$ and $m_{\mathrm{Y'}} = 170\,\mathrm{GeV}$ at resonance masses of 3 TeV (left) and 5 TeV (right), using a Double Crystal Ball function.

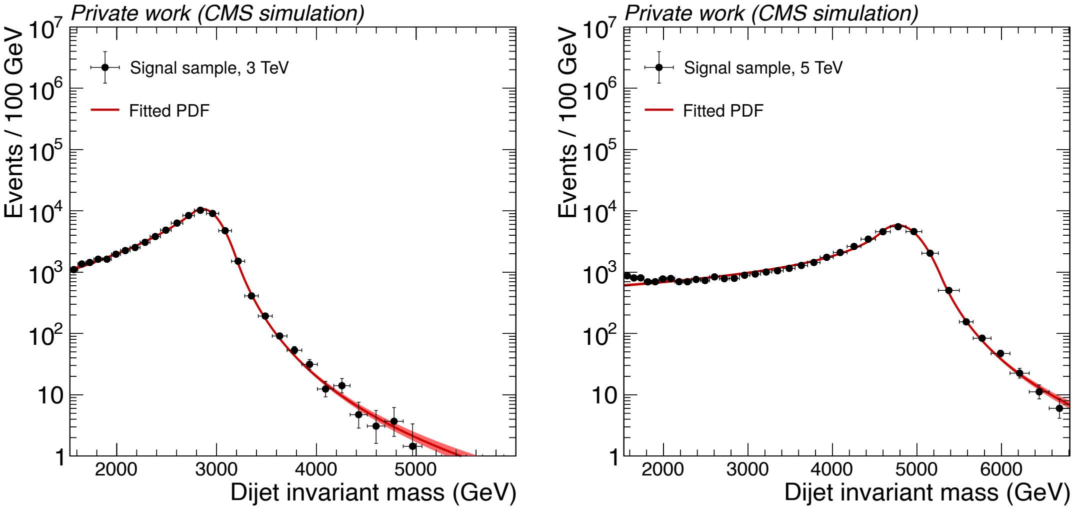

Figure B.2: Signal shape template obtained from fitting the W′ → B′t → bZt signal MC simulation samples with a daughter mass of $m_{\mathrm{B'}} = 400\,\mathrm{GeV}$ at resonance masses of 3 TeV (left) and 5 TeV (right), using a Double Crystal Ball function.

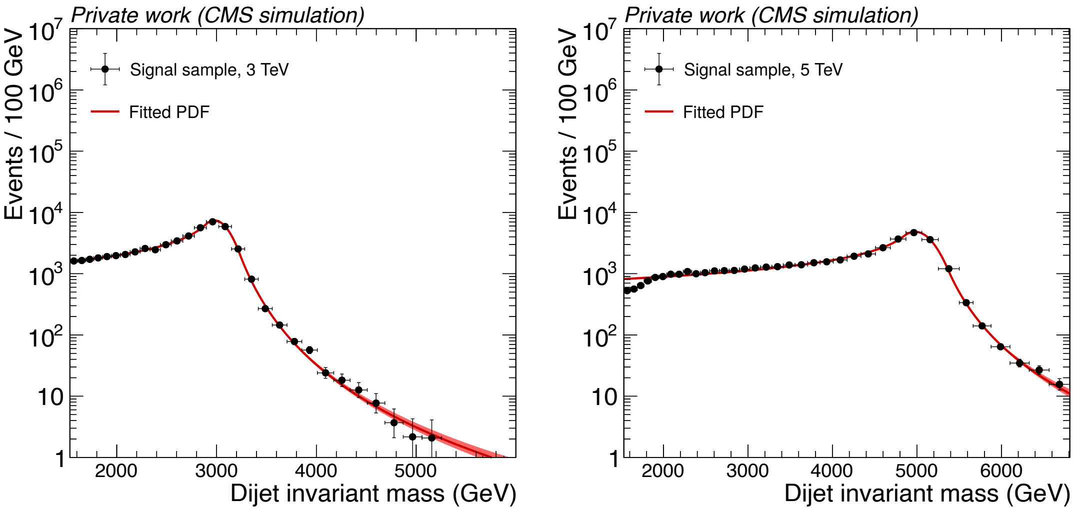

Figure B.3: Signal shape template obtained from fitting the $W_{KK} \to RW \to 3W$ signal MC simulation samples with a daughter mass of $m_R = 400\,\text{GeV}$ at resonance masses of 3 TeV (left) and 5 TeV (right), using a Double Crystal Ball function.

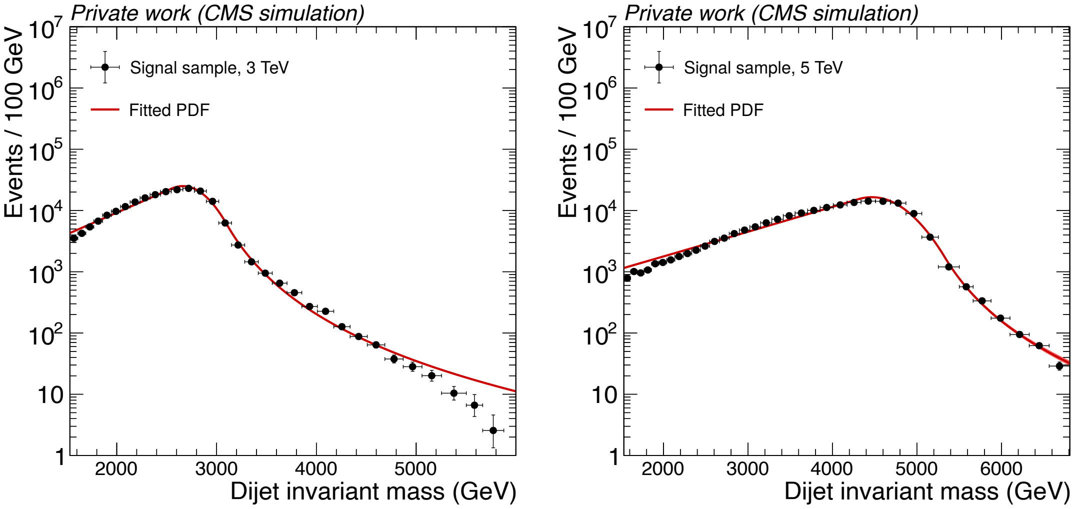

Figure B.4: Signal shape template obtained from fitting the $Y \to HH \to 4t$ signal MC simulation samples with a daughter mass of $m_H = 400\,\text{GeV}$ at resonance masses of 3 TeV (left) and 5 TeV (right), using a Double Crystal Ball function.

