Application of Machine Learning to Supersymmetric Models at Collider Experiments

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Lars Gerrit Bickendorf
aus
Köln

Bonn, Juli 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter / Betreuer:	Prof. Dr. Manuel Drees
Gutachter:	Prof. Dr. Herbert Dreiner
Tag der Promotion:	04.09.2024
Erscheinungsjahr:	2024

Acknowledgements

First, I would like to express my gratitude to my advisor, Manuel Drees, for his invaluable insights and kind, hands-off attitude, which inspired my pursuit of particle physics as presented in this thesis.

I extend my thanks to Herbi Dreiner for becoming the second referee and for his enjoyable, light-hearted approach to teaching physics. I also thank Prof. Desch and Prof. Klein for serving on my committee as the experimentalist and computer science member, respectively.

I am grateful to various administrative members of the institute, particularly Petra Weiss, Patricia Zündorf, Christa Börsch, Andreas Wißkirchen, and Dominik Köhler. Dominik, despite being a doctoral student, was essential in keeping the BCTP running smoothly, contributing to a productive period.

For the collaboration on resonant anomaly detection, I thank David Shih, Claudius Krause, and Gregor Kasieczka. Though the project took time to complete, it was both fun and fascinating.

I appreciate Bardia Najjari Farizhendi and Rahul Mehra for encouraging me to join the research group, which turned out to be a great decision. Being in the same group as Lina and meeting her in the corridor or at Christmas parties has always been a joy.

Moritz Wolter, Julian (et. al.) Günther, and Marc Vaisband kindly reviewed various chapters of my thesis, helping to improve the text.

I am thankful for my parents, Magdalene and Ralph, (and their cats) for their unwavering support in my studies of physics.

Lastly, I would like to thank Sophie for her steadfast support since the beginning of this journey. She stood by me, especially during my long-standing disagreement with an event generator, encouraging me to persevere.

Abstract

In this thesis, we study the application of modern machine learning methods to searches for supersymmetric models of physics beyond the Standard Model.

In recent years, resonant anomaly detection methods, such as CATHODE, have gained much attention. Using weakly supervised learning, these methods are built to be signal-model agnostic. The main advantage is that they are not only sensitive to a specific signal model, the analysis is tailored to, but cover a potentially much larger region of the parameter space. These methods are most often demonstrated on signal models that contain purely localized features.

However, the well-motivated R-parity conserving minimally supersymmetric Standard Model is often found at the tails of distributions of features such as p_T^{miss} or H_T. Pair produced gluinos with the decay chain $\widetilde{g} \to q\bar{q}\widetilde{\chi}_2^0(\widetilde{\chi}_2^0 \to X\widetilde{\chi}_1^0)$ with X either the Z or Higgs boson, light $\widetilde{\chi}_1^0$ and small mass splitting between $\widetilde{g}$ and $\widetilde{\chi}_2^0$ will be used to demonstrate CATHODEs sensitivity. We, for the first time, demonstrate that CATHODE is only slightly less sensitive than multiple dedicated searches while covering multiple signal models simultaneously.

This method can not uncover all signal models. For example the R-parity violating scalar top quark decay $\tilde{t} \to t\widetilde{\chi}_1^0(\widetilde{\chi}_1^0 \to qqq)$ with weak scale $\widetilde{\chi}_1^0$ and sub-TeV $\tilde{t}$ fails to produce features that CATHODE can reliably be applied to. For this signal model, we build a supervised classifier. We utilize recent innovations in computer vision, such as CoAtNet and MaxViT, that apply the self-attention mechanism to images. We represent calorimeter towers and tracks of jets as 2D images and show that the transformer-based classifiers outperform more classical convolutional neural networks in using the jet substructure to predict whether a given jet is neutralino-initiated or not. We show that replacing a CNN with MaxViT excludes up to 100 GeV of additional scalar top mass at 95% C.L. in a simple mock analysis for 100 GeV neutralinos.

Contents

CHAPTER 1

Introduction

Particle physics is the study of the most fundamental building blocks of our universe. Our current best understanding is captured in the Standard Model of particle physics (SM). With this model and the machinery of quantum field theory, it is possible to predict the interactions of particles at extremely small scales. At the Large Hadron Collider (LHC), ordinary matter in the form of hadrons is collided with extremely high energies. The collisions produce particle sprays that are analyzed with statistical methods to infer the parameters of the theoretical model, which agree with the predictions of the Standard Model to great precision [1]. Despite this, the Standard Model is incomplete. The overwhelming amount of matter present in the universe can not be described within the framework of the Standard Model, hinting at the existence of particle dark matter whose precise nature is yet to be understood [2]. Additionally, the energy scale of the Standard Model is surprisingly small compared to the scale at which gravity has to be included in the quantum theory. This is the scale at which we expect the SM to break down. Moreover, one of the particles, the Higgs boson, is sensitive to the presence of hypothetical more massive particles [3]. The fact, that we do not see the effects of this sensitivity is, from the theoretical frame, unnatural and somewhat unaesthetic. Motivated by this, one may introduce extensions that describe physics beyond the Standard Model. Supersymmetry is a popular and elegant way to alleviate the aforementioned problems. Supersymmetry postulates a symmetry that implies the existence of either bosonic partners of SM fermions or fermionic partners of SM bosons. These additional partner particles are searched for by collider experiments at the LHC, such as CMS and ATLAS [4]. This far, only null results were achieved, leading to ever higher lower bounds on the mass of the hypothetical supersymmetric particles.

To probe even higher masses, two paths can be taken, which are mutually non-exclusive. First, one may collide particles at higher energies E, which is related to the potentially created mass m via $E = mc^2$. Increasing the energy by a lot, however, is very expensive and needs a complete overhaul of the particle collider. Upgrading the Large Electron-Positron Collider (LEP) to the LHC was in no small part done because the Higgs boson was very likely to be found with the new energy – due

to theoretical considerations [5, 6], it could not have been much heavier than the observed 125 GeV. This guarantee does not exist with the supersymmetric partners. It could be that the signatures are already hiding in our data, or that the masses are out of reach for the coming decades.

The second strategy to find new physics at slightly higher masses is simply more data. Potentially we already produce some supersymmetric particles [1], although in quantities that are not significant enough yet. Increasing the amount of data takes time. Current analyses at CMS and ATLAS use the Run 2 dataset at 137 fb^{-1} of integrated luminosity [7], while the LHC has delivered a factor of two more to this day. To make the most of the available data, one needs to apply ever more sophisticated analysis strategies.

Vast amounts of data are precisely the context, in which machine learning excels. This will be the approach we will cover in this text. First, we will focus on resonant anomaly detection. This technique requires a new physics signal to produce a strongly localized (resonant) excess in some feature that is extracted from particle collisions. The aim will be to use the fact that this feature can be used to construct signal-enriched and signal-depleted sets of collision events that allow to find the signal in an overwhelming amount of Standard Model background events. The main advantage of this approach is, that it can be built very signal-model-agnostically. Therefore, it covers many signal-model hypotheses simultaneously. We will demonstrate that one of the recently proposed techniques, CATHODE [8], is more general than it has been previously shown. This allows uncovering supersymmetric signals without explicitly restricting the method to a specific signal model. We will also show that CATHODE is not the ultimate approach applicable to all signal models that produce resonances. To amplify the statistical significance of another supersymmetric model we demonstrate, that recent improvements in computer vision can be translated into physics analyses. Particle detectors can be understood as very large cameras, providing pictures of the moment after the particle collision. Image recognition can be used to find signal-like images which amplifies the discovery potential of new physics.

Because the topic of this thesis is physics and machine learning, we introduce both concepts in separate chapters. We start by giving an overview of the Standard Model of particle physics and its supersymmetric extension in Chapter 2. In Chapter 3 we review the machine learning techniques that will be used in the main text. The application of resonant anomaly detection to models that populate the tail of the distribution of some feature is shown in Chapter 4. Since not all signal models that produce resonant features can be found this way, we introduce modern computer vision techniques to high energy physics and show superior performance compared to more established techniques in Chapter 5. We present an overall discussion and conclusion in Chapter 6. In Appendix A we show additional information about the resonant anomaly detection methods. The second appendix, Chapter B, contains additional studies on the computer vision application of Chapter 5.

[1] This is also motivated due to naturalness arguments

CHAPTER 2

Theoretical Overview

In this chapter, we review the theoretical basics that are needed to follow the treatment in the following.

2.1 Standard Model of Particle Physics

The SM is the quantum field theory that describes all elementary particles and their interactions, the strong, weak, and electromagnetic forces. Gravity is explicitly left out since its effects are negligible at the energy scales we can probe at collider experiments.

2.1.1 Gauge Fields

We start with the gauge fields of the Standard Model. These stem from the local gauge symmetry of the three factors [9–13]

$$SU(3)_S \times SU(2)_L \times U(1)_Y \,, \tag{2.1}$$

which induce three gauge fields with spin $S = 1$, namely B_μ, W_μ, and G_μ. The various representations

Field	Lorentz-Rep.	Coupling constant	Y	SU(2)-Rep.	SU(3)-Rep.
B_μ	$(1/2, 1/2)$	g'	0	**1**	**1**
$\sigma_i W^i_\mu$	$(1/2, 1/2)$	g	0	**3**	**1**
$T_i G^i_\mu$	$(1/2, 1/2)$	g_s	0	**1**	**8**

Table 2.1: Gauge field content of the Standard Model. σ_i and T_i denote the generators of the adjoint representation of $SU(2)$ and $SU(3)$ respectively.

of the gauge groups these fields belong to are shown in table 2.1. The field strength tensor a gauge

field A with coupling constant $\tilde{g}$ is defined as

$$A_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + i\tilde{g}\left[A_\mu, A_\nu\right]. \tag{2.2}$$

Given this, the kinetic term of the Lagrangian is

$$\mathcal{L}_{\text{gauge kin.}} = -\frac{1}{2}\operatorname{Tr} G_{\mu\nu}G^{\mu\nu} - \frac{1}{2}\operatorname{Tr} W_{\mu\nu}W^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{2.3}$$

where the trace runs over the $SU(3)_S$ and $SU(2)_L$ indices. Since the B_μ field corresponds to the Abelian gauge group whose commutator vanishes in equation 2.2, it does not have a self-interaction term. Notably, G_μ and W_μ interact with themselves, leading to three and four-gauge boson interactions.

Since the Standard Model undergoes spontaneous symmetry breaking, only the Quantum Chromodynamics (QCD) gauge group factor $SU(3)_S$ will stay unbroken. The corresponding gauge field, the gluon, therefore stays massless.

2.1.2 Fermions

Field	Lorentz-Rep.	Y	SU(2)-Rep.	SU(3)-Rep.
$L_i = \begin{pmatrix}\nu_{eL}\\ e_L\end{pmatrix}, \begin{pmatrix}\nu_{\mu L}\\ \mu_L\end{pmatrix}, \begin{pmatrix}\nu_{\tau L}\\ \tau_L\end{pmatrix}$	$(1/2, 0)$	-1	**2**	**1**
$l_{Ri} = e_R, \mu_R, \tau_R$	$(0, 1/2)$	-2	**1**	**1**
$Q_i = \begin{pmatrix}u_{Li}\\ d_{Li}\end{pmatrix}$	$(1/2, 0)$	$1/3$	**2**	**3**
u_{Ri}	$(0, 1/2)$	$4/3$	**1**	**3**
d_{Ri}	$(0, 1/2)$	$-2/3$	**1**	**3**

Table 2.2: Fermionic content of the Standard Model as Weyl spinors [14]. The index i denotes generations from 1 to 3. The subscripts L and R are meant as implicit chiral projections if the fields are taken to be Dirac fermions. A potential right-handed neutrino is not part of the standard model.

With the introduction of local gauge transformations, one has to introduce the covariant derivative, which can be written as [14]

$$\mathcal{D}_\mu = \partial_\mu - ig'B_\mu\frac{Y}{2} - igW_\mu - ig_sG_\mu. \tag{2.4}$$

The gauge field terms are only present if the field it is applied to transforms non-trivially under the respective transformation. The kinetic terms for the fermionic field content of the Standard Model also

induce the fermion-gauge boson interactions via the covariant derivative. This term can be written as [14]

$$\mathcal{L}_{\text{fermion kin.}} = i\bar{\psi}\gamma^{\mu}\mathcal{D}_{\mu}\psi\,, \tag{2.5}$$

where the various fermions that replace ψ are given in table 2.2. Since both the Standard Model and the supersymmetric extension introduced in the following sections are inherently chiral theories, it would be advantageous to use the two-component [15, 16] notation of fermionic fields if one were to carry out computations by hand.

2.1.3 Higgs Mechanism

Field	Lorentz-Rep.	Y	SU(2)-Rep.	SU(3)-Rep.
$\Phi = \begin{pmatrix} \phi^{+} \\ \phi^{0} \end{pmatrix}$	$(0,0)$	1	**2**	**1**

Table 2.3: Additional Higgs sector of the Standard Model with one Higgs doublet that will give mass to both up and down type quarks simultaneously.

Because a Dirac mass term for the fermions

$$m\bar{\psi}\psi = m\left(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L\right) \tag{2.6}$$

would be gauge-variant for the $SU(2)_L$ factor, this is not a viable approach to give the observed masses to the fermions in the Standard Model. Additionally, the Z and W bosons are both massive, which is impossible in a pure gauge theory. To remedy this, the Brout-Englert-Higgs mechanism [17–22] (Higgs mechanism for short) adds a $SU(2)$-doublet Lorentz scalar field Φ with quantum number shown in table 2.3. This introduces an additional Lagrangian term [14]

$$\mathcal{L}_{\text{Higgs}} = \left(\mathcal{D}_{\mu}\Phi\right)^{\dagger}\left(\mathcal{D}^{\mu}\Phi\right) - \mu^2\Phi^{\dagger}\Phi - \lambda\left(\Phi^{\dagger}\Phi\right)^2 \tag{2.7}$$

and Yukawa coupling to the fermions of the form

$$\mathcal{L}_{\text{Yukawa}} = -y^{d}_{ij}\bar{Q}_i\Phi d_{Rj} - y^{u}_{ij}\bar{Q}_i\widetilde{\Phi}u_{Rj} - y^{l}_{ij}\bar{L}_i\Phi l_{Rj} + \text{h.c.}\,, \tag{2.8}$$

with $\widetilde{\Phi} = i\sigma_2\Phi^{*}$. Here, i and j are generation indices. For $\mu^2 < 0$ and $\lambda > 0$, one obtains the famous Mexican-hat-like Higgs potential with a minimum at $\Phi \neq 0$. When the Higgs condenses, it obtains a vacuum expectation value v that spontaneously breaks the $SU(2)_L$ group. After fixing the gauge, the Higgs field can be written as

$$\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h \end{pmatrix} \tag{2.9}$$

where h is a real scalar field. This leads to a new form of equation 2.7, eliminating μ^2 in favour of v [23]:

$$\begin{aligned}\mathcal{L}'_{\text{Higgs}} =& \frac{1}{2}\partial_\mu h \partial^\mu h + \frac{1}{8} g^2 (v+h)^2 \left(W^1_\mu W^{1\mu} + W^2_\mu W^{2\mu}\right) \\ &+ \frac{1}{8}(v+h)^2 \begin{pmatrix} W^3_\mu & B_\mu \end{pmatrix} \begin{pmatrix} g^2 & -g'g \\ -g'g & g'^2 \end{pmatrix} \begin{pmatrix} W^{\mu 3} \\ B^\mu \end{pmatrix} \\ &- \lambda h^2 \left(v + \frac{1}{2}h\right)^2 .\end{aligned} \tag{2.10}$$

The first line contains the kinetic term for the Higgs boson and, defining $W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \mp i W^2_\mu)$, a mass term for the electrically charged W boson with $M_W = \frac{gv}{2}$. The second line can be diagonalized by a change of basis with an orthogonal matrix parametrized by the Weinberg angle θ_W. This leads to the new basis states $Z_\mu = \cos\theta_W W^3_\mu - \sin\theta_W B_\mu$, known as the Z boson with mass $M_Z = \frac{v}{2}\sqrt{g^2 + g'^2} = M_W/\cos\theta_W$, and the still massless photon $A_\mu = \sin\theta_W W^3_\mu + \cos\theta_W B_\mu$. The Fermi constant inferred from the muons lifetime can be used to fix the value of v via [4]

$$v = \left(\sqrt{2} G_F\right)^{-1/2} \approx 246\,\text{GeV}\,, \tag{2.11}$$

while the Higgs mass of $m_h \approx 124\,\text{GeV}$ [24] fixes the remaining parameters μ^2 and λ of the Higgs sector. After all this, the gauge group of the Standard Model is spontaneously broken into

$$SU(3)_S \times SU(2)_L \times U(1)_Y \to SU(3)_S \times U(1)_{\text{EM}}. \tag{2.12}$$

This has consequences for the fermionic sector. For N generations, the Yukawa couplings y_d, y_u and y_l are $N \times N$ complex matrices. The structure of the Lagrangian allows a rotation of the left- and right-handed leptons by unitary $U(N)$ matrices V^l and U^l via $L_i \to V^l_{ij} L_j$ and $l_{Ri} \to U^l_{ij} l_{Rj}$. The rotations can be chosen to diagonalize the Yukawa couplings, or phrased in another way, the Yukawa matrix y_l can be chosen as a diagonal matrix without loss of generality. As a consequence, the gauge eigenstates coincide with the mass eigenstates. The charged leptons obtain a mass of

$$m_i = \frac{v y_i}{\sqrt{2}}\,, \tag{2.13}$$

where y_i are the diagonal elements of the Yukawa matrix. In other words, the coupling to the Higgs field determines the mass of the leptons. Because there exists no right-handed neutrino in the Standard Model, there is no Yukawa matrix that would generate a mass term for the neutrinos.

The story is not quite as straightforward with the quarks. To rotate the basis of a general complex Yukawa matrix such that it becomes diagonal with nonnegataive diagonal entries, one needs the freedom to rotate by two unitary matrices. In the lepton sector, we only have one Yukawa matrix

and as such, can choose to rotate the lepton doublets and singlets. Therefore the gauge and mass eigenbase become aligned and one does not have to choose one over the other. This is different in the quarks sector, as we have two Yukawa matrices. The doublets would have to produce rotations that diagonalize both matrices, which is not possible in general. Therefore, one has to choose between the gauge and mass eigenbase for the quark sector.

Neglecting $SU(2)$ invariance, as it is broken anyway, we are allowed to choose four unitary matrices that rotate the left, right, up-type and down-type quarks, such as [14]

$$\begin{aligned} u_{Li} &\to V^u_{ij} u_{Lj} & d_{Li} &\to V^d_{ij} d_{Lj} \\ u_{Ri} &\to U^u_{ij} u_{Rj} & d_{Ri} &\to U^d_{ij} d_{Rj}. \end{aligned} \tag{2.14}$$

With these, it is possible to diagonalize the Yukawa matrices by a simple change of basis and obtain masses for the quarks $M_{u/d} = V^\dagger_{u/d} y^{u/d} U_{u/d}$. For three generations, the mass eigenstates are called up, charm and top for the up-type quarks and down, strange and bottom for the down-type quarks. This comes at a price, as now the interactions with the W-bosons are no longer diagonal in the generation. It takes the form

$$-\frac{g}{\sqrt{2}} \bar{u}_i \gamma_\mu P_L \left(V_{\mathrm{CKM}}\right)_{i,j} d_j W^{+\mu} + h.c.\,, \tag{2.15}$$

where the Cabibbo-Kobayashi-Maskawa (CKM) matrix V_{CKM} is defined as $V^u V^{d\dagger}$. Since the 3×3 CKM-matrix is unitary, it can be parametrized by three angles and six phases, five of which can be set to zero by using $U(1)$ symmetries. The last phase that we cannot get rid of is responsible for the CP-violation of the Standard Model and only appears if there are more than two quark generations.

2.2 Problems with the Standard Model

The Standard Model has been probed with remarkable precision, reaching experimental and theoretical uncertainties of the order of one part per trillion for the magnetic moment of the electron [25]. Although these predictions are reassuring, the Standard Model is not perfect, and observations show that the Standard Model cannot be the end of the story. We now review some of the aspects that might hint at Physics beyond the Standard Model (BSM) with varying degrees of urgency. This list is by no means complete.

Gauge Coupling Unification

An intrinsic feature of quantum field theories is the running of coupling constants, i.e., any coupling g is not a constant, but changes as a function of the energy scale $g(Q)$ that is probed. For the Standard Model, the three fine structure constants $g^2/4\pi$ due to the three gauge factors get close to each other at $Q \sim O(10^{13} - 10^{17})$GeV [26]. The fact that any pair of couplings meet below the Planck scale hints at the possibility that this is not entirely accidental. Since the Standard Model already contains the

unification of electromagnetic and weak interactions into the $SU(2) \times U(1)$ gauge group, this is a compelling reason to search for grand unified theories that force the three couplings to meet at the corresponding scale exactly [27].

Missing Dark Matter Candidate

Dark matter is a term to collectively describe a form of matter that interacts exceptionally weakly –or not at all– with the electromagnetic field, thus appearing dark in astronomy. Observations of the rotation curves of galaxies [28] and galaxy clusters [29, 30] show a large discrepancy between the total amount of gravitating matter and visible matter.

Further evidence can be seen if one directly compares the distribution of luminous and gravitating matter after multiple galaxies pass each other. This was done for the bullet cluster 1E 0657–558 [31] by comparing an x-ray image of the merger by the Chandra observatory with a map of its gravitational potential derived from weak gravitational lensing effects. During the collision, the stars are too sparse to interact much between the mergers, while the hot intergalactic gas (which is seen in the x-ray observation) experiences friction. The gravitational map shows that the majority of the mass distribution is not affected by the collision and is consequently separated from the gas. This is one of the major hurdles in explaining gravitational observations by alternate theories of gravity instead of the dark matter paradigm, and also provides insights into the dark matter self-interaction cross sections [32]. Within the ΛCDM model, the matter content of the universe contains 84% dark matter [33]. The Standard Model lacks a convincing particle candidate to explain dark matter. The only hypothesis for dark matter within the Standard Model and ΛCDM are primordial black holes [2, 34] although the evidence is still inconclusive.

Hierarchy Problem

The Standard Model describes the fundamental particle interactions remarkably well. Neutrino masses aside, there is no clear sign of additional physics beyond the Standard Model besides what is known at the moment. However, one cannot expect the Standard Model to be the theory of everything since it cannot be reconciled with a quantum theory of gravity in this state. One expects quantum gravity to play a role at roughly the Planck scale $M_P = \sqrt{\hbar c^5 / G} \approx 1.22 \times 10^{19}$ GeV with Newtons constant G[35]. The mere fact that the weak scale and the Planck scale are separated by a factor of $M_P / M_W \sim 10^{18}$ appears highly unnatural in itself. But this is only the first half of the argument. This hierarchy of masses gets into even more unnatural territories if one assumes there are additional, more massive particles that communicate with the particle content of the Standard Model. As we have seen in the previous section, the masses of the electroweak gauge bosons are determined by the Higgs vacuum expectation value, which in turn depends on $-\mu^2$ which is set by the Higgs mass m_H. The

one fermion loop correction from the coupling $-\lambda_f \Phi \bar{f} f$ takes the form [3]

$$\Delta m_H^2 = -\frac{|\lambda_f|^2}{8\pi^2}\Lambda^2 + ..., \tag{2.16}$$

where Λ is a cutoff that regularizes the loop-integral, commonly interpreted as the scale at which new physics joins the dynamics. If there is no new physics at energies below the Planck scale, this implies $\Lambda \sim M_P$ which requires an enormous amount of fine-tuning of the ultraviolet parameters to keep the electroweak sector at its tiny mass. If there were new heavy particles beyond the Standard Model μ^2 becomes usually quadratically sensitive to those large masses. In either case, the smallness of μ^2 compared to the Planck or BSM scale is surprising and therefore unnatural if no symmetry protects the parameter from these large contributions. This effect is called the hierarchy problem. The Standard Model fermions do not have this behavior because the chiral symmetry protects the masses from quadratic corrections, such that only logarithmic corrections appear.

2.3 Supersymmetry

We start this section by observing an intriguing detail in the way the loop corrections to the Higgs mass from fermions and bosons differ. If the Higgs field couples to a massive BSM complex scalar field with coupling term $-\lambda_S \Phi^\dagger \Phi |S|^2$ the resulting correction takes the form

$$\Delta m_H^2 = +\frac{\lambda_S}{16\pi^2}\Lambda^2 + \dots . \tag{2.17}$$

This closely resembles the form of equation 2.16 up to a factor of -2 when $\lambda_S = |\lambda_f|^2$. When our theory contains two sets of S and the coupling constants are indeed related to that of the fermions, the quadratic sensitivity of the Higgs mass to higher mass scales cancels out and only the logarithmic terms hidden in the eclipses remain. A theory will not contain the relation between fermions and bosons by chance, so we need a symmetry to relate the two. This relation is called Supersymmetry (SUSY) [36–38]. The Coleman-Mandula-theorem [39] severely limits the possibility of transforming fermions into bosons (and vice versa) in an interacting theory. Supersymmetry avoids one of the core assumptions by having Lie superalgebras instead of normal Lie algebras, as used by the Standard Model. Furthermore, the Haag-Łopuszański-Sohnius theorem [40] implies that the superalgebra furnishing supersymmetry is the only realization of a non-trivial extension of the Poincaré algebra.

We will not delve deeper into the mathematical foundation of supersymmetry and rather just state how to formulate a $N = 1$ supersymmetric field theory (i.e. one that is generated by a single set of supersymmetry generators), that will contain the known Standard Model. Particles that are for the Standard Model the representations of the Poincaré group must now be contained in the irreducible representations of the supersymmetry algebra paired with their superpartners. These partners differ by half a unit of spin but sit in the same representations of the gauge groups. A chiral supermultiplet

consists of a fermion represented by a Weyl spinor ψ and its bosonic scalar superpartner ϕ.

To build a manifestly supersymmetric field theory one defines the superpotential W, which is a holomorphic function of the scalar fields which enters the dynamics via

$$\begin{aligned} W^{ij} &= \frac{\delta^2 W}{\delta\phi_i \delta\phi_j}, \\ W^{i} &= \frac{\delta W}{\delta\phi_i}. \end{aligned} \tag{2.18}$$

With these definitions, the chiral part of the Lagrangian is [1] [3]

$$\begin{aligned} \mathcal{L}_{\text{chiral}} =& - D^{\mu}\phi^{*} D_{\mu}\phi + i\psi^{\dagger}\bar{\sigma}^{\mu} D_{\mu}\psi \\ & - \frac{1}{2}\left(W^{ij}\psi_i\psi_j + W^{*ij}\psi_i^{\dagger}\psi_j^{\dagger}\right) - W^{i}W_i^{*}, \end{aligned} \tag{2.19}$$

where $\sigma^0 = \bar{\sigma}^0 = I$, the Pauli matrices $\sigma^i = -\bar{\sigma}^i$ and D_μ is the gauge covariant derivative. To add gauge interactions, we consider the vector supermultiplet, that contains a gauge field A^a_μ, its fermionic superpartner λ^a_μ which both transform under the adjoint representation of the group with generators T^a and gauge coupling g. The kinetic terms for the gauge supermultiplet are [3]

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu a} + i\lambda^{\dagger a}\bar{\sigma}^{\mu} D_{\mu}\lambda_a. \tag{2.20}$$

The complete Lagrangian[2] can be written as

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{\text{chiral}} + \mathcal{L}_{\text{gauge}} - \sqrt{2}g\left(\phi^{*}T^{a}\psi\right)\lambda_a \\ & - \sqrt{2}g\lambda^{\dagger a}\left(\psi^{\dagger}T^{a}\phi\right) - \frac{1}{2}g^2\left(\phi^{*}T^{a}\phi\right)^2. \end{aligned} \tag{2.21}$$

2.3.1 Minimally Supersymmetric Standard Model

We now turn our attention to the minimal field content and minimal superpotential to realize the Standard Model as part of a supersymmetric theory, the Minimal Supersymmetric Standard Model (MSSM)[3, 36–38, 41, 42]. The supermultiplets that incorporate the Standard Model fermions are shown in table 2.4 and those that incorporate the Standard Model gauge bosons are shown in table 2.5.

The Yukawa interactions between the Higgs sector and the Standard Model fermions that lead to the masses after the electroweak symmetry breaking need to be induced by the superpotential. Since it needs to be holomorphic, one Higgs field can only give masses to either the up-type or down-type quarks[3]. Therefore, the Higgs sector has to be extended by an additional Higgs doublet which is

[1] Here the auxiliary F-fields that need to be added to close the SUSY algebra off-shell are already integrated out. This term does play an important role in a mechanism that beaks supersymmetry.

[2] Again, the auxiliary D-fields are already integrated out but also play an important role in supersymmetry breaking

[3] Two Higgs doublets are also needed for gauge anomaly cancellation

Superfield	SM Field	Superpartner
$\widehat{L}_i$	$L_i = \begin{pmatrix} \nu_{Li} \\ e_{Li} \end{pmatrix}$	$\widetilde{L}_i = \begin{pmatrix} \widetilde{\nu}_{Li} \\ \widetilde{e}_{Li} \end{pmatrix}$
$\widehat{\bar{e}}_i$	$e^{\dagger}_{Ri}$	$\widetilde{e}^{*}_{Ri}$
$\widehat{Q}_i$	$Q_i = \begin{pmatrix} u_{Li} \\ d_{Li} \end{pmatrix}$	$\widetilde{Q}_i = \begin{pmatrix} \widetilde{u}_{Li} \\ \widetilde{d}_{Li} \end{pmatrix}$
$\widehat{\bar{u}}_i$	$u^{\dagger}_{Ri}$	$\widetilde{u}^{*}_{Ri}$
$\widehat{\bar{d}}_i$	$d^{\dagger}_{Ri}$	$\widetilde{d}^{*}_{Ri}$

Table 2.4: Supermultiplet content of the MSSM that contains the Standard Model fermions. The superpartners are spin-0 complex scalars. Gauge representations are the same as the ones shown in table 2.2. The names of the scalar superpartners of the fermions are usually prepended by an s, i.e., the scalar superpartner of a quark is a squark.

Superfield	SM Field	Superpartner
$\widehat{B}$	B_μ	$\widetilde{B}$
$\widehat{W}^i$	W^i_μ	$\widetilde{W}^i$
$\widehat{G}^i$	G^i_μ	$\widetilde{g}^i$

Table 2.5: Supermultiplet content of the MSSM that contains the Standard Model gauge vector fields. The superpartners are spin-1/2 fermions. Gauge representations are the same as the ones shown in table 2.1. The fermionic superpartners of bosonic SM fields are named with an -ino appended to the name, i.e., the fermionic partner of the W-boson is the wino.

shown in table 2.6. The superpotential is given by [3]

$$W_{\text{MSSM}} = -(y_d)_{ij}\widehat{Q}_i\widehat{H}_d\widehat{\bar{d}}_j + (y_u)_{ij}\widehat{Q}_i\widehat{H}_u\widehat{\bar{u}}_j - (y_l)_{ij}\widehat{L}_i\widehat{H}_d\widehat{\bar{e}}_j + \mu\widehat{H}_u\widehat{H}_d \tag{2.22}$$

which reproduces the Yukawa couplings of the Standard Model from equation 2.8 due to the W^{ij} term in the chiral Lagrangian in equation 2.19.

2.3.2 Soft SUSY Breaking

A universe with unbroken supersymmetry is very different from the world we observe, since there is no sign of the superpartners with the same gauge couplings and mass. Consequently, supersymmetry (if realized in nature) has to be broken, leading to the apparent absence at the energies currently being probed with high-energy colliders. Since there exist many mechanisms that lead to

Superfield	SM Field	Superpartner	Y	SU(2)-Rep.	SU(3)-Rep.
$\widehat{H}_u$	$\begin{pmatrix} H_u^+ \\ H_u^0 \end{pmatrix}$	$\begin{pmatrix} \widetilde{H}_u^+ \\ \widetilde{H}_u^0 \end{pmatrix}$	$\frac{1}{2}$	**2**	**1**
$\widehat{H}_d$	$\begin{pmatrix} H_d^0 \\ H_d^- \end{pmatrix}$	$\begin{pmatrix} \widetilde{H}_d^0 \\ \widetilde{H}_d^- \end{pmatrix}$	$-\frac{1}{2}$	**2**	**1**

Table 2.6: Superfield content of the MSSM that is necessary for breaking the electroweak symmetry by the Higgs mechanism. Notice, that for the SM we only need one Higgs doublet, of which only one degree of freedom propagates as the Higgs boson. The fermionic fields are called Higgsinos.

spontaneous supersymmetry breaking [43–48], we parametrize our ignorance of the actual mechanism by introducing explicit supersymmetry breaking terms to the Lagrangian. These couplings have positive mass dimension [3] to retain the cancellation of quadratic divergencies to the scalar particles, a desirable trait for the scalar Higgs. For the MSSM, these terms are given by an additional Lagrangian

$$\begin{aligned}
\mathcal{L}_{\text{soft}} = &-\frac{1}{2}\left(M_1\widetilde{B}\widetilde{B} + M_2\widetilde{W}\widetilde{W} + M_3\widetilde{g}\widetilde{g} + \text{h.c.}\right) \\
&- \widetilde{Q}_i^*\left(m_{\widetilde{Q}}^2\right)_{ij}\widetilde{Q}_j - \widetilde{L}_i^*\left(m_{\widetilde{L}}^2\right)_{ij}\widetilde{L}_j \\
&- \widetilde{u}_{Ri}^*\left(m_{\widetilde{u}}^2\right)_{ij}\widetilde{u}_{Rj} - \widetilde{d}_{Ri}^*\left(m_{\widetilde{d}}^2\right)_{ij}\widetilde{d}_{Rj} - \widetilde{e}_{Ri}^*\left(m_{\widetilde{e}}^2\right)_{ij}\widetilde{e}_{Rj} \\
&- m_{H_d}^2 H_d^* H_d - m_{H_u}^2 H_u^* H_u - \left(bH_u H_d + \text{h.c.}\right) \\
&- \left(\left(T_d\right)_{ij}\widetilde{Q}_i H_d \widetilde{d}_{Rj}^* + \left(T_u\right)_{ij}\widetilde{Q}_i H_u \widetilde{u}_{Rj}^* + \left(T_e\right)_{ij}\widetilde{L}_i H_d \widetilde{e}_{Rj}^* + \text{h.c.}\right).
\end{aligned} \tag{2.23}$$

The fourth line is responsible for giving the Higgs sector a nontrivial vacuum, which is responsible for spontaneously breaking the electroweak symmetry. All in all, the terms shown in equation 2.23 and the supersymmetric Lagrangian introduce 124 real physical parameters that contain the 18 real parameters[4] of the Standard Model [52, 53]. In phenomenological studies, one often uses a restricted set of parameters to avoid an investigation involving 105 parameters.

Interestingly, the particle content of the MSSM can lead to gauge coupling unification when the sparticles are not too heavy [54].

2.3.3 Mass Mixing in the MSSM

Given viable soft-breaking parameters, the Higgs potential leads to electroweak symmetry breaking. Instead of a single vacuum expectation value, we now obtain one vacuum expectation value for each Higgs doublet $\langle H_u \rangle = v_u$ and $\langle H_d \rangle = v_d$. After the Nambu-Goldstone bosons have been absorbed into

[4] This includes the QCD angle θ_{QCD} whose value agrees with 0 [49–51].

the W and Z bosons, there are five real degrees of freedom left. The corresponding mass eigenstates (after diagonalizing the mass matrix through two-dimensional rotations) are two neutral CP-even scalars h and H, one of which is the observed 125 GeV Higgs boson. In addition, there is a CP-odd neutral scalar A and a charged scalar $H^{\pm}$.

Furthermore, both neutral higgsinos, the bino, and wino now mix into the mass eigenstates called neutralinos $\widetilde{\chi}_i^0$ with $i = 1, 2, 3, 4$. By definition, the neutralinos are sorted in ascending order of the mass. The mass matrix in the gauge eigenstate is [3]

$$\mathcal{L}_{\text{Neu.}} = -\frac{1}{2}\begin{pmatrix}\widetilde{B} & \widetilde{W} & \widetilde{B}_d^0 & \widetilde{H}_u^0\end{pmatrix}\begin{pmatrix} M_1 & 0 & \frac{-g'v_d}{\sqrt{2}} & \frac{g'v_u}{\sqrt{2}} \\ 0 & M_2 & \frac{gv_d}{\sqrt{2}} & \frac{-gv_u}{\sqrt{2}} \\ \frac{-g'v_d}{\sqrt{2}} & \frac{gv_d}{\sqrt{2}} & 0 & -\mu \\ \frac{g'v_u}{\sqrt{2}} & \frac{-gv_u}{\sqrt{2}} & -\mu & 0 \end{pmatrix}\begin{pmatrix}\widetilde{B} \\ \widetilde{W} \\ \widetilde{B}_d^0 \\ \widetilde{H}_u^0\end{pmatrix} + c.c., \tag{2.24}$$

which is diagonalized into the $\widetilde{\chi}_i^0$ mass eigenstates by a unitary matrix. Therefore, the choice of M_1, M_2 and μ determines whether the lightest supersymmetric particle is bino-, wino- or Higgsino-like.