B.4 Effect of Systematic Uncertainties

While the proof-of-concept studies in Sec. 6 do not consider systematic uncertainties, they are an essential part of any realistic analysis. In the context of the CMS dijet search, Sec. 7, the effect of systematic uncertainties on CATHODE were evaluated for the first time, as described in Sec. 7.5. A particular focus was the impact on the exclusion limits, as there a particular signal model was targeted and thus all uncertainties in the modeling of the process are relevant, as discussed in Sec. 7.6.3.

In order to capture the effect of these modeling uncertainties, the 95% CL upper limits from Sec. 7.10.2 are shown with and without accounting for the uncertainty in the signal normalization when applying CATHODE in Fig. B.5 and Fig. B.6 for signals with resonance masses of 3 TeV and 5 TeV, respectively. The same comparison is shown with the CATHODE-b analysis strategy in Fig. B.7 and Fig. B.8.

As expected, the inclusion of systematic uncertainties leads to a degradation of the exclusion limits, as the signal strength is allowed to vary within the uncertainties. Across all different signals, it seems the average effect of the uncertainties is roughly a 20% decrease in sensitivity, for both CATHODE and CATHODE-b. In the most severely affected signals, the sensitivity loss reaches up to 50%, i.e., an upper limit that is twice as high as without systematic uncertainties. In a few cases, the inclusion of uncertainties appears to have no effect. A reason for this is the procedure of always quoting the maximum between the injected and excluded cross section. The uncertainty only affects the latter, thus if the former is sufficiently larger, it can hide the effect on the fitted signal strength.

It should be noted that the uncertainties discussed here only apply to the limit setting procedure, because a specific signal model has been used for training the classifier. The discovery potential, which is achieved through the significance calculation after a purely data-driven CATHODE training, is not affected by these uncertainties. In fact, this is in contrast to a supervised search, where the significance would usually be computed after selecting the signal candidates using a signal-trained classifier, which is subject to modeling uncertainties.

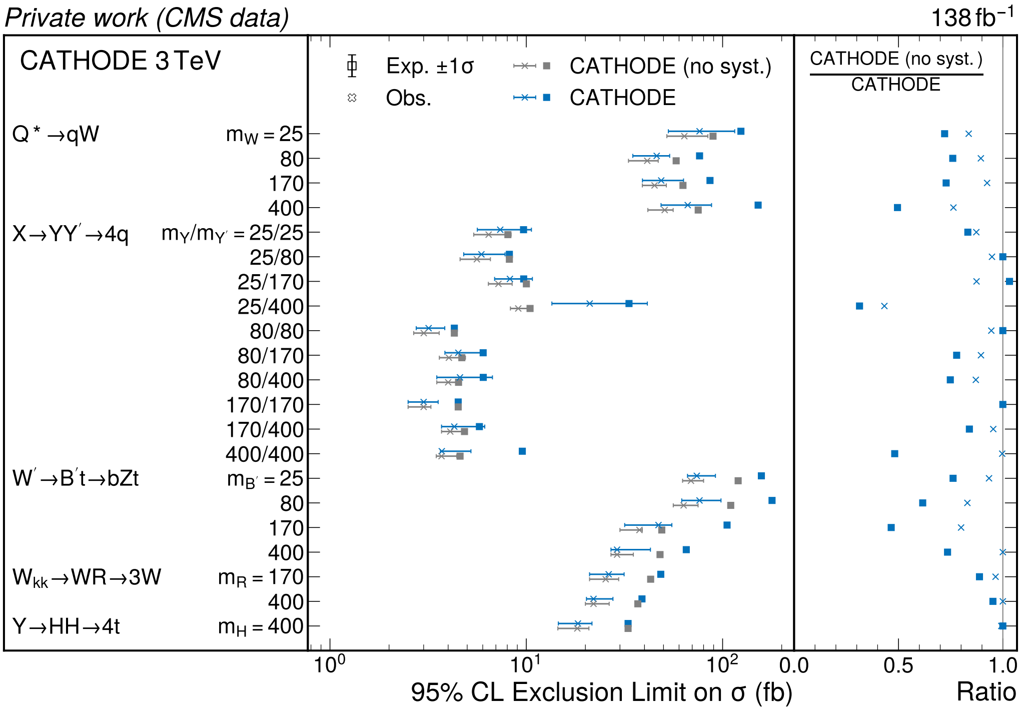

Figure B.5: Comparison of the CATHODE observed and expected upper 95% CL cross section limits before and after the inclusion of systematic uncertainties in the signal strength, for signals with a resonance mass of 3 TeV.

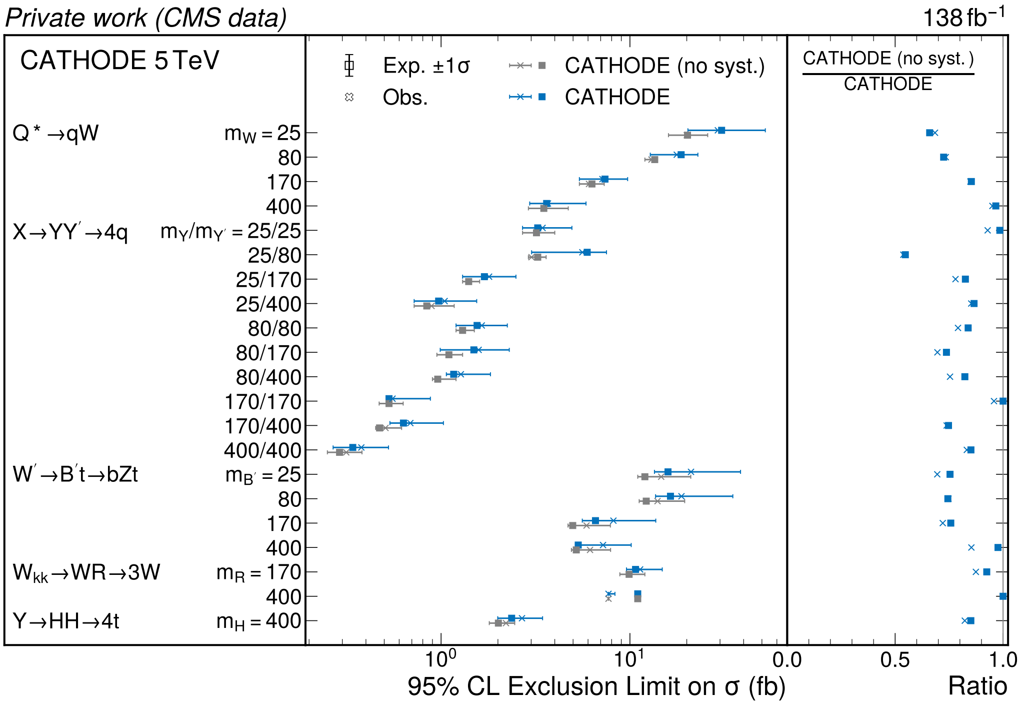

Figure B.6: Comparison of the CATHODE observed and expected upper 95% CL cross section limits before and after the inclusion of systematic uncertainties in the signal strength, for signals with a resonance mass of 5 TeV.

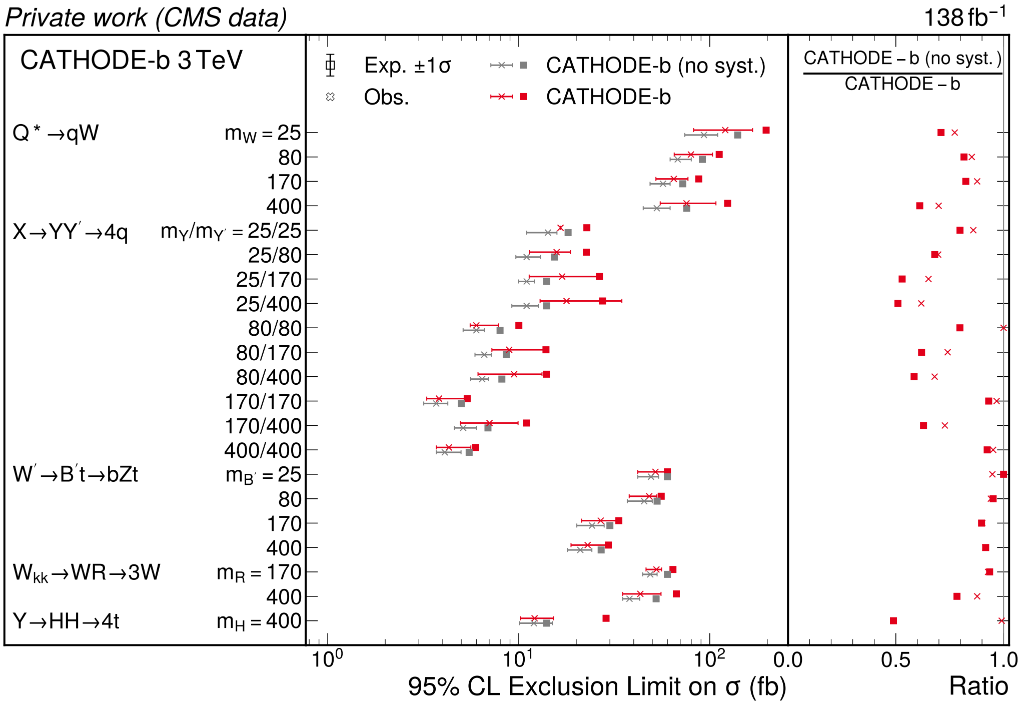

Figure B.7: Comparison of the CATHODE-b observed and expected upper 95% CL cross section limits before and after the inclusion of systematic uncertainties in the signal strength, for signals with a resonance mass of 3 TeV.

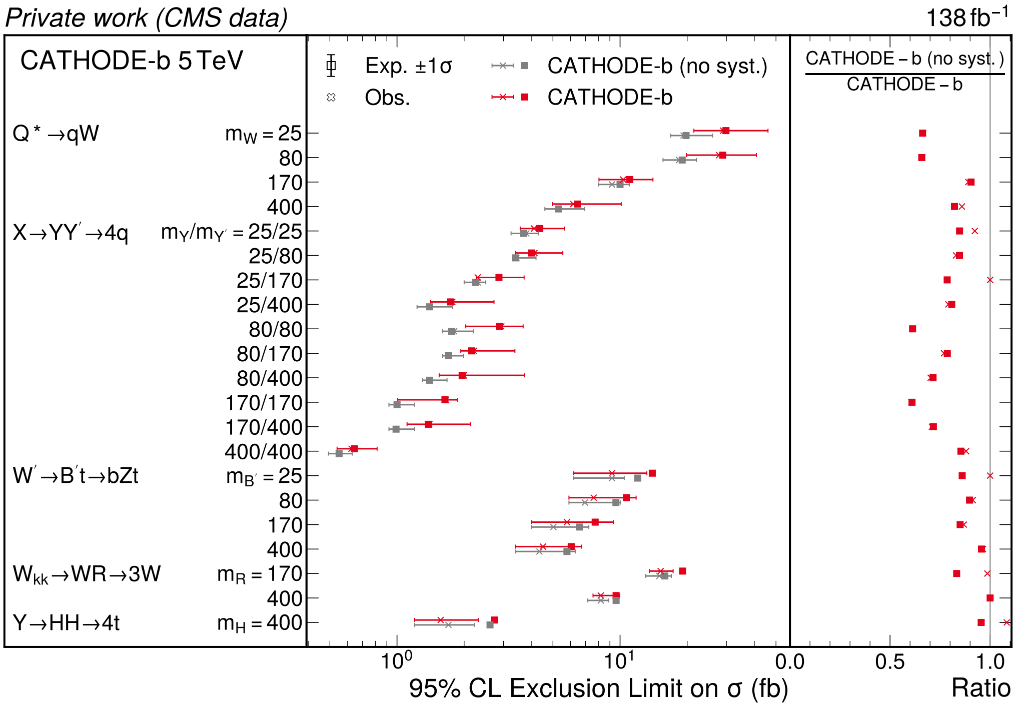

Figure B.8: Comparison of the CATHODE-b observed and expected upper 95% CL cross section limits before and after the inclusion of systematic uncertainties in the signal strength, for signals with a resonance mass of 5 TeV.

B.5 Background Sculpting

Section 6.3 discusses the issue of background sculpting for CATHODE when the auxiliary features are correlated with the resonant feature on the LHCO R&D dataset. A particular focus there lies on the *extrapolation* of classifier models trained on the signal region phase space to the sidebands, which differs with increasing correlations. Here, we extend this discussion to the significantly larger MC background simulation dataset from the CMS anomalous dijet search, with a focus on disentangling sculpting sources *within the signal region.*

A key advantage of the CMS MC background simulation dataset, introduced in Sec. 7.2, is the substantially larger number of events available for the classifier training, which thus approximates more closely the likelihood ratio estimator, as discussed in the Appendix A.3. This enables a more detailed study of the effects that arise from training on an imperfect CATHODE background template. We choose the lowest signal region from 1.65 to 2.017 TeV for this study, as it is the most populated bin and thus has the largest number of classifier training events and simultaneously the smallest training sample for the conditional normalizing flow that learns the background template. Moreover, we train the classifier with a learning rate of 10^{-4} instead of 10^{-3} and using early stopping with a patience of 10 epochs to ensure that the classifier reaches full convergence (instead of ending the training at 100 epochs). Otherwise, the CATHODE implementation follows the description in Sec. 7.3, including the double k-folding scheme. Since the dataset contains no signal, an ideal CATHODE background template would be indistinguishable from the "data" (i.e., the proxy data from a background-only simulation), and thus the classifier should not be able to consistently agree on anomalous phase space regions.