The other mixing of interest here is the top squark mixing. In the gauge basis, this is [55]

$$\mathcal{L} = -\begin{pmatrix}\widetilde{t}_L^* & \widetilde{t}_R^*\end{pmatrix}\begin{pmatrix} m_{\widetilde{t}_L}^2 & \Delta^\dagger \\ \Delta & m_{\widetilde{t}_R}^2 \end{pmatrix}\begin{pmatrix}\widetilde{t}_L \\ \widetilde{t}_R\end{pmatrix}, \tag{2.25}$$

with

$$\begin{aligned} \tan\beta &= v_u/v_d \\ m_{\widetilde{t}_L}^2 &= m_{\widetilde{Q}_3}^2 + m_t^2 + \left(\frac{1}{2} - \frac{2}{3}\sin^2\theta_W\right) m_Z^2 \cos 2\beta \\ m_{\widetilde{t}_R}^2 &= m_{\widetilde{u}_3}^2 + m_t^2 + \frac{2}{3}\sin^2\theta_W m_Z^2 \cos 2\beta \\ \Delta &= vT_{u3}\sin\beta - m_t\mu^*\cot\beta. \end{aligned} \tag{2.26}$$

The off-diagonal entry Δ is responsible for the mixing, raising one eigenvalue (i.e., physical mass) and lowering the other. Therefore, most models predict the mixing to be most pronounced in the heaviest third-generation sfermions, leading to rather light scalar top or scalar bottom masses. The mixing in the other two generations is regarded as approximately absent.

2.3.4 R-Parity

The MSSM we have introduced so far obeys an interesting symmetry. Consider assigning a baryon number B to all particles, such that $\widehat{Q}_i$ carries baryon number $B = 1/3$, $\widehat{\bar{u}}_{Ri}$ and $\widehat{\bar{d}}_{Ri}$ carry $B = -1/3$ and all other fields $B = 0$. Additionally, one may introduce the lepton number L such that $\widehat{L}_i$ carries

lepton number $L = 1$, $\widehat{\bar{e}}_{Ri}$ carries $L = -1$ and all other fields $L = 0$. In the Standard Model, both numbers are conserved independently at tree level, simply because renormalizable violating terms are not permitted without adding additional fields. With these two quantities, one may introduce R-parity, which is a Z_2 symmetry defined as

$$P_R = (-1)^{3(B-L)+2s} \ , \tag{2.27}$$

where s is the spin. Under this symmetry, all Standard Model fields carry $P_R = +1$ while the superpartners, differing only by $\Delta s = 1/2$, carry $P_R = -1$. All terms in the MSSM Lagrangian and the soft supersymmetry breaking Lagrangian are symmetric under R-parity. If we promote R-parity to a symmetry our theory should respect rather than an accidental symmetry, there are no additional allowed interactions besides the ones induced by the superpotential given in equation 2.22. As a consequence, all interactions must involve an even number of superpartners. As a corollary, the lightest supersymmetric particle (LSP) is necessarily stable and might be a candidate for dark matter that the Standard Model misses.

2.3.5 R-Parity Violating MSSM

In the development of the Standard Model, it has been fruitful to write down all allowed interactions and introduce symmetries to avoid others. We should do the same with the MSSM. If R-parity is not conserved, there are additional allowed terms in the superpotential. The part that violates lepton number by one unit is

$$W_{\Delta L=1} = \frac{1}{2}\lambda^{ijk}\widehat{L}_i\widehat{L}_j\widehat{\bar{e}}_k + \lambda^{'ijk}\widehat{L}_i\widehat{Q}_j\widehat{\bar{d}}_k + \epsilon^i\widehat{L}_i\widehat{H}_u \tag{2.28}$$

and an additional term that violates baryon number by one unit is

$$W_{\Delta B=1} = \frac{1}{2}\lambda^{''ijk}\widehat{\bar{u}}_i\widehat{\bar{d}}_j\widehat{\bar{d}}_k. \tag{2.29}$$

where the indices i, j, k are generation indices. Here anti-symmetrization over color (i.e., contraction with the totally antisymmetric tensor in color space) is implied; hence, the coupling λ'' has to be antisymmetric in the last two indices. Therefore, there are in general nine independent coupling constants λ''_{ijk}.

With these terms enabled, there are numerous tightly constrained new processes. The most famous of these allows one up- and one down-quark inside the proton to fuse into a (virtual) squark via $\lambda^{''11i}$, which decays via $\lambda^{'11i}$ into a positron and $\bar{u}$. Therefore, the proton decays into a positron and π^0. The lifetime of the proton is measured to be larger than $O(10^{32})$ years[56], which puts severe constraints on the couplings.

The absence of R-parity conservation also has consequences on the experimental signatures. With R-parity conservation, the end of any sparticle decay chain is the absolutely stable Lightest

Supersymmetric Particle (LSP) that is favored to be neutral, which leaves the detector undetected. Therefore, the universal signature is large missing momentum. Once the LSP decays, for example via the operator in equation 2.29, this completely changes as the previous missing momentum now gets deposited into the large hadronic activity in hadron collider experiments. Many exclusion bounds on sparticle masses are therefore considerably weaker.

CHAPTER 3

Machine Learning

In this section, we cover the basis of the machine-learning algorithms that were used in this text. We focus on an intuitive understanding from a physicist's point of view rather than formal results. Since the tasks covered in this thesis are classification problems, we will focus on them. The aim will be to take input features $\mathbf{x} \in \mathcal{X}$ and correctly map them onto the target variable $y \in \mathcal{Y}$ via some function $\mathbf{h} : \mathcal{X} \to \mathcal{Y}$ that needs to be found. In the context of supervised learning, we have access to a training set made of n pairs $(\mathbf{x}_i, y_i)$ that can be used to construct $\mathbf{h}$ where the index i enumerates the pairs. One way to construct $\mathbf{h}$ is analogous to function fitting. One chooses a sufficiently complex function with parameters θ for which the problem is solved for some value of these parameters, even though these are not known beforehand. To calculate how well the function predicts the right output on an example $\mathbf{x}_i$, given some θ one introduces a cost or loss function such as

$$l(\mathbf{x}_i, y_i, \theta) = \left(\mathbf{h}(\mathbf{x}_i; \theta) - y_i\right)^2 . \tag{3.1}$$

To find the optimal θ, the expectation value of equation 3.1 is minimized numerically over all training examples.

3.1 Optimizers

One way to find the optimal θ is by gradient descent. Computing the full gradient is prohibitively expensive for large training sets and complex models. To remedy this, one uses mini-batch-based stochastic gradient descent, which takes a randomized subset of n_b training examples called mini-batch for each step. The batch size n_b is a hyperparameter that has to be chosen depending on the computational resources at hand. After an initial guess, θ will be updated from timestep τ to timestep $\tau + 1$ with the gradient

$$G_j(\mathbf{x}_i) = \frac{\partial}{\partial \theta_j} l(\mathbf{x}_i, y_i, \theta) , \tag{3.2}$$

where θ_j is the j'th component of θ and the rule

$$\theta_j(\tau+1) = \theta_j(\tau) - \frac{\eta}{n_b}\sum_b^{n_b} G_j(\mathbf{x}_b)\,, \tag{3.3}$$

where the index b enumerates all examples $\mathbf{x}$ in a given mini-batch. η is called the learning rate, which controls the stepsize along the gradient. The learning rate is a hyperparameter and as such will not be optimized with gradient descent during the learning stage, but has to be chosen beforehand. Especially for small batches, the gradient tends to have strong fluctuations, which renders the optimization procedure unstable. This can be avoided by using a technique called Momentum [57] that changes the update rule to

$$V_j(\tau) = \alpha V_j(\tau-1) - \frac{\eta}{n_b}\sum_b^{n_b} G_j(\mathbf{x}_b) \tag{3.4}$$

$$\theta_j(\tau+1) = \theta_j(\tau) + V_j(\tau). \tag{3.5}$$

Analogously to the mass in physical momentum, the hyperparameter $\alpha \in [0,1)$ controls how much inertia the gradient step carries, i.e. how much it resists the change due to the gradient at this step. The optimizer that is used in this text, Adam [58], is a slight modification of this.[1] For this, two new hyperparameters $\beta_1, \beta_2 \in [0,1)$ with typical values of 0.9 and 0.999, respectively, are introduced. The update rule is as follows:

$$g_j(\tau) = \frac{1}{n_b}\sum_b^{n_b} G_j(\mathbf{x}_b) \tag{3.6}$$

$$M_j(\tau) = \beta_1 M_j(\tau-1) + (1-\beta_1) g_j(\tau) \tag{3.7}$$

$$V_j(\tau) = \beta_2 V_j(\tau-1) + (1-\beta_2) g_j(\tau)^2 \tag{3.8}$$

$$\widehat{M}_j(\tau) = \frac{1}{1-\beta_1^\tau} M_j(\tau) \tag{3.9}$$

$$\widehat{V}_j(\tau) = \frac{1}{1-\beta_2^\tau} V_j(\tau) \tag{3.10}$$

$$\theta_j(\tau+1) = \theta_j(\tau) - \eta \frac{\widehat{M}_j(\tau)}{\sqrt{\widehat{V}_j(\tau)} + \epsilon}\,, \tag{3.11}$$

where $g_j(\tau)^2$ and $\sqrt{\widehat{V}_j(\tau)}$ are evaluated elementwise, ϵ regularizes the expression for small V and is typically set to 10^{-8} and $M_j(0)$ and $V_j(0)$ are set to zero. The hatted definitions are added to correct the bias in the early steps introduced by the zero initialization. Similar to Momentum, M_j captures the

[1] Technically it can be understood as a combination of another optimization algorithm, RMSProp[59] and Momentum [57].

mean[2] of the gradient, suppressing fluctuations. In contrast to this, V_j captures the variance, which is a measure of how much the gradient fluctuates. Combining both in the update step for θ_j we can think of Adam as gradient descent with momentum, with an adaptive learning rate $\sim 1/\sqrt{V_j}$, which takes large steps when the fluctuations of the gradient are small and vice versa. This effectively gives every parameter θ_i an independently adapted learning rate.

3.2 Feed Forward Neural Network

Now that we are equipped with the tool to find the optimal parameters θ, we turn our attention to how to build the function $\mathbf{h}$. One recipe to construct a sufficiently complex $\mathbf{h}$ is to build it as a fully connected feed-forward network [60]. A single layer of this network can be defined by

$$\widehat{\mathbf{h}} = W\mathbf{x} + \mathbf{b} \tag{3.12}$$

$$\mathbf{h} = \sigma\left(\widehat{\mathbf{h}}\right) . \tag{3.13}$$

Here, $\mathbf{x} \in \mathbb{R}^{n_i}$ is the n_i dimensional input, while $\mathbf{h}$ is the n_o dimensional output of the layer. $W \in \mathbb{R}^{n_i, n_o}$ is called the weight matrix, and $\mathbf{b} \in \mathbb{R}^{n_0}$ is the bias vector. σ is the element-wise applied non-linear activation function. With $\hat{h}_m$ being the m'th component of $\widehat{\mathbf{h}}$, this activation function is most often one of the following:

$$\mathrm{ReLU}\left(\widehat{\mathbf{h}}\right)_m = \max(\hat{h}_m, 0) \tag{3.14}$$

$$\mathrm{ReLU6}\left(\widehat{\mathbf{h}}\right)_m = \min(\max(\hat{h}_m, 0), 6) \tag{3.15}$$

$$\mathrm{Sigmoid}\left(\widehat{\mathbf{h}}\right)_m = \frac{1}{1 + \exp(-\hat{h}_m)} \tag{3.16}$$

$$\mathrm{Softmax}\left(\widehat{\mathbf{h}}\right)_m = \frac{\exp \hat{h}_m}{\sum_j \exp(\hat{h}_j)} . \tag{3.17}$$

These activation functions have simple and computationally cheap derivatives, which is useful for the calculation of the gradient. Typically, both W and $\mathbf{b}$ are combined into θ from the previous section to be optimized. $\mathbf{h}$ can either be the target y or be the input to another layer, in which case $\mathbf{h}$ is called the hidden representation. Such networks with at least one hidden layer are historically called Multilayer Perceptron (MLP). A simple example is shown in figure 3.1. Even though this construction seems ad hoc, there are formal results that put the technique on solid foundations. Especially, various forms of the Universal Approximation Theorem [61, 62] prove the existence of ReLU-activated networks that approximate well-behaved functions arbitrarily well.

[2] Strictly speaking, this is not the mean due to the powers of $\beta_{1/2}$ but the intuition still holds.

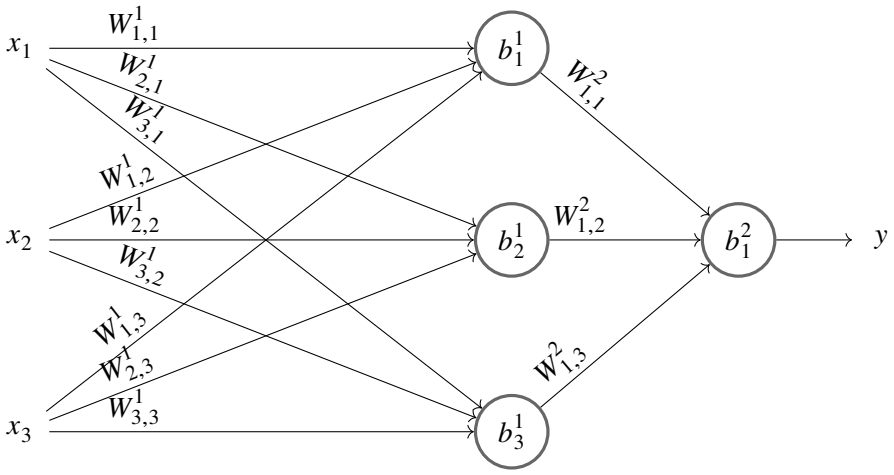

Figure 3.1: Representation of a simple two-layer fully connected feed-forward network that takes three inputs x_1, x_2, x_3 to calculate the output y. The circles denote the neurons that add the bias term b^i_j to the sum of the inputs and apply the activation function. Arrows denote the flow of values, which get multiplied by the weight denoted by $W^i_{j,m}$.

3.3 Gradient Boosted Decision Trees

Another powerful method to construct the function **h** are Classification and Regression Trees (CART) [63]. As the name implies, decision trees can be used for regression and classification tasks. The structure is as follows: The set of inputs starts at the root node of the tree. From here, the tree consists of nodes that take one of the input features, which is used to decide to which of multiple subsets (in our case, two) a given example belongs. These subsets are passed along to the next corresponding node. This procedure is repeated until a stopping criterion is reached. At the end of a path through the tree, there are leaves that carry a leaf index j and a weight w_j that represent the output. For classification, these leaves correspond to the inferred class a given example belongs to. For regression, the weight is the discretized inferred output. An example of a decision tree is shown in figure 3.2. Note, that even a regression tree can be used for binary classification when a logistic function is applied afterward, as is the case in logistic regression. Since we do not use vanilla decision trees, we will not cover the algorithms used to construct them. Due to their intuitive structure, these trees are very interpretable, although they usually perform weakly. This can be improved upon by gradient boosting.

The specific implementation of gradient tree boosting we will use in later chapters is XGBoost [64] which we will summarize in the following. Instead of proposing a single architecture and simply finding the best parameters, this approach builds the desired function iteratively with

$$h(\mathbf{x}_i) = \sum_{n=1}^{N} f_n(\mathbf{x}_i)\,, \tag{3.18}$$

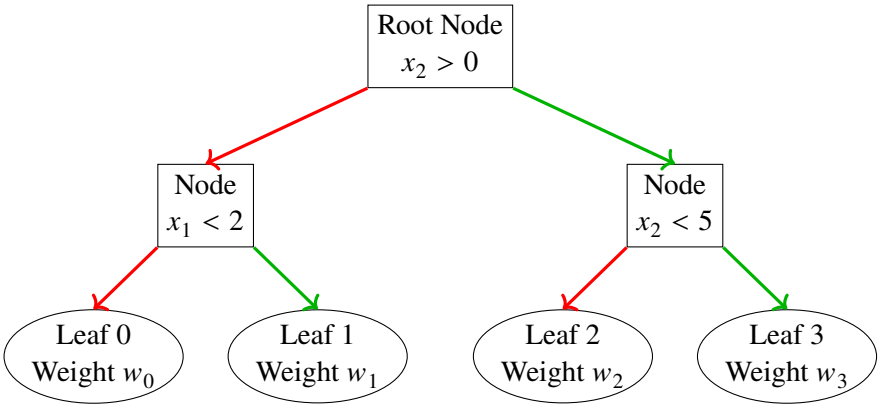

Figure 3.2: A short example of a decision tree. The tree starts at the root node. Examples that pass the decision flow along the green arrows, while examples that fail the decision flow along the red lines. As a concrete example, $x = (4, 2)$ ends in leaf 3 and is assigned the weight w_3.

where the functions f_n are weak learners, i.e., learners that perform relatively poorly on their own on the task at hand. In our case, these learners are simple regression trees. Each partial sum aims to fit the desired function as well as possible, such that each additional term only needs to fit the residuals of the previous terms.

To formalize how to grow the tree, one needs to define an objective. Since the tree growing does not have a natural cutoff that regularizes its growth, it tends to grow until it learns the noise present in the data, which leads to poor generalization. To remedy this, the objective contains, in addition to the loss function, regularization terms that penalize overly complex trees. Formally, this can be written as

$$\mathcal{L}(h) = \sum_i l(\hat{y}_i, y_i) + \sum_n \omega(f_n) \tag{3.19}$$

$$\omega(f_n) = \gamma T + \frac{1}{2}\lambda \sum_t^T |w_t|^2 , \tag{3.20}$$

where $\hat{y}_i$ is the inferred output for $\mathbf{x}_i$, i runs over the trainings samples, $l(\cdot, \cdot)$ is the loss function, T is the number of leaves in the tree f_n and w_t are the leaf-weights. γ and λ are hyperparameters that control how much the model is regularized. For growing the trees to optimize this objective, XGBoost approximates the loss function with the second-order Taylor expansion at each point of the expansion in equation 3.18. After m terms are considered in equation 3.18 and the prediction up until this point

is $\hat{y}_i^m$, the expansion can be written as

$$\mathcal{L}' = \sum_i l(\hat{y}_i^m, y_i) + \omega(f_m) \tag{3.21}$$

$$= \sum_i l(\hat{y}_i^{m-1} + f_m(\mathbf{x}_i), y_i) + \omega(f_m) \tag{3.22}$$

$$= \sum_i l(\hat{y}_i^{m-1}, y_i) + \frac{\partial l(\hat{y}_i^{m-1}, y_i)}{\partial \hat{y}^{m-1}} f_m(\mathbf{x}_i) \tag{3.23}$$

$$+ \frac{1}{2} \frac{\partial^2 l(\hat{y}_i^{m-1}, y_i)}{\partial (\hat{y}^{m-1})^2} f_m^2(\mathbf{x}_i) + \omega(f_m) + O\left(f_m^3(\mathbf{x}_i)\right) \tag{3.24}$$

$$\approx \sum_i g_i f_m(\mathbf{x}_i) + \frac{1}{2} h_i f_m^2(\mathbf{x}_i) + \omega(f_m)\,, \tag{3.25}$$

where all terms that do not depend on the structure of the m-th tree are dropped. The g_i and h_i are given by

$$g_i = \frac{\partial l(\hat{y}_i^{m-1}, y_i)}{\partial \hat{y}^{m-1}} \tag{3.26}$$

$$h_i = \frac{\partial^2 l(\hat{y}_i^{m-1}, y_i)}{\partial (\hat{y}^{m-1})^2}\,. \tag{3.27}$$

Again, enumerating the leaves by t and defining I_t as the set of indices i of inputs x_i that belong to the t-th leaf this can be written as

$$\mathcal{L}' = \sum_t^T \left(\sum_{i \in I_t} g_i w_t + \left(\frac{1}{2} \sum_{i \in I_t} h_i + \lambda \right) w_t^2 \right) + \gamma T\,. \tag{3.28}$$

This can be used to obtain the weights that extremize this expression by setting the derivative to zero, given that each term can be considered independently:

$$w_t = -\frac{\sum_{i \in I_t} g_i}{\sum_{i \in I_t} h_i + \lambda}\,, \tag{3.29}$$

which leads to the optimal objective

$$\mathcal{L}' = -\frac{1}{2} \sum_t^T \frac{\left(\sum_{i \in I_t} g_i\right)^2}{\sum_{i \in I_t} h_i + \lambda} + \gamma T\,. \tag{3.30}$$

This expression can now be used to grow the tree. Consider a node in a tree under construction, that contains the indices I and a split into indices I_L and I_R for the left and right leaf respectively is

proposed. The reduction in the objective function due to the additional split can be written as

$$\text{Gain} = \frac{1}{2}\left(\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right) - \gamma\,. \tag{3.31}$$

One then finds the feature and split value with the greatest Gain and adds this split to the tree if the objective is indeed reduced. Alternatively, one terminates the tree in a leaf when the penalty of the added split γ is greater than the improvement, a technique called pre-pruning. Notably, a feature is only selected to construct an additional node if the split improves the model. In other words, the method is inherently insensitive to the inclusion of useless features.

3.4 Convolutional Neural Network

Not all objects have a sensible representation as a fixed-size vector $\mathbf{x}$, which a simple MLP can handle. Images, for example, have numerous drawbacks in this representation. In a flattened representation, the notion of neighboring pixels, or objects in the image being close is hidden and would need to be learned by the network. Additionally, single pixels carry only a small amount of information on their own. A better approach is to keep the two-dimensional nature, such as in Convolutional Neural Networks (CNN)[65, 66]. For this, one regards an image as a matrix $I \in \mathbb{R}^{n_c, n_w, n_h}$, where n_c is the number of channels (e.g., red, green, blue for RGB-images), n_w, n_h the number of pixel columns and rows. Next, one considers a learnable matrix called kernel $K \in \mathbb{R}^{n_c, k_w, k_h}$ with k_w, k_h kernel columns and rows. This kernel is convolved [3] with the image to construct the output C at position i, j as

$$C(i,j) = \sum_{i_c}^{n_c}\sum_{i_w}^{k_w}\sum_{i_h}^{k_h} I(i_c, i+i_w, j+i_h)K(i_c, i_w, i_h) + b\,, \tag{3.32}$$

where b is a bias parameter. This is passed afterward through an activation function. In other words, this operator slides the kernel over the spatial dimensions of the image and sums the element-wise product between the image and the kernel. Given one only allows convolutions where the kernel is entirely contained in the image, the output C (also called feature map) of this operation has shrunken in size and has dimension $(n_w - k_w + 1) \times (n_h - k_h + 1)$. One technique to avoid this spatial shrinking is padding, where the image gets expanded by additional values before the convolution so that the size stays fixed. The channel dimension is always collapsed into a single value at each position. This collapse results usually in information loss. To avoid this n_o independent kernels are used to construct the output with n_o feature maps. This way, each kernel could learn to find a different property of the image.

[3] Technically, what is implemented in various machine learning libraries is cross-correlation instead of proper convolution, which differs by flipping the kernel.

This convolution operator has several properties (also called inductive biases), that aid in learning the important features such that the model generalizes well onto unseen data [57]. For one, the operation is inherently local if the filters are spatially much smaller than the entire image. Since many features of an image are also local, this allows to share a relatively small number of weights across the entire space while still being useful. For example, small features such as edges can be found in only a small region of the image, without the need for global connections. Also, the operation is translation equivariant, i.e. a spatial shift in the input image results only in a spatial shift of the feature map.

To build a classifier for images, such as the one we will encounter later in this text, one usually stacks multiple CNN layers, feeding the feature maps after applying an activation function into subsequent layers. The last feature map is flattened and fed into a classification network which gets expensive if the feature map is too large. To reduce the spatial dimensions one can either use convolution with strides, where the indices i, j in equation 3.32 are not traversed consecutively but only values that are different by a fixed step size, or a pooling layer. For the latter, a pooling operation (often the maximum or average) is applied on a sliding window across the spatial dimensions. The dimensions of the image can effectively be cut down when this sliding is stridden without the introduction of additional learnable parameters. For example, a square window of width two that gets applied with stride two cuts both spatial dimensions in half. Often both reduction techniques are applied simultaneously.

3.5 Attention Mechanism

Now we turn our attention to another mechanism that will be used in this thesis, the attention mechanism. Attention was originally designed for use in natural language processing but gained wide popularity in a wide range of applications[67]. We briefly summarize how the attention mechanism works following the treatment of ref. [68]. For this, let us regard the input of the mechanism X as a vector of m n-dimensional vectors, also called tokens. These tokens may be embedded words in the context of natural language processing, pixels with multiple channels, or flattened patches of images in the context of computer vision. The aim is to gain information that uses all tokens on an equal footing no matter how far (first vs. last word in a sentence, rightmost top corner vs. leftmost bottom corner of an image) the tokens are from each other, without relying on excessively large amounts of involved neurons. First, the elements of X are multiplied by learnable matrices W^V, W^K and $W^Q \in \mathbb{R}^{n,d}$ to form V, K, Q called value, key and query. Note that, in general, W^V may have another second dimension. Next, the scalar product of all rows of Q with all rows of K builds a matrix that the softmax function is applied to. The resulting matrix is used to mix and reweight the tokens stored in rows of V, depending on the representation learned by the matrices. The full result can be expressed using matrices by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V\,. \tag{3.33}$$

The mechanism is also shown diagrammatically in figure 3.3.

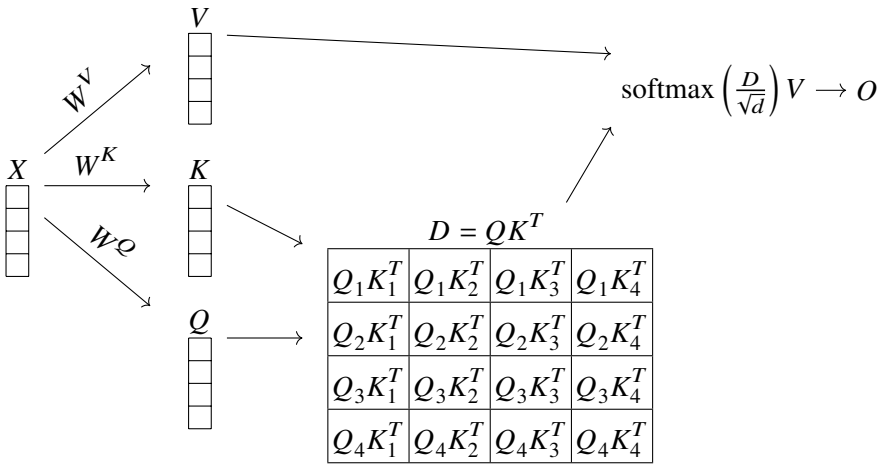

Figure 3.3: Sketch how the attention mechanism works. The input X is a vector or tokens. The learnable matrices W^V, W^K and W^Q transform each token into a d-dimensional representation that built V, K and Q. The product $D = QK^T \in \mathbb{R}^{d,d}$ are scaled by $1/\sqrt{d}$ and fed into the softmax function, which builds the weights by which V is multiplied, which furnishes the output O.

This particular variant of the mechanism is called Scaled Dot-Product Self-Attention and outputs m d-dimensional vectors. Note here that the number of learnable parameters is independent of the number of tokens. Also, note that the calculation of the scalar products gets expensive for a large number of tokens. For n tokens, there are n^2 scalar products that need to be computed. Especially in images, where the number of pixels grows quadratically with the size of the image, the vanilla attention mechanism gets unmanageably costly when applied directly to pixels of large images. Sometimes it is useful to use n_h distinct matrices W^V, W^K and W^Q. The n_h attention outputs are called heads and are concatenated. The now large contextualized output vectors are as a last step transformed to be again n-dimensional by another learnable matrix. The whole mechanism is then called Multi-Head Self-Attention. This allows a single layer to extract information from n_h different projections in a single layer.

3.6 Self-Attention Applied to Images

Now that we have covered how Multi-Head Self-Attention works, we now introduce the concept of transformer. As the name suggests, the mechanism aims to transform sequences into more informative states. For this, we follow the treatment of the transformer-encoder of ref. [68]. This structure is built of several identical layers. Each layer consists of a Multi-Head Self-Attention module with a skip connection and layer-normalization [69]. Skip connections take the input of whatever they are built around and add it to the output. This is done to stabilize training by avoiding vanishing gradients. These tokens are then fed separately through a fully connected feed-forward network, again with skip connections and layer normalization. With each layer, the tokens pick up more context, i.e.

information depending on the entire set of tokens.

As mentioned earlier, applying self-attention and therefore transformers on large images is prohibitively costly when applied directly to all pixels in an image, even though the number of parameters may stay manageable compared to applying a fully connected network directly to pixels. To circumvent this, multiple techniques break the image into smaller chunks and apply self-attention to these. We will only focus on image classification architectures, as these are what is covered in the remaining text.

3.6.1 Vision Transformer

The Vision Transformer (ViT) [70] breaks the image into a grid of $P \times P$ patches. Each patch is flattened to $P^2 \cdot C$ dimensional vectors, where C is the number of channels the image had. These are then linearly projected into D dimensional vectors to dial the desired size. These vectors are subsequently added to a learnable position encoding. This allows the next step to be informed about the position of each patch since the encoding only depends on the position of the path within the image. A single additional learnable `cls`-token is added to the list of vectors. The results are then fed into the transformer, which consists of multiple transformer-encoder layers. After the transformer layers, only the transformed `cls`-token is fed into the classification head. This way, the input features for the MLP are not biased to a single patch, as would be the case if one were to simply choose another output token for classification.

As noted by the original publication, this architecture only performs better than CNNs for large datasets, because it lacks some of the inductive biases mentioned in Chapter 3.4, that help to generalize from the limited training data. Several techniques exist that introduce back the inductive biases into the transformer approach.

3.6.2 CoAtNet

One of the techniques we will use later in this text was introduced in Combination of Depthwise **Co**nvolution and self-**At**tention (CoAtNet) [71]. First, the authors propose aggressively shrinking the large input image by strided convolutions or pooling operations such that the size is manageable for the transformer stage. The key idea is a technique called relative attention. The version used by CoAtNet adds a weight to the attention matrix before the application of softmax that depends on the relative position between the tokens. This can be written schematically as

$$y_i = \sum_j \frac{\exp(q_i k_j^T + \omega_{i-j})}{\sum_l \exp(q_i k_l^T + \omega_{i-l})} v_i \,, \tag{3.34}$$

where q_n, k_n and v_n are query key and value vectors at the position n and ω are the weights encoding the relative position, hence in itself being translation equivariant. ω is learnable and has size $(2H-1) \times (2W-1)$ for input images of height H and width W. Depending on the relative size of the

dot-product and ω this allows the relative attention mechanism to either enforce the global receptive field of the vanilla attention with small values of ω_{ij} or focus on local fields by learning larger values of ω_{ij} for small distances between tokens.

3.6.3 MaxViT

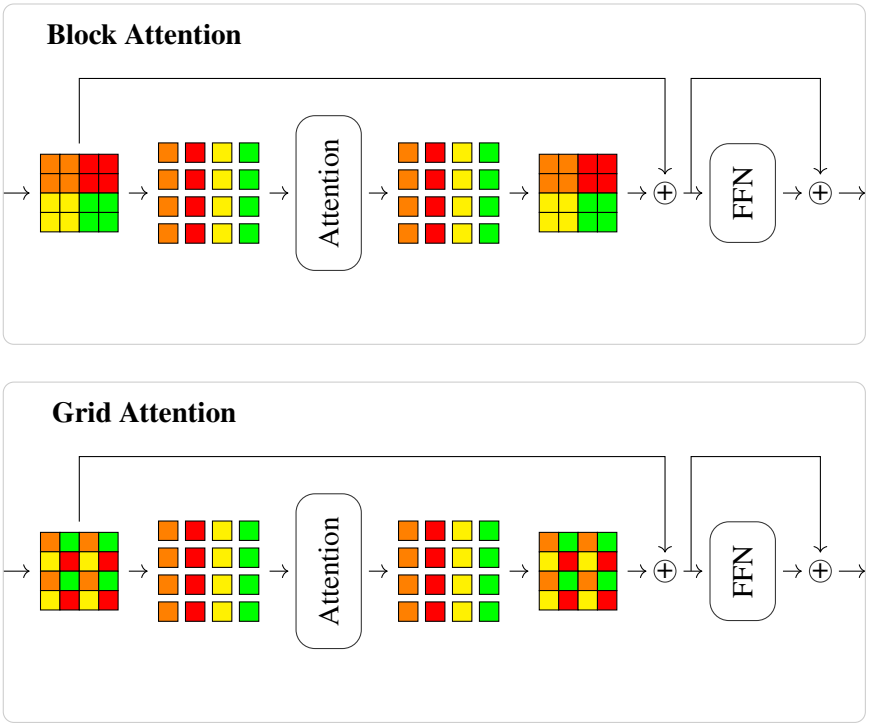

Figure 3.4: Block and grid attention as used by MaxViT. For block attention the image is split into regular-sized blocks within each attention is applied, as shown in the upper panel. For grid attention (shown in the lower panel) the image is split into a regular grid. Attention is applied to each subgrid.

The second technique we will use is the Multi-AXis VIsion Transformer (MaxViT) [72]. The main way this accomplishes applying attention to images is in two stages. First, the image is split into square blocks. Within each block, one applies the attention mechanism, which the authors call Block Attention, after which the image is recombined. This essentially applies the mechanism only locally with high resolution. In a second stage, the image is split into a regular uniform grid such that the attention mechanism is applied to each grid, which the authors call Grid Attention [4]. The specific attention mechanism used here is the same relative attention used by CoAtNet. This allows the attention mechanism to use global interactions on the lower-resolution patches. Both steps are illustrated in figure 3.4.

[4] An interesting fact from the intersection of physics and computer science: This splitting into grids and blocks is done using Einstein summation [73]

The reason why this works is the following: As mentioned before, the computational cost of the attention mechanism scales as n^2 for n tokens. For a square $N \times N$ image, with each pixel representing a token, the cost scales as N^4. On the contrary, block attention with block size $B \times B$ costs constantly B^4 computations but has to be computed N^2/B^2 times. Therefore, another factor of N^2/B^2 computation is saved.

To build a MaxViT-layer, a MBConv-block [74] is prepended to the block attention, followed by grid attention. Multiple MaxViT-layers are stacked to form a MaxViT-block. Only the convolution of the first layer in each block uses stride two to reduce the spatial size. The number of MaxViT-layers in each MaxViT-block is a hyperparameter.

The full classification network, as proposed by the authors, is built as follows: First, the image is passed through two convolution layers with 3×3 kernels and stride two and one, respectively. This is followed by four MaxViT-blocks. The output tokens of the last block are average-pooled and fed through a simple MLP for classification.

3.7 Density Estimation

Density estimation is another class of problems we encounter in this thesis. For this, we want to model or estimate the Probability Density Function (PDF) $p(\mathbf{x})$ from only a finite number of independent and identically distributed samples $\mathbf{x}$.

3.7.1 Kernel Density Estimation

For the density estimation of a univariate PDF, we do not even need to use machine learning. A popular, non-parametric method to solve this problem is Kernel Density Estimation (KDE). Given a set of n independent and identically distributed samples x_i from the PDF $p(x)$ one may estimate $p(x)$ by [75]:

$$\hat{p}(x) = \frac{1}{nh} \sum_{i}^{n} K\left(\frac{x - x_i}{h}\right) . \tag{3.35}$$

Here, K is a non-negative and normalized function called the kernel, and h is called the bandwidth. A common choice for the kernel is the unit normal distribution

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x}{2}\right) . \tag{3.36}$$

The bandwidth h needs to be tuned to a usable value. If h is too large, the estimated distribution is overly smooth, and details are washed out. If h is too small, the estimated distribution behaves irregularly and models artifacts due to the finite number of samples. Notice that if we bin the samples x_i into bins of width h and choose a uniform distribution as the kernel, this technique closely resembles normalized histograms. There are also multivariate kernel density estimation techniques (e.g., see ref. [76]), but we will not pursue them further.

3.7.2 Normalizing Flow

To estimate a multivariate distribution, we will use another technique called normalizing flow [77, 78]. For this treatment, we follow ref. [79]. This aims at modeling the target distribution $p_x(\mathbf{x})$ of a continuous D dimensional random variable $\mathbf{x}$ parametrically by transforming the samples $\mathbf{u}$ of a base distribution $p_u(\mathbf{u})$ using a transformation T such that $\mathbf{x} = T(\mathbf{u})$. If one demands T to be invertible and both T and its inverse T^{-1} to be differentiable, one finds by using a change of variable

$$p_x(\mathbf{x}) = p_u(\mathbf{u}) |\det J_T(\mathbf{u})|^{-1} , \tag{3.37}$$

where J_T is the Jacobian matrix corresponding to the transformation T. In practice, one usually chooses $p_u(\mathbf{u})$ to be a simple tractable distribution such as a multivariate normal and instead of just a single T one decomposes T into K transformations $T_K \circ ... \circ T_1$. With the definition $\mathbf{u} \equiv \mathbf{z}_0$ and $\mathbf{x} \equiv \mathbf{z}_K$ we can write the intermediate $\mathbf{z}_k$ as

$$\mathbf{z}_k = T_k(\mathbf{z}_{k-1}) , \tag{3.38}$$

with $k = 1...K$.

A sample from the simple distribution $p_u(\mathbf{u})$ *flows* through all K transformations via the $\mathbf{z}_k$'s into the target space. Inversely given the samples $\mathbf{x}$ and traversing the transformations in the opposite direction gets *normalized* into the simpler distribution $p_u(\mathbf{u})$ thus the name normalizing flow.

Framed in this context, the aim is to parametrize the transformations T_i by some sufficiently expressive model and find the parameters that solve the problem. As before, these parameters are found by minimizing a cost function. One way to write the cost function is the forward Kullback-Leibler divergence between the model and the true target distribution, which measures, informally speaking, the difference between two probability distributions. Given a set of N samples, sampled from the true target distribution $\mathbf{x}_n$ this can be approximated up to terms that do not depend on the model parameters as

$$\mathcal{L} \approx -\frac{1}{N} \sum_{n=1}^{N} \log p_u \left(T^{-1}(\mathbf{x}_n) \right) + \log |\det J_{T^{-1}}(\mathbf{x}_n)| . \tag{3.39}$$

Since, from a practical point of view, we want to arrive at the parameters that minimize equation 3.39 via gradient descent, the cost function should be computationally easy to evaluate. The logarithm of the determinant in the last term decomposes into a sum of the logarithms of the determinants of the partial transformations T_k, though the determinants are still potentially costly.

It is particularly easy to calculate the determinant of a matrix if it is triangular. In this case, the log-determinant is simply the sum of the logarithms of the diagonal entries.

To see how this is useful, we focus on a model with a single transformation T from $\mathbf{u}$ to $\mathbf{x}$. Let us denote $\mathbf{x}_{1:i}$ as the vector built from the first through i'th component of $\mathbf{x}$. The chain rule of probability

states

$$p(\mathbf{x}) = \prod_{i=1}^{D} p(x_i|\mathbf{x}_{1:i-1}) \,, \tag{3.40}$$

where $p(x_1|\mathbf{x}_{1:0}) \equiv p(x_1)$ and $p(x_i|\mathbf{x}_{1:i-1})$ are probabilities of x_i conditioned on the previous $i-1$ dimensions. The Jacobian is always triangular if T_i (the i'th component of T) depends only on $\mathbf{u}_{1:i}$ or $\mathbf{u}_{i:D}$. This is the point of attack affine autoregressive flows take [80]. For this, one defines the transformation as

$$x_i = u_i \exp \alpha_i(\mathbf{x}_{1:i-1}) + \mu_i(\mathbf{x}_{1:i-1}) \,, \tag{3.41}$$

where α_i and μ_i are neural networks. For all α_i and μ_i, this expression is invertible. In this parametrization, the log-determinant is simply the sum of the α_i which is computationally cheap. The version of normalizing flow we will use later in this text is the Masked Autoregressive Flow (MAF) [81]. This aims to model the $2 \cdot D$ functions α_i and μ_i in a single pass using masking in the form introduced by Masked Autoencoder for Distribution Estimation (MADE) [82]. For this, one first builds a fully connected neural network with D input and D output neurons and a number of hidden layers with more neurons. It takes $\mathbf{u}$ as input and applies a mask on the weight matrices such that any given output neuron is only connected to a desired subset of input nodes. An algorithm to produce this mask efficiently is given in the original MADE publication. This way, it produces the entire set of D functions α_i or μ_i in a single forward pass while it still can be efficiently represented for evaluation on a GPU.

CHAPTER 4

Resonant Anomaly Detection

The results of this chapter have been published at:
Combining resonant and tail-based anomaly detection
Gerrit Bickendorf, Manuel Drees, Gregor Kasieczka, Claudius Krause, and David Shih
Phys. Rev. D 109, 096031 - Published 23 May 2024

In many scenarios of physics beyond the Standard Model, there is at least one new particle that decays visibly. Usually, a search for these signals boils down to searching for a localized excess compared to the Standard Model in a mass-like feature m. To calculate the significance of such an excess, some searches rely on simulated Standard Model background events. This is suboptimal, because from the computational cost of sufficiently accurate modeling of the hard process and showering to the non-perturbative nature of hadronization and the finely-grained simulation of the detector response (to mention a few), numerous steps add uncertainties to the simulations. Even if the background template is constructed in a data-driven way, these searches usually assume a specific signal model. By optimizing selection cuts to maximize the discovery potential of a specific signal, these strategies have low sensitivity to other models.

New physics might hide in the data collected by the experiments, only slightly out of reach of our classical searches. Therefore, it is imperative to also employ strategies that are less model-specific. Classical bump-hunts in a feature with a smooth non-zero background shape fit this description. Put simply, one could divide the range of m into a Signal Region (SR) and Side Band (SB), interpolate a smooth function from the SB into the SR and calculate p-values for all potential locations of the resonance. This is rather insensitive because one relies on only this feature, such that the signal has to have a large cross section to significantly impact the p-value. Recent machine learning methods can be used to suppress the background efficiently without exact knowledge of the signal model. For this, the methods use auxiliary features x that might be different between the background and a potential signal model to generate a signal-enriched subset of events. These methods can be summarized by the umbrella term *resonant anomaly detection* [8, 83–102].

4.1 Overview of Resonant Anomaly Detection

As implied by the Neyman-Pearson lemma [103] the optimal technique to classify a given event as either signal- or background-like constructs the likelihood ratio $p_{\text{si}}(x)/p_{\text{bg}}(x)$ where p_{bg} is the distribution of the true background and $p_{\text{si}}(x)$ is the distribution of the signal. In physics, we do not have access to either distribution exactly, even though we can approximately construct $p_{\text{bg}}(x)$, for example, using the SB, while $p_{\text{si}}(x)$ is supposed to be left unfixed in favor of being signal model agnostic.

On the other hand the proxy task

$$R(x) = \frac{p_{\text{data}}(x)}{p_{\text{bg}}(x)}\,, \tag{4.1}$$

where p_{data} is the distribution of the data, is available. When the distribution of the background and signal is known, the data distribution can be written as

$$p_{\text{data}}(x) = (1-\epsilon)p_{\text{bg}}(x) + \epsilon p_{\text{si}}(x)\,, \tag{4.2}$$

where ϵ is the signal strength parameter, depending on the signal process cross section. With this equation 4.1 can be rewritten as

$$R(x) = (1-\epsilon) + \epsilon\frac{p_{\text{si}}(x)}{p_{\text{bg}}(x)}\,, \tag{4.3}$$

which is monotonically related to the likelihood ratio between signal and background and hence still Neyman-Pearson optimal. This is the key observation that is exploited in the following.

Next, we will give an overview of some resonant anomaly detection methods that use this or a similar approach. This is by no means an exhaustive treatment since it is an active field.

CWoLa Hunting

Perhaps the most immediate application of a similar technique is in Classification Without Labels (CWoLa)[104]. For this, one builds two datasets with features x, one inside the SR and one inside the SB. Even though the SB is not completely signal-free, it still contains fewer signal events, such that the likelihood ratio between SR and SB $p_{\text{SR}}(x)/p_{\text{SB}}(x)$ is still monotonically increasing with $p_{\text{si}}(x)/p_{\text{bg}}(x)$ as described in the original publication [104]. This approach depends on the assumption that for pure background events the SR and SB are indistinguishable, i.e. $p_{\text{SB}}(x) = p_{\text{SR}}(x)$.

This implies that special attention has to be paid to select and modify the features x such that they are uncorrelated with the feature m. Otherwise, the classifier will pick up the correlation with m and simply learn the definition of the SR and SB while ignoring the difference between both sets due to the signal presence in the SR. This method has already been applied at ATLAS [105] on the decay $A \to BC$ with $m_A = 1\,\text{TeV} \gg m_B \sim m_C \sim 100\,\text{GeV}$ and B and C are reconstructed as large-radius jets. Also, the CMS measurement of the $t\bar{t}b\bar{b}$ production cross section used CWoLa [106].

Tag N' Train

The Tag N' Train (TNT) [107] method is closely related to CWoLa. For CWoLa one needs to define a signal-enriched and signal-depleted sample by hand. In general, one might not know how to design these. For Tag N' Train this is done automatically using a tagger that might perform poorly on its own, followed by training a CWoLa-like anomaly detector. One example application is the search for an anomalous process that shows up in two jets. To keep the training of the tagger data-driven, one possible architecture is an autoencoder that is trained completely unsupervised on only one jet of the event. For this autoencoder, one represents the jet constituents (either calorimeter cells, tracks or particle candidates) as an image, treating the deposited energy as the brightness and discretizing the detector coordinates ϕ and η to a regular grid[1], regarding these as pixel coordinates. One autoencoder architecture [108–111] uses convolution and pooling layers to reduce the image size. The image is then flattened and fed through multiple dense layers, further reducing the number of neurons. After the desired latent space dimension is reached, the architecture is built inversely such that the output is the same size as the input image. One then trains the autoencoder to reconstruct the input image. Since the latent space is small, the exact identity transformation cannot be learned, and –informally speaking–, the autoencoder has to focus on reconstructing the most probable image. Since one assumes the majority of jets are from background processes, the autoencoder should reconstruct these images better than it does for signal images, which it encounters less frequently. Thus, the reconstruction loss can be seen as an anomaly score. By cutting on this loss, the dataset is split into a signal-enriched and signal-depleted subset. The second jet is then used to train a CWoLa-like classifier. In principle, this can again be iterated, splitting the dataset using the CWoLa anomaly score and training another CWoLa-like classifier on the first jet, which was used by the autoencoder. However, the performance tends to plateau rather quickly.

Note, that the assumption of two anomalous jets not only reduces signal model independence, but might be problematic for models where the potentially anomalous jets cannot be reliably separated. If non-anomalous jets fit the selection criterion, this might dilute the anomalous jets, even though one of the jets is correctly identified.

ANODE

Another approach is Anomaly Detection with Density Estimation (ANODE) [112]. This immediately tackles equation 4.1 by first learning the conditional density of events $p(x|m)$ in the SB and interpolating it into the SR. If all goes well, this can be seen as an estimate of p_{bg} in the SR. Additionally, a second conditional density $p_{\text{data}}(x|m)$ is estimated using the data in the SR. Both density estimators are chosen as normalizing flows, as described in the earlier chapter. Combining both, one can calculate the likelihood ratio

$$R(x|m) = \frac{p_{\text{data}}(x|m)}{p_{\text{bg}}(x|m)}\,, \tag{4.4}$$

[1] More detail on how to obtain images from jets will be provided in the next chapter

which can be evaluated on the samples to assign an anomaly score to each event. A cut on this score can then be used to amplify the signal-to-background ratio. One clear advantage over CWoLa is that it does not break down when x and m are slightly correlated because the learned conditional density interpolates the correlations from the SB into the SR.

CATHODE

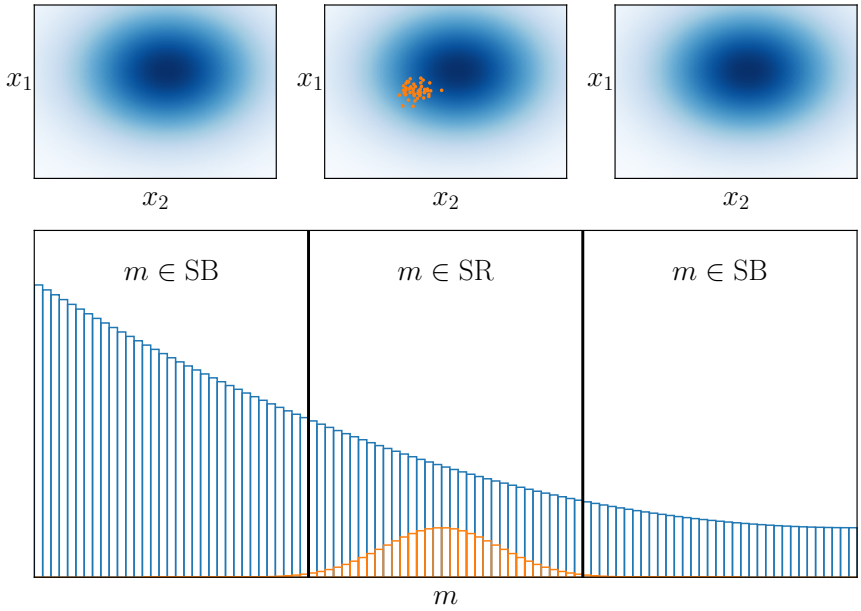

Figure 4.1: Sketch of CATHODE. The signal shown in orange forms a bump in the resonant feature m, while the background shown in blue is smooth. One defines the SR, which contains almost all signal events. The density of the auxiliary features x, shown in the upper panels, is learned in the SB and interpolated into the SR. Overdensities in the SR data compared to the interpolated background template are then associated with anomalous events. Figure inspired by ref. [8].

Another improvement of ANODE is Classifying Anomalies THrough Outer Density Estimation (CATHODE) [8]. This will be the technique we will follow further in this chapter and will be described in more detail later on. Similar to ANODE, one first learns $p(x|m)$ in the SB. Next, artificial samples from this estimate are generated in the SR. If all goes well and the signal region indeed contains almost all signal events, the artificial samples should approximately follow $p_{\text{bg}}(x)$. Correlations between x and m are also modeled by the density estimator. Therefore, it is possible to apply CWoLa between real SR events and artificial samples without careful consideration of possible slight correlations. The CWoLa classifier can then be used to assign anomaly scores to the SR events, which can then be cut on. A sketch of the distributions is shown in figure 4.1. We will describe the method in more detail later in this chapter. The whole method works by learning directly from the data. The training and model selection of both the density estimation and classification are completely agnostic of any signal

truth label.

LHC Olympics 2020

Most of these techniques have only been shown to perform well in a limited set of signal scenarios. Of special note here are the LHC Olympics 2020 [113] which posed a community challenge to seed new developments in data-driven approaches to find new physics model-agnostically. The challenge consists of three datasets with unknown signal production processes, and participants were tasked with uncovering new physics signals from an overwhelming QCD background. One signal model was the production of a Z' boson with decay $Z' \to X(\to qq)Y(\to qq)$ with $m_{Z'} = 3\,500\,\text{GeV}$, $m_X = 500\,\text{GeV}$ and $m_Y = 100\,\text{GeV}$. This model was never hidden, such that it is used to develop new approaches. Many techniques have been compared on this de-facto benchmark dataset. CATHODE has been shown to work well on this dataset by using the two leading jets J_1 and J_2 [8]. The feature that the SR is defined on is the dijet invariant mass m_{JJ} while the additional features are the jet mass m_{J_1}, the difference between both jet masses $m_{J_2} - m_{J_1}$ and the two n-subjettiness variables $\tau_{21}^{J_1}$ and $\tau_{21}^{J_2}$. The signal is always found in the bulk of the distribution of the features x, never at the far end of a tail.

4.2 Combining Resonant and Tail-based Anomaly Detection

Many BSM physics scenarios at the TeV-scale leave their signature at the tails of distributions such as p_T^{miss}, H_T or M_{eff}. It has never been explicitly shown that these signatures can be effectively leveraged by anomaly detection methods such as CATHODE. As was covered in Chapter 2.3.4, many processes within the MSSM will produce events with large p_T^{miss} since the LSP, often assumed to be the lightest neutralino $\widetilde{\chi}_1^0$, leaves the detector undetected. The neutralino is produced either directly or indirectly at the end of a decay chain of heavier sparticles. To show that CATHODE is indeed capable of uncovering such signatures, we show its performance for a well-motivated MSSM model. For this, we consider gluino pair production for which the two gluinos undergo a cascade decay first into a light quark plus off-shell squark, which immediately decays into another quark and a neutralino $\widetilde{\chi}_2^0$. The $\widetilde{\chi}_2^0$ subsequently decays into the LSP $\widetilde{\chi}_1^0$ plus X, where X can be a Z or Standard Model Higgs-boson or even the non-standard Higgs boson of the extended Higgs sector of the MSSM, as covered in Section 2.3.3.

Our treatment should be contrasted with existing machine-learning-based approaches to supersymmetric scenarios, that are fully supervised (e.g,[114–118]).

The case where X is the Z boson has been searched for by CMS [119] with a traditional, cut-based analysis. The full process is shown in figure 4.2. Here, the gluino and $\widetilde{\chi}_2^0$ have a small mass splitting such that the two quarks per gluino will produce relatively soft jets. When the $\widetilde{\chi}_1^0$ and X are relatively light compared to the $\widetilde{\chi}_2^0$ this process will contain p_T^{miss} and two highly boosted Zs. When the Zs decay hadronically (as is the case in 69.9% of cases [4]) this leaves two highly boosted large-radius jets.

The CMS search defined a signal region requiring the mass of the leading and subleading AK08

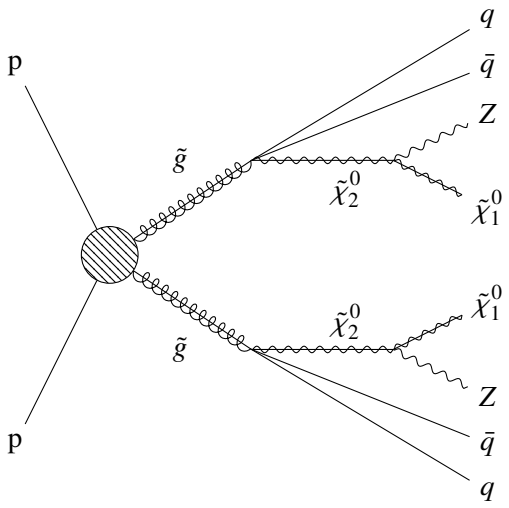

Figure 4.2: Diagram of the signal process $pp \to \widetilde{g}\widetilde{g}$ with $\widetilde{g} \to q\bar{q}\widetilde{\chi}_2^0$, $\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0$

jets to lie in $m \in [70\,\text{GeV}, 100\,\text{GeV}]$. The background was estimated in two data-driven steps, which are shown in more detail in our recreation in Appendix A.1. First, the total number of expected background events in the SR B_{norm} is found by interpolating the sideband in the leading AK08 jet into the SR (the subleading AK8 jet was required to be in the SR). The shape of the p_T^{miss} distribution is determined by the distribution where neither jet lies inside the SR. Normalizing this shape to B_{norm} made it possible to derive exclusion bounds on the gluino production cross section from exclusive p_T^{miss} bins. This procedure can be seen as a low-resolution p_T^{miss}-density estimation. The probing of the p_T^{miss} spectrum with signal regions defined by the jet mass can also be done with CATHODE. The clear advantage is, that this works automatically without explicitly stating that the signal is found at large p_T^{miss}. For this, we will use the mass of the hardest large-radius jet as the resonant feature. The additional features x include the mass of the second hardest large-radius jet, H_T and p_T^{miss}. Both H_T and p_T^{miss} will be found at the tail of the distributions for this signal. p_T^{miss} is essential to suppress the Standard Model background from Z/W + jets with hadronically decaying Z/W.

4.3 Data

Since all the methods described here (both the CMS search and CATHODE) fully rely on data for estimating backgrounds (aka are “fully data-driven"), the simulation data we generate here is meant to play the role of real data, and all background estimates and significances etc. we derive are meant to illustrate the result one would get applying these methods to collider data. There will be no events generated here that play the role of simulations at the LHC.

For Standard Model background data, we take into account the three largest contributions of background events to the CMS search, arising from Z + jets, W + jets and $t\bar{t}$+ jets. W and Z events were generated with one to four additional final state partons while $t\bar{t}$ were generated with up to 3

additional partons.

All events are generated with MadGraph5_aMC@NLO 3.2.0 [120] with $\sqrt{s} = 13\,\text{TeV}$. The NNPDF3.1LO PDF-set [121] is used throughout. At the generator level, a minimum H_T cut of 250 GeV is imposed.

For the benchmark signal (to be used to compare the performance of the CMS search vs. the CATHODE method), we follow the CMS search and generate gluino pair production (with zero to two additional partons), with subsequent cascade decay $pp \to \widetilde{g}\widetilde{g}, \widetilde{g} \to q\bar{q}\widetilde{\chi}_2^0, \widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0$ where the neutralino $\widetilde{\chi}_2^0$ is the next-to-lightest supersymmetric particle (NLSP) and $\widetilde{\chi}_1^0$ is the LSP. The mass splitting between the gluinos and NLSP is set to 50 GeV while the LSP-mass is 1 GeV. This results in soft jets from the first step of the decay and a highly boosted Z-boson. The LSP escapes the detector and contributes large amounts of missing energy.

Later we will also consider decays of $\widetilde{\chi}_2^0$ to $X\widetilde{\chi}_1^0$ where the X is either a Standard Model Higgs boson or a new Higgs boson with mass besides 125 GeV, like the new Higgs bosons in supersymmetric extensions of the Standard Model. The Standard Model Higgs boson decays in $\sim 58\%$ of cases to $b\bar{b}$ while for the latter case, we set the branching ratio to 100%.

Gluinos are decayed spin-uncorrelated with Madspin [122] to $q\bar{q}\widetilde{\chi}_2^0$ via an off-shell squark and subsequently $\widetilde{\chi}_2^0 \to X\widetilde{\chi}_1^0$. Showering is done using Pythia 8.306 [123] with MLM merging. Pythia-Tune CP5 was used for background events while CP2 [124] was used for the signal samples. The number of background events in each channel is scaled to match their respective next-to-leading-order cross sections [125]. Detector effects are simulated using Delphes 3.5.0 [126] with the `delphes_card_CMS.tcl` detector card modified to account for the lepton isolation criterion. Particles are clustered into jets using the anti-k_T clustering algorithm with cone-radius parameter $R = 0.4$ for AK4 jets and $R = 0.8$ for AK8 jets. To be considered, jets have to have $p_T > 30$ and $|\eta| < 2.4$.

The following selection criteria are imposed for both the classical CMS recast and the dataset for CATHODE:

1. $N_{\text{AK4 jet}} \geq 2$
2. $p_T^{\text{miss}} > 300\,\text{GeV}$
3. $H_T > 400\,\text{GeV}$, where $H_T = \sum_{\text{AK4 jets}} |\vec{p}_T|$
4. $|\Delta\phi_j, \vec{H}_T^{\text{miss}}| > 0.5(0.3)$ for the first two (up to next two) AK4 jets, where $\vec{H}_T^{\text{miss}} = -\sum_{\text{AK4 jets}} \vec{p}_T$
5. no isolated photon, electron or muon candidate with $p_T > 10\,\text{GeV}$ with isolation variables $I < 0.1, 0.2$ and $1.3\,\text{GeV}/p_T + 0.005$ for isolated electron, muon and photon respectively
6. no isolated track with $m_T = \sqrt{2p_T^{\text{track}} p_T^{\text{miss}}(1 - \cos(\phi^{\text{miss}} - \phi^{\text{track}})} < 100\,\text{GeV}$ and $p_T > 5\,\text{GeV}$ for tracks identified as an electron/muon or else 10 GeV.

Selection	W	Z	$t\bar{t}$
Baseline selection	73790	25725	7906
$m_{J_1} \in [70\,\text{GeV}, 100\,\text{GeV}]$	5936	2401	1320
CMS-SUS-19-013 [119] signal region	420	237	153

Table 4.1: Number of events passing each selection-requirement for $\mathcal{L}_{\text{int}} = 300\,\text{fb}^{-1}$

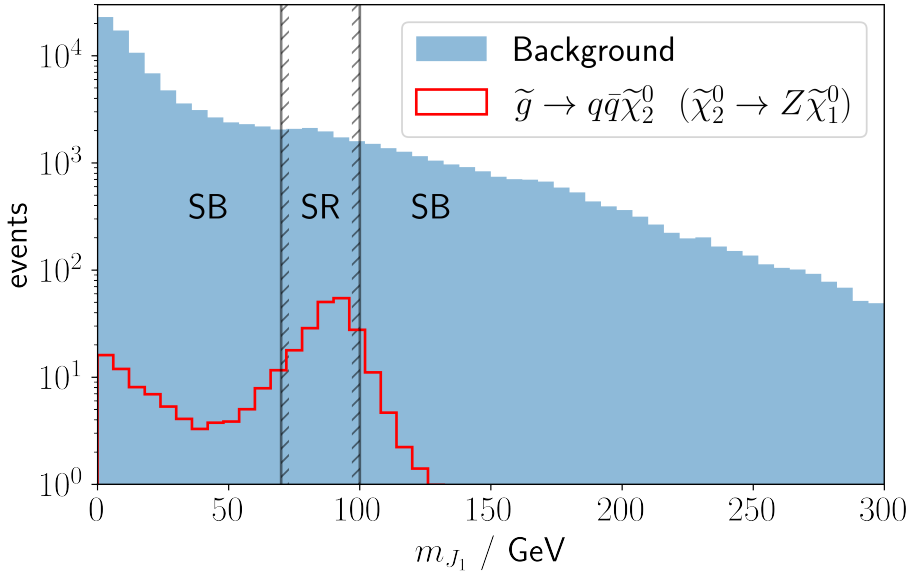

Figure 4.3: Distribution of the resonant feature m_{J_1} for background and signal events in the SB and the SR. The signal corresponds to $m_{\tilde{g}} = 1\,700\,\text{GeV}$. The distributions are scaled to $\mathcal{L}_{\text{int}} = 300\,\text{fb}^{-1}$.

7. at least two AK8 jets with $p_T > 200\,\text{GeV}$

The number of background events that pass this baseline selection is shown in the first line of table 4.1. In total, the dataset is composed of 107,421 background events corresponding to $\mathcal{L}_{\text{int}} = 300\,\text{fb}^{-1}$ after cuts 1-7. Signal events are injected according to the gluino-pair production cross section.

Figure 4.3 shows that the feature m_{J_1} is smooth for the background, while it is resonant for the signal. (Hadronically decaying W's and $Z's$ are eliminated by the requirements on p_T^{miss}.) This is a necessary feature for the application of the CATHODE method employed in Section 4.4. Figures 4.4 and 4.5 show that the signal of new physics is found on the tail of the p_T^{miss}-distribution, while the background peaks at lower p_T^{miss}. We will show that the powerful discriminator p_T^{miss} can be leveraged by CATHODE even though the signal is found on the tail of the distribution.

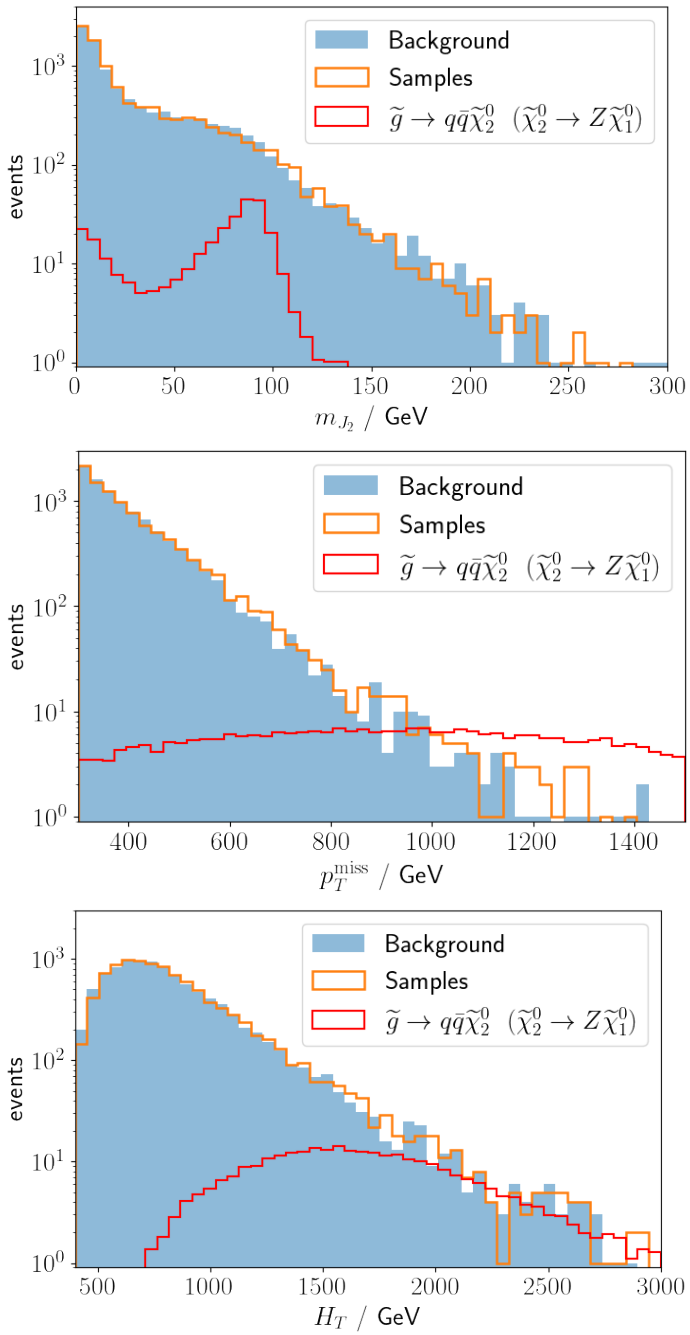

Figure 4.4: Comparison of the signal and background distribution inside the signal region and the artificial samples. The artificial samples will be discussed in the next section. The signal corresponds to $m_{\tilde{g}} = 1\,700\,\text{GeV}$. The distributions are scaled to $\mathcal{L}_{\text{int}} = 300\,\text{fb}^{-1}$.

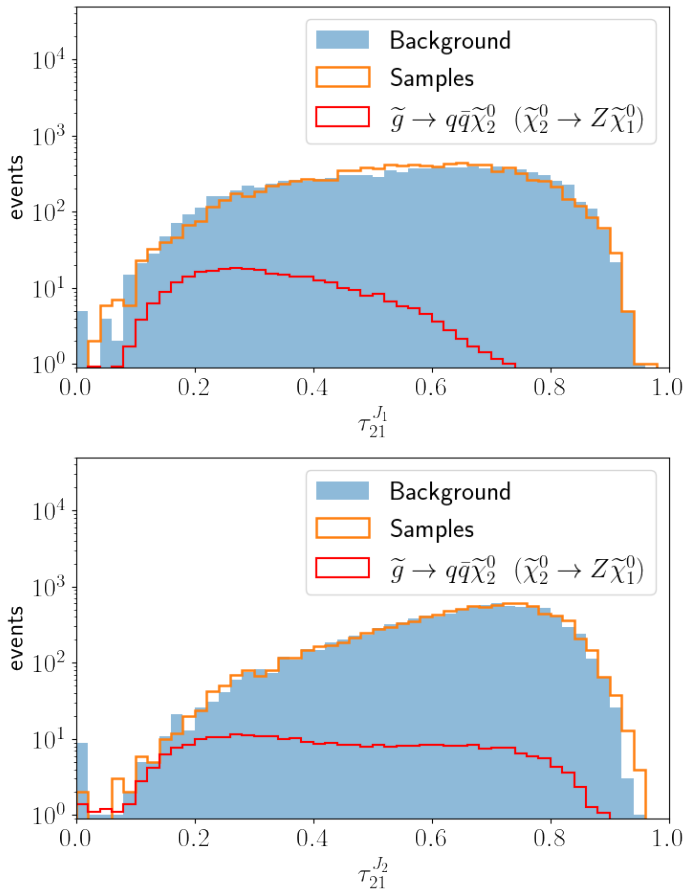

Figure 4.5: Comparison of the signal and background distribution inside the signal region and the artificial samples. The artificial samples will be discussed in the next section. The signal corresponds to $m_{\tilde{g}} = 1\,700\,\text{GeV}$. The distributions are scaled to $\mathcal{L}_{\text{int}} = 300\,\text{fb}^{-1}$.

4.4 CATHODE

Here we recap the main points of the inner workings of Classifying Anomalies THrough Outer Density Estimation CATHODE (for more detail see [8]). In this study, the events are represented as the tuple m_{J_1} and x with

$$x = \left(m_{J_2}, p_T^{\text{miss}}, H_T, \tau_{21}^{J_1}, \tau_{21}^{J_2}\right) , \tag{4.5}$$

where J_1, J_2 are the leading/subleading AK8 jets and $\tau_{21} = \tau_2/\tau_1$ is the ratio of n-subjettiness variables [127]. To compare the technique to the classical search more directly, we also consider the reduced set of features

$$x = \left(m_{J_2}, p_T^{\text{miss}}, H_T\right) \tag{4.6}$$

so that CATHODE only gets to use the same information. We use a slightly modified version of the original repository[2] to allow for any dimension for x.

4.4.1 Data Preparation and Density Estimation

First, one defines the SR as an interval in m_{J_1} where the signal is expected to be concentrated, similar to a classical bump hunt. The complement of the SR defines the SB. As in any bump hunt, the SR window has to account for the position and the width of the signal bump. Because the reconstructed jet mass is not distributed symmetrically around the mass m of the mother particle (which is the Z, the Higgs or a BSM Higgs here) we chose the parameterization

$$m_{J_1} \in [m\left(1 - \frac{4}{3}\sigma_m\right), m\left(1 + \frac{2}{3}\sigma_m\right)] . \tag{4.7}$$

We estimate the mass resolution to $\sigma_{\text{m}} = 15\%$ and round the window to the closest GeV. The lower sideband extends to $m_{J_1} = 0$ while the upper sideband is only limited by the phase space.

Events in the SB are partitioned into a training set (75%) used for the actual training and a validation set (25%), used to select the models used in the next steps. To address the finite number of real SB events, we use four-fold cross-validation such that we get four datasets with non-overlapping validation sets. The data is transformed (preprocessed) for easier learning by shifting and scaling the observables in x to fit the interval $(0, 1)$, then applying a logit tranformation[3], and again shifting and scaling to unit standard deviation and zero mean.

For density estimation, a MAF is used with affine transformations [81], as described in Section 3.7.2. The MAF constructs invertible transformations with tractable Jacobians that map a simple multidimensional distribution (e.g., multiple Gaussians as is considered here) to the target density, in this case, the conditional probability $p_{\text{data}}(x|m_{J_1} \in SB)$. The MAF uses 15 MADE [82] blocks to learn the transformations. The number of events it is trained on depends on the signal region but is typically of the order of 10^5 for $\mathcal{L} = 300\,\text{fb}^{-1}$.Training is done with the hyperparameters listed in table 4.2.

After training, model states at the ten epochs with the lowest validation loss are selected for the sampling step.

4.4.2 Sampling SR Events

The next step aims to sample synthetic events inside the SR using the four density estimates of the last step. KDE with a Gaussian kernel and a bandwidth of 0.01 is used to model the m_{J_1} distribution inside the SR. This is then used to sample $N = 1\,000$ events from each of the ten density-estimator-model-states, which are combined, shuffled and split between the training set (60%) and validation set (40%) for the next step. The training and validation sets of all four density estimators are combined,

[2] https://github.com/HEPML-AnomalyDetection/CATHODE

[3] $\text{logit}(x) = \ln \frac{x}{1-x}$

Hyperparameter	Value
`optimizer`	Adam
`epochs`	100
`learning_rate`	10^{-4}
`batch_norm`	`true`
`batch_norm_momentum`	1
`batch_size`	256

Table 4.2: Parameters of the density estimator

respectively, to form the synthetic dataset with a total of 40 000 events. Compared to the roughly 10 000 real events in the SR (see table 4.1 second line) this is intentionally oversampled to improve the classification performance [8], as we will show in the following section. The synthetic background events and the real SR events are then standardized in the SR without the logit transformation.

The distributions of the synthetic events are shown in orange in figures 4.4 and 4.5. In all our models, the signal is located in a resonance in m_{J_2} and in the tail of the p_T^{miss} distribution. The density estimation has to model the shape reasonably well so that this powerful classification feature can be leveraged. This is accomplished successfully, as shown in figure 4.4 and 4.5.

4.4.3 Classifier and Anomaly Detection

Now a classifier is trained on both the synthetic and real SR datasets to distinguish the sampled events, which should follow the background distribution, from the real events, which additionally might contain events following the signal distribution.