Figure B.9 shows the resulting background template samples (orange) in comparison with the "data" (light blue) in the resonant feature, $m = m_{jj}$, and the four auxiliary features, $x = \left(m_{j1}, \Delta m_j, \tau_{41,j1}, \tau_{41,j1}\right)$. As the histograms overlap for a large part of the phase space, their difference is best seen in the lower panel, which shows the per-bin ratio of the two normalized histograms for the auxiliary features in orange. (Since the agreement is expected to be good in m because of the samples following a direct KDE fit, the first lower panel shows instead the ratio between the "data" before and after the anomaly selection in dark blue.) In particular, the two mass-based features are subject to mismodeled peak-like structures[15]. The most anomalous percentile of "data" events according to the CATHODE classifier is shown in the upper panel in dark blue. In all auxiliary features, the resulting distribution differs visibly from the "data" before selection. In the mass-based features, the anomalous events seem to follow an oscillatory pattern with peaks that are well aligned with the mismodeled peaks from the background template. This is an indication that the classifier correctly identifies these differences between the two training samples and the likelihood ratio estimator is accordingly larger where the "data" phase space is more populated than the background template. Moreover, the selection seems to result in a slight change of m_{jj} distribution, from a smoothly exponentially falling shape to more long-ranged oscillations than one would expect from pure statistical fluctuations. This can also be seen in the lower panel where the ratio before and after the selection shows a trend where higher values of m_{jj} are more likely to be selected as anomalous. This type of sculpting is a direct consequence of correlations between x and m, as the visible change of shape in x translates to a change in the m distribution. In this case, the correlations are only minor and thus the resulting sculpting in m is likely small enough for this analysis. However, a more correlated choice of x might be substantially affected by the observed patterns in the auxiliary features.

To confirm the hypothesis that the most anomalous percentile is related to the mismod-

[15] Because of their positions similar to SM particle masses, an early suspicion has been that the peaks might arise due to the small contributions of more complex background processes, such as top quarks. However, a CATHODE training on a pure QCD MC simulation resulted in peaks at the same positions.

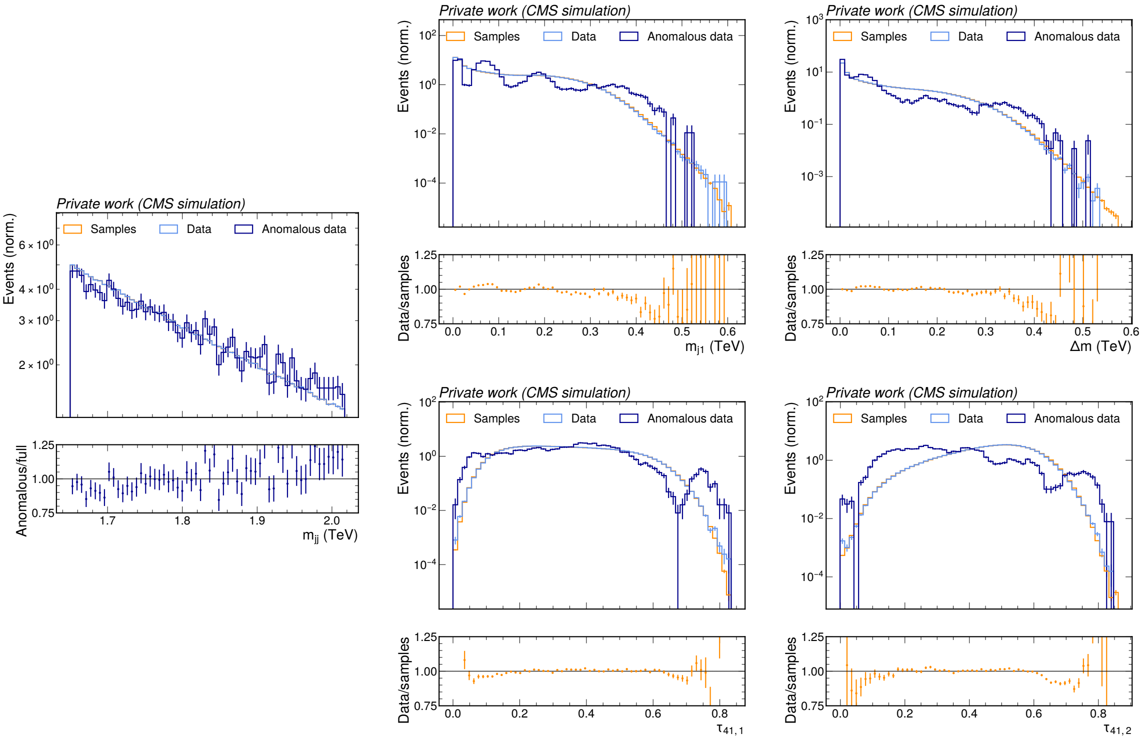

Figure B.9: Comparison of background template samples (orange), “data” (light blue) and the most anomalous 1% of events (dark blue) in all features after CATHODE was trained to full classifier convergence with a low learning rate on the MC background simulation dataset, using a signal region of 1.65 to 2.017 TeV. The bottom panel shows the ratio between the normalized distributions of the simulated “data” events before and after the anomaly selection for m_{jj} and between the “data” and the CATHODE background template for the other features.

eled regions of the background template, Fig. B.10 (top set of plots) resolves the modeling imperfections in terms of two-dimensional correlations. The pulls, defined as the per-bin difference between the two normalized histograms divided by their total statistical uncertainty, are shown as a heatmap such that red regions correspond to statistically significant data overdensities over the background template and blue regions relate to a significant overestimation of the background. Evidently, the background template is not perfect and could be improved by a better generative model. In comparison, the classifier predictions are shown in the same two-dimensional feature space in the lower set of plots. In this heatmap representation, more anomalous predictions are denoted in red and more anti-anomalous regions are blue. Comparing the two sets of plots, the same regions that are overdense in the data also result in a higher classifier prediction. This confirms that the classifier correctly learns the likelihood ratio estimator between the two training samples, which are subject to differences localized in phase space.