The classifier is a fully-connected neural network, which consists of 3 hidden layers with 64 nodes and ReLU activation each, and it is optimized using the hyperparameters given in table 4.3. Because the datasets are imbalanced between artificial and real data, a weight is assigned such that both classes contribute equally to the loss.

Since, in a realistic example, the number of events to train and validate on is limited, we employ an additional step of four-fold cross-validation. The real SR data is partitioned into four subsets of equal size. In each subset, one-quarter of the real events are held back as a test set for anomaly detection, while the remaining 75% are split between the training set (60%) and the validation set (40%). The synthetic background events are also split into train/val sets with the same proportions. After training, the ten model states with the lowest validation loss are selected and evaluated on the test set. The predicted labels are then averaged over the models and assigned as anomaly scores to the events. This is repeated for the next quarter of the SR data, and so on, until every event in the SR is assigned an anomaly score.

To reduce the statistical effects of severely over- and under-performing models, each dataset is

Hyperparameter	Value
`optimizer`	Adam
`epochs`	100
`learning_rate`	10^{-3}
`batch_size`	128

Table 4.3: Parameters of the classifier

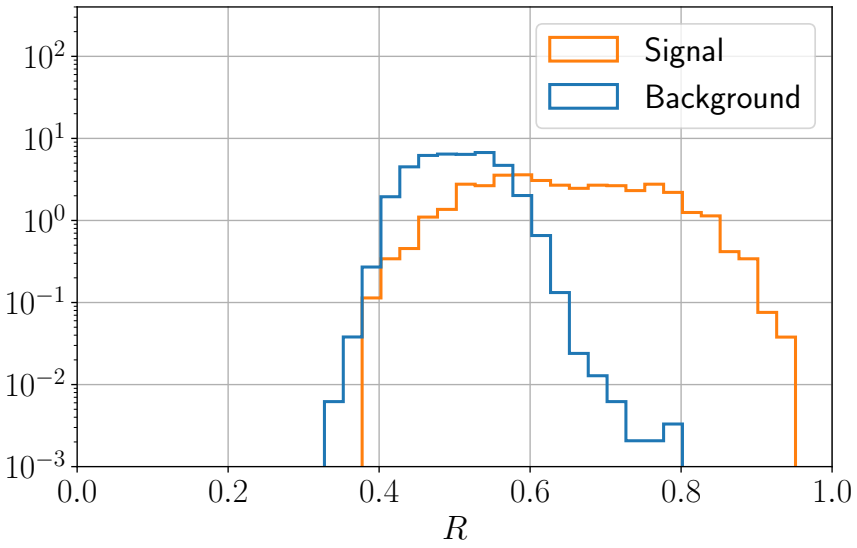

Figure 4.6: Normalized distributions of the anomaly score R of the signal and background processes. The signal corresponds to the average distribution of ten independent injections with $m_{\tilde{g}} = 1\,900\,\text{GeV}$.

shuffled 5 times to allow different selections. Then the entire process of the preceding paragraph is repeated to produce 5 different anomaly scores. All 5 anomaly score assignments are averaged to produce a final, more robust score.

Finally, to even out the influence of signal-event selection, everything is repeated ten times with differing independent sets of signal-events. In all the results we report below, we will report the mean and standard deviation of these ten different trials.

The signal-to-background ratio is improved by cutting on the anomaly score above a critical value R_c. Figure 4.6 shows the distributions of the anomaly score R for the signal and background. No additional selections are performed.

In a real application, one would also evaluate the anomaly score on the sideband. Since the classifier is only trained on the x features, it should be applicable for every resonant feature value, whether inside the SR or the SB. For this, an additional subset of SB events has to be set aside, similarly to the SR strategy. This is only possible if x and m are fairly uncorrelated, which has to be validated.

For any given R_c a smooth function is either fit only to the SB or the entire m space, which models the expected background distribution. For this, CATHODE mustn't sculpt features in the otherwise smooth background distribution in m, which would be validated, e.g. on simulated background events. Standard statistical inference techniques can be used from this point onward to construct p-values. However, this is beyond the scope of this work. Instead, the performance is evaluated using the nominal significance

$$Z = S/\sqrt{B} \tag{4.8}$$

with S (B) the number of signal (background) events after imposing this cut. This makes use of the truth labels which an experiment would have to replace by other means of background estimation. One still has to choose a strategy to set R_c. In the following, we will show the signal significance with R_c set to maximize Z with at least 5 background events left to show the best performance one could hope for. Since a real application does not have access to the truth labels, this is not immediately applicable. To show a more realistic method we also show the performance where R_c is set so that 1% of SR-events pass the cut while also containing at least 5 background events.

4.5 Results

4.5.1 Nominal Signal Model

We first turn our attention to the nominal signal model with the decay $\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0$. This is the signal model the dedicated CMS search [119] was aimed at.

Hyperparameter N_{Sample}

First, we show that $N_{\text{Sample}} = 10000$ artificial events sampled from each of the four density estimators is indeed a sensible choice. Note, that N_{Sample} is the number of samples per density estimator, of which there are four. We fix the gluino mass to $m_{\widetilde{g}} = 1\,700\,\text{GeV}$ while using the full feature set shown in equation 4.5. We scan N_{Sample} from 40 to $4 \cdot 10^6$. Class weights are used to scale the total weight of the real events and the synthetic events to be the same, mending the class imbalance. This is done by weighting the loss per sample by the class weight. This way, we are free to choose N_{Sample} without biasing the classifier towards a single class. The results are shown in figure 4.7. We see that the naive choice of the same number of artificial samples as there are real samples is suboptimal. The value of $N_{\text{Sample}} = 10\,000$ lies at the beginning of a plateau. Larger values do not improve the anomaly detection performance meaningfully but come at a cost of substantially longer training duration. The factor of four between real data and artificial samples is similar to the point that saturates CATHODEs performance in the original publication on the LHC-Olympics dataset [8].

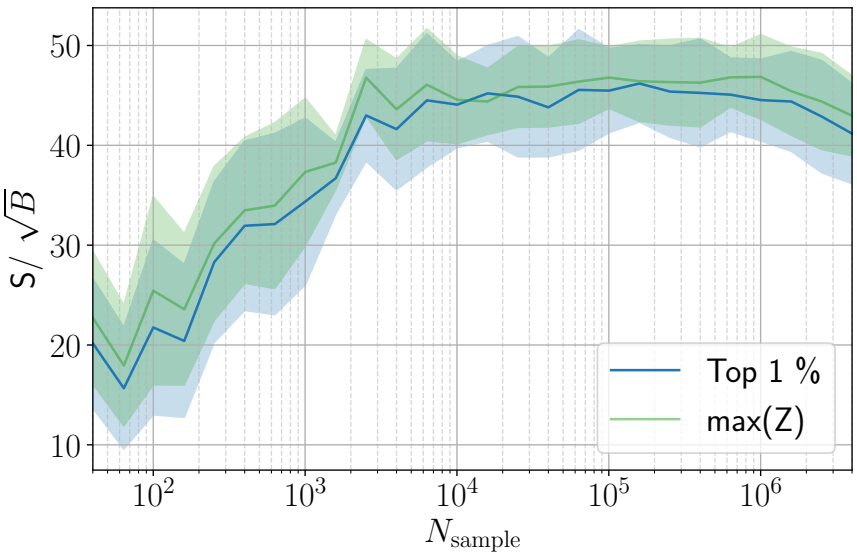

Figure 4.7: Dependence of $S/\sqrt{B}$ on N_{Sample}. The signal model is a 1 700 GeV gluino.

Three Features

We continue by using the limited feature set $x = \left(m_{J_2}, p_T^{\text{miss}}, H_T\right)$ so CATHODE does not have access to more information than the classical search. To compare with CMS we calculate the signal significance for events inside the signal region $m_{J_1/J_2} \in [70\,\text{GeV}, 100\,\text{GeV}]$ with the b-veto mentioned in Appendix A.1 applied. Since the search gets most of its sensitivity from the highest p_T^{miss}-bins we apply an additional cut $p_T^{\text{miss}} > 800\,\text{GeV}$.[4] This leads to roughly the same number of events as when only the top 1% of events are kept for CATHODE. For a gluino-mass with a sizeable cross section like 1 700 GeV, the classical search yields on average for ten independent signal injections Z = 20. Using CATHODE with three features, the significance is on average Z = 34 ± 2.

CATHODE outperforms the classical approach, even though CATHODE is more model-agnostic. The reason is that the classical approach, being cut-based, misses correlations between the features that the multivariate classifier of CATHODE can pick up.

To confirm this, we also investigated the sensitivity of a fully supervised approach, using the same classifier architecture and hyperparameters as that of CATHODE. The training data for the fully supervised classifier consists of an additional 300 fb^{-1} background events and 10,000 signal events. 60% of this dataset is used in training, while the remaining 40% is used as a validation set to select the best-performing model. Evaluating this classifier again by selecting only the top 1% of anomaly scores results in a significance of on average Z = 33 ± 4. We conclude that CATHODE is saturating the performance of the fully supervised classifier for this amount of signal (unsurprisingly, since this

[4] Technically, the original CMS search uses p_T^{miss}-bins, and most of the sensitivity comes from the three highest bins, 800–1000 GeV, 1000–1200 GeV and larger than 1200 GeV, where the background is comparable or subdominant to the signal hypothesis. To get a fair comparison with CATHODE, we replace this with a single cut.

is a lot of signal), and that the deep neural networks of both CATHODE's classifier and the supervised classifier can leverage correlations to improve the signal significance significantly over the classical approach.

Five Features

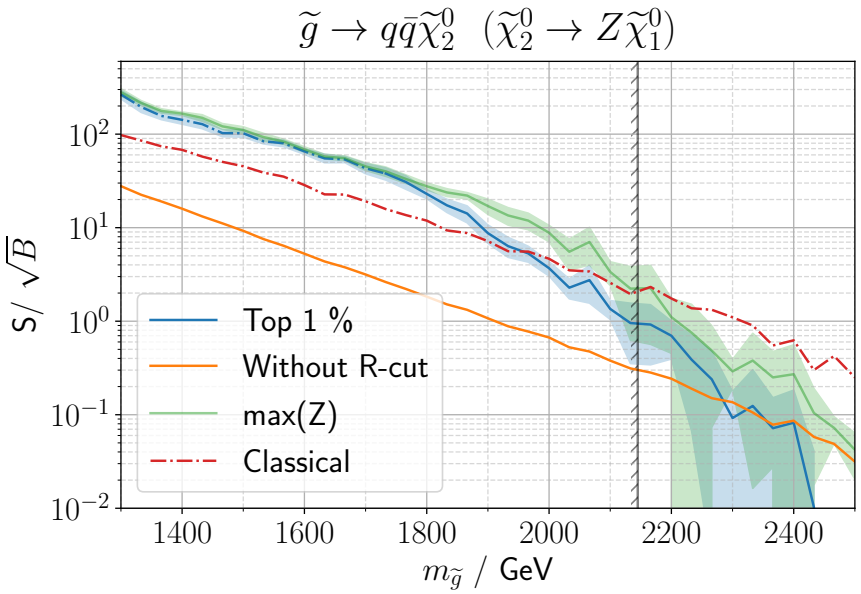

Figure 4.8: Sensitivity of CATHODE and the classical strategy. The signal window is set as $m_{J_1} \in [70\,\text{GeV}, 100\,\text{GeV}]$. For the blue line, R_c is set to allow 1% of events to pass this cut while the orange line omits the cut completely. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections. The dot-dashed part of the blue line represents parameter points where R_c has to be lowered to allow 5 background events. The vertical black line at 2 145 GeV indicates gluino-mass that is excluded at 95% confidence level by our 300 fb^{-1} recreation of the dedicated search [119]. The red dot-dashed line is calculated using the classical strategy with $m_{J_1/J_2} \in [70\,\text{GeV}, 100\,\text{GeV}]$, $p_T^{\text{miss}} > 800\,\text{GeV}$ and the b-veto.

From now on, we will use the five features $\left(m_{J_2}, p_T^{\text{miss}}, H_T, \tau_{21}^{J_1}, \tau_{21}^{J_2}\right)$ because the subjettiness-variables τ_{21} are useful discriminants. Figure 4.8 shows CATHODE's performance compared to the classical strategy. We see that in the relevant region at high gluino masses, the conservative cut on R (allowing only the top 1% to pass) reaches only slightly weaker results. We identify the mass where the signal significance is Z = 1.645 [5] with the expected 95% limit on the mass in a real application [128]. The conservative cut on R alone excludes gluino masses up to $m_{\widetilde{g}} = 2\,066\,\text{GeV}$. This is only slightly weaker than the expected excluded mass of $m_{\widetilde{g}} < 2\,145\,\text{GeV}$ for a dedicated search at this integrated luminosity. This is expected because a model-specific search will be fine-tuned to the specific process, while CATHODE is intentionally kept more general. CATHODE's strength lies in this generalization,

[5] The 95% one-sided normal quantile

as it can detect different models without the need to tweak the approach, as we will show in the following sections.

4.5.2 Alternate Signal Model: Decays to SM Higgs

Now we turn our attention to another model, where the neutralinos decay via $\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0$ where h is the 125 GeV Standard Model Higgs boson. All that has to be done for CATHODE is to select a new signal window around 125 GeV. A scan over the gluino-mass is shown in figure 4.9. A b-jet selection criterion would be beneficial in this case, but we omit this to keep CATHODE as general as possible. Even without the b-tag, CATHODE still generates a sizable signal-significance for gluino masses comparable to the expected excluded value. While the dedicated search is expected to exclude gluino masses below 2 355 GeV, CATHODE with the 1% cut reaches Z > 1.645 for all masses up to 2 233 GeV. With the best possible cut on R, this can be pushed to 2 300 GeV. As expected, CATHODE results in slightly weaker bounds. The opportunity cost of this is significantly lower than a specialized search. The only change in the approach is the choice of the signal region. The intended use of CATHODE scans the signal region over the entire mass range, such that both the decay to Z and Higgs bosons would be included automatically in this strategy without any extra considerations.

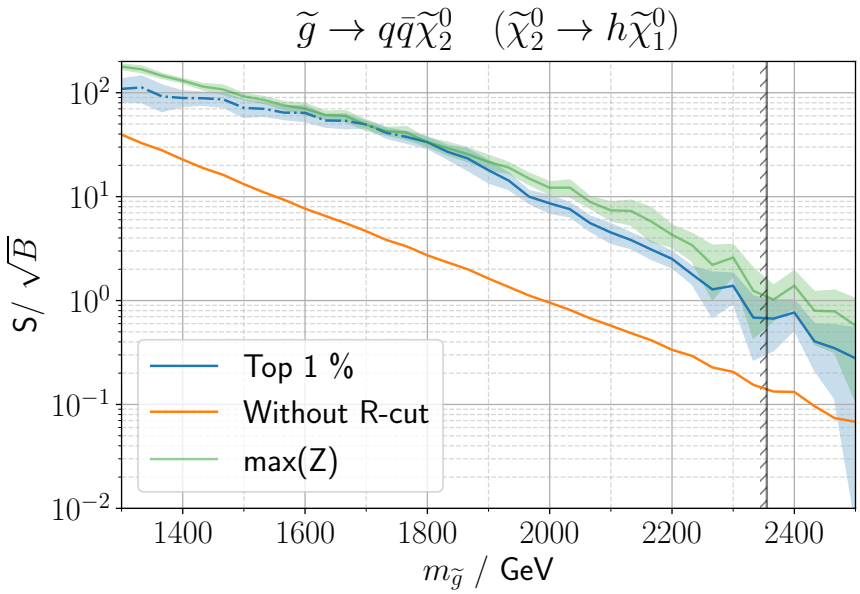

Figure 4.9: CATHODE's performance for $\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0$. The signal window is set as $m_{J_1} \in [100\,\text{GeV}, 140\,\text{GeV}]$. For the blue line, R_c is set to allow 1% of events to pass this cut while the orange line omits the cut completely. The dot-dashed part of the blue line represents parameter points where R_c has to be lowered to allow 5 background events. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections. The vertical black line at 2 355 GeV indicates gluino-mass that is expected to be excluded by rescaling the (expected) limit from a dedicated CMS search for this decay [129] from 137 fb^{-1} to 300 fb^{-1} integrated luminosity. There is no red line corresponding to the classical search (as in Fig. 4.8) because we did not perform a detailed recast of [129].

4.5.3 Alternate Signal Model: Mixed Z/h Decays

Setting the branching ratio of the $\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0$ or $\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0$ decays to 100% is a rather unnatural choice. Therefore, we also show CATHODE's performance for a model where both branching ratios are 50%. This time, the anomaly-detection has to find two bumps simultaneously. For this, we chose the signal window to contain both resonances: $m_{J_1} \in [70\,\text{GeV}, 140\,\text{GeV}]$. The results of a scan over the gluino masses are shown in figure 4.10. This time CATHODE seems to outperform the extrapolated bound from the dedicated search [130]. The extrapolation from 35.9 fb^{-1} to 300 fb^{-1} integrated luminosity is quite far and should be interpreted with care. The dedicated search classifies events in 0,1 and 2 Higgs categories using b-tags. The signal model populates all categories simultaneously. The approach using CATHODE only uses a single signal region without further input being necessary to generate these results.

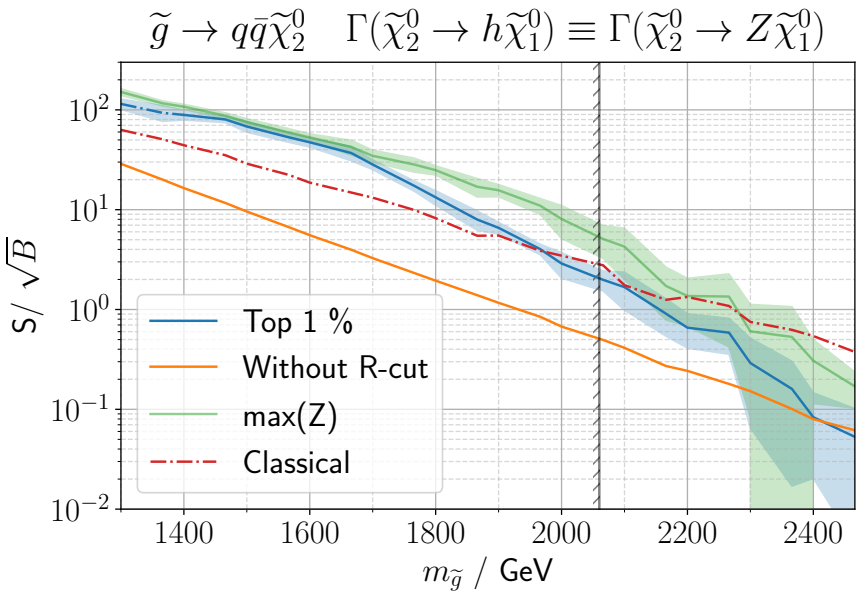

Figure 4.10: Sensitivity of CATHODE and the classical strategy. The signal window is set as $m_{J_1} \in [70\,\text{GeV}, 140\,\text{GeV}]$. For the blue line, R_c is set to allow 1% of events to pass this cut while the orange line omits the cut completely. The dot-dashed part of the blue line represents parameter points where R_c has to be lowered to allow 5 background events. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections. The vertical black line at 2 060 GeV indicates gluino-mass that is expected to be excluded by rescaling the expected excluded cross section obtained by the dedicated CMS search for this decay [130] from 35.9 fb^{-1} to 300 fb^{-1} integrated luminosity.

In figure 4.11 we show that CATHODE is indeed capable of recovering both bumps corresponding to the decay into Z and Higgs bosons respectively.

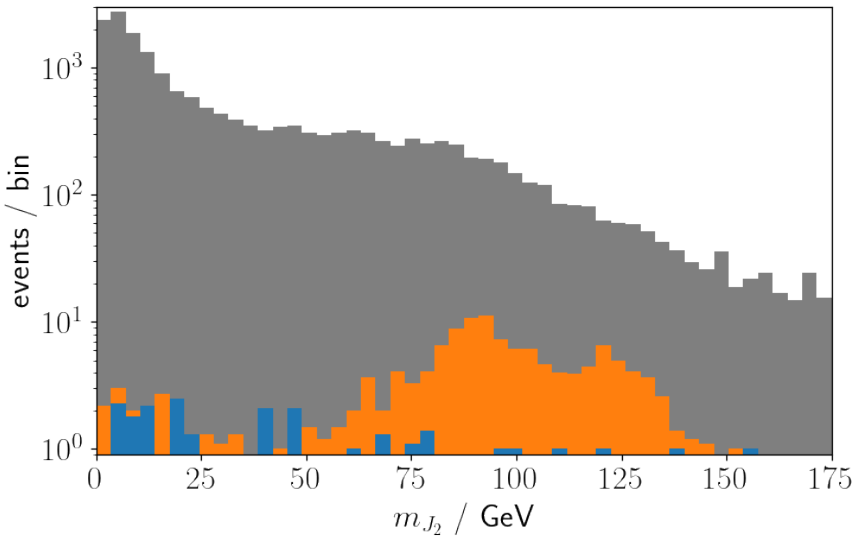

Figure 4.11: The distribution of the data inside the signal region before the anomaly score cut is shown in gray. After selecting the top 1% of events in the SR the remaining signal events are shown in orange while the remaining background events are shown in blue. The signal corresponds to $m_{\tilde{g}} = 1\,700\,\text{GeV}$.

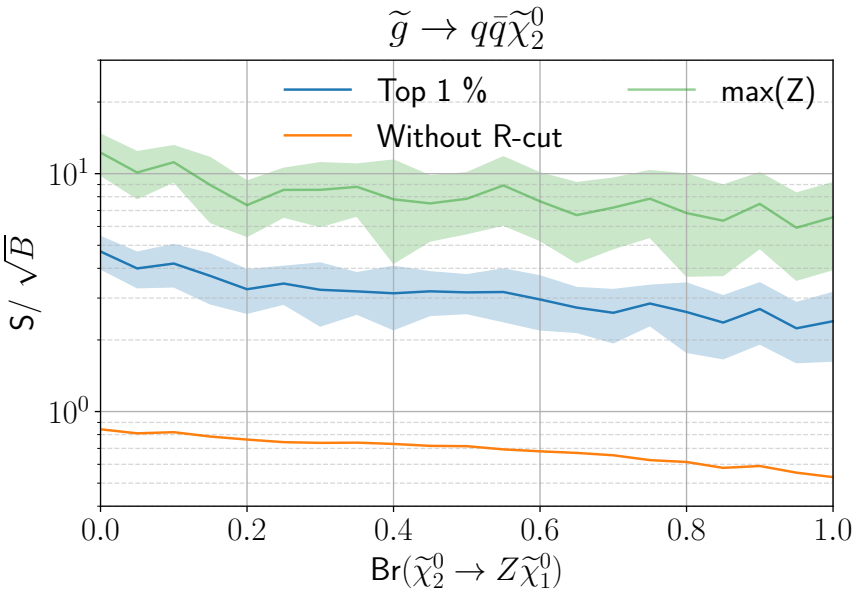

Figure 4.12: Sensitivity of CATHODE for varying branching ratios to Z bosons for $m_{\tilde{g}} = 2\,000\,\text{GeV}$. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections.

Figure 4.12 shows that CATHODE is very robust against changes in branching ratios. We vary the branching ratio $\text{Br}(\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0)$ with $\text{Br}(\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0) = 1 - \text{Br}(\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0)$ and calculate the significance. Regardless of the branching ratio, the multiplicative gain of significance by applying the

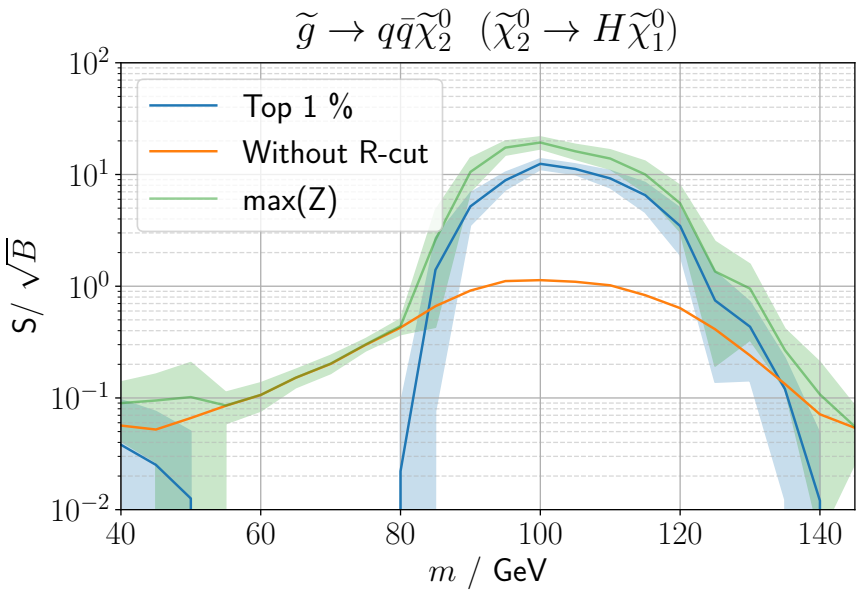

Figure 4.13: Significance for a parameter scan over the mass hypothesis in steps of 5 GeV, when the mass is not known a priori. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections. Masses are chosen as $m_{\widetilde{g}} = 2\,000$ GeV and $m_H = 100$ GeV.

technique is always between 5 and 6. This shows the real strength of the CATHODE approach over the dedicated searches [119, 129, 130]. With the enlarged SR that covers both decay modes, CATHODE only needs to be trained once, independent of the assumption on the BRs, compared to performing a dedicated analysis for each BR assumption.

4.5.4 Alternate Signal Model: Decays to BSM Higgs

Until now, we applied CATHODE only to models where the position of the bump is known beforehand. But one strength of the technique is that we don't even need to know that. To discuss this further we now focus on another model that induces the neutralino decay $\widetilde{\chi}_2^0 \to H\widetilde{\chi}_1^0$ where H is one of the additional Higgs-bosons introduced by the (N)MSSM that has a mass different from 125 GeV. Because the decay of H depends on the specific implementation of SUSY-breaking parameters, we set the branching ratio $BR(H \to b\bar{b}) = 100\%$. To find the signal, CATHODE is applied to different signal regions given by varying mass hypotheses m in equation 4.7, scanning the entire mass range in discrete steps and the signal significance is determined. To demonstrate this, we chose $m_H = 100$ GeV and $m_{\widetilde{g}} = 2\,000$ GeV and show the result in figure 4.13. Once the signal window has a significant overlap with the signal-bump, the signal significance gets sufficiently improved to show the presence of anomalous events. In a real application, this would then warrant further investigation with a dedicated search.

Finally, we show how wide the possible choice of m_H is that CATHODE can still help to find in our dataset with the given choice of features. For this purpose, we perform a parameter-scan over m_H

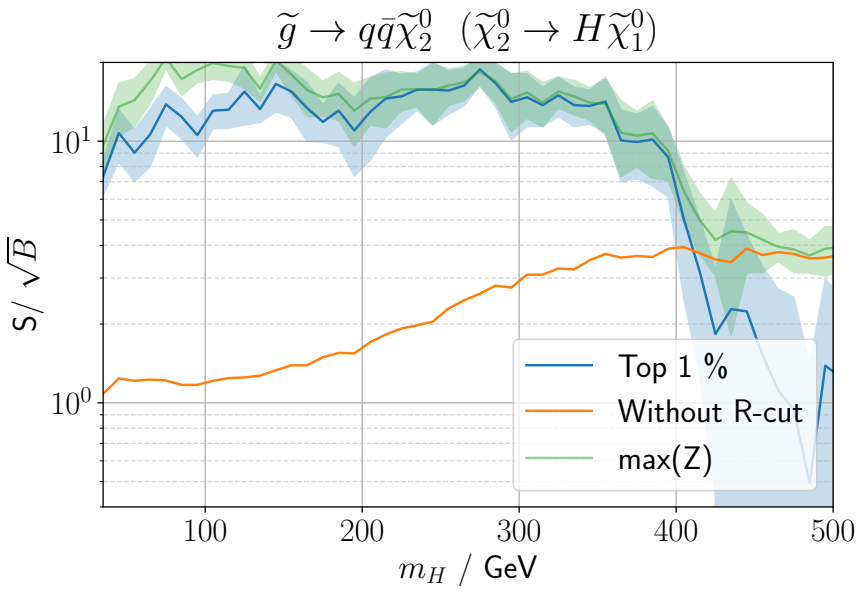

Figure 4.14: Parameter-scan of m_H with $m_{\widetilde{g}} = 2\,000$ GeV to show which signals CATHODE can help find in the dataset. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections.

from 35 GeV to 515 GeV in steps of 10 GeV, as shown in figure 4.14. The method reliably reaches signal significances of order ten up to $m_H \sim 350$ GeV without using b-tags as otherwise powerful discriminators.

4.6 Summary

In this chapter, we have shown how recently developed techniques for weakly-supervised resonant anomaly detection can be easily extended to cover anomalies that also are located at the tails of distributions. This situation commonly arises in well-motivated weak-scale scenarios such as SUSY, where the cascade decays of heavier BSM particles can produce resonances such as Z's and Higgs bosons while simultaneously populating the tails of features such as p_T^{miss} and H_T. As long as the signal is localized in one feature where the background is smooth, resonant anomaly detection can be brought to bear on these additional features in order to enhance the sensitivity to the signal.

As a proof-of-concept demonstration, we applied the state-of-the-art anomaly detection method CATHODE [8] to the SUSY scenario $pp \to \widetilde{g}\widetilde{g}, \widetilde{g} \to q\bar{q}\widetilde{\chi}_2^0, \widetilde{\chi}_2^0 \to X\widetilde{\chi}_1^0$ where X is either a Z boson, Standard Model Higgs, or an additional (N)MSSM Higgs boson. Despite being model-agnostic, we showed that the CATHODE method is competitive with existing, dedicated, cut-based searches [119, 129, 130], because — being inherently multivariate — it takes advantage of correlations between features. Moreover, whereas each decay scenario required a separate, optimized analysis, CATHODE — being model-agnostic — is able to simultaneously target them all.

We considered two different feature sets for the CATHODE algorithm, as shown in eqs. (4.5) and (4.6). These were motivated by the SUSY scenarios we considered, and it would be interesting to generalize our study beyond these feature sets, both to increase the degree of model-agnosticism of the method and possibly to enhance the sensitivity to the SUSY signals considered here. For example, our benchmark signals all come with ~ 4 additional jets from the gluino decay, and their detailed kinematic distributions (instead of just the aggregate feature H_T) may offer additional discriminating power versus the QCD background. Adding features related to additional jets in the event may also give us more sensitivity to spectra not explicitly considered here, for example, where the NLSP mass is not so close to the gluino mass. As long as $m_{LSP} + m_Z \ll m_{\tilde{g}}$, the Z will still be boosted, but the extra jets will get harder as m_{LSP} moves away from $m_{\tilde{g}}$.

When trying to incorporate more features into our approach, we noticed that the performance decreased as the classifier became more prone to overfit. Each additional feature has to contribute an overhead of additional discriminating information to cancel this decrease before it improves the performance. Ideally, one would use a large and therefore general feature set to ensure that many possible signal models produce discriminating differences in the distributions. Necessarily, this implies that some features are distributed similarly in background and signal processes and are consequently irrelevant to amplifying the signal. This aspect of the CATHODE approach has been studied while our work was in progress. To make CATHODE more robust against irrelevant features, it was shown to be beneficial to replace the fully connected neural network in the classification step with a gradient-boosted decision tree [131]. As we covered previously in Chapter 3.3, Gradient Boosted Decision Tree (GBDT)s can be built to only add splits along a feature if the loss decreases by a minimum amount. Therefore, irrelevant features will most often be ignored, as long as the GBDT does not overfit. The performance itself is relatively insensitive to the replacement when only informative features are selected for the LHC Olympics dataset.

Although the performance studies on simulated data are promising, this might not carry over to real collision data. Applying the statistical inference machinery after anomaly scores have been assigned is non-trivial. Instead of simply calculating $S/\sqrt{B}$ with truth level information, one needs to calculate p-values another way; for example, a one-dimensional bump hunt after a cut on the anomaly score R. This might suffer systematic errors if the cut sculpts the background inside the signal region, therefore falsely inferring a bump. Background sculpting due to correlations between m and R can be avoided when the anomaly score and thus its cut is calculated in the latent space of the normalizing flow [98], i.e., where the background samples follow a multidimensional normal distribution. Since here all samples follow the same distribution, independent of the conditional m, a cut does not lead to a sculpted background in m.

Even though the approach is as model-independent as possible, one would still derive exclusion limits for specific models, given that no significant deviation from the background-only hypothesis is found. In contrast to classical cut-based searches, where the signal efficiency ϵ_S is often only given by the selection cuts, this will be more intricate here. Since the classifier only learns to differentiate the artificial samples from the signal region if a signal is present, its performance depends on the amount

of signal present. Hence, the signal efficiency is a function of the signal production cross section, i.e. $\epsilon_S(\sigma_S)$. To set an upper limit on σ_S, one therefore needs to inject signal events into the real collision data and retrain CATHODE for different numbers of injected signals. One then needs to find the number of signal events that result in the desired p-value.

Among other anomaly detection methods, CATHODE has already been applied to real data by CMS[132]. This shows that resonant anomaly detection is not a mock analysis-only exercise, but also sparks the interest of the experimentalist community.

CHAPTER 5

Learning to see R-parity violating scalar top decays

The results of this chapter have been published at:
Learning to see R-parity violating scalar top decays
Gerrit Bickendorf, Manuel Drees
Phys. Rev. D 110, 056006 - Published 4 Sep 2024

Although resonant anomaly detection methods show impressive results, they are not the ultimate technique applicable in every scenario. While constructing the auxiliary features x for the CATHODE method is relatively straightforward, finding a suitable m is not always easy. A signal region has to be relatively narrow to avoid large errors from far interpolations of the conditional feature m by the normalizing flow when m and x are slightly correlated. The classifier will simply pick up these inaccuracies and therefore universally correctly classify synthetic events without amplifying only the signal events. Also, a strong correlation between m and x degrades the anomaly detection performance, even though the density estimation step improves this degradation compared to directly applying CWoLa hunting. This has to be considered when constructing suitable input features. Also, m has to be smoothly distributed for the background as the core concept of CATHODE. When a Standard Model resonance is present in the interesting region of m and it cannot be removed by preselection cuts (as was the case in the previous chapter, where the p_T^{miss} requirement eliminated hadronic Z and W boson decays) this will not work as intended. If the Standard Model resonance is broad, this can invalidate the technique on a large chunk of m parameter space. This will be precisely the setting we will encounter next.

For this, we will stay within the MSSM but drop the R-parity conservation, which otherwise leads to stable LSPs that leave the experiment undetected, which in turn can be leveraged in searches by exploiting the large resulting p_T^{miss} [133], just as in the previous chapter.

Once the RPC assumption is dropped, searches involving large p_T^{miss} often become insensitive. As covered in Section 2.3.5, in the context of the R-parity violating (RPV) MSSM, new terms are added to the superpotential that break lepton- or baryon-number conservation [134]. These additional terms imply that drastically different search strategies are needed [135], especially when prompt hadronic decays of supersymmetric particles become drowned by the hadronic activity inside the detector.

At a hadron collider like the LHC, the production of strongly interacting superparticles has the largest cross section for a given mass. Among these, the stops – the superpartners of the top quark – are often assumed to be the lightest. On the one hand, for equal squark masses at some very high (e.g. Grand Unified or Planckian) energy scale, renormalization group effects reduce the masses of the stops; mixing between the masses of the $SU(2)$ doublet and singlet stops will reduce the mass of the lightest eigenstate even more [3]. On the other hand, simple naturalness arguments [136–138] prefer stop squarks to be not too heavy, but allow much heavier first and second-generation squarks. This motivates the analysis of scenarios where the mass of the lighter stop squark lies well below those of the other strongly interacting superparticles.

The same naturalness arguments also prefer rather small supersymmetric contributions to the masses of the Higgs bosons. In most (though not all [139]) versions of the MSSM, this implies rather light higgsinos, typically below the stop. Since the mass splitting between the three higgsino-like mass eigenstates is small, they all behave similarly if the LSP is higgsino-like. In particular, in the kind of RPV scenario we consider, all three states would lead to very similar "fat jets" when produced in stop decays; the recognition of such jets by exploiting recent developments in computer vision is one of the central points of this chapter, which would apply equally to all three higgsino states. However, about half of all stop decays would then produce a bottom, rather than a top, together with a higgsino, thereby complicating the analysis of the remainder of the final state. Moreover, higgsinos being $SU(2)$ doublets have a sizeable direct production rate. Their non-observation therefore leads to significant constraints on parameter space, especially (but not only) if the bino has mass comparable to or smaller than the higgsinos [129, 140–144].

In order to avoid such complications, we consider the pair production of scalar top quarks, which decay to top quarks plus two neutralinos with unit branching ratio. The neutralinos in turn decay promptly via the UDD $R-$parity breaking term, which is fairly difficult to constrain [145].

This scenario contains multiple resonant features that one might hope to exploit with CATHODE. The obvious choice is the jet mass of one of the hard (fat) jets produced by the neutralinos. This proves futile because the signal process also produces a top quark pair. Therefore, the irreducible background for this process will contain the Standard Model pair-produced top quarks. These will produce a broad hadronic resonance in the jet mass distribution around 172 GeV which hinders CATHODEs application in the immediate vicinity. Furthermore, the hardest jet in the signal model will not necessarily be purely initiated by the neutralino. Since the top quarks recoil against the neutralinos, these will also often produce the hardest jet. Combined with the imperfect reconstruction of all three neutralino-induced partons in a single jet, this leads to signal contamination at jet masses far from the neutralino mass. Therefore, we will leave resonant anomaly detection and tackle this signature using

supervised learning.

Since each neutralino may form a (fat) jet, one can use the substructure to differentiate it from background processes [146]. Let us briefly review some jet-tagging machine-learning algorithms that use the substructure.

5.1 Overview of Jet Tagging Using Machine Learning

Here we give a partial overview of recent jet-tagging architectures. The aim will always be to use the jet substructure to classify it between two or more classes, without the need to construct physics-informed features by hand. In the description, we omit the various preprocessing steps used to make training more efficient as well as implementation details and focus on broad ideas. The interested reader may refer to the original publications.

TopoDNN

The most immediate representation of the jet substructure is in a (e.g., p_T-) ordered list of massless constituents, characterized by the tuple (p_T, η, ϕ) as studied by ref. [147]. One builds a vector of a fixed number of constituents by zero-padding whenever there are fewer constituents in the jet, which can be fed directly to a fully connected neural network. This has been studied by ATLAS [148] using topological clusters [149] (hence the name) for W-boson and top quark tagging, which performed well in suppressing the background.

CNN

One may represent the jets as images, where the spatial dimensions are simply the detector coordinates $\eta - \phi$. These images may be represented with multiple channels (i.e., base colors in normal images) by using calorimeter p_T, potentially split between hadronic and electromagnetic calorimeter, track p_T, number of tracks or muons within each pixel. Convolutional neural networks (CNN) have already been demonstrated to produce good results on these images [150–158]. This technique has been used by ATLAS in electron identification [159] or quark-gluon jet discrimination [160] on simulations. CMS studied a CNN on these images for top-tagging [161] and found a good agreement between the distribution of tagging score on simulated data and real data, demonstrating that the simulation is faithful enough for computer vision application. Similarly, the Deep Underground Neutrino Experiment (DUNE) studied CNNs for tagging neutralino flavors, although the images, in that case, were not in $\eta - \phi$ coordinates but rather in detector wire-number vs. time coordinates, where the *brightness* is the collected charge [162].

Energy Flow Network And Particle Flow Network

Observables at collision experiments can be seen as functions of sets of detected particles. These functions can be generalized and decomposed into [163]

$$O = F\left(\sum_{i}^{M} \Phi(p_i)\right), \tag{5.1}$$

where F and Φ are functions and p_i is a parametrization (e.g. p_T, η, ϕ, particle ID, charge etc.) of the M particles that contribute to O. The function Φ maps each particle separately to the latent space. F maps the permutation-invariant latent space representation onto the target O. The observable might be per event (e.g., H_T) or per reconstructed object, like jet-p_T. One machine learning approach is to represent both functions F and Φ by sufficiently expressive neural networks. This is the intuition of Particle Flow Networks. One can make the observables infrared- and collinear-safe by replacing $\Phi(p_i) \to p_{Ti}\Phi(p_i/E_i)$, which is called the Energy Flow Network. The observable is simply the tagger for jet-tagging where p_i are the jet constituents.

ParticleNet

ParticleNet [164] represents the jet as an unordered set of its constituents, i.e. a cloud. This particle cloud can be represented as a graph with constituents as vertices and edges to the k nearest neighboring vertices. The information contained in this graph can be further processed by a convolution operation called EdgeConv [165]. Given the vertices x_i the EdgeConv operation of ParticleNet can be written as

$$x_i' = \frac{1}{k}\sum_{k} \mathbf{h}(x_i, x_k - x_i), \tag{5.2}$$

where the index k runs over all k-nearest neighbors and $\mathbf{h}$ is represented by a simple MLP. The EdgeConv block takes coordinates and features as input and outputs a transformed cloud with the same number of points. To build ParticleNet, multiple EdgeConv blocks are stacked, where the coordinates and features are the same as the output of the previous block. The first EdgeConv block calculates the k-nearest neighbors using the geometric $\eta - \phi$ distance. After all EdgeConv lavers, the outputs are average-pooled and fed into an MLP for classification. ParticleNet has successfully been used by CMS to constrain the quartic HHVV coupling [166] and for the observation of the $Z \to cc$ decay [167].

LorentzNet

LorentzNet [168] tackles jet tagging from another perspective which is derived from an intuitive argument: If the underlying quantum field theory respects Lorentz-symmetry, why should the tagging architecture not do the same?

Let us denote the Lorentz-invariant inner product between the Lorentz-vectors u and v as $\langle u, v\rangle$ and the Lorentz norm as $||u|| = \sqrt{\langle u, u\rangle}$. We represent the N constituent particles of a jet as the

corresponding four-momentum $x_i = (E_i, p_{xi}, p_{yi}, p_{zi})$ as well as several Lorentz scalars s_i such as the mass. The scalars can be freely passed through a linear layer to the embedded representation h_i without breaking Lorentz symmetry. To build a Lorentz group equivariant block, the authors propose several steps. From the inputs x_i and h_i one builds the Lorentz scalar

$$m_{ij} = \phi_e \left(h_i, h_j, \psi \left(||x_i - x_j|| \right)^2, \psi \left(\langle x_i, x_j \rangle \right) \right) , \tag{5.3}$$

where ϕ_e is an MLP and $\psi(x) = \text{sign}(x) \log(|x| + 1)$.This is used to reweight the Lorentz vectors into a new vector that transforms properly under Lorentz transformations via

$$x_i' = x_i + c \sum_j \phi_x \left(m_{ij} \right) x_i , \tag{5.4}$$

where ϕ_x is another MLP and a Lorentz scalar, and c is a hyperparameter. The scalars get reweighted similarly via

$$h_i' = h_i + \phi_h(h_i, \sum_j \phi_m(m_{ij}) m_{ij}) , \tag{5.5}$$

where ϕ_h and ϕ_m are MLPs with the output of the latter being restricted to $[0, 1]$. Similarly to a transformer, this block is repeated until the desired depth is achieved. Afterward, the h's are average-pooled and fed through a classification MLP. This performs impressively well, as it contains a powerful theoretical inductive bias.

Particle Transformer

The last architecture we will cover here is the Particle Transformer [169] (ParT). This approach again takes the features of all N particles inside the jet as input. Additionally, one constructs interactions, i.e. additional features that depend on the $N \times N$ pairings of particles (e.g., the invariant mass or ΔR). Then, two MLPs are used to embed the features into a $N \times d$-dimensional matrix $\mathbf{x}$ and the interactions into a $N \times N \times d'$ matrix $\mathbf{U}$, which will be held fixed across all layers. The vectors x are fed into the Particle Attention Block, for which the multi-head self-attention mechanism introduced in 3.5 is slightly modified. The attention weights are calculated via

$$\text{P-MHA}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k} + \mathbf{U})V , \tag{5.6}$$

which allows the attention weights to use handcrafted physics-informed interactions without learning them. This block is repeated L times, transforming the matrix $\mathbf{x}$ into a more contextualized state. The authors propose two blocks of Class attention for classification. For this, a class token x_{class} is introduced. Mixing the class token with $\mathbf{x}$ via attention allows it to pick up the information necessary for classification. Only the transformed class token is fed through the last layer, the classification MLP. The ParT has sparked the interest of the ATLAS collaboration and has been studied by them on

simulations and compared to ParticleNet, Energy Flow Network and Particle Flow Network [170].

5.2 Vision Transformers on Jets

In recent years, computer vision techniques have improved drastically with novel approaches such as the vision transformer [70]. In standardized computer vision tasks, these models have been shown to outperform CNN-based models for large datasets [70–72]. Fortunately, generating large sets of simulated events is relatively cheap in particle physics, which motivates the use of these new techniques. As we have covered for the ParT, transformers have already been applied to classification in particle physics scenarios [169, 171–178], although these focus on representing the jet as a set of particles instead of an image.

In this chapter, we for the first time apply two modern transformer-based computer vision techniques to find neutralinos from scalar top quark decays and compare the results to a classical CNN to see if the gain in performance translates to detector images. Using GBDT, we combine the data from both neutralinos tagged in this way and add further high-level features to construct our final event classifier.

5.3 Signal Model

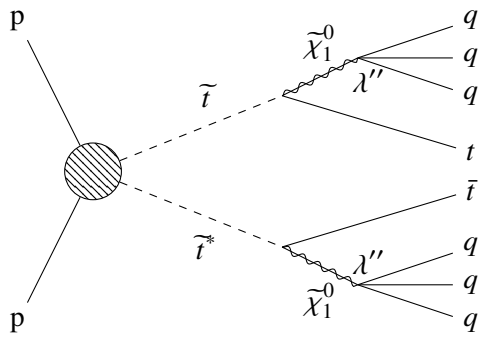

Figure 5.1: Stop pair production with each stop decaying to a top quark and a neutralino. The neutralinos decay via the RPV UDD operator with nonzero λ''.

We will work with breaking parameters, so that the lighter mass eigenstate of the top squark $\tilde{t}_1$ contains mainly the right-handed top squark which decays promptly into a top quark and the bino-like neutralino $\tilde{\chi}_1^0 \equiv \tilde{\chi}$. We consider all other scalar quarks to be decoupled. In order to avoid the constraints of missing E_T-based searches, we add the baryon number violating term shown in equation 2.29 to the superpotential.

When $i = 3$, this would allow the stop to decay directly to two lighter quarks, which has already been extensively studied [179–182]. A coupling with $i \neq 3$ allows even a light neutralino to decay into three quark jets via an off-shell squark. The process we are interested in is shown in figure 5.1.

We also note that a mostly $\tilde{t}_R$ eigenstate decaying into a bino-like neutralino produces a predominantly right-handed top quark. The same is true for a $\tilde{t}_L$ decaying into a neutral higgsino. In contrast, a $\tilde{t}_L$ decaying into a bino or a $\tilde{t}_R$ decaying into a neutral higgsino would produce a left-handed top quark. Since we do not try to reconstruct the polarization of the top (anti)quark in the final state, all four reactions would have very similar signatures and could be treated with the methods developed in this chapter. However, a light neutral higgsino implies the existence of a nearly mass-degenerate charged higgsino (and of a second neutral higgsino), thereby reducing the branching ratio for $\tilde{t} \to t + \tilde{\chi}$ decays. Moreover, by $SU(2)$ invariance, a mostly $\tilde{t}_L$ stop eigenstate would be close in mass to $\tilde{b}_L$, leading to additional signals from $\tilde{b}_L$ pair production. By focusing on a mostly $\tilde{t}_R$ lighter stop and a bino-like LSP we avoid these complications.

5.4 Data Generation and Preselection

For baseline selections, we follow roughly the CMS search for this signal process [183]. We impose the following preselection cuts:

1. One muon with $p_T > 30\,\text{GeV}$ or electron with $p_T > 37\,\text{GeV}$ and $|\eta| < 2.4$.

2. The lepton must be isolated within a cone radius depending on the p_T of the lepton as

$$R = \begin{cases} 0.2 & p_T < 50\,\text{GeV} \\ 10\,\text{GeV}/p_T & 50\,\text{GeV} < p_T < 200\,\text{GeV} \\ 0.05 & p_T > 200\,\text{GeV} \end{cases}$$

   Together with the first cut, this isolation requirement implies that in almost all events the lepton originates from the semileptonic decay of one of the top (anti)quarks in the final state. These two cuts satisfy the requirements of the single lepton trigger. Note that the events must contain exactly one such isolated lepton; this largely removes Z+ jets backgrounds.

3. We define "AK04 jets" via the anti-k_T jet clustering algorithm with distance parameter $R = 0.4$, requiring $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ for each jet. We demand that the event contains at least 7 such AK04 jets, at least one of which is b-tagged. We note that our signal events contain at least two b (anti)quarks from top decay. Moreover, even if both t and $\bar{t}$ decay semi-leptonically, signal events contain 8 energetic quarks even in the absence of QCD radiation. They should therefore pass this cut with high efficiency, except for very light neutralinos where several of their decay products might end up in the same (quite narrow) AK04 jet. On the other hand, SM $t\bar{t}$ events with one top decaying semi-leptonically contain only 4 hard quarks. Hence, at least three additional jets would have to be produced by QCD radiation, significantly reducing the $t\bar{t}$ background, and reducing the W+ jets background even more.

4. $H_T > 300\,\text{GeV}$, where H_T is the scalar sum of the transverse momenta of all AK04 jets. This cut is mostly effective against $W, Z+$ jets backgrounds.

5. At least one combination of b-tagged jet and isolated lepton must have an invariant mass between 50 GeV and 250 GeV. Most events where the lepton and the b quark originate from the decay of the same t quark pass this cut, which helps to further reduce the $W+$ jets background.

6. At least one AK08 jet (defined with distance parameter $R = 0.8$), with $p_T > 100\,\text{GeV}$. We will later try to tag these "fat jets" as coming from neutralino decay. However, a boosted, hadronically decaying top (anti)quark can also produce such a jet. We will also consider even fatter jets. Since (nearly) all particles inside an AK08 jet will end up inside the same jet if $R > 0.8$ is used in the jet clustering, while these fatter jets will contain additional "nearby" particles, they will automatically also have $p_T > 100\,\text{GeV}$.

After these cuts, the remaining background is almost exclusively due to top quark pair production as can be seen in the original CMS publication [183]. In our simulation, we therefore only consider this background process.

For the signal model, we set the masses of squarks (except that of the stop), gluinos, wino- and higgsino-like neutralinos to 5 TeV. We only set one RPV coupling to be nonzero, $\lambda''_{223} = -\lambda''_{232} = 0.75$; this leads to prompt neutralino $\tilde{\chi} \to csb$ decay even if the exchanged squark has a mass of 5 TeV, $\tau_{\tilde{\chi}} \sim 10^{-18}\,\text{s} \cdot [m_{\tilde{\chi}}/(100\,\text{GeV})]^{-5}$. We scan over the stop mass from $m_{\tilde{t}} = 700\,\text{GeV}$ to $m_{\tilde{t}} = 1\,200\,\text{GeV}$ in steps of 25 GeV. We also scan over the neutralino mass from $m_{\tilde{\chi}} = 100\,\text{GeV}$ to $m_{\tilde{\chi}} = 500\,\text{GeV}$ in steps of 10 GeV.

Background and signal events are simulated using MadGraph5_aMC@NLO 3.2.0 [120]. The $t\bar{t}$ background is generated with between 0 and 3 additional matrix element partons while the signal events contain up to 2 additional partons. The NNPDF3.1 PDF-set [184] is used. We use Pythia 8.306 [185] for parton showering and hadronization; background events are showered with the CP5 tune while signal events are showered with the CP2 tune [186]. Events with different matrix element level final state parton multiplicities are merged with the MLM prescription [187], in order to avoid double counting events where the parton shower produces additional jets. Finally, detector effects are simulated with the CMS card of Delphes 3.5.0 [188, 189].

5.5 Preprocessing

The main novelty of this chapter is the adoption of very recent computer vision techniques to tag the hadronically decaying neutralinos. To that end, we first have to translate the simulated detector data into images.

The objects we are interested in are jets clustered with the anti-k_T (AK) jet algorithm as implemented by the FASTJET package [190]. Choosing the optimal distance parameter R for a given purpose can be somewhat nontrivial. A small value of R means that most particles inside a sufficiently hard

jet originated from the same parton, but some of the energy of that parton might not be counted in this jet due to final state showering. On the other hand, a large R likely leads to jets that capture all daughter particles while also muddying the waters by including unrelated objects, e.g. from initial state showering. One can use the fact that the decay products of a resonance with a fixed mass m and transverse momentum p_T spread roughly like $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} \propto m/p_T$ and the typical energy scale of the process to arrive at a *best guess* for an optimal R parameter. This can be aided by the use of jet clustering algorithms with variable R (e.g.,[191, 192]). In the case at hand, this optimal value of R would depend on both the stop and the neutralino mass. We, therefore, do not work with a single fixed value of R, but instead, we will cluster each event using several values of R, and ensemble the resulting jet images to get a better per-event classification. Because we consider rather large neutralino masses, $m_{\tilde{\chi}} \geq 100\,\text{GeV}$, we consider AK08 ($R = 0.8$), AK10 ($R = 1.0$) and AK14 ($R = 1.4$) jets. This also allows us to keep the technique general, i.e. to use the same algorithm over the entire parameter space. Recall that the resulting fat jet has to satisfy $p_T > 100\,\text{GeV}$ and $|\eta| < 2.4$.

In order to get images out of the jets, we now consider the calorimeter towers and tracks as jet constituents in the (η, ϕ) plane. As in the construction of top taggers [153] we will not use the energy E of the calorimeter towers directly but rather opt for the transverse energy $E_T = E/\cosh\eta$. The relevant features are more readily learned by the classifier if we normalize the coordinates. First, we calculate the E_T weighted center of the calorimeter towers via

$$\overline{\eta} = \frac{\sum_i E_{Ti}\eta_i}{\sum_i E_{Ti}}\,, \tag{5.7}$$

$$\overline{\phi} = \frac{\sum_i E_{Ti}\phi_i}{\sum_i E_{Ti}}\,; \tag{5.8}$$

here the sums run over all the constituents of a given fat jet. We then shift the coordinates $\eta_i \to \eta_i - \overline{\eta}$ and $\phi_i \to \phi_i - \overline{\phi}$ so that the image is centered on the origin. Next, we rotate the coordinate system around the origin so that the calorimeter tower with the highest E_T points vertically from the origin. We use the last degree of freedom to make sure that the calorimeter tower with the second highest E_T lies in the right half of the coordinate, by flipping along the vertical axis if necessary.

Next, we pixelate the coordinates to a 0.04×0.04 grid. The brightness/intensity of each pixel is given as the measured E_T. We use three channels, corresponding to E_T in the Electromagnetic Calorimeter (ECAL), Hadron Calorimeter (HCAL), and p_T of the tracks, analogous to three color channels in classical images.[1] We divide each pixel by the maximal value found in this image, so that each intensity is between 0 and 1. This makes learning more efficient. It also removes information about the p_T and mass of the jet which are powerful discriminators. We partly remedy this by giving the classifier the mass of the fat jet as another input; this will be explained in more detail in a following section. We also note that we will later introduce additional high-level features to our final classifier,

[1] In principle, the tracks have a much higher resolution compared to the calorimeter towers and could thus be pixelated into a finer grid. However, we do not expect these very fine details to improve the discrimination between signal and background. We therefore use the same grid spacing for all three channels.

which will reintroduce information about the overall E_T scale of the event.

In the last preprocessing step, we crop the image to a square centered around the origin with side lengths chosen as 64 pixels for AK08 and AK10 jets and 128 pixels for AK14 jets. This size is chosen to contain most of the constituents while also being a power of two which aids in the application of the computer vision techniques. The resulting images after all preprocessing steps, averaged over the entire event sample, are shown in figure 5.2. These average images look quite similar for signal and background, at least to the human eye; however, taking the difference between the average images does reveal some differences.

Moreover, there is more information available to the computer vision techniques than can be displayed in the figure. For instance, the number of non-zero pixel values is useful for classification. On average, the signal images contain more non-zero pixels than the the background images do. Just cutting on this quantity allows, for one set of parameters, to reach an accuracy of 74% when applied to a sample containing an equal number of signal and background events. Of course, our final classifier should perform much better than this.[2]

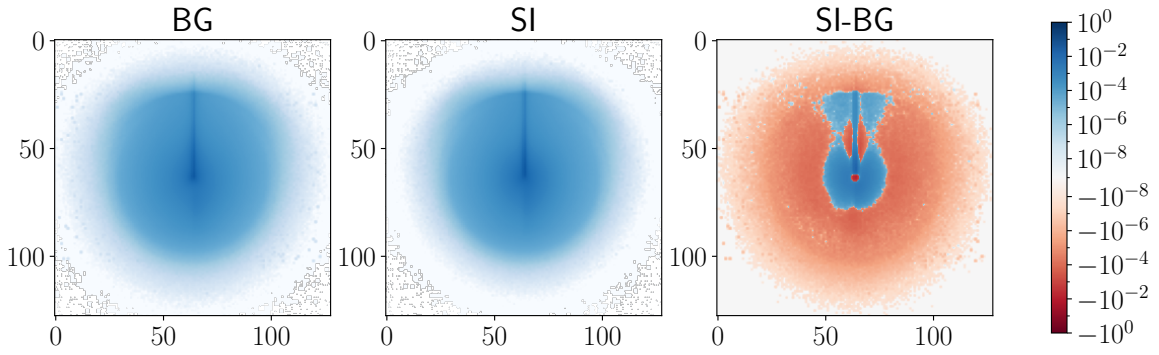

Figure 5.2: Signal and background AK14 jet image averaged over the entire training dataset. All three channels are aggregated by summation. The rightmost plot shows the difference between the average signal and the average background jet image. Signal events are more concentrated at the origin while background jets are more spread out.

5.6 Architectures

In this section we describe how we process a single fat jet, the goal being to distinguish jets due to the three-body decay of neutralinos from SM background ("LSP tagging"). At the core is one of three architectures adapted from computer vision, and described in more detail in Chapter 3. We only describe the chosen hyperparameters in the following. In all three cases, the output of this architecture

[2] We note in passing that this multiplicity does contain some information on the hardness of the event since it correlates positively with both the mass and the transverse momentum of the fat jet. Hence, the normalization step described above does not completely remove the information on these quantities. However, these dependencies are only logarithmic and subject to large event-by-event fluctuations. Explicitly adding the jet mass as an input variable can therefore still be expected to aid in the classification task.

is concatenated with the measured jet mass and fed into the same multilayer perceptron classification network. It is built from a dense layer with 256 neurons followed by another dense layer with 128 neurons which connect to two output neurons. Between all three layers, the ReLU activation function is used. The two neurons of the last layer are passed into the softmax activation function, such that the output can be interpreted as the predicted probability of the image belonging to either the signal or the background; since these probabilities should add to 1. In the following, we denote this by *MLP Head*. All architectures are built and trained within the PyTorch [193] deep learning library.

5.6.1 CNN

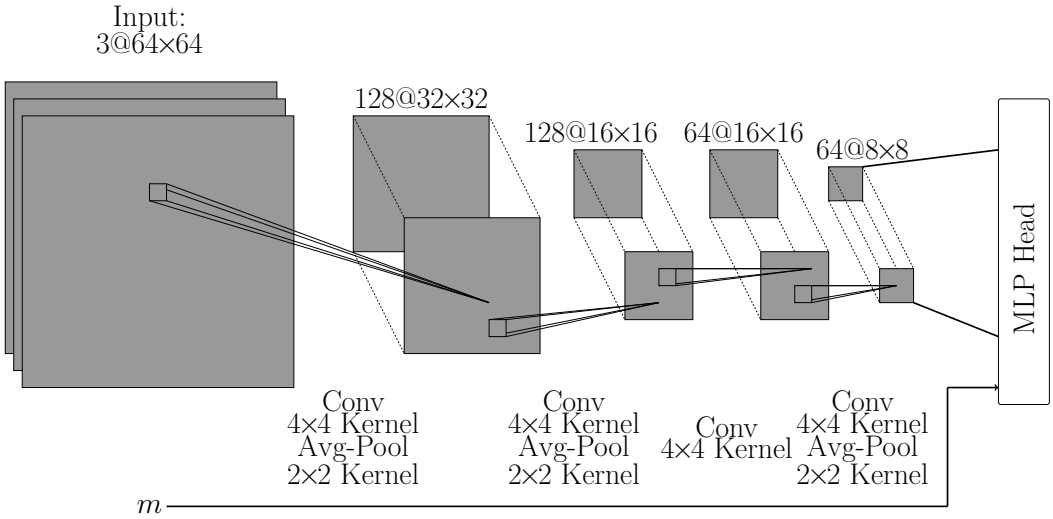

Figure 5.3: Architecture of the CNN for AK08 and AK10 jets.

The first architecture is a (comparatively) simple CNN, described in Section 3.4. We follow loosely an existing model used for top tagging [150]. The first layers of the CNN are two blocks, each containing a convolutional layer with 128 kernels of size 4×4, stride 1, zero padding to keep the image dimensions, ReLU activation function and average pooling with kernel size 2 and stride 2. This halves the spatial image dimensions. Next, we apply the same block with only 64 kernels and without pooling. The last convolution block contains again 64 kernels but this time with the pooling operation.

In order to put AK14 images of size 128×128 on the same footing as the smaller AK08 and AK10 images we repeat the last block one more time. The output of shape $64 \times 8 \times 8$ is then flattened into 4096 features and fed into the MLP head.

The full architecture for AK08 and AK10 jets is shown in figure 5.3. As already noted, this kind of architecture is already being used for similar tasks; it serves as our baseline, against which we compare the more advanced architectures described in the following two subsections.

5.6.2 CoAtNet

As mentioned in Chapter 3.6 transformer architectures have been successfully applied to images. Among the first applications is the vision transformer, further described in Section 3.6.1, although in our tests, the vanilla ViT indeed performs poorly, so we will only show the performance in a limited fashion in Appendix B.3.

Instead, the first transformer-based model we will pursue here is CoAtNet, described in Section 3.6.2. The model we use here is constructed in five stages. The first stage consists of three convolution layers with 3×3 kernels, where the first has a stride of 2. This halves the spatial resolution of the input image. This is followed by two stages of three MBConv blocks [74] which are computationally cheaper while maintaining most of the performance of full convolutional layers. In both stages, the first layers perform downsampling again with a stride size of 2. By now the width and height of the input image are shrunk by a factor of 2^3 so global attention is feasible even for the large AK14 jet. Thus the last two stages consist of five and two transformer blocks respectively. In each transformer, 2D relative attention is used. These steps were performed using the publicly available code[3]. Finally, we average pool the outputs and feed the 768 features into the classification head.

5.6.3 MaxViT

The third architecture we will use is the MaxViT [72] described in Chapter 3.6.3.

For our application, we chose the publicly available implementation that is included in PyTorch[4] with the only change being the replacement of the MLP head that takes the 256 output features. Concretely, the first stage consists of two convolutional layers with 64 3×3 kernels each, where the first has stride 2, reducing the spatial dimensions by half. This is followed by three stages with two MaxViT blocks each. The convolution is strided with size 2 for the first block of each stage. The partitions are of size 4×4 each. The MaxViT stages have 64,128 and 256 channels respectively. Self-attention uses 32-dimensional heads.

Note that in the meantime, the main idea of MaxViT, the multi-axis self-attention, has been combined with ParT, to form a particle multi-axis transformer [194]. This does, similarly to ParT, use the particle cloud representation of the jet.

5.7 Dataset Creation

In this section, we describe how the dataset used to train the LSP taggers is defined. In most events, there is more than one fat jet that passes the selection criteria. Since we investigate pair production, not all the information useful for event classification can be expected to be contained in the hardest jet. It is therefore expected to be useful to combine information from more than one jet into the analysis. Wide jets that are produced from the $\tilde{\chi}$ decays are expected to be hard because of the large stop

[3] https://github.com/chinhsuanwu/coatnet-pytorch
[4] https://github.com/pytorch/vision/blob/main/torchvision/models/maxvit.py

mass. Therefore, the two jets with the largest p_T are expected to be signal-enriched. The preselection requirements imply that one top quark from stop decay generally will decay semileptonically. However, the second top quark might decay fully hadronically, resulting in a third wide jet with large p_T. Since the top quarks and the LSPs have very similar p_T distributions, the third largest-p_T jet may also well be from an LSP.

In order to design taggers that perform well on all three leading jets, and hence for a wide range of p_T, we, therefore, include samples of all three leading fat jets in our training dataset in the ratios that the respective number of jets are present in the full events. To this end, we add up to three fat jets present in an event as images to the dataset. Of course, an event may also contain only one or two such jets; in fact, this is generally the case for background events. We generate as many events as required to reach the desired size of the training set for each jet size.

5.8 Training the LSP taggers

We start by verifying that the different PYTHIA tunes that we adopted do not significantly influence our results. To this end, we train the CNN model described in sec. 5.6.1 to differentiate not between signal and background samples but between background events generated with the CP2 tune and background samples generated with the CP5 tune. We combine 1 000 000 jet images generated with each tune into a dataset and split it equally between training and test sets for each jet size. The initial learning rate η_L is chosen as $5 \cdot 10^{-4}$. This value worked best in tests of the LSP taggers. At the end of each epoch, the learning rate is lowered by a factor of 0.7 and the entire training dataset is shuffled. The batch size is 64. We minimize the averaged binary cross-entropy loss

$$l = -\frac{1}{N} \sum_{i}^{N} y_i \ln x_i + (1 - y_i) \ln(1 - x_i) \,, \tag{5.9}$$

where the index i runs through all $N = 64$ images in the batch, y_i is the true label and x_i is the predicted label. Adam [58] is chosen as the optimizer. All taggers are trained for a total of 15 epochs.

The minimum validation losses for AK08, AK10 and AK14 jets are found to be 0.6917, 0.6918 and 0.6920, respectively. When the classifier is tasked with assigning the label 0 to the first class (e.g. CP2 tune) and the label 1 to the second class (CP5 tune) and the classifier is perfectly confused (i.e. unable to distinguish between the classes), it will assign labels close to 0.5 regardless of the true class. The binary cross entropy per image is then $\ln(2) = 0.6931$. Evidently, our observed losses are only very slightly below the value expected for a classifier that learns nothing. We therefore conclude that the difference in PYTHIA tunes can be neglected in the following.

We now turn to the actual training of the taggers to select LSP-like fat jets. Signal samples are generated for $\tilde{\chi}$ masses between 100 GeV and 500 GeV in steps of 10 GeV, and for stop masses between 700 GeV and 1 200 GeV in steps of 25 GeV. For each combination of stop and neutralino mass, we take 4750 sample images from $\tilde{t}_1 \tilde{t}_1^*$ signal events. In order to generate an almost pure signal sample for

training, we only include images of jets that are within $\Delta R < 0.5$ of a parton level $\tilde{\chi}$. Since we want the LSP tagger to work for all combinations of $m_{\tilde{t}_1}$ and $m_{\tilde{\chi}}$, we combine all $41 \times 21 \times 4750 = 4\,089\,750$ images into a single training set. We take the same number of background images, 4 089 750, from $t\bar{t}$+jets events.

Finally, we split the 8 179 500 images into 5 725 650 images for training and 2 453 850 images for validation. After training, the model state at the epoch with the lowest validation loss is selected to define the tagger.

5.9 Results for Neutralino Taggers

In order to compare the performance of our classifiers, we neglect any systematic uncertainties and define the signal significance Z as

$$Z = \frac{S}{\sqrt{B}} = \frac{\epsilon_S}{\sqrt{\epsilon_B}} \cdot \frac{\sigma_S}{\sqrt{\sigma_B}} \sqrt{\mathcal{L}_{\text{int}}}\,, \tag{5.10}$$

where S and B are the numbers of signal and background samples passing a cut (e.g., on the value of an output neuron of the MLP), $\epsilon_{S/B}$ is the selection efficiency of this cut, $\sigma_{S/B}$ is the fiducial production cross section and $\mathcal{L}_{\text{int}}$ is the integrated luminosity of the dataset considered. Instead of comparing signal significances directly, we compare the significance improvement $\epsilon_S/\sqrt{\epsilon_B}$, which captures the gain due to the sophisticated event classifiers and is independent of the assumed luminosity.

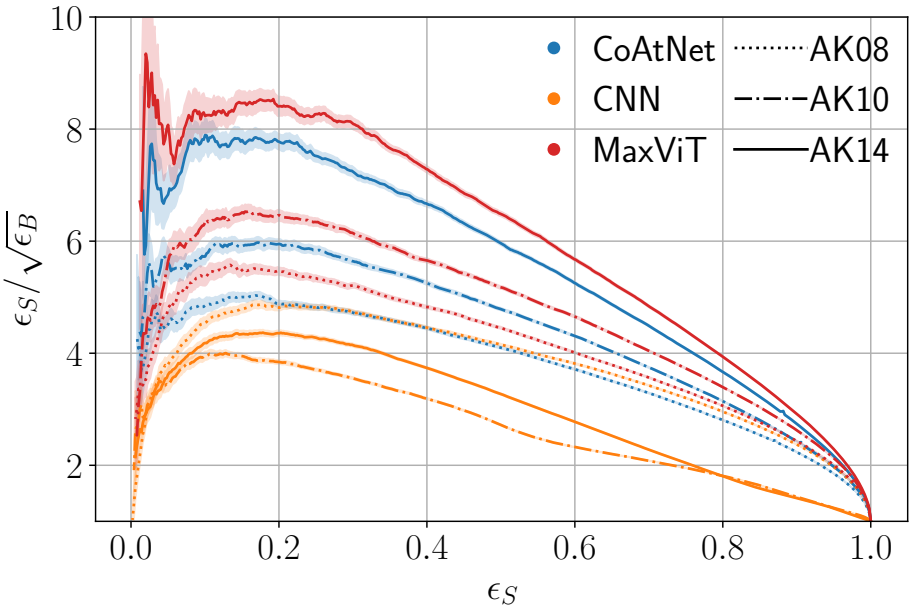

Figure 5.4: Significance improvement curves for all three neutralino taggers for all single jet samples in the test dataset. The shaded regions show one bootstrapped standard deviation with 100 rounds of bootstrapping.

Figure 5.4 shows the performance of all neutralino taggers on the entire test dataset, i.e. with all signal masses present and with the three leading jets mixed as mentioned in sec. 5.8. As a working point for the following analysis, we choose the cut on the MLP output neuron such that $\epsilon_S = 0.3$. Even lower values of ϵ_S can still increase ϵ_S/ϵ_B, but the significance improvement is already close to the maximum at the chosen point. Moreover, for smaller ϵ_S the background efficiency ϵ_B becomes so small that the statistical uncertainty on the accepted background becomes sizeable, despite the large number of generated background events.

Both CoAtNet and MaxViT showed superior performance in classical image classification tasks compared to CNN-based models, as reported in the respective original publications. We expect this to carry over to jet classification. Indeed, this is the case here and both models outperform the classical CNN by up to a factor of 2 for AK14 jets. The most performant classifiers are the transformer-based models trained on the large radius jets. These large jets still contain the entire narrow jets from small LSP masses, while the small jets might miss important features for larger neutralino masses. We also observe that MaxViT performs slightly better than CoAtNet, as is the case in the original MaxViT publication. Evidently, improvements in modern computer vision translate well to the classification of jet images. Even the worst transformer-based model (i.e. AK08 CoAtNet) matches the best CNN. Interestingly, despite the transformer models showing a clear hierarchy, performing better on larger jets, this is not the case for the CNN, which performs best for AK08 jets.

So far, we have considered the classification of singlet jets. In the next section, we will show how this can be used for event classification.

5.10 Boosted Classifiers

As previously mentioned, our signal model always produces two neutralinos that subsequently decay into three quarks (plus possible gluons from final state radiation). It is therefore sensible to combine multiple jet images into our predictions. To this end, we apply one of our LSP taggers described above on the three leading fat jets in an event. From now on, we will drop the merging requirement since it is not meaningful anymore. The three resulting MLP outputs are used as inputs for a GBDT classifier. If an event contains less than 3 fat jets with $p_T > 100\,\text{GeV}$, we assign the label -1 for the missing jets. The GBDT is implemented using the XGBoost [64] package. We use 120 trees with a learning rate of 0.1, with the other hyperparameters left unchanged at the default values. In order to train the GBDT and calculate its results, we use 3000 and 2500 events, respectively, for each combination of stop and LSP masses. This corresponds to a total of 2 583 000 signal events for training and 2 152 500 signal events for evaluation. We again generate an equal number of $t\bar{t}$+ jets background events.