For a different angle in understanding the impact of the visible discrepancy between the background template and the actual "data" distribution, one can train an idealized anomaly detector, as in Sec. 6.3. Here, we split the MC background simulation dataset into two equal parts, reserving one as the new proxy data and the other half as the background template. The resulting distribution of most anomalous events according to this classifier is shown in Fig. B.11. While the background template now trivially aligns with the "data" in all features, there is still a visible change of pattern in the auxiliary features. Independently of any discrepancies between the training samples, the classifier learns *some* function of the auxiliary features and basing a selection on this function results in a change of input distribution. This change still translates to a slight slope in m_{jj}, as seen in the lower left-most panel.

To further investigate the impact of the background template mismodeling, which is the primary difference between CATHODE and the IAD, one can further assess what the classifier ensemble in each of the five k-folds selects as most anomalous. To this end, Fig. B.12 and Fig. B.13 show the most anomalous 1% of events according to CATHODE and the IAD, respectively, for each of the k-folds. The lower panels again show the ratio between histograms before and after selection for m_{jj} and between the "data" and the background template for the other features, but now split into separate k-folds. For CATHODE, the distributions of the most anomalous events are very similar among the five k-folds, each roughly aligning with the data-to-sample ratios shown in the lower panels, in case of the mass-based features. In the IAD case, on the other hand, one observes a substantially larger variance between the distributions of the most anomalous events in the different k-folds, seemingly because the learning process between the two identical sets is more random. This shows why the change of x shape and thus the potential effect on m is more severe when the background template is imperfect: the sculpted k-folds do not average out to something less sculpted, but instead the sculpting is amplified.

The inherent randomness of what an ensemble of classifiers learns when tasked to distinguish two identical sets, can be further studied by repeating the training. In order to perform this test on the same number of training samples as CATHODE, we construct a modified IAD implementation by mixing the "data" events and the background template of CATHODE, randomly shuffling, and splitting again into the same proportions as in the original training. This mixture of "data" and background template events is identical in both of the two sets. Figure B.14 shows the most anomalous 1% of events according to this IAD in four separate runs, each with a newly shuffled separation into proxy data and background template. Not only are they different from the unselected proxy data distribution (black), they also vary between the different trainings. However, their average (gray) does not resemble the unselected distribution either. Controlling the effect of the inherent randomness of a weakly supervised training on the selected feature distribution thus remains a non-trivial problem.

As an additional test, we perform the same study but with a constant partition of the IAD

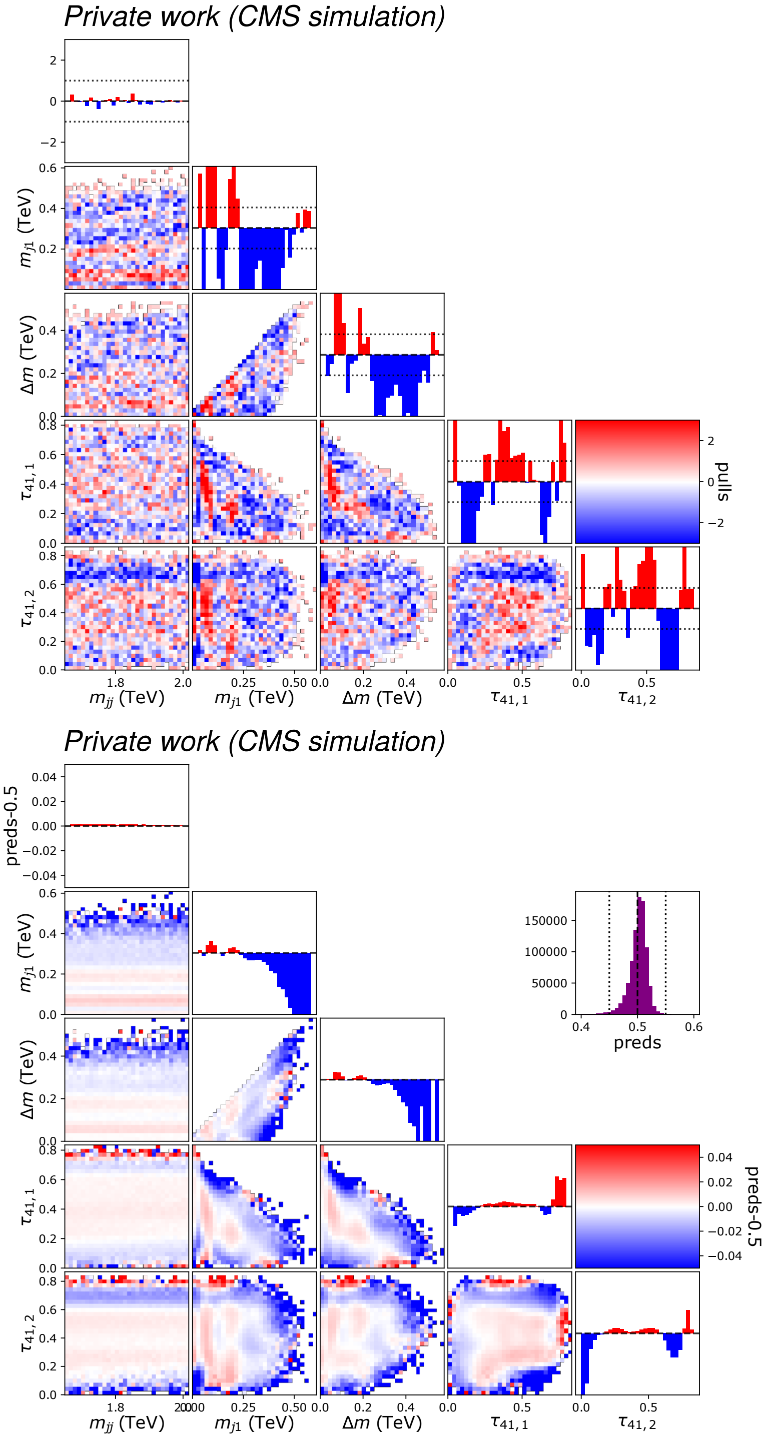

Figure B.10: The top set of plots shows the pulls between the MC background simulation data and the CATHODE background template, i.e., the per-bin difference between the two normalized histograms divided by their total statistical uncertainty. The plots along the diagonal are the one-dimensional marginal distributions of pulls in each individual feature, and the off-diagonal plots show the two-dimensional correlations. The bottom set of plots shows the classifier predictions as a function of the features, again in one dimension along the diagonal and in two dimensions off-diagonal. The central value of 0.5 is subtracted from the classifier predictions to show the deviations from the expected background.