Figure 5.5 shows the significance improvement after a cut on the signal probability given by the GBDT. The difference in performance between the two transformer-based models has shrunk significantly for all jet sizes, especially for AK10 and AK08 jets. Comparing this with figure 5.4 the gain by combining the three jets is not very large. Note, that the merging requirement is now dropped. If we calculate the significance improvement for only the jet with the highest p_T without requiring

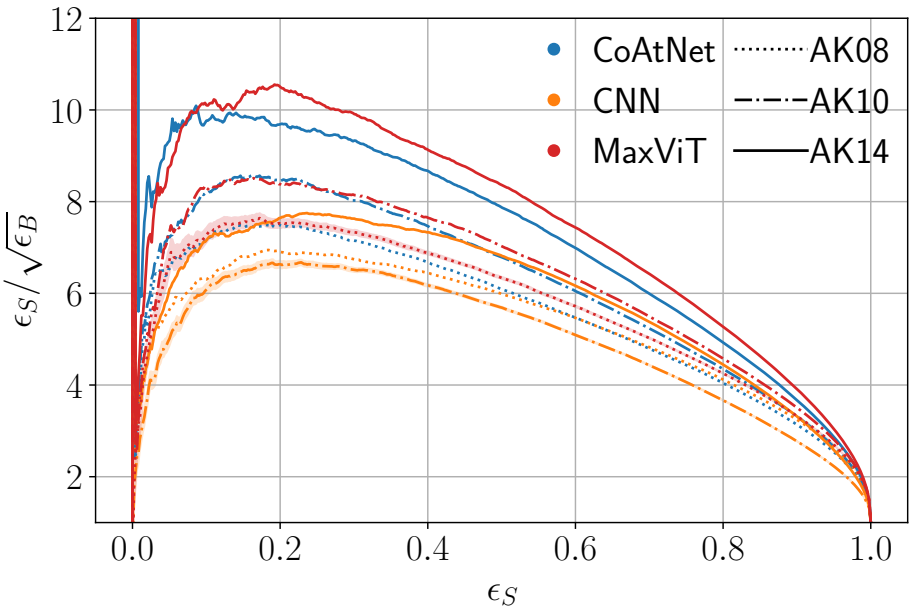

Figure 5.5: Significance improvement curves for all GBDT classifiers built to combine the LSP tagger outputs for the three highest p_T jets. The shaded regions are one bootstrapped standard deviation.

it to be close to a (truth-level) LSP, MaxViT reaches 6.79 ± 0.05 at $\epsilon_S = 0.3$. Comparing this with 9.92 ± 0.12 for the same base model after combining the LSP tagger output for the three hardest jets shows an improvement of almost 50%, equivalent to doubling the integrated luminosity in equation 5.10. The CNN now also works best with AK14 jets, even though the AK08 version is still better than the AK10 version, contrary to the hierarchy of the other models.

Overall, the level of improvement between the results of figure 5.5, which use information from up to three jets per event, and figure 5.4 for single jets, might seem somewhat disappointing. After all, in the absence of QCD radiation, a $\tilde{t}_1\tilde{t}_1^*$ signal event contains two signal jets plus one fat background jet from the hadronically decaying top quark, whereas a generic $t\bar{t}$ event with one top quark decaying semileptonically contains only a single background fat jet. In such a situation, simply requiring at least one fat jet to be tagged as signal would increase the signal efficiency (for $\epsilon_S \gg \epsilon_B$) from ϵ_S to $1 - (1 - \epsilon_S)^2$ while the background efficiency remains unchanged. Recall, however, that we require each event to contain at least seven AK04 jets. This greatly reduces the $t\bar{t}$ background, since at least three additional partons need to be emitted for the event to pass this cut; on the other hand, it also means that background events frequently contain several fat jets, in which case a simple single tag requirement would not increase the significance. In any case, as noted above, there is a significant improvement in performance when information of the three leading fat jets is combined using a GBDT; of course, the GBDT output is not equivalent to simply demanding a fixed number of jets in a given event being tagged as LSP-like.

In figure 5.6 we show how the performance of the GBDT depends on the LSP mass. For small

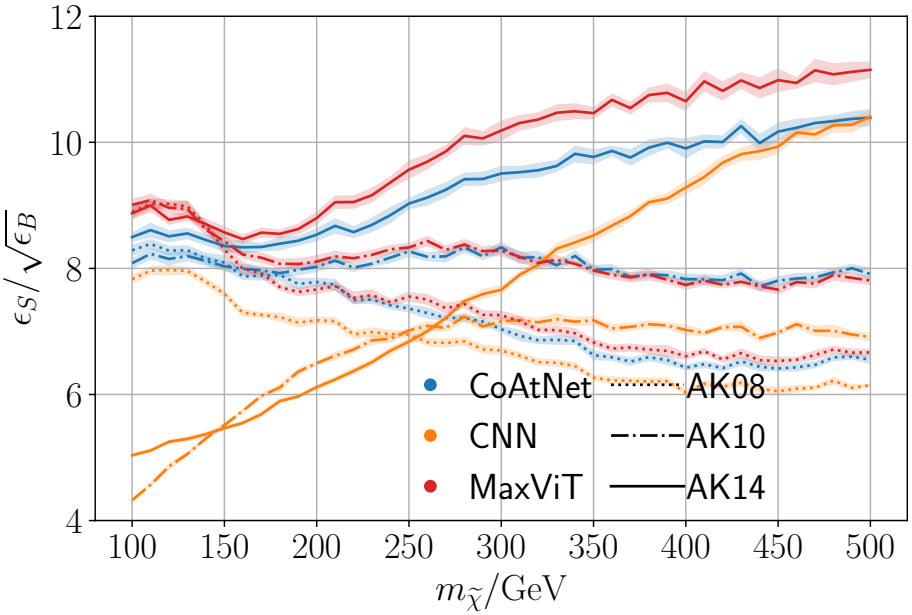

Figure 5.6: Significance improvements depending on the LSP mass for all GBDT classifiers built to combine the LSP tagger outputs for the three highest p_T jets. The cut on the GBDT output has been set such that $\epsilon_S = 0.3$ for each given LSP mass. The shaded regions are one bootstrapped standard deviation.

masses, the two transformer-based models perform comparably for all three jet sizes. Here, the decay products are usually contained even in the AK08 jet so all three jet sizes contain the necessary information for our task. Evidently, the transformer networks are able to filter out the noise from particles not related to LSP decay that are present in the AK10 and AK14 jets, while the simpler CNN cannot; hence the GBDT using the CNN applied to AK10 or AK14 jets performs relatively poorly for small LSP mass. On the other hand, for LSP masses above 200 GeV, the GBDT performs significantly worse when used on the smaller jets, which no longer contain all particles originating from LSP decay.

We also note that using the CNN applied to AK14 jets performs far worse than the other models for small LSP mass, but matches the performance of the CoAtNet-based model for $m_{\tilde{\chi}}$ between 450 and 500 GeV. This curve also shows the strongest LSP mass dependence. We will revisit this point later in this chapter.

Finally, while the MaxViT architecture with AK10 and AK14 jets again shows the best overall performance, the resulting $\epsilon_S/\sqrt{\epsilon_B}$ shows a shallow minimum at $m_{\tilde{\chi}} \simeq m_t$. For a given p_T, fat jets originating from LSP and top decay will then have similar overall features, and the additional information about the jet mass will not provide further benefit. Moreover, recall that in our scenario, the LSP decay products contain exactly one b-quark, just like nearly all jets from top decay. Nevertheless, the model performs quite well even in this difficult mass region. Presumably, it exploits the fact that top decays into three quarks proceed via two 2-body decays with a color singlet on-shell W boson in the intermediate state, whereas the LSP decays via the exchange of a (far) off-shell squark, therefore

Model	AK14	Combined jets
CoAtNet	9.32 ± 0.10	9.63 ± 0.10
MaxViT	9.91 ± 0.11	10.09 ± 0.12
CNN	7.62 ± 0.06	9.16 ± 0.09

Table 5.1: Significance improvement, $\epsilon_S/\sqrt{\epsilon_B}$, for $\epsilon_S = 0.3$ when only using AK14 jets (second column), and when combining the LSP tagger outputs on AK08, AK10 and AK14 jets using a larger GBDT (third column). The uncertainties are bootstrapped standard deviations.

Jet	Combined	Best single model
AK08	7.53 ± 0.06	7.34 ± 0.06 (MaxViT)
AK10	8.94 ± 0.09	8.15 ± 0.07 (MaxViT)
AK14	10.51 ± 0.11	9.91 ± 0.11 (MaxViT)

Table 5.2: Significance improvement with $\epsilon_S = 0.3$ when feeding the outputs of all three LSP taggers simultaneously to the GBDT, keeping the jet definition fixed. For comparison, the third column shows the significance improvement for the MaxViT-based model, which performs best for all three jet sizes. The uncertainties are bootstrapped standard deviations.

leading to a slightly different jet-pattern.

At this point, we still have nine predictions for each event (the output of three architectures applied to AK08, AK10 and AK14 jets). Of course, these nine numbers are highly correlated. Nevertheless, a further improvement of the performance might be possible by either combining results from different jet definitions within a given architecture or vice versa. Comparing these results might also allow us to infer in which aspect a single model has room for improvements that might be gained by another architecture.

We start by combining LSP tagger outputs for different jet sizes. We show the results in table 5.1 and compare the performance to that of the best single jet definition, which is achieved for AK14 jets, as we saw in figure 5.5. Evidently, the improvement is barely statistically significant for the two transformer-based models. These models extract most of the useful information from the images of the large AK14 jets, even when there is a lot of clutter present. The improvement is larger for the CNN-based classifier, which, however, still performs somewhat worse than the other models. It seems to benefit from the multiple jet definitions intended to extract high-level features, such as the mass in classical applications. In particular, the combination allows to compensate for the degraded performance when using the large jets for LSP mass below 250 GeV by information from the AK08 jets, which is more useful in this parameter region, as we saw in figure 5.6.

Next, we combine the outputs of different LSP taggers into a single GBDT, for fixed jet definition.

The results are shown in table 5.2. This time, the combination leads to a slight but significant improvement over the best single model (the one based on MaxViT). This shows that even though the transformer-based models perform almost equally well for the AK14 jets while the CNN is noticeably weaker, each model misses complementary information that the GBDT can combine into a stronger classifier.

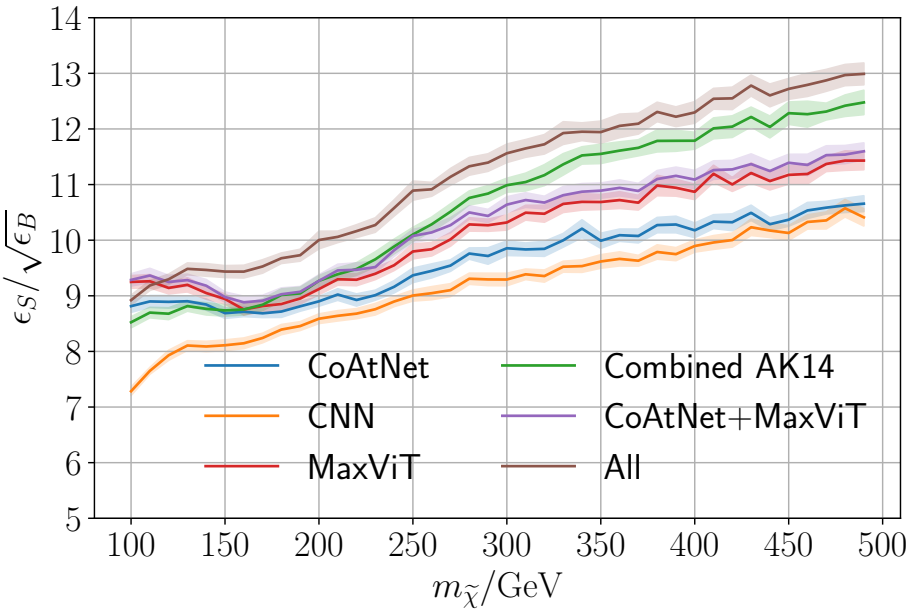

Figure 5.7: Significance improvement as a function of the LSP mass for GBDT classifiers built by combining the output of different LSP taggers, with $\epsilon_S = 0.3$ in each case. The shaded regions are one bootstrapped standard deviation. The curves labeled CoAtNet (blue), CNN (orange) and MaxViT (red) show the performance of GBDTs built from combining the jet sizes for the given model, as in the third column of table 5.1. The green curve is for the GBDT that uses the outputs of all LSP taggers, but only for the AK14 jets, as in the third row of table 5.2. The purple line results from combining both transformer-based LSP taggers for all jet sizes, while the brown line is for a GBDT that combines all LSP taggers and all jet sizes.

In figure 5.7 we show how the performance of various strategies to combine LSP taggers varies with the neutralino mass. Combining all transformer-based predictions into a single GBDT does not show any significant improvement over the performance of the MaxViT-based tagger. This indicates that these models use the same features of the jet images and do not find complementary information. The combination of all CNN predictions is comparable to the weaker transformer-based model, CoAtNet, for LSP mass above 200 GeV, while MaxViT is still more sensitive for all LSP masses.

Because our LSP taggers generally perform best on AK14 jets, we also show the combination of all three architectures using only AK14 jets, as in the last row of table 5.2. Comparing to figure 5.6, we see that for LSP mass below $\sim$ 160 GeV this combination does not further improve on the MaxViT-based model applied to AK14 jets. Between $\sim$ 160 GeV and $\sim$ 300 GeV, the performance closely follows that of the two combined transformer models shown in purple. Since we already showed that one does

not gain much combining the CoAtNet and MaxViT models, this shows that the CNN does not yield useful information in this region of parameter space, either.

However, as we saw in figure 5.6, the CNN-based model applied to AK14 jets improves more with increasing LSP mass than the transformer-based models do, even matching CoAtNet at 500 GeV. The combination profits from this fact and outperforms the GBDTs in the > 300 GeV range, using only input from the transformer-based LSP taggers. This shows that the CNN learns something about the sample that the other models miss.

Finally, we show the result of a GBDT that is trained on the LSP tagger outputs of all three models and all three jet sizes and thus has 27 inputs in total for each event. Compared to the AK14-only case, this does benefit from the inclusion of smaller jets, in particular at smaller LSP masses where the AK08 and AK10 jets already capture most LSP decay products. For larger LSP masses, the performance is only slightly better than that of the AK14-only case.

These various comparisons show that for the given signal process, the largest improvement in significance $\epsilon_S/\sqrt{\epsilon_B}$ is achieved by the transformer-based models applied to AK14 jets. Both models capture details of the jet images that the CNN misses. Nevertheless, also feeding the output of the CNN-based LSP tagger into a larger GBDT leads to a further slight improvement in the performance. This indicates that one might be able to find new architectures that perform even better than MaxViT.

5.11 Adding High-Level Features

The cuts discussed in Section 5.4 are only preselections. They ensure that the event passes the single lepton trigger and contains at least one fat jet to which the LSP tagger can be applied. They also reduce the background, but even after including information from the LSP tagger, these cuts are not likely to yield the optimal distinction between signal and background. A full event has additional features that allow to define additional, potentially useful cuts, even if they may show some correlation with the output of the LSP tagger.

In particular, so far the only dimensionful quantities we used in the construction of our classifier are the masses of the hardest three fat jets, which we use as input to the LSP tagger. We therefore now introduce additional input variables for the final GBDT: the sum of the masses of all AK14 jets [195],

$$M_J = \sum_{\text{AK14}} m \,, \tag{5.11}$$

and the total missing transverse momentum p_T^{miss}. In addition, we use the total number N_j of all AK04 jets, as well as the scalar sum H_T of their transverse momenta.

Moreover, information about the angular separation of the jets might be helpful. Inspired by ref. [183] we capture this information via the Fox-Wolfram moments [196] H_l, defined by

$$H_l = \sum_{i,j=1} \frac{p_{Ti} p_{Tj}}{\left(\sum_k p_{Tk}\right)^2} P_l(\cos\Omega_{ij}) \,; \tag{5.12}$$

here i, j, k run over all AK04 jets in the event, p_{Ti} is the p_T of the i.th jet, P_l is the l'th Legendre polynomial and

$$\cos\Omega_{ij} = \cos\theta_i \cos\theta_j + \sin\theta_i \sin\theta_j \cos(\phi_i - \phi_j) \tag{5.13}$$

is the cosine of the opening angle between the jets i and j. We show the distribution of all features in Appendix B.1.

We combine these features into two sets: the small set DS1= $[p_T^{\text{miss}}, H_T, M_J, N_j]$, which includes the most commonly used features for new physics searches in hadronic final states and a slightly larger set DS2, which also includes the second to sixth Fox-Wolfram moments. We combine these features with the output of the LSP tagger based on MaxViT applied to AK14 jets (i.e., the most performant single model) and derive predictions with a similar GBDT as before.

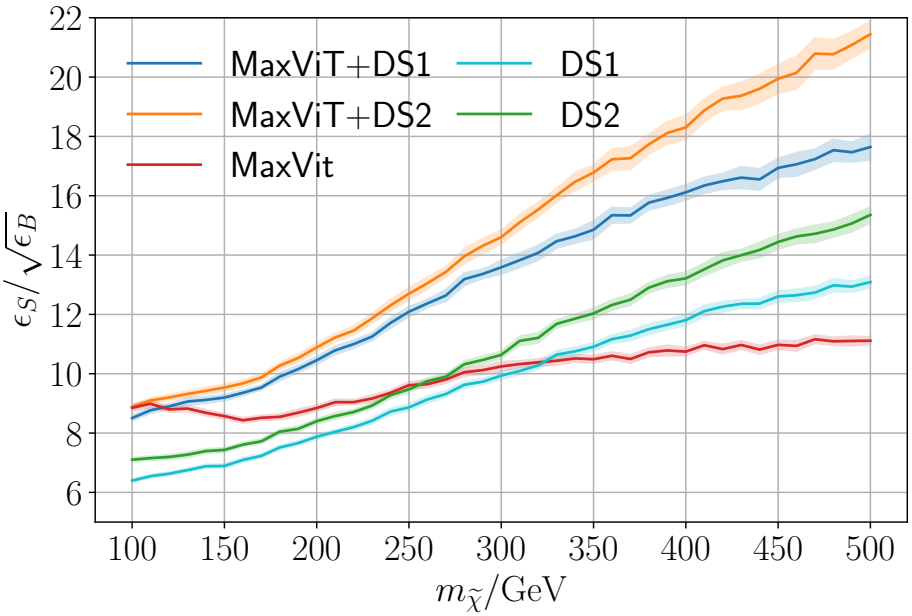

Figure 5.8: Significance improvements as a function of the LSP mass for various GBDT classifiers. In all cases, the cut on the GBDT output has been set such that $\epsilon_S = 0.3$ for each given LSP mass. The upper two curves show results from classifiers that combine the output of the MaxViT-based LSP tagger applied to AK14 jets with additional kinematical features. The feature set DS1 contains $[p_T^{\text{miss}}, H_T, M_J, N_j]$ while DS2 contains, in addition, the second to sixth Fox-Wolfram moments. For comparison, the red curve id obtained when using only LSP tagger information, as in figure 5.6, while the lower blue and green curves are for GBDTs that only use kinematical information. The shaded regions are one bootstrapped standard deviation.

Results are shown in figure 5.8. We see that even GBDT classifiers that only use the kinematic information of sets DS1 or DS2 are quite capable of separating signal from background, especially for larger LSP masses; this reconfirms the usefulness of these variables for new physics searches at the LHC. In fact, for LSP mass above 300 GeV, these classifiers even outperform the GBDT that only uses information from the MaxViT-based LSP tagger. On the other hand, except for $m_{\tilde{\chi}} = 100$ GeV, adding kinematic information to the output of the LSP tagger clearly improves the performance of the event

classifier, indicating that the Fox-Wolfram moments prove useful for LSP mass above 250 GeV or so.

Conversely, adding information from the LSP tagger to the purely kinematic variables raises the significance improvement by an amount that is nearly independent of the LSP mass. We expect the gain of performance to be even larger when compared to a classical selection based purely on kinematical cuts.

5.12 Application at 137 fb^{-1}

We are now ready to discuss how the different classifiers fare, in terms of the reach in stop mass for exclusion or discovery. Here we set the integrated luminosity to $\mathcal{L}_{\text{int}} = 137\,\text{fb}^{-1}$, as in the original CMS publication [183]. For simplicity, we ignore the systematic uncertainty on the signal, as well as the uncertainty from the finite size of our Monte Carlo samples. The former is much less important than the systematic error on the background estimate, and the latter should be much smaller than the statistical uncertainty due to the finite integrated luminosity. The $t\bar{t}$ background is normalized to the next to leading order production cross section [120]. The simulated stop pair samples are normalized to NLO + NLL accuracy [197]. This corresponds to 273 084 background events and a stop mass-dependent number of signal events. We calculate exclusion limits from the expected exclusion significance [198]:

$$Z_{\text{excl}} = \left[2S - 2B \ln\left(\frac{B+S+x}{2B}\right) - \frac{2B^2}{\Delta_B^2} \ln\left(\frac{B-S+x}{2B}\right) - (B+S-x)\frac{B+\Delta_B^2}{\Delta_B^2} \right]^{1/2}, \tag{5.14}$$

where B and S are the expected number of background and signal events, Δ_B is the absolute systematic uncertainty on the background, and

$$x = \sqrt{(S+B)^2 - \frac{4SB\Delta_B^2}{B+\Delta_B^2}}\,. \tag{5.15}$$

We chose $\Delta_B = 0.06B$, as described below. Z_{excl} is the expected number of standard deviations with which the predicted signal S can be excluded if the background-only hypothesis, described by the background B, is correct; note that $Z_{\text{excl}} \to S/\sqrt{B+\Delta_B^2}$ if $B \gg S$. This quantity is computed for every combination of stop and LSP masses introduced in sec. 5.4, using four different event classifiers.

The results are shown in figure 5.9. We again define signal-like events through a cut on the GBDT output corresponding to $\epsilon_S = 0.3$. The top-left frame is for a GBDT that uses only kinematic information about the AK04 jets, as in the green curve of figure 5.8. Associating the contour along $Z_{\text{excl}} = 1.645$ with the 95% confidence level exclusion bounds of this "traditional" analysis, we find an expected exclusion reach in $m_{\tilde{t}_1}$ of about 740 GeV for an LSP mass of 100 GeV This is rather close to the expected reach of about 710 GeV for the same LSP mass achieved in the CMS search,[5] which is

[5] We note in passing that the actual CMS limit on the stop mass is only 670 GeV for this LSP mass, due to a small (not

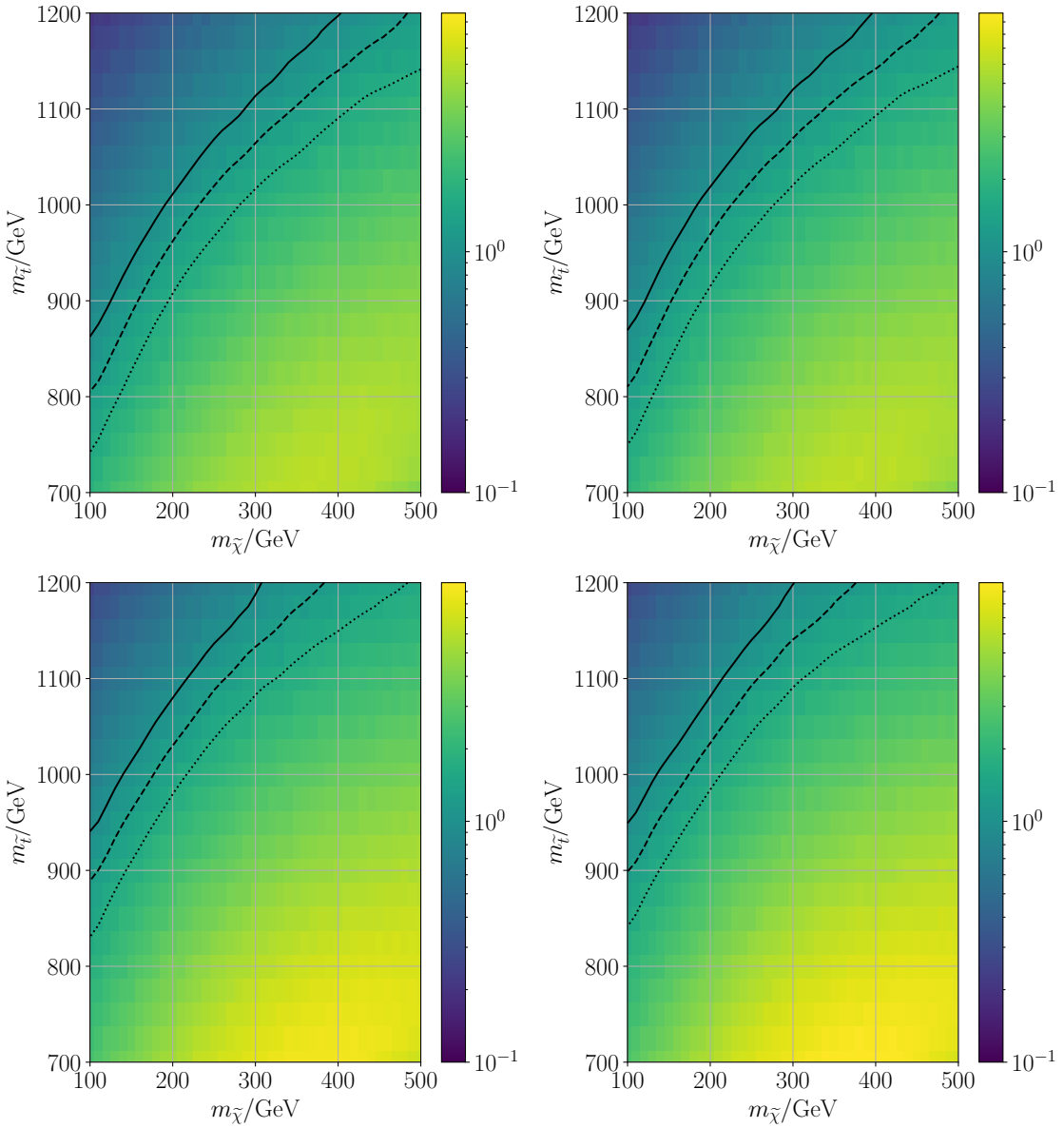

Figure 5.9: Exclusion significance Z_{excl} defined in equation 5.14 for an integrated luminosity of 137 fb^{-1}. In all cases, the cut on the GBDT output has been chosen such that the signal efficiency $\epsilon_S = 0.3$, and $\Delta_B = 0.06B$. The top-left frame is for a GBDT using kinematical information only, corresponding to the green curve in figure 5.8. The other three frames are for GBDTs that also use the output of LSP taggers applied to AK14 jets, based on the CNN (top-right), on CoAtNet (bottom-left) and on MaxViT (bottom-right). Solid, dashed and dotted lines denote contour lines corresponding to a signal significance of 1, 1.281 and 1.645 respectively. These are smoothed by a Gaussian filter with standard deviations 10 GeV and 25 GeV on the neutralino mass and stop mass axis respectively, applied to the logarithm of the signal significances.

based on a neural network (NN) "trained to recognize differences in the spatial distribution of jets and decay kinematic distributions" [183]. Unfortunately, they don't show results for other LSP masses. This agreement is not accidental; we chose the systematic background uncertainty, $\Delta_B = 0.06B$, accordingly. Presumably, even closer agreement would have been possible for somewhat larger Δ_B. However, it would then be significantly larger than the actual systematic error on the background estimate found by CMS, which is below 5%. We note that for $\Delta_B^2 \gg B$ the significance scales $\propto 1/B$, rather than $\propto 1/\sqrt{B}$, if Δ_B is a fixed percentage of B. A larger Δ_B, therefore, increases the relative improvement in reach achieved by including information from one of our LSP taggers; recall that this leads to a significant improvement of $\epsilon_S/\sqrt{\epsilon_B}$, and hence to an even bigger improvement in ϵ_S/ϵ_B.

The other three frames show results for GBDTs that also use the outputs of an LSP tagger as input variables; we apply this tagger to the three leading AK14 jets. We see that the simpler CNN-based tagger (top right) increases the reach in stop mass only by less than 10 GeV. Recall from figure 5.6 that the CNN tagger applied on AK14 jets does not perform well for small LSP mass. For larger LSP mass, and hence larger angular spread of the LSP decay products, the kinematic information on the AK04 jets, many of which are components of AK14 jets, already seems to capture much of the physics found by the CNN. Recall that the kinematic GBDT includes information on the angular separation of these jets via the Fox-Wolfram moments of equation 5.12.

In contrast, using the transformer-based LSP taggers does improve the reach considerably. As before, MaxViT (bottom right) performs slightly better than CoAtNet (bottom left); the reach in stop mass increases by 100 GeV for $m_{\tilde{\chi}} = 100$ GeV, and by about 60 GeV for $m_{\tilde{\chi}} = 500$ GeV. This again indicates that the kinematic information on the AK04 jets allows some effective LSP tagging for large LSP masses.

For stop masses in the interesting range, the $\tilde{t}_1\tilde{t}_1^*$ production cross section of [197] can be roughly parameterized as

$$\sigma(pp \to \tilde{t}_1\tilde{t}_1^*) \simeq 0.08 \text{ pb} \cdot \left(\frac{m_{\tilde{t}_1}}{700 \text{ GeV}}\right)^{-7.8} . \tag{5.16}$$

Increasing the reach from 740 to 840 GeV (for $m_{\tilde{\chi}} = 100$ GeV) thus corresponds to reducing the bound on the stop pair production cross section by a factor of ~ 2.7. Note that the limit setting procedure is quite nonlinear because the background falls by nearly two orders of magnitude when $m_{\tilde{t}_1}$ is increased from 700 to 1 200 GeV while keeping $\epsilon_S = 0.3$ fixed.

5.13 Summary

The large hadronic activity in pp collisions makes the search for physics beyond the Standard Model in purely hadronic processes at the LHC especially challenging. This problem can be mitigated by the use of sophisticated analysis methods. In particular, jet substructure has proved a powerful discriminator between various production processes.

statistically significant) excess of events.

In this chapter, we studied the feasibility of applying modern computer vision techniques in detecting RPV stop decays. As a benchmark, we use $\tilde{t}_1$ pair production, where each stop decays to a top and a neutralino LSP which subsequently decays via the UDD operator to three quarks. For not-too-small mass splitting between the stop and the LSP, the decay products of the latter tend to reside in a single fat (e.g., AK14) jet. One can build images from the constituents of such jets by using the angle ϕ and pseudorapidity η as spatial positions and deposited energy into the detector as pixel intensity. One can then use computer vision techniques on this representation to build classifiers (" LSP taggers") that aid in amplifying the signal process.

In recent years, transformer-based architectures have been shown to improve on the performance of more classical convolutional neural network-based structures in standard classification tasks. We study how well these novel architectures work on jet images by training LSP taggers based on MaxViT, CoAtNet and a CNN architecture. The training is done on single-jet images. We then combine the output of the LSP tagger applied to the three jets with the highest p_T, using a gradient-boosted decision tree into a more robust classification score. We find that the CNN-based tagger improves the statistical significance of the signal by a factor between 5 and 10 for fixed signal efficiency $\epsilon_S = 0.3$, the exact factor depending on the neutralino mass and the definition of the fat jets. In contrast, the transformer-based models lead to an improvement factor between 8 and 11, outperforming the CNN over the entire parameter space. We also combine the predictions of all architectures for each jet size separately and find a modest improvement, hinting that even the transformer-based models do not use the entire information present in the images; hence an investigation of further improvements of the architecture might be worthwhile.

Since the kinematic preselection cuts are not optimized for sensitivity, we also use high-level features such as Fox-Wolfram-moments, p_T^{miss}, H_T, M_J and N_j as inputs to a GBDT, in combination with the output of one of our LSP taggers. This leads to a total gain of sensitivity by a factor of 20 for 500 GeV LSPs, on top of the effect due to the acceptance cuts.

Finally, we estimate the reach in stop and LSP mass that could be expected from the full run-2 dataset. We chose the systematic uncertainty on the background such that a GBDT that only uses kinematic information on AK04 jets leads to a reach (for LSP mass of 100 GeV) similar to that found by CMS [183]. Additionally, using the output of the relatively simple CNN-based LSP tagger then leads to almost no further improvement of the reach. By instead using the MaxViT-based tagger, one can improve the reach by 100 GeV (60 GeV) for neutralino masses of 100 GeV (500 GeV), under the assumption that the relative size of the systematic uncertainty remains the same. This corresponds to a reduction of the bound on the stop pair production cross section by up to a factor of 2.7.

We conclude that LSP taggers built on modern transformer-based neural networks hold great promise in searches for supersymmetry with neutralino LSP where $R-$parity is broken by the UDD operator. This result can presumably be generalized to models with different LSP, e.g., a gluino decaying via the same operator, or a slepton decaying into a lepton and three jets via the exchange of a virtual neutralino.

In fact, it seems likely that these advanced techniques can also be used to build improved taggers

for boosted, hadronically decaying top quarks or weak gauge or Higgs bosons. We did not attempt to construct such taggers ourselves, since this field is already quite mature. Convincing progress would therefore have to be based on fully realistic detector-level simulations, for which we lack the computational resources. Moreover, a careful treatment of systematic uncertainties would be required, which ideally uses real data. However, we see no reason why the improvement relative to CNN-based taggers that we saw in our relatively simple simulations should not carry over to fully realistic ones.

CHAPTER 6

Conclusion and Outlook

The Standard Model, in its current form, is remarkably predictive and accurately describes a wide range of physical phenomena. Its symmetries and anomaly cancellations also present an elegant theoretical framework. Despite these strengths, the Standard Model has limitations that suggest there is yet more to discover. Notably, it lacks a convincing dark matter candidate, and its cosmological predictions do not match the universe we observe today.

The latest analyses from CMS and ATLAS, based on the LHC Run 2 dataset with an integrated luminosity of $137\,\text{fb}^{-1}$, have so far yielded null results, gradually constraining the parameter space for new physics. Including data from the ongoing Run 3, the LHC has now accumulated approximately $300\,\text{fb}^{-1}$ of integrated luminosity. The upcoming High-Luminosity LHC aims to increase this further by a factor of ten [199], thanks to significant upgrades in experimental apparatus. While the number of recorded events will continue to rise steadily, substantial increases in energy are unlikely in the near future. Consequently, it is crucial to make optimal use of the extensive data available. Machine learning, which thrives on large datasets, stands to provide significant benefits to physics research, thanks to the unprecedented quantities of data generated by the LHC experiments.

The next major step toward discovering new physics lies in employing more sophisticated analysis strategies that leverage machine learning. This thesis contributes to this goal by demonstrating that recent innovations in resonant anomaly detection techniques are more broadly applicable than previously shown. We have demonstrated that even features localized at the tails of distributions can be used to reliably and signal-model-agnostically detect new physics signals. This implies that even the R-parity conserving minimal supersymmetric Standard Model that often produces large missing momentum can potentially be found this way. We explicitly demonstrated that the technique uncovers multiple different supersymmetric signal models with a general approach, while only being marginally less sensitive than multiple dedicated searches. This demonstrates the value that machine learning techniques can bring for resonant anomaly detection, which will be crucial going forward in the development of new, broader analysis strategies

Not all interesting models that produce hadronic resonances can be uncovered using this method. Therefore, further innovations besides the field of anomaly detection will need to be developed. We have found that recent advances in computer vision techniques can be effectively translated to physics, creating a symbiotic relationship between the two fields and leading to increasingly sensitive analysis strategies for high-energy collider datasets. That way, even notoriously difficult signal models have a chance to be uncovered.

Looking ahead, maximizing the discovery potential of the true underlying model of nature will require simultaneous improvements in experimental techniques and the development of more sophisticated analysis strategies.

APPENDIX A

Additional Studies on CATHODE

A.1 Recreating CMS-SUS-19-013

The recreation of CMS-SUS-19-013 [119] follows the most important analysis steps of the original publication. The number of events is set to the integrated luminosity of $\mathcal{L}_{\text{int}} = 137\,\text{fb}^{-1}$. First, a set of remaining cuts are applied to select Z-candidates, then the background estimation is recreated before the statistical analysis is performed. The following cuts are applied to select hadronically decaying Z bosons:

8. Softdropped $m_{\text{jet}} \in [40\,\text{GeV}, 140\,\text{GeV}]$ of the 2 highest p_T AK8 jets
9. $\Delta R_{Z,b} > 0.8$ for the second highest p_T AK8 jet Z and any b-tagged jet where the angular separation is defined as $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$

The resulting p_T^{miss}-spectrum is shown in figure A.1 which agrees with the spectrum shown in the original publication within uncertainties.

The background estimation consists of the normalisation and the shape estimation. The SR is defined as $m_{\text{jet}} \in [70\,\text{GeV}, 100\,\text{GeV}]$. First, one demands the subleading AK8-jet to be in the SR. Then a linear function is fitted to the m_{jet} spectrum of the leading AK8 jet outside its SR. The nominal yield $\mathcal{B}_{\text{norm}}$ is obtained by integrating the linear function in the SR. The statistical error of the yield is obtained from the spread of pseudo-experiments sampled from the fit. In addition to the linear function, Chebychev functions up to the fourth order are fitted. The largest deviation of the nominal yield is then assigned as an additional uncertainty.