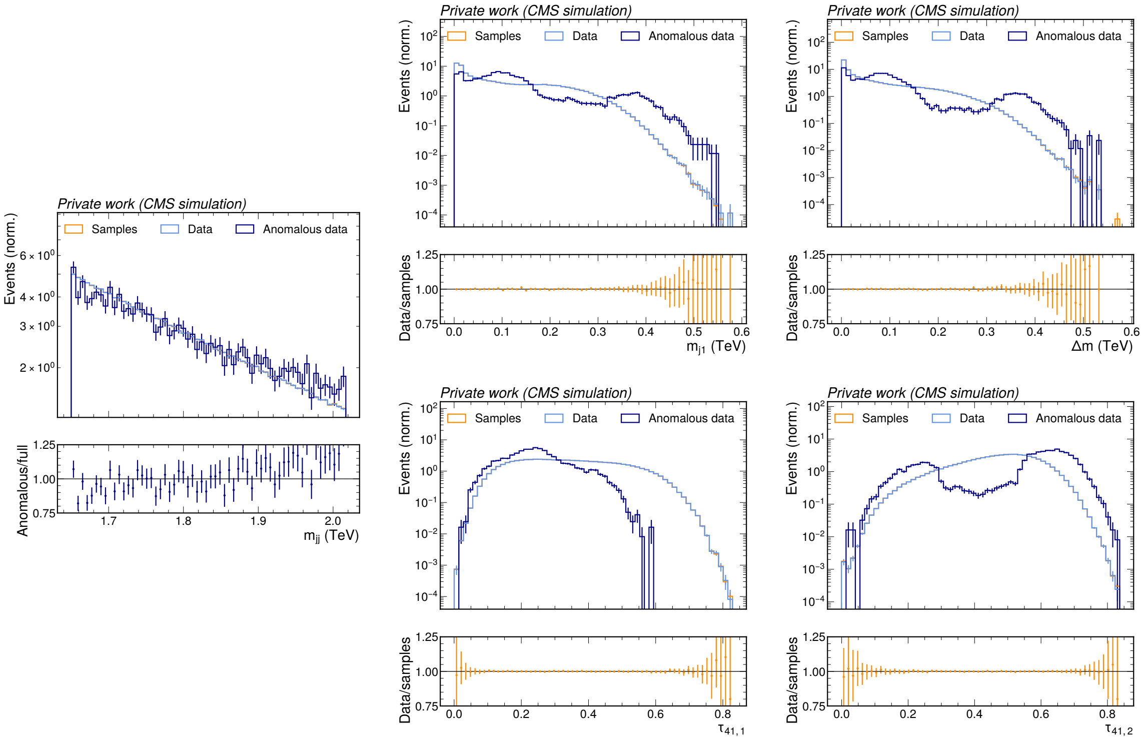

Figure B.11: Analogous to Fig. B.9, but for an idealized anomaly detector where the background template is constructed from half of the MC background simulation dataset and the other half is used as “data”.

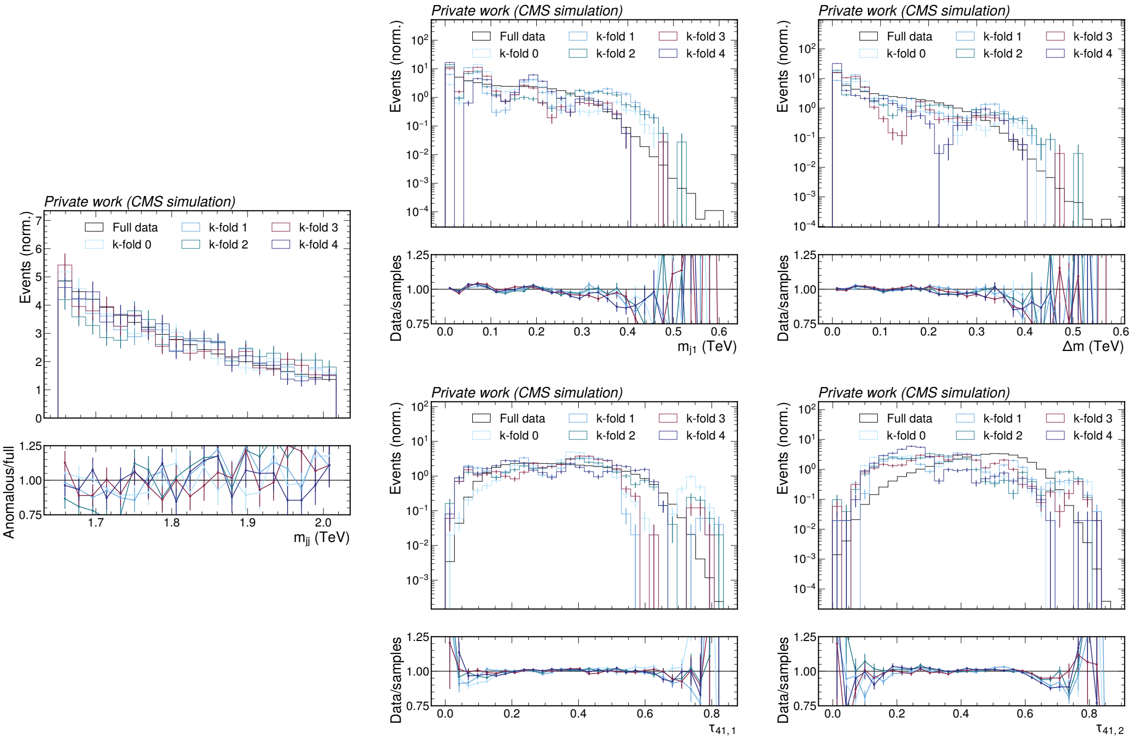

Figure B.12: Similar to Fig. B.9, but showing the most anomalous 1% of events according to CATHODE separately for the five k-folds, and in comparison to the full MC background simulation dataset (black). The bottom panel displays the ratio between the normalized distributions of the simulated “data” events before and after the anomaly selection in m_{jj}, and between “data“ and the CATHODE background template before selection in the other features, separately for each k-fold.

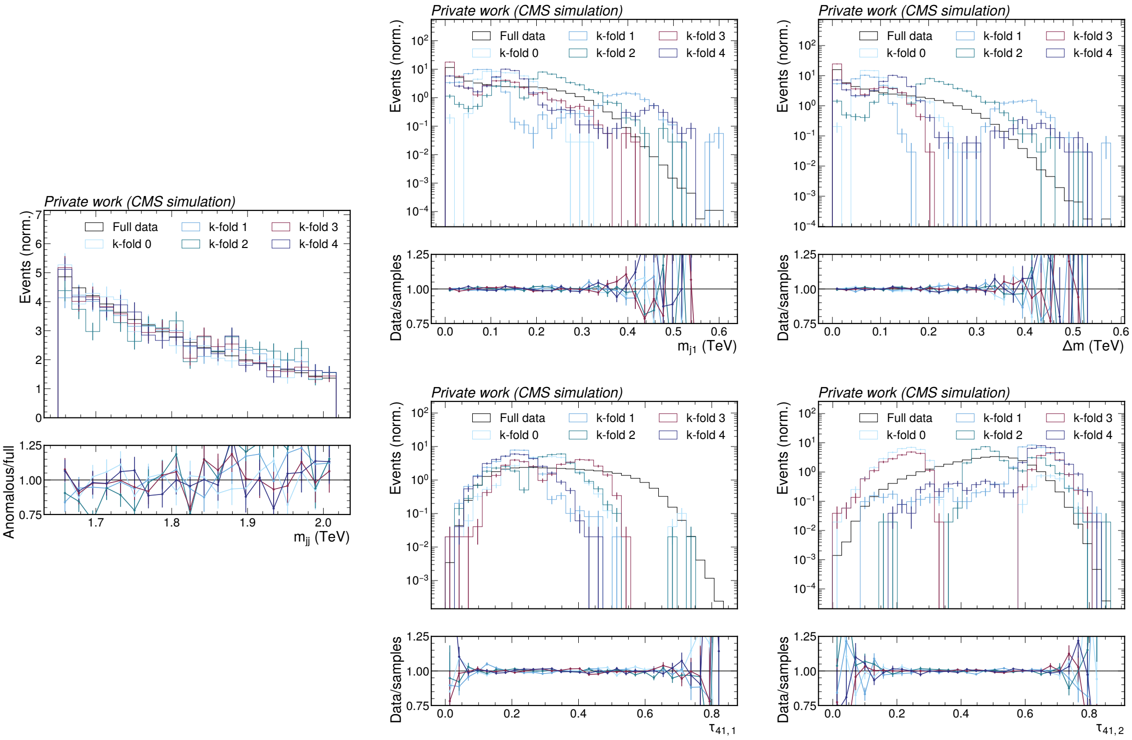

Figure B.13: Analogous to Fig. B.12, but for an idealized anomaly detector where the background template is constructed from half of the MC background simulation dataset and the other half is used as “data”.

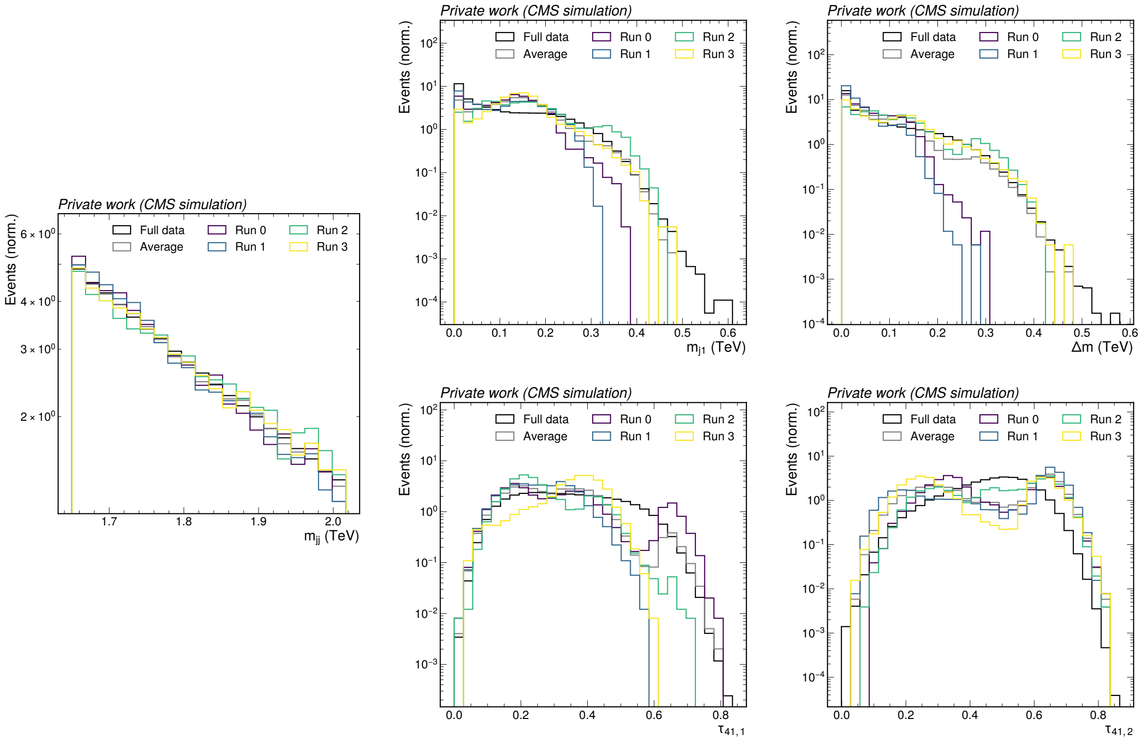

Figure B.14: Most anomalous 1% of events according to a sample-enhanced idealized anomaly detector, trained on a shuffled partition of the MC background simulation dataset and the CATHODE background template. Shown are four trainings, each with a new reshuffling of the dataset before separating into “data” and background template, as well as the full MC background simulation dataset (black).

dataset, i.e., with each classifier run only differing by the random initialization of the network weights but not by the training data. The resulting distributions of the most anomalous events are shown in Fig. B.15. In this case, they are strikingly similar among the four training runs. This is a strong indication that the classifier ensembles are overfitting on the statistical fluctuations of the training data, which is the primary source of the randomness.