The background p_T^{miss}-shape is obtained by the SB with both AK8 jets outside the SR. The content of the ith p_T^{miss} bin is denoted as N_i^{SB}. The transfer factor from the SB to the SR is then calculated as

$$\mathcal{T} \equiv \frac{\mathcal{B}_{\text{norm}}}{\sum_i N_i^{\text{SB}}} = 0.206 \pm 0.023. \tag{A.1}$$

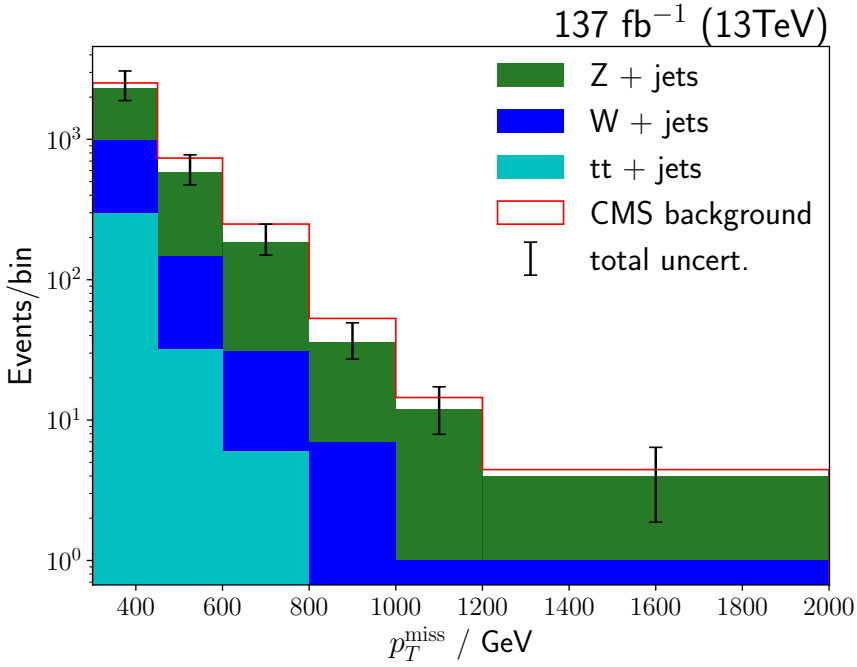

Figure A.1: p_T^{miss} spectrum of the three leading background processes. The background of the same three processes from the CMS-publication is shown in red. The variation of cross section due to changing the energy-scale by a factor of 1/2 and 2 as computed by MadGraph is assigned as a systematic uncertainty and added to the statistic errors in quadrature and shown as error bars.

which agrees with the original publication within uncertainties. The expected background in bin i is

$$\mathcal{B}_i = \mathcal{T} N_i^{\text{SB}}. \tag{A.2}$$

RooStats [200] is used for statistical modeling. It takes N_i^{SB} with statistical errors, $\mathcal{T}$ and $\Delta\mathcal{T}$ to model the background in the SR with uncertainties. The signal model contains signal events that pass all cuts and is rescaled to the approximate NNLO+NNLL cross section [201]. The overall uncertainty of the cross section is applied to all signal bins. The resulting statistical model is evaluated with the CL_s [202] approach and the asymptotic form of the onesided profile likelihood teststatistic. This is used to obtain the 95% CL. cross section. The limits for the integrated luminosity $\mathcal{L}_{\text{int}} = 137\,\text{fb}^{-1}$ are shown in figure A.2 and for $\mathcal{L}_{\text{int}} = 300\,\text{fb}^{-1}$ in figure A.3 . We use the latter dataset for the application of the ML-technique since the accuracy is greatly improved with more datapoints to learn on while in reach for the collider in the near future.

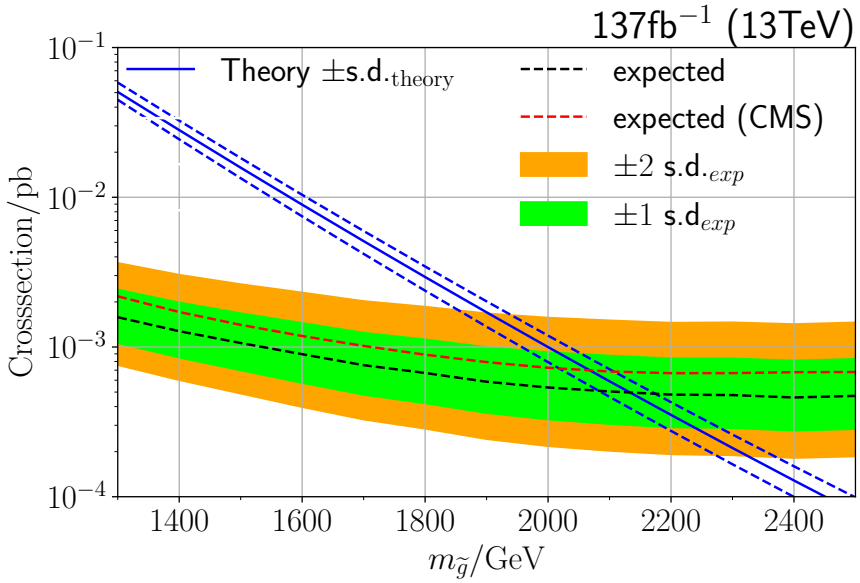

Figure A.2: Recreation of CMS-SUS-19-013 [119]. The red dashed line denotes the expected limits of original CMS-search. The black dashed line shows the expected limits of the recreation.

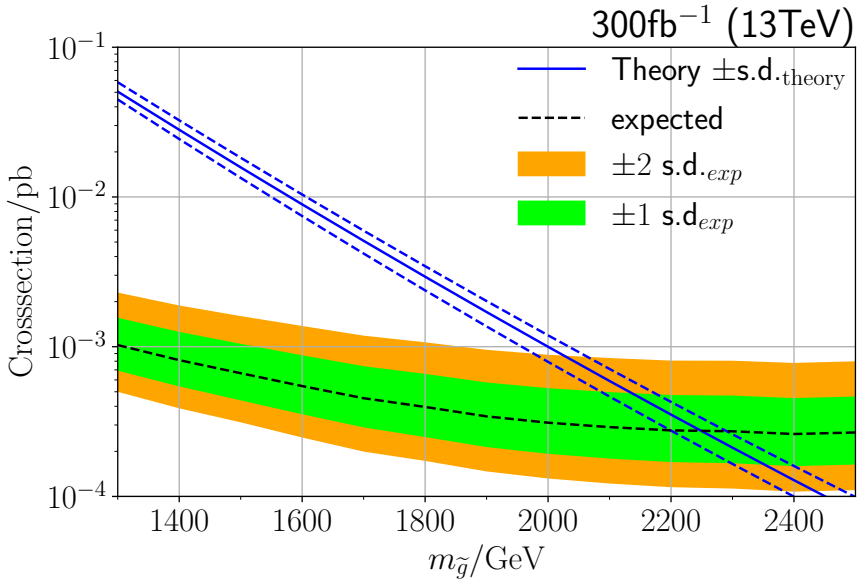

Figure A.3: Results of the classical search for 300 fb^{-1} integrated luminosity

A.2 Comparison with Idealistic Methods

We now compare CATHODEs performance to that of overly idealistic methods. First of all, we focus on a fully supervised approach, learning to distinguish signal and background events directly.

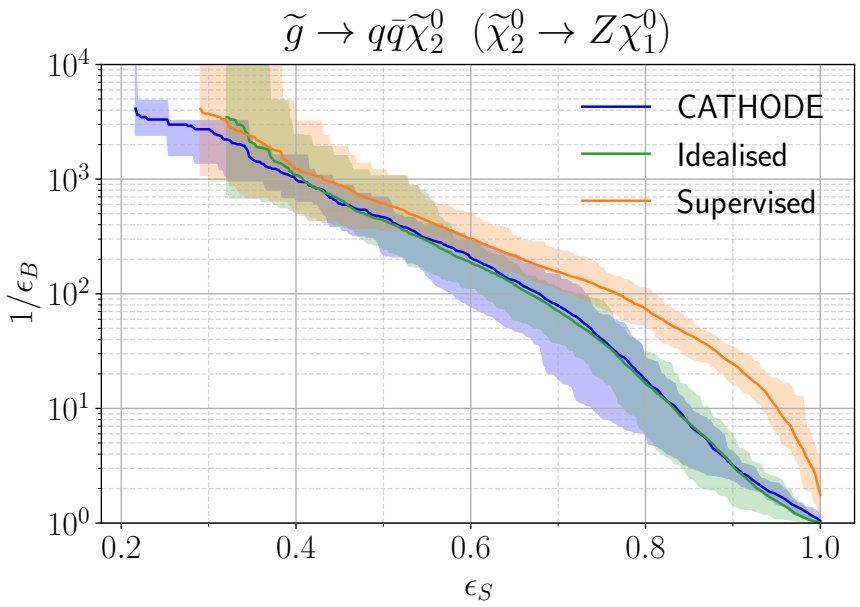

Figure A.4: Background rejection $1/\epsilon_B$ as a function as the signal efficiency ϵ_S. Solid lines indicate the mean. The shaded bands range from the minimal to the maximal value. The CATHODE models shown use 10 000 artificially sampled background events in total inside the signal region.

Because the number of background events is limited [1] and generating more events is computationally prohibitively expensive, we need to take care to make the most of the available data. With the limited training data, we use a gradient-boosted decision tree as implemented by scikit-learn [203]. We choose a maximum number of 150 boosting rounds and leave the remaining hyperparameters at their default values. Input data is not standardized because gradient-boosted decision trees are more robust in this regard than the fully connected neural network used in CATHODE.

We use the same signal model with decay $\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0$ and fix the gluino mass to $m_{\widetilde{g}} = 1\,700\,\text{GeV}$.

Similar to the previous section, we define the test dataset as $\mathcal{L}_{\text{int.}} = 300\,\text{fb}^{-1}$ of signal and background events within the signal region defined by equation 4.7. We use the four-fold cross-validation procedure described above by splitting the dataset into four equally sized parts.

For the supervised model, we combine three of the subsets with additional 10 000 background and 17 500 signal events inside the SR to form the training set. This way, we have roughly the same number of signal and background events in the training set. Contrary to CATHODE, we use the truth labels target labels.

Similarly to ref. [8] we also compare to an idealized anomaly detector. For this, the classification step of CATHODE does not try to tell apart real data from synthetically generated background events but from real background events. This idealized step can be thought of as the limit where the estimated background distribution models the real background distribution perfectly. To train this, we combine

[1] We generated a total of $\mathcal{L}_{\text{int.}} = 618\,\text{fb}^{-1}$ background events, allowing us to estimate various statistical uncertainties.

the three previously mentioned subsets with 10 000 additional background events. This time, the target labels if a given event comes from real data or the background-only simulation. The only way to tell these apart is to find the signal events that are only in one class, which is the whole principle behind CATHODE.

For both the supervised model and the idealized anomaly detector, this is done four times to assign predicted labels to the entire set of real data. Additionally, we reshuffle the dataset ten times and average the ten predicted labels to gain more stable predictions. We repeat this 25 times to calculate statistical uncertainties by sampling new datasets from the available simulated events.

We compare this to CATHODE with a total of 10 000 artificially sampled background events. This way the comparison to the idealized anomaly detector is fair. The results are shown in figure A.4. The fully supervised method performs the best on the entire range of signal efficiency. This is expected because it immediately solves the problem we are testing, i.e. separating signal from background and not the proxy task the anomaly detector methods are trained on. The idealized anomaly detector and CATHODE almost perform the same, a property that has already been observed by the original CATHODE publication. The fact that our signal lies at the tail of both p_T^{miss} and H_T does not degrade its performance, especially the density estimation step.

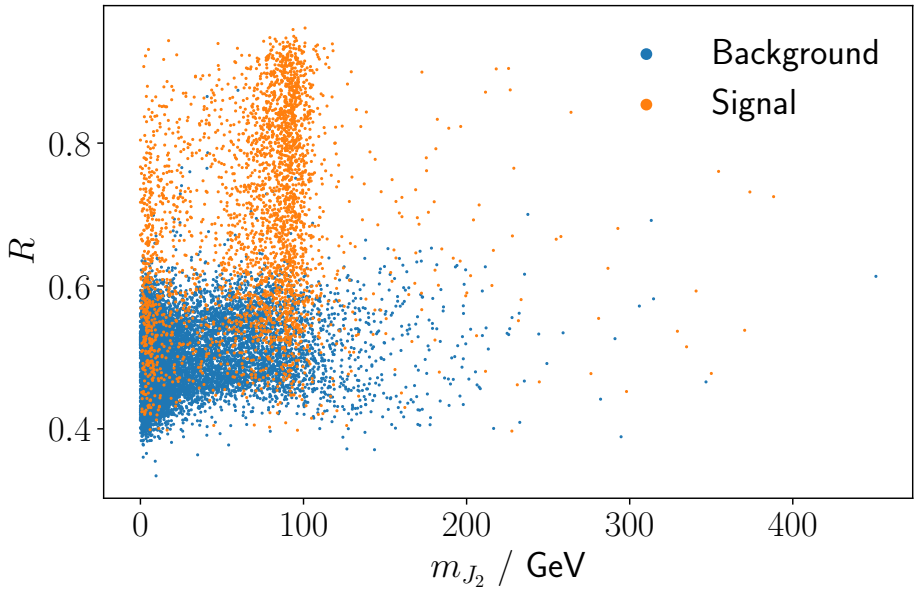

Figure A.5: Correlation between the anomaly score R and m_{J_2}.

A.3 Correlations of Anomaly Scores and Features

In figure A.5, A.6 and A.7 we show the correlation between the anomaly score R and the features CATHODE uses. The signal model is $\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0$ with $m_{\widetilde{g}} = 1\,700\,\text{GeV}$. Since we always use the same background events, we average the anomaly score of background events over the five reshuffled datasets and the ten independent signal injections. For better visibility, we combine the ten independent signal injections into a total of 3109 signal events in the signal region. The anomaly score of these is the average of the five assignments to the reshuffled datasets.

In figure A.5 we note visually that CATHODE does not sculpt the background in m_{J_2}. Even though the signal is predominantly found in $m_{J_2} \in [70\,\text{GeV}, 100\,\text{GeV}]$ it does not assign higher anomaly scores to background events in this region. Therefore, one could even use m_{J_2} in a bump-hunt, although with limited signal model generalizability. In figure A.6 it seems CATHODE learns a cut on H_T of ~ 1 TeV for events with $R > 0.65$. Making precise statements from these figures is impossible because CATHODE also uses correlations between the features that can not be shown here.

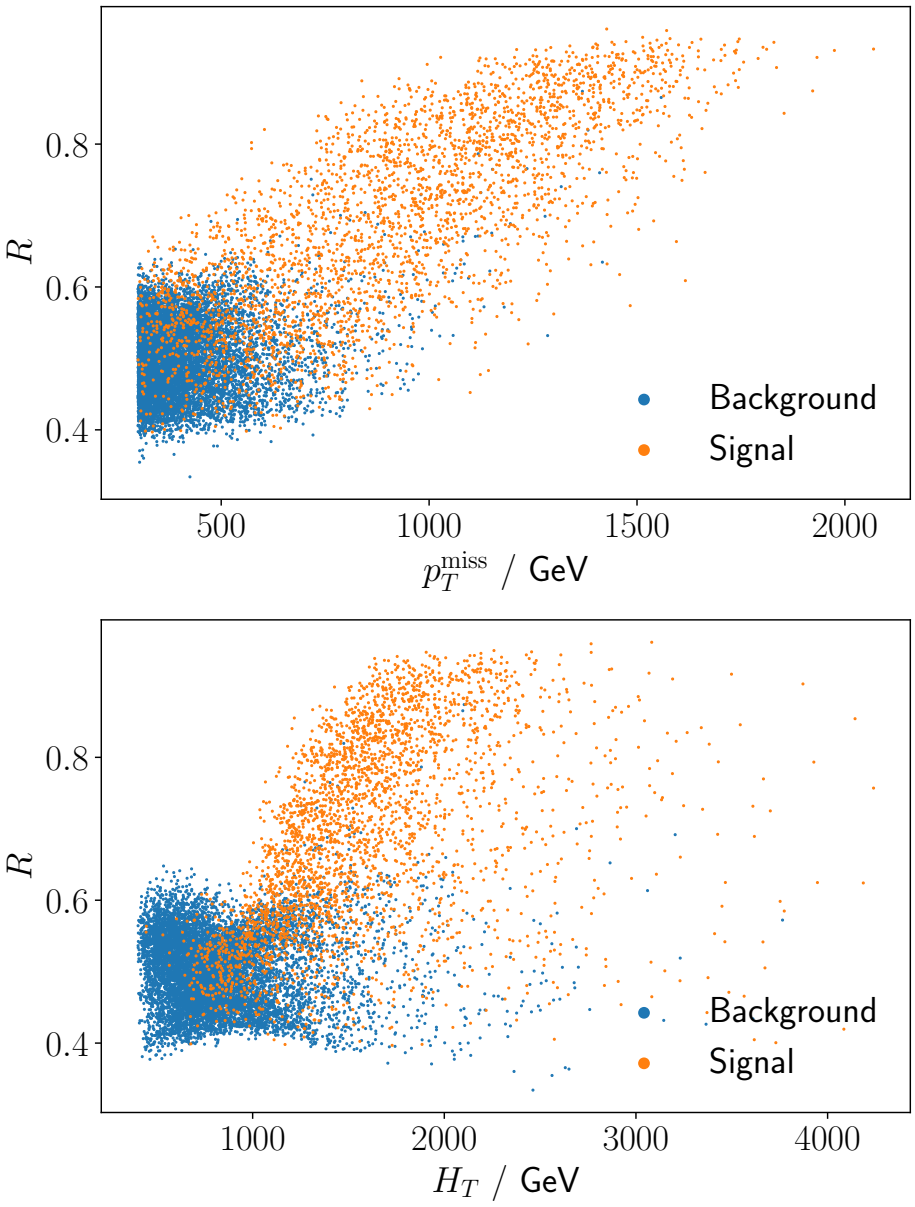

Figure A.6: Correlation between the anomaly score R and p_T^{miss} and H_T.

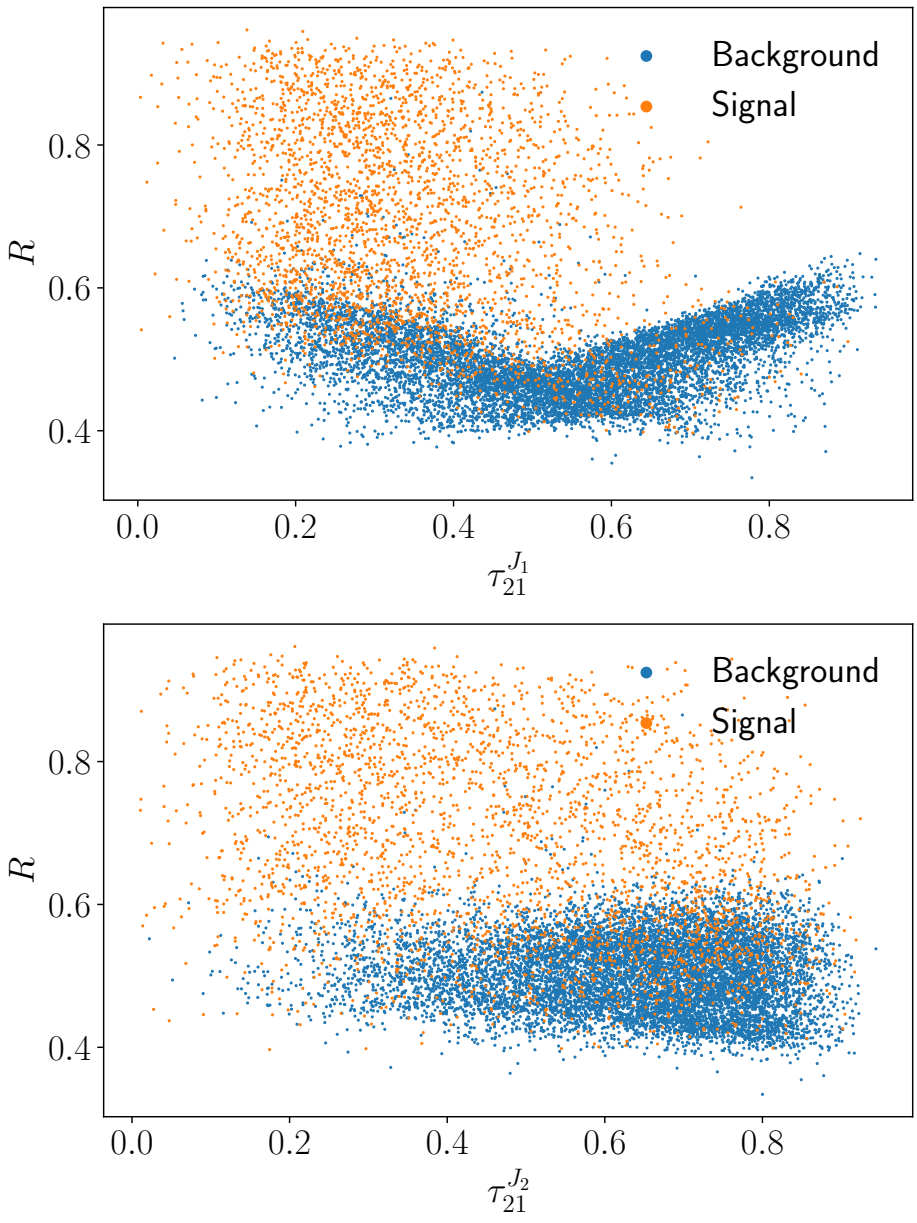

Figure A.7: Correlation between the anomaly score R and $\tau_{21}^{J_1}$ and $\tau_{21}^{J_2}$.

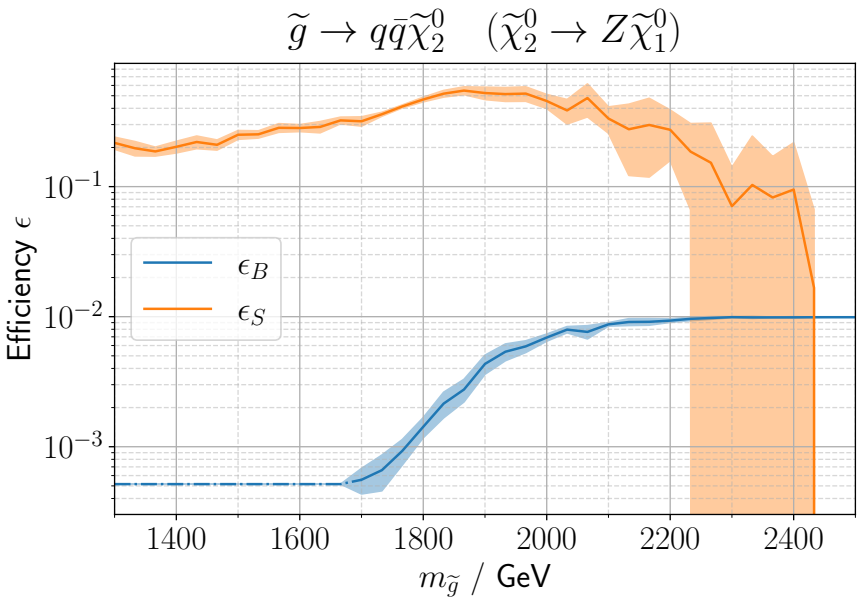

Figure A.8: Signal and background efficiencies of the model with decay $\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0$ with a cut on the anomaly score R_c such that 1% of events pass the selection. The dot-dashed part of the blue line represents parameter points where R_c has to be lowered to allow five background events. The shaded region shows one standard deviation around the mean efficiency.

A.4 Signal and Background Efficiencies

In figures A.8 and A.9 we show the signal and background efficiencies ϵ_S and ϵ_B respectively, that correspond to the results shown in Section 4.5. Over a large section of parameter space CATHODE retains more than 10% of signal events when the cut on the anomaly score R_c is chosen conservatively to only pass 1% of all events. In an application with real data, this selection value would have to be chosen with more careful consideration.

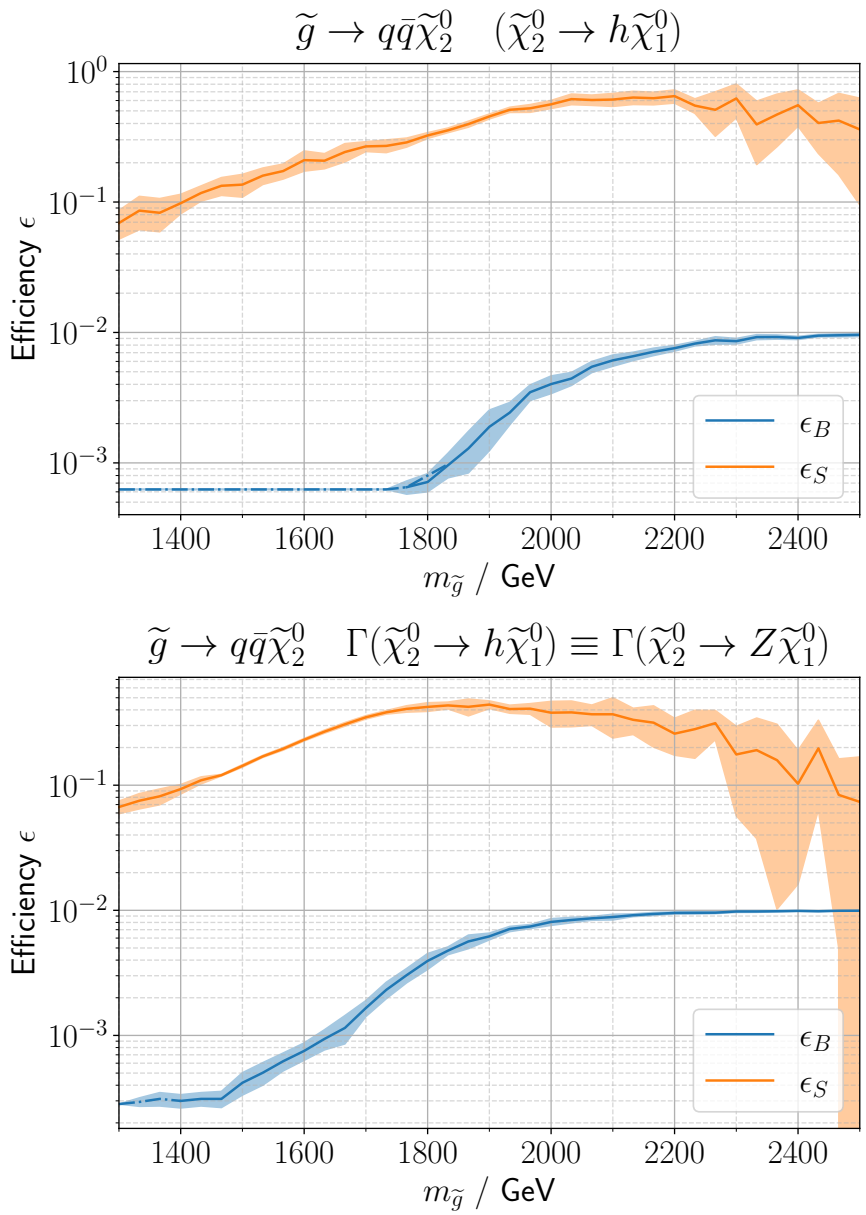

Figure A.9: Signal and background efficiencies of the model with decay $\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0$ and $\mathrm{Br}(\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0) \equiv \mathrm{Br}(\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0)$ with a cut on the anomaly score R_c such that 1% of events pass the selection. The dot-dashed part of the blue line represents parameter points where R_c has to be lowered to allow five background events. The shaded region shows one standard deviation around the mean efficiency.

A.5 ROC-Curves

In the figures A.10 and A.11 we show Receiver Operating Characteristic (ROC) curves, i.e. background suppression, as a function of the signal-efficiency of our benchmark models. We also observe a common feature of anomaly detection techniques. With rising signal cross section the classifier learns to separate background from signal-like events better. At the same time, larger signal cross sections correspond to smaller gluino masses, which in turn lead to less expressive features. Both effects combined lead to intermediate gluino masses having the largest background suppression at the same signal efficiency compared to small masses with large cross sections or large masses with very obvious signatures, especially in the decay to Z and Standard Model Higgs bosons. We also observe in the bottom right figure that for low and high Higgs masses the background rejection is noticeably weaker than for intermediate masses. For light Higgs masses, the jets are too similar to background jets, while high Higgs masses lead to wide jets that get reconstructed incorrectly.

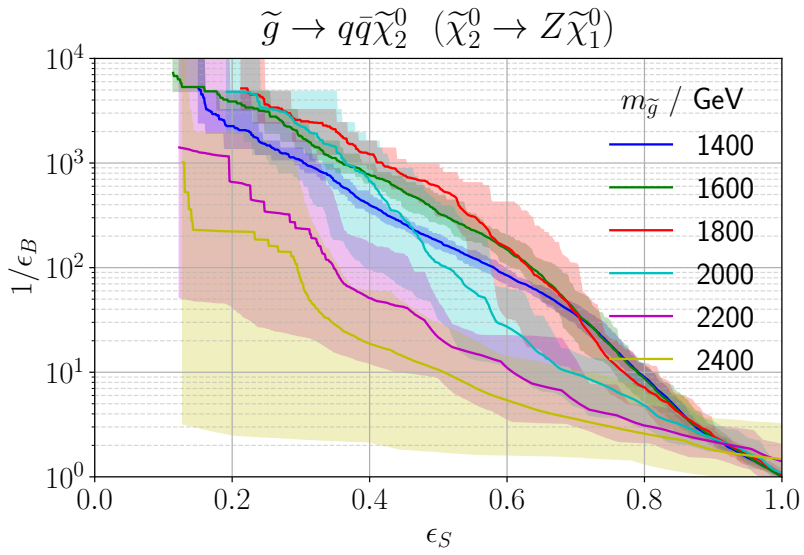

Figure A.10: ROC-curves of the $\tilde{\chi}_2^0 \to Z\tilde{\chi}_1^0$. Solid lines denote the mean value and shaded regions show the span between the minimum and maximum values obtained from ten different signal injections.

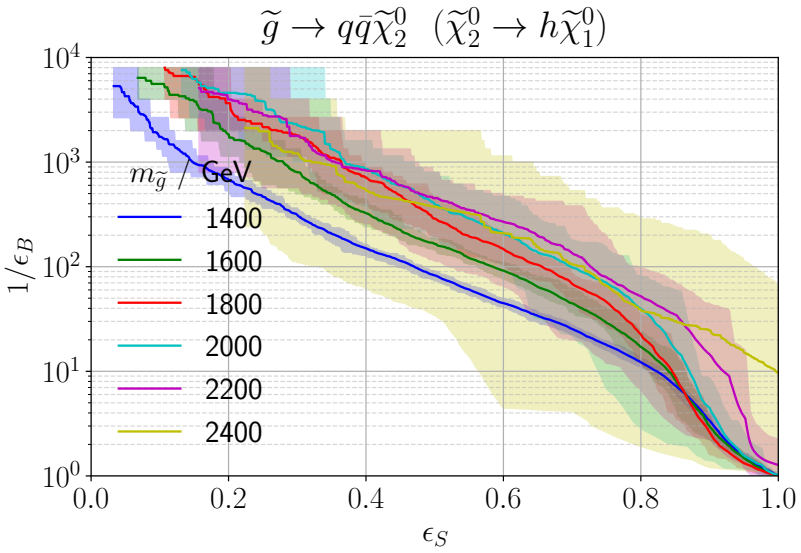

Figure A.11: ROC-curves of the $\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0$ models. Solid lines denote the mean value and shaded regions show the span between the minimum and maximum values obtained from ten different signal injections.

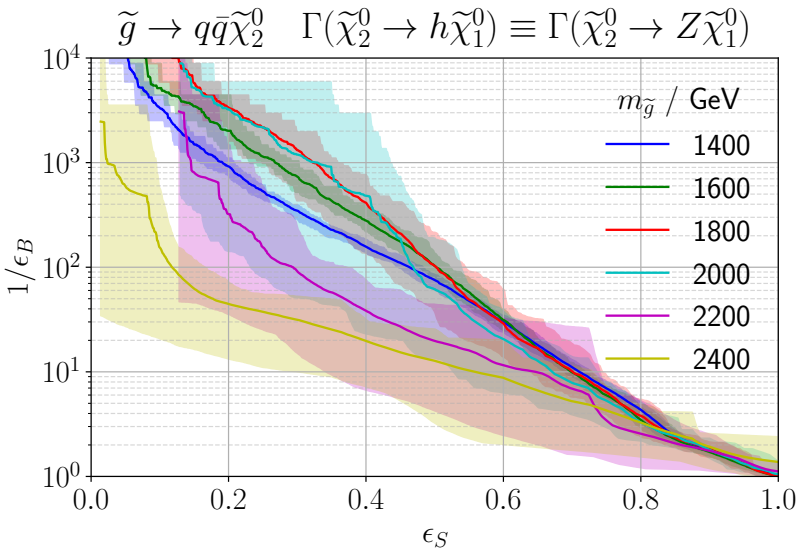

Figure A.12: ROC-curves of the $\mathrm{Br}(\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0) \equiv \mathrm{Br}(\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0)$ models. Solid lines denote the mean value and shaded regions show the span between the minimum and maximum values obtained from ten different signal injections.

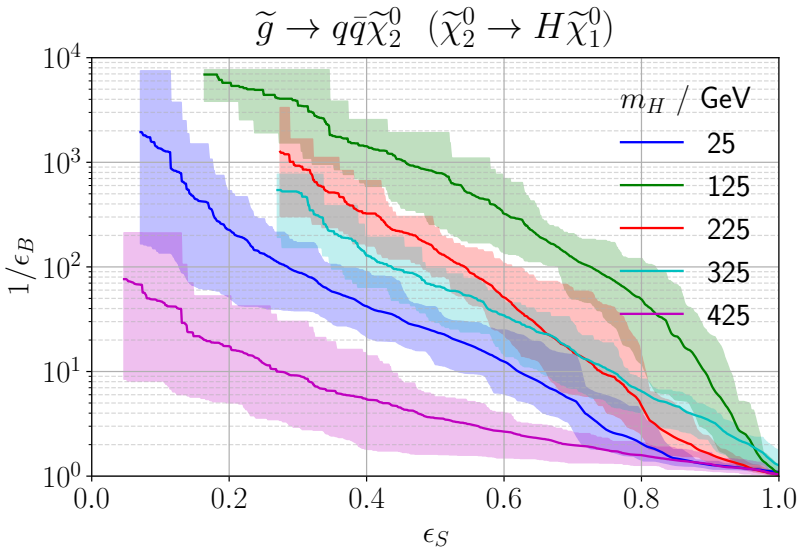

Figure A.13: ROC-curves of the $\widetilde{\chi}_2^0 \to H\widetilde{\chi}_1^0$ models. Solid lines denote the mean value and shaded regions show the span between the minimum and maximum values obtained from ten different signal injections.

APPENDIX B

Additional Studies on the Stop Pair Search

B.1 Additional Features

Here we show the additional features for the datasets DS1 and DS2 in Section 5.11. This is shown for all 6 844 975 background events and 9 500 signal events each. H_T is shown in figure B.1 while p_T^{miss}, M_j and N_j are shown in figure B.2. Although the Fox-Wolfram moments shown in figures B.3 and B.4 are only subtly different between most signal and background events, it proved still useful to include these in the larger dataset.

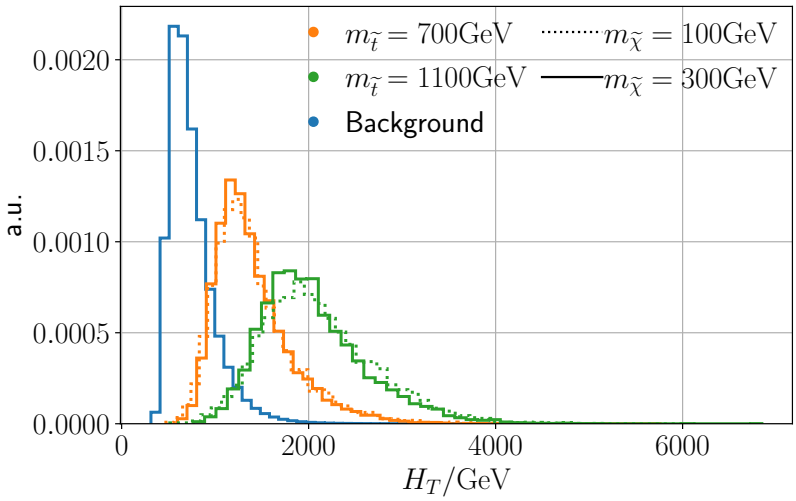

Figure B.1: Distribution of H_T in background events and four signal models as defined in Section 5.4. All histograms are normalized to unit area.