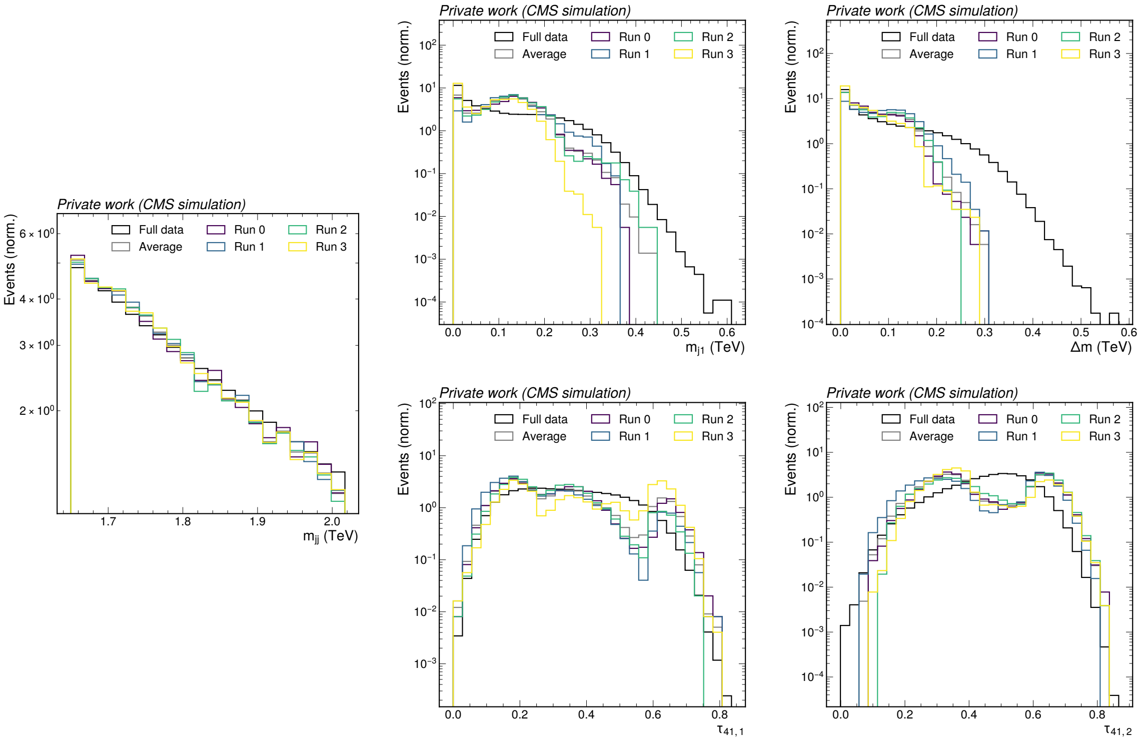

Figure B.15: Analogous to Fig. B.14, but each training is performed on the same shuffled partition of the MC background simulation dataset, thus only differing by a different random initialization of classifier network weights.

As a final test, we investigate the effect of not training the classifier at all. An untrained classifier ensemble is expected to be fully defined by the random initialization of the network weights and is thus neither affected by a mismodeled background template nor by the statistical fluctuations of the training data. The resulting distributions of the most anomalous events between four independently initialized classifier ensembles are shown in Fig. B.16. Still, the resulting feature distributions are different from the unselected "data" events, even more so than previously seen with CATHODE or the IAD. The classifier models remain a deterministic function of the auxiliary feature input, and thus a selection based on this function will always result in a change of the input distribution. The absence of training cannot avoid this, and instead we observe a visible sculpting effect in m_{jj}.

In the studies above, we have demonstrated that the background shape in the auxiliary features is subject to change whenever a selection is made on the classifier output as it is a deterministic function of its inputs. Any correlations with the resonant feature will translate to some degree of background sculpting in that distribution as well. This change of shape can be more or less pronounced, depending on which patterns the classifier finds to be discriminative between the data events and the background sample. The ensemble average is affected more systematically if a consistent mismodeling of certain phase space regions in the background

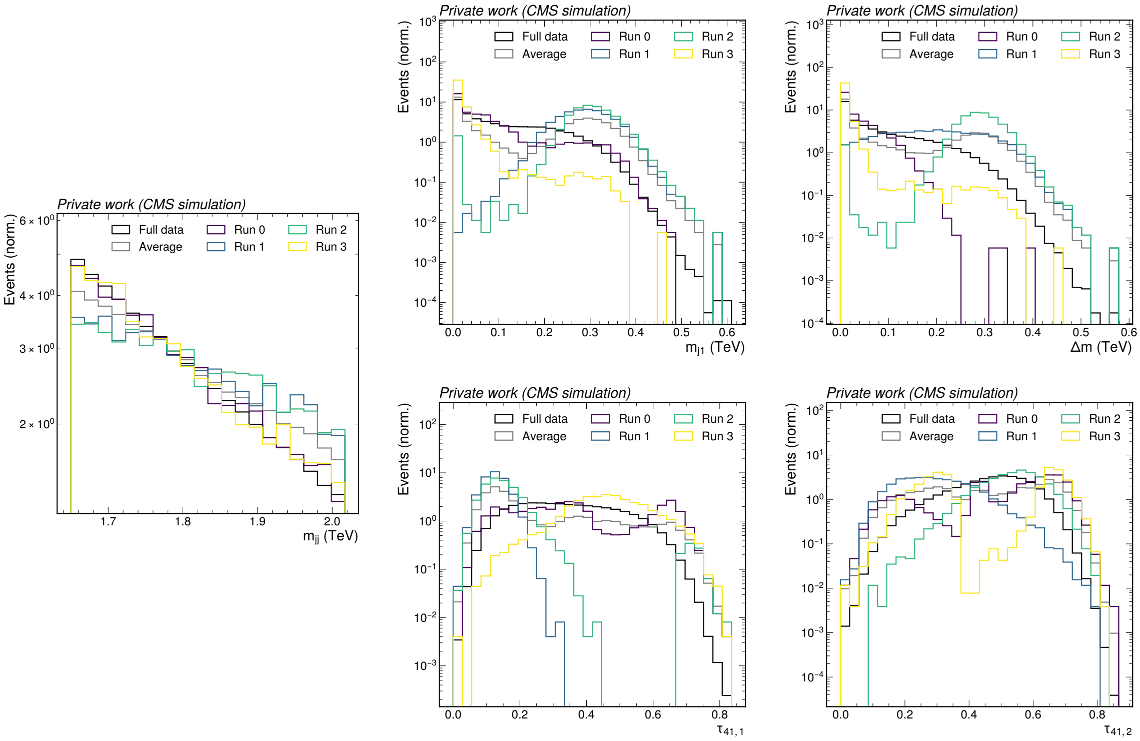

Figure B.16: Analogous to Fig. B.14, but the classifier has not been trained, and thus the runs are each fully defined by the random initialization of the network weights.

template is present. Even in the absence of such a systematic discrepancy, the classifier might overfit on the statistical fluctuations. While the background sculpting might not be dominant in the case of relatively small correlations, as seen in the case of the CMS anomalous dijet search, it might substantially challenge the background estimation strategy in the presence of more severe correlations. As demonstrated in Sec. 6.3, a safe mitigation strategy is to decorrelate the auxiliary and resonant features completely. An example for this type of approach is the LaCATHODE method, proposed in the context of this thesis.