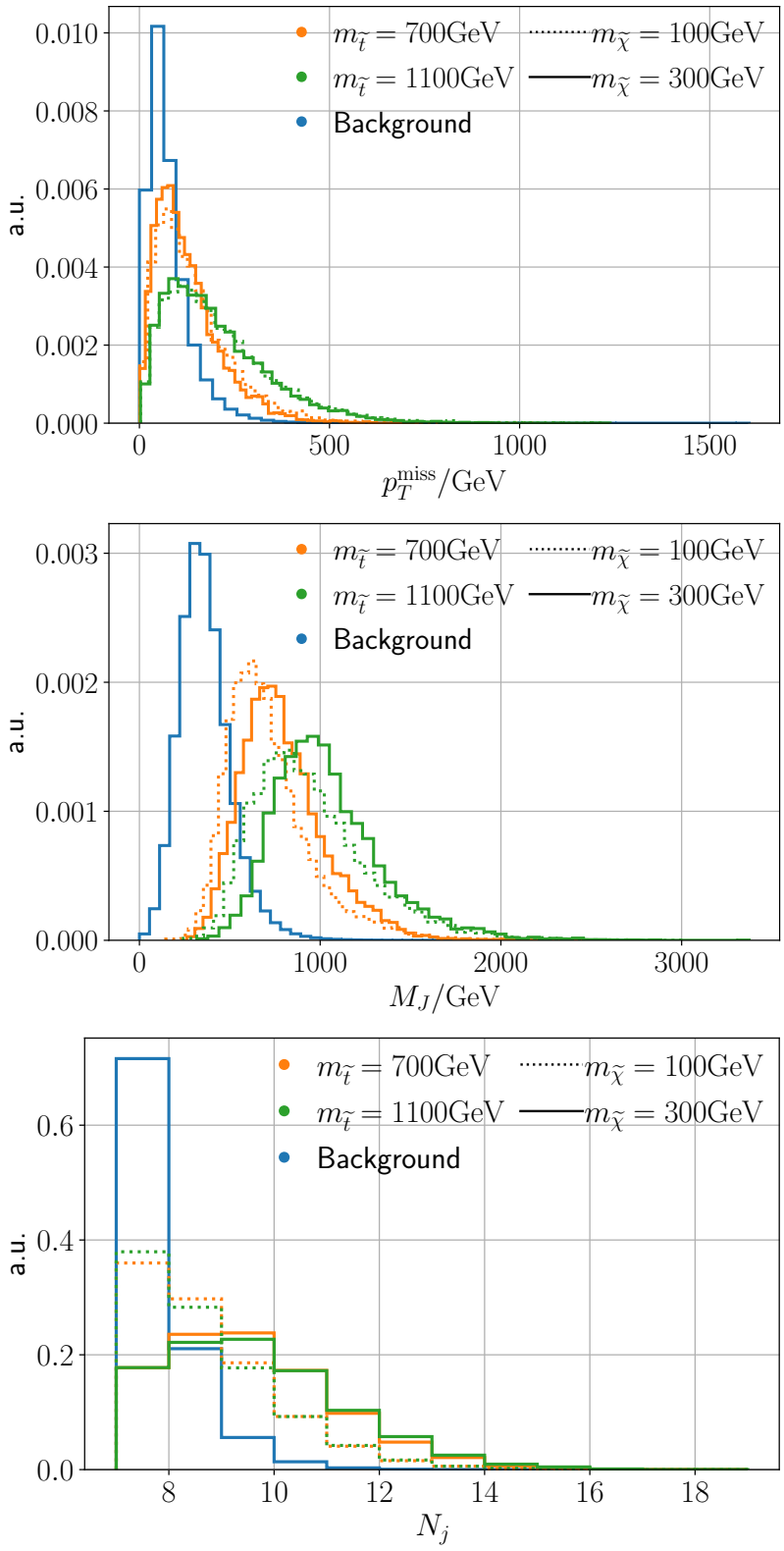

Figure B.2: Distribution of p_T^{miss}, M_J, N_j in background events and four signal models as defined in Section 5.4. All histograms are normalized to unit area.

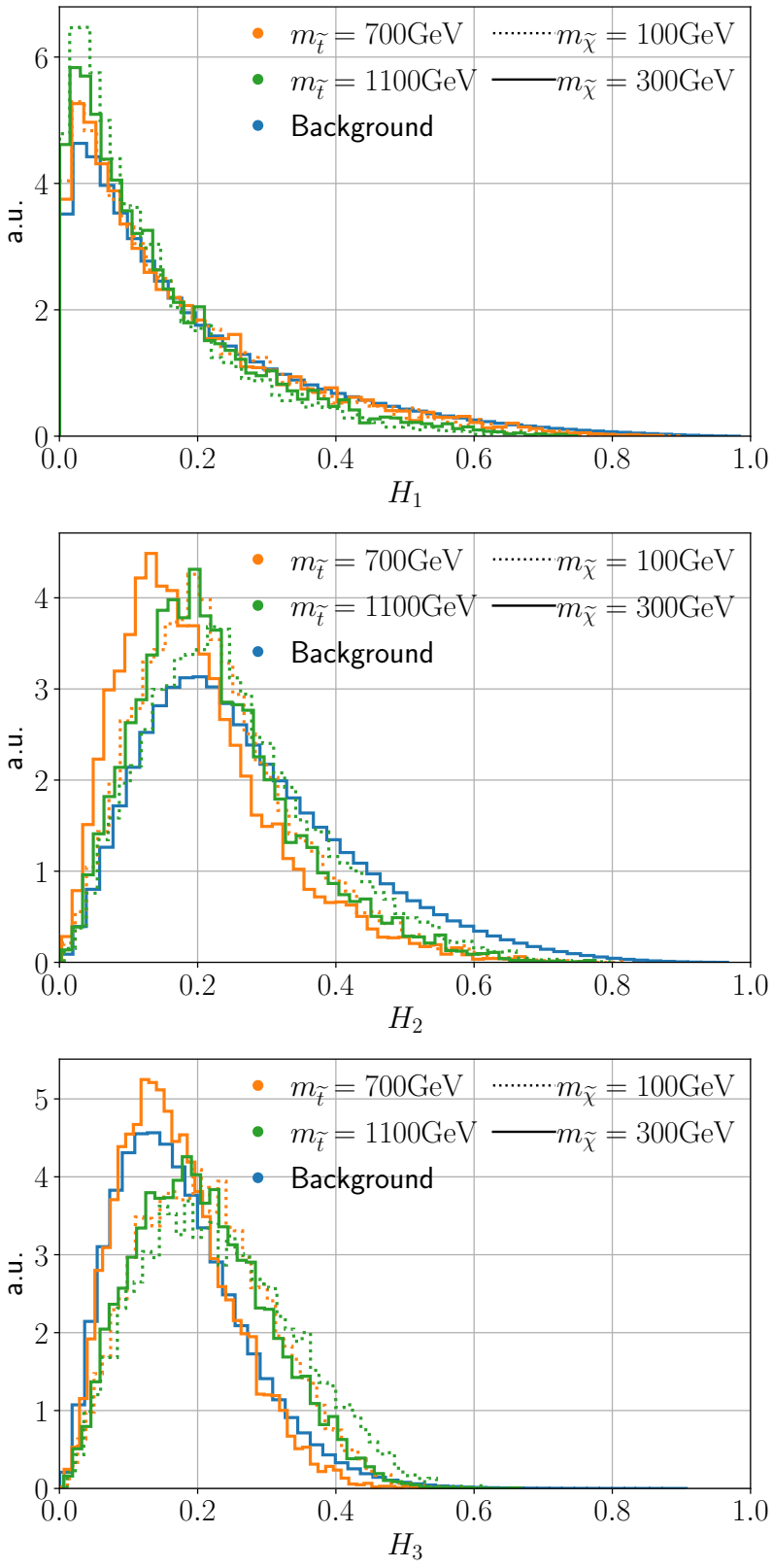

Figure B.3: Distribution of H_1, H_2, H_3 in background events and four signal models as defined in Section 5.4. All histograms are normalized to unit area.

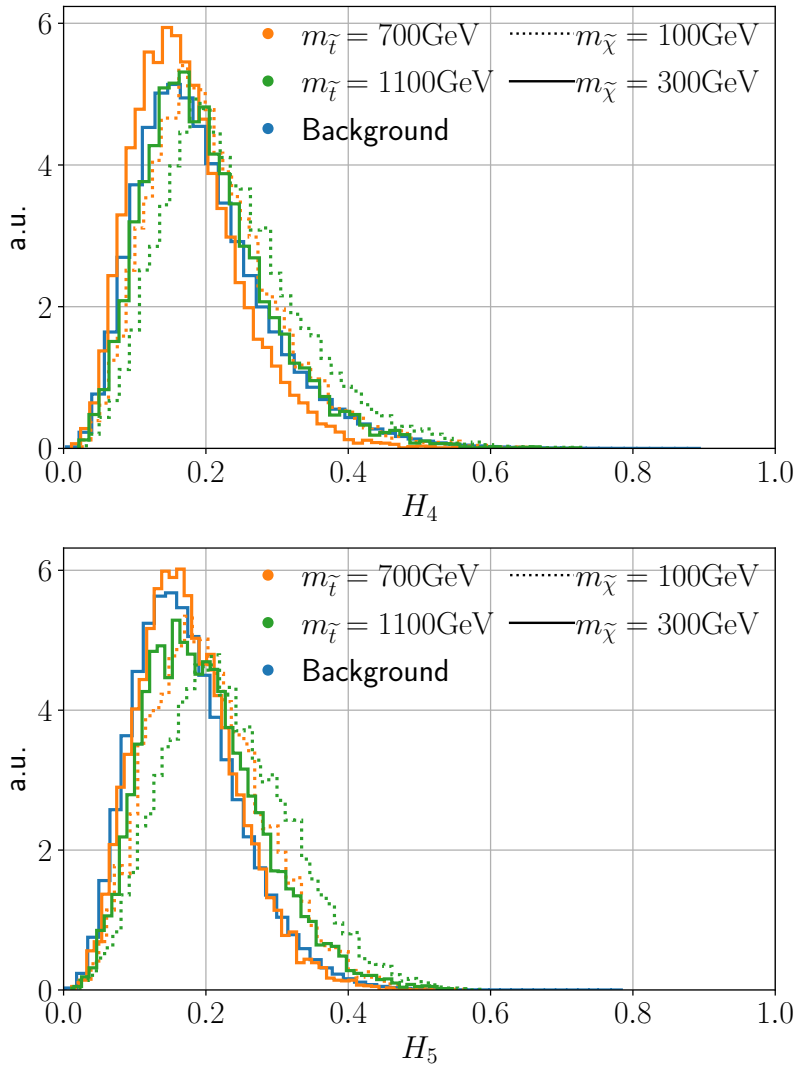

Figure B.4: Distribution of H_4, H_5 in background events and four signal models as defined in Section 5.4. All histograms are normalized to unit area.

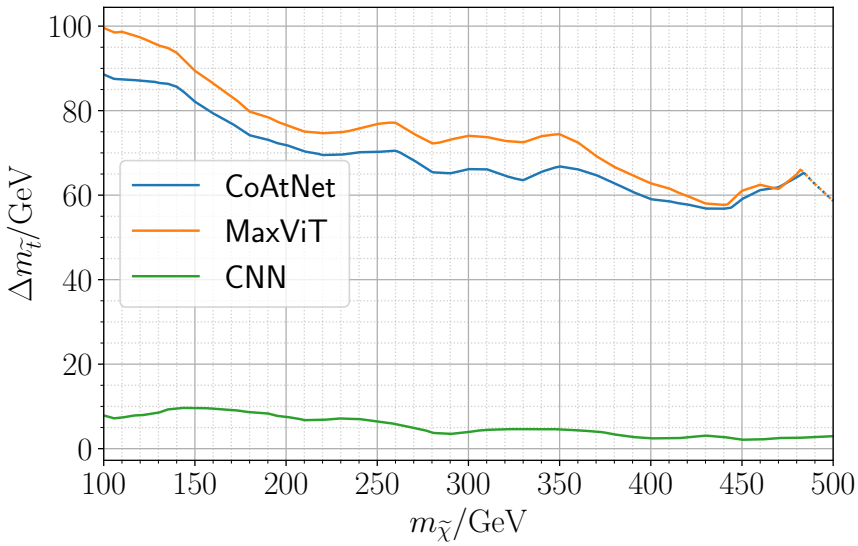

Figure B.5: Additional excluded stop masses Δm at 95% C.L. as a function of the neutralino mass $m_{\tilde{\chi}}$. The dotted line denotes stop masses where the kinematic GBDT excludes all probed stop masses. Therefore, the line is the worst case estimate.

B.2 Excluded Stop Masses

Here we show the contour of additionally excluded stop masses extracted from figure 5.9 for better comparison. The additional reach is shown in figure B.5.

B.3 Vanilla Vision Transformer

Here we demonstrate that the vanilla vision transformer [70], described in Section 3.6.1 will not yield competitive results when applied as a neutralino tagger. We use the publicly available code [1]. For this, we use 16×16 pixel patches. These patches are flattened into $n = h/16 \cdot w/16$ tokens of length $c \cdot 16^2$ where c is the number of input channels. The dimension of the tokens is $D = 256$, achieved by multiplying with a learnable matrix. The transformer stage consists of six transformer encoder layers. Each transformer layer contains 16 64-dimensional attention-heads. The MLP layer of the transformer has a single hidden layer with 1024 neurons. The `cls`-token is concatenated with the mass m and fed into the classification MLP-head identically to the other techniques. After the same training routine as the other techniques, the results are shown in figure B.6. Its performance is only slightly better than the CNN shown in figure 5.4 while MaxViT is far stronger. Therefore, we did not pursue this technique further.

[1] https://github.com/lucidrains/vit-pytorch

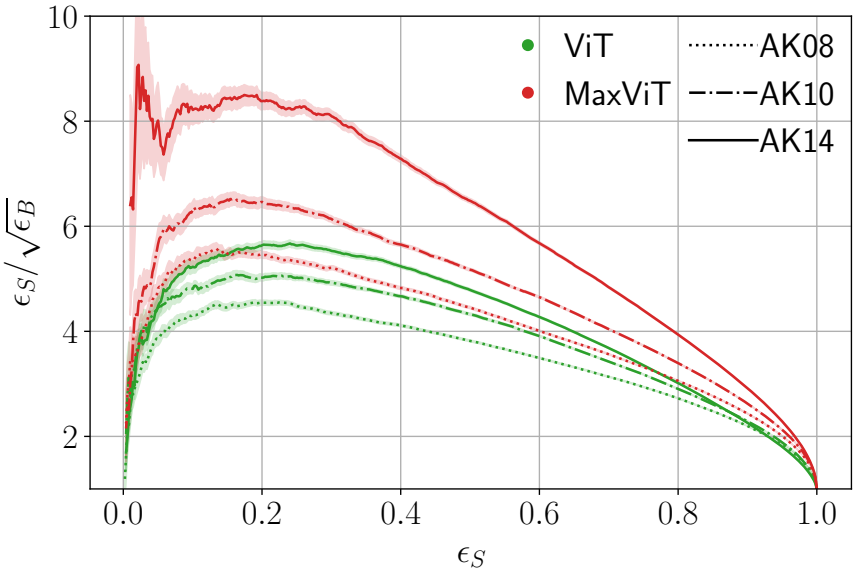

Figure B.6: Significance improvement curves for the vanilla vision transformer (ViT) compared to MaxViT for all single jet samples in the test data set. The shaded regions show one bootstrapped standard deviation.

Bibliography

[1] J. Erler and M. Schott,
Electroweak precision tests of the Standard Model after the discovery of the Higgs boson,
Progress in Particle and Nuclear Physics **106** (2019) 68, ISSN: 0146-6410,
URL: http://dx.doi.org/10.1016/j.ppnp.2019.02.007 (cit. on p. 1).

[2] G. Bertone, D. Hooper and J. Silk, *Particle dark matter: evidence, candidates and constraints*,
Physics Reports **405** (2005) 279, ISSN: 0370-1573,
URL: http://dx.doi.org/10.1016/j.physrep.2004.08.031 (cit. on pp. 1, 8).

[3] S. P. Martin, *A Supersymmetry Primer*, World Scientific, 1998, eprint: hep-ph/9709356,
URL: http://dx.doi.org/10.1142/9789812839657_0001 (cit. on pp. 1, 9–13, 56).

[4] R. L. e. a. Workman, *Review of Particle Physics*, PTEP **2022** (2022) 083C01
(cit. on pp. 1, 6, 35).

[5] B. W. Lee, C. Quigg and H. B. Thacker,
Weak interactions at very high energies: The role of the Higgs-boson mass,
Phys. Rev. D **16** (5 1977) 1519,
URL: https://link.aps.org/doi/10.1103/PhysRevD.16.1519 (cit. on p. 2).

[6] G. Ridolfi, *Search for the Higgs boson: theoretical perspectives*, 2001,
arXiv: hep-ph/0106300 [hep-ph], URL: https://arxiv.org/abs/hep-ph/0106300
(cit. on p. 2).

[7] R. Bruce, N. Fuster-Martínez, A. Mereghetti, D. Mirarchi and S. Redaelli,
Review of LHC Run 2 Machine Configurations, (2019) 187,
URL: https://cds.cern.ch/record/2750415 (cit. on p. 2).

[8] A. Hallin et al., *Classifying anomalies through outer density estimation*,
Phys. Rev. D **106** (2022) 055006, arXiv: 2109.00546 [hep-ph]
(cit. on pp. 2, 31, 34, 35, 40, 42, 44, 51, 86).

[9] M. Y. Han and Y. Nambu, *Three-Triplet Model with Double* SU(3) *Symmetry*,
Phys. Rev. **139** (4B 1965) B1006,
URL: https://link.aps.org/doi/10.1103/PhysRev.139.B1006 (cit. on p. 3).

[10] O. W. Greenberg,
Spin and Unitary-Spin Independence in a Paraquark Model of Baryons and Mesons,
Phys. Rev. Lett. **13** (20 1964) 598,
URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.598 (cit. on p. 3).

[11] S. L. Glashow, *The renormalizability of vector meson interactions*,
Nuclear Physics **10** (1959) 107, ISSN: 0029-5582,
URL: https://www.sciencedirect.com/science/article/pii/0029558259901968
(cit. on p. 3).

[12] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. C **680519** (1968) 367
(cit. on p. 3).

[13] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19** (21 1967) 1264,
URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264 (cit. on p. 3).

[14] C. G. Tully, *Elementary Particle Physics in a Nutshell*, Princeton University Press, 2011,
ISBN: 9780691131160,
URL: http://www.jstor.org/stable/j.ctvcmxp3f (visited on 28/06/2024)
(cit. on pp. 4, 5, 7).

[15] H. K. Dreiner, H. E. Haber and S. P. Martin, *Two-component spinor techniques and Feynman rules for quantum field theory and supersymmetry*, Physics Reports **494** (2010) 1,
ISSN: 0370-1573, URL: http://dx.doi.org/10.1016/j.physrep.2010.05.002
(cit. on p. 5).

[16] S. P. MARTIN,
"TASI 2011 Lectures Notes: Two-Component Fermion Notation and Supersymmetry",
The Dark Secrets of the Terascale, WORLD SCIENTIFIC, 2013 199, ISBN: 9789814390163,
URL: http://dx.doi.org/10.1142/9789814390163_0005 (cit. on p. 5).

[17] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*,
Phys. Rev. Lett. **13** (9 1964) 321,
URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.321 (cit. on p. 5).

[18] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*,
Phys. Rev. Lett. **13** (16 1964) 508,
URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508 (cit. on p. 5).

[19] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble,
Global Conservation Laws and Massless Particles, Phys. Rev. Lett. **13** (20 1964) 585,
URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.585 (cit. on p. 5).

[20] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*,
Phys. Lett. **12** (1964) 132 (cit. on p. 5).

[21] P. W. Higgs, *Spontaneous Symmetry Breakdown without Massless Bosons*, Phys. Rev. **145** (4 1966) 1156, URL: https://link.aps.org/doi/10.1103/PhysRev.145.1156 (cit. on p. 5).

[22] T. W. B. Kibble, *Symmetry Breaking in Non-Abelian Gauge Theories*, Phys. Rev. **155** (5 1967) 1554, URL: https://link.aps.org/doi/10.1103/PhysRev.155.1554 (cit. on p. 5).

[23] F. Halzen and A. D. Martin, *QUARKS AND LEPTONS: AN INTRODUCTORY COURSE IN MODERN PARTICLE PHYSICS*, 1984, ISBN: 978-0-471-88741-6 (cit. on p. 6).

[24] G. Aad et al., *Combined Measurement of the Higgs Boson Mass from the $H \to \gamma\gamma$ and $hh \to ZZ^* \to 4l$ decay channels with the ATLAS detector using $\sqrt{s} = 7$TeV Collision Data*, Physical Review Letters **131** (2023), ISSN: 1079-7114, URL: http://dx.doi.org/10.1103/PhysRevLett.131.251802 (cit. on p. 6).

[25] X. Fan, T. G. Myers, B. A. D. Sukra and G. Gabrielse, *Measurement of the Electron Magnetic Moment*, Phys. Rev. Lett. **130** (2023) 071801, arXiv: 2209.13084 [physics.atom-ph] (cit. on p. 7).

[26] A. G. Dias and V. Pleitez, *Grand unification and proton stability near the Peccei-Quinn scale*, Physical Review D **70** (2004), ISSN: 1550-2368, URL: http://dx.doi.org/10.1103/PhysRevD.70.055009 (cit. on p. 7).

[27] M. E. Peskin, *Supersymmetry in Elementary Particle Physics*, 2008, arXiv: 0801.1928 [hep-ph], URL: https://arxiv.org/abs/0801.1928 (cit. on p. 8).

[28] A. Borriello and P. Salucci, *The dark matter distribution in disc galaxies*, Monthly Notices of the Royal Astronomical Society **323** (2001) 285, ISSN: 1365-2966, URL: http://dx.doi.org/10.1046/j.1365-8711.2001.04077.x (cit. on p. 8).

[29] F. Zwicky, *On the Masses of Nebulae and of Clusters of Nebulae*, apj **86** (1937) 217 (cit. on p. 8).

[30] F. Zwicky, *The Redshift of Extragalactic Nebulae*, (2023) (cit. on p. 8).

[31] D. Clowe et al., *A Direct Empirical Proof of the Existence of Dark Matter**, The Astrophysical Journal **648** (2006) L109, URL: https://dx.doi.org/10.1086/508162 (cit. on p. 8).

[32] G. Madejski, "Recent and Future Observations in the X-ray and Gamma-ray Bands: Chandra, Suzaku, GLAST, and NuSTAR", *AIP Conference Proceedings*, AIP, 2005, URL: http://dx.doi.org/10.1063/1.2141828 (cit. on p. 8).

[33] N. Aghanim et al., *Planck2018 results: VI. Cosmological parameters*, Astronomy ; Astrophysics **641** (2020) A6, ISSN: 1432-0746, URL: http://dx.doi.org/10.1051/0004-6361/201833910 (cit. on p. 8).

[34] S. Bird et al., *Did LIGO Detect Dark Matter?*, Physical Review Letters **116** (2016), ISSN: 1079-7114, URL: http://dx.doi.org/10.1103/PhysRevLett.116.201301 (cit. on p. 8).

[35] E. Tiesinga, P. Mohr, D. Newell and B. Taylor, *CODATA Recommended Values of the Fundamental Physical Constants: 2018*, en, (2021), URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931443 (cit. on p. 8).

[36] J. Wess and B. Zumino, *Supergauge transformations in four dimensions*, Nuclear Physics B **70** (1974) 39, ISSN: 0550-3213, URL: https://www.sciencedirect.com/science/article/pii/0550321374903551 (cit. on pp. 9, 10).

[37] H. Nilles, *Supersymmetry, supergravity and particle physics*, Physics Reports **110** (1984) 1, ISSN: 0370-1573, URL: https://www.sciencedirect.com/science/article/pii/0370157384900085 (cit. on pp. 9, 10).

[38] M. Drees, R. Godbole and P. Roy, *Theory and Phenomenology of Sparticles*, World Scientific, 2005, URL: https://www.worldscientific.com/doi/abs/10.1142/4001 (cit. on pp. 9, 10).

[39] S. Coleman and J. Mandula, *All Possible Symmetries of the S Matrix*, Phys. Rev. **159** (5 1967) 1251, URL: https://link.aps.org/doi/10.1103/PhysRev.159.1251 (cit. on p. 9).

[40] R. Haag, J. T. Łopuszański and M. Sohnius, *All possible generators of supersymmetries of the S-matrix*, Nuclear Physics B **88** (1975) 257, ISSN: 0550-3213, URL: https://www.sciencedirect.com/science/article/pii/0550321375902795 (cit. on p. 9).

[41] H. Haber and G. Kane, *The search for supersymmetry: Probing physics beyond the standard model*, Physics Reports **117** (1985) 75, ISSN: 0370-1573, URL: https://www.sciencedirect.com/science/article/pii/0370157385900511 (cit. on p. 10).

[42] P. Fayet and S. Ferrara, *Supersymmetry*, Physics Reports **32** (1977) 249, ISSN: 0370-1573, URL: https://www.sciencedirect.com/science/article/pii/0370157377900667 (cit. on p. 10).

[43] I. Jack and D. Jones, *Non-standard soft supersymmetry breaking*, Physics Letters B **457** (1999) 101, ISSN: 0370-2693, URL: http://dx.doi.org/10.1016/S0370-2693(99)00530-4 (cit. on p. 12).

[44] P. Fayet and J. Iliopoulos,
Spontaneously broken supergauge symmetries and goldstone spinors,
Physics Letters B **51** (1974) 461, ISSN: 0370-2693,
URL: https://www.sciencedirect.com/science/article/pii/0370269374903104
(cit. on p. 12).

[45] P. Fayet, *Supergauge invariant extension of the Higgs mechanism and a model for the electron and its neutrino*, Nuclear Physics B **90** (1975) 104, ISSN: 0550-3213,
URL: https://www.sciencedirect.com/science/article/pii/0550321375906367
(cit. on p. 12).

[46] L. O'Raifeartaigh, *Spontaneous symmetry breaking for chirals scalar superfields*,
Nuclear Physics B **96** (1975) 331, ISSN: 0550-3213,
URL: https://www.sciencedirect.com/science/article/pii/0550321375905854
(cit. on p. 12).

[47] C. R. Nappi and B. A. Ovrut, *Supersymmetric extension of the SU(3)×SU(2)×U(1) model*,
Physics Letters B **113** (1982) 175, ISSN: 0370-2693,
URL: https://www.sciencedirect.com/science/article/pii/037026938290418X
(cit. on p. 12).

[48] M. A. Luty, *2004 TASI Lectures on Supersymmetry Breaking*, 2005,
arXiv: hep-th/0509029 [hep-th], URL: https://arxiv.org/abs/hep-th/0509029
(cit. on p. 12).

[49] C. A. Baker et al.,
Improved Experimental Limit on the Electric Dipole Moment of the Neutron,
Physical Review Letters **97** (2006), ISSN: 1079-7114,
URL: http://dx.doi.org/10.1103/PhysRevLett.97.131801 (cit. on p. 12).

[50] J. M. Pendlebury et al.,
Revised experimental upper limit on the electric dipole moment of the neutron,
Physical Review D **92** (2015), ISSN: 1550-2368,
URL: http://dx.doi.org/10.1103/PhysRevD.92.092003 (cit. on p. 12).

[51] B. Graner, Y. Chen, E. G. Lindahl and B. R. Heckel,
Reduced Limit on the Permanent Electric Dipole Moment of 199Hg,
Physical Review Letters **116** (2016), ISSN: 1079-7114,
URL: http://dx.doi.org/10.1103/PhysRevLett.116.161601 (cit. on p. 12).

[52] S. Dimopoulos and D. Sutter, *The supersymmetric flavor problem*,
Nuclear Physics B **452** (1995) 496, ISSN: 0550-3213,
URL: http://dx.doi.org/10.1016/0550-3213(95)00421-N (cit. on p. 12).

[53] H. E. Haber, *The status of the minimal supersymmetric standard model and beyond*, Nuclear Physics B - Proceedings Supplements **62** (1998) 469, Proceedings of the Fifth International Conference on Supersymmetries in Physics, ISSN: 0920-5632, URL: https://www.sciencedirect.com/science/article/pii/S0920563297006889 (cit. on p. 12).

[54] N. Polonsky, *Supersymmetry Structure and Phenomena*, 2001, arXiv: hep-ph/0108236 [hep-ph], URL: https://arxiv.org/abs/hep-ph/0108236 (cit. on p. 12).

[55] C. Han, J. Ren, L. Wu, J. M. Yang and M. Zhang, *Top-squark in natural SUSY under current LHC run-2 data*, The European Physical Journal C **77** (2017), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-017-4662-7 (cit. on p. 13).

[56] S. Geer et al., *Search for anti-proton decay at the Fermilab anti-proton accumulator*, Phys. Rev. D **62** (2000) 052004, arXiv: hep-ex/9908036 (cit. on p. 14).

[57] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, http://www.deeplearningbook.org, MIT Press, 2016 (cit. on pp. 18, 24).

[58] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2017, arXiv: 1412.6980 [cs.LG] (cit. on pp. 18, 67).

[59] T. Tieleman and G. Hinton, *Lecture 6.5-rmsprop, coursera: Neural networks for machine learning*, University of Toronto, Technical Report, 2012 (cit. on p. 18).

[60] S. J. Prince, *Understanding Deep Learning*, The MIT Press, 2023, URL: http://udlbook.com (cit. on p. 19).

[61] A. Pinkus, *Approximation theory of the MLP model in neural networks*, Acta Numerica **8** (1999) 143 (cit. on p. 19).

[62] S. Park, C. Yun, J. Lee and J. Shin, *Minimum Width for Universal Approximation*, 2020, arXiv: 2006.08859 [cs.LG] (cit. on p. 19).

[63] L. Breiman, J. Friedman, C. Stone and R. Olshen, *Classification and Regression Trees*, Taylor & Francis, 1984, ISBN: 9780412048418, URL: https://books.google.de/books?id=JwQx-WOmSyQC (cit. on p. 20).

[64] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, ACM, 2016, eprint: 1603.02754, URL: http://dx.doi.org/10.1145/2939672.2939785 (cit. on pp. 20, 69).

[65] Y. LeCun et al., *Backpropagation Applied to Handwritten Zip Code Recognition*, Neural Computation **1** (1989) 541 (cit. on p. 23).

[66] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner,
Gradient-based learning applied to document recognition,
Proceedings of the IEEE **86** (1998) 2278 (cit. on p. 23).

[67] M.-H. Guo et al., *Attention mechanisms in computer vision: A survey*,
Computational Visual Media **8** (2022) 331, ISSN: 2096-0662,
URL: http://dx.doi.org/10.1007/s41095-022-0271-y (cit. on p. 24).

[68] A. Vaswani et al., *Attention Is All You Need*, 2017, arXiv: 1706.03762 [cs.CL]
(cit. on pp. 24, 25).

[69] J. L. Ba, J. R. Kiros and G. E. Hinton, *Layer Normalization*, 2016,
arXiv: 1607.06450 [stat.ML] (cit. on p. 25).

[70] A. Dosovitskiy et al.,
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021,
arXiv: 2010.11929 [cs.CV] (cit. on pp. 26, 60, 101).

[71] Z. Dai, H. Liu, Q. V. Le and M. Tan,
CoAtNet: Marrying Convolution and Attention for All Data Sizes, 2021,
arXiv: 2106.04803 [cs.CV] (cit. on pp. 26, 60).

[72] Z. Tu et al., *MaxViT: Multi-Axis Vision Transformer*, 2022, arXiv: 2204.01697 [cs.CV]
(cit. on pp. 27, 60, 66).

[73] D. G. a. Smith and J. Gray,
opt_einsum - A Python package for optimizing contraction order for einsum-like expressions,
Journal of Open Source Software **3** (2018) 753,
URL: https://doi.org/10.21105/joss.00753 (cit. on p. 27).

[74] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen,
MobileNetV2: Inverted Residuals and Linear Bottlenecks, 2019,
arXiv: 1801.04381 [cs.CV] (cit. on pp. 28, 66).

[75] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006,
URL: https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/ (cit. on p. 28).

[76] T. A. O. Brien, K. Kashinath, N. R. Cavanaugh, W. D. Collins and J. P. O. Brien,
A fast and objective multidimensional kernel density estimation method: fastKDE,
Computational Statistics and Data Analysis **101** (2016) 148, ISSN: 0167-9473, URL:
https://www.sciencedirect.com/science/article/pii/S0167947316300408
(cit. on p. 28).

[77] I. Kobyzev, S. J. Prince and M. A. Brubaker,
Normalizing Flows: An Introduction and Review of Current Methods,
IEEE Transactions on Pattern Analysis and Machine Intelligence **43** (2021) 3964,
ISSN: 1939-3539, URL: http://dx.doi.org/10.1109/TPAMI.2020.2992934
(cit. on p. 29).

[78] D. J. Rezende and S. Mohamed, *Variational Inference with Normalizing Flows*, 2016,
arXiv: 1505.05770 [stat.ML] (cit. on p. 29).

[79] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed and B. Lakshminarayanan,
Normalizing flows for probabilistic modeling and inference, J. Mach. Learn. Res. **22** (2021),
ISSN: 1532-4435 (cit. on p. 29).

[80] C.-W. Huang, D. Krueger, A. Lacoste and A. Courville, *Neural Autoregressive Flows*, 2018,
arXiv: 1804.00779 [cs.LG] (cit. on p. 30).

[81] G. Papamakarios, T. Pavlakou and I. Murray,
Masked Autoregressive Flow for Density Estimation, 2017,
URL: https://arxiv.org/abs/1705.07057 (cit. on pp. 30, 41).

[82] M. Germain, K. Gregor, I. Murray and H. Larochelle,
MADE: Masked Autoencoder for Distribution Estimation, (2015),
URL: https://arxiv.org/abs/1502.03509 (cit. on pp. 30, 41).

[83] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman and D. Shih,
Machine Learning in the Search for New Fundamental Physics, (2021),
arXiv: 2112.03769 [hep-ph] (cit. on p. 31).

[84] J. H. Collins, K. Howe and B. Nachman,
Anomaly Detection for Resonant New Physics with Machine Learning,
Phys. Rev. Lett. **121** (2018) 241803, arXiv: 1805.02664 [hep-ph] (cit. on p. 31).

[85] T. Aarrestad et al., *The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider*,
SciPost Phys. **12** (2022) 043, arXiv: 2105.14027 [hep-ph] (cit. on p. 31).

[86] J. H. Collins, K. Howe and B. Nachman,
Extending the search for new resonances with machine learning,
Phys. Rev. D **99** (2019) 014038, arXiv: 1902.02634 [hep-ph] (cit. on p. 31).

[87] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*,
Phys. Rev. D **101** (2020) 075042, arXiv: 2001.04990 [hep-ph] (cit. on p. 31).

[88] A. Andreassen, B. Nachman and D. Shih,
Simulation Assisted Likelihood-free Anomaly Detection, Phys. Rev. D **101** (2020) 095004,
arXiv: 2001.05001 [hep-ph] (cit. on p. 31).

[89] O. Amram and C. M. Suarez,
Tag N' Train: a technique to train improved classifiers on unlabeled data,
JHEP **01** (2021) 153, arXiv: 2002.12376 [hep-ph] (cit. on p. 31).

[90] G. Aad et al., *Dijet resonance search with weak supervision using* $\sqrt{s} = 13$ *TeV pp collisions in the ATLAS detector*, Phys. Rev. Lett. **125** (2020) 131801, arXiv: 2005.02983 [hep-ex] (cit. on p. 31).

[91] K. Benkendorfer, L. L. Pottier and B. Nachman,
Simulation-assisted decorrelation for resonant anomaly detection,
Phys. Rev. D **104** (2021) 035003, arXiv: 2009.02205 [hep-ph] (cit. on p. 31).

[92] G. Stein, U. Seljak and B. Dai, "Unsupervised in-distribution anomaly detection of new physics through conditional density estimation",
34th Conference on Neural Information Processing Systems, 2020,
arXiv: 2012.11638 [cs.LG] (cit. on p. 31).

[93] S. E. Park, D. Rankin, S.-M. Udrescu, M. Yunus and P. Harris,
Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge,
JHEP **21** (2020) 030, arXiv: 2011.03550 [hep-ph] (cit. on p. 31).

[94] J. H. Collins, P. Martín-Ramiro, B. Nachman and D. Shih,
Comparing weak- and unsupervised methods for resonant anomaly detection,
Eur. Phys. J. C **81** (2021) 617, arXiv: 2104.02092 [hep-ph] (cit. on p. 31).

[95] J. F. Kamenik and M. Szewc, *Null hypothesis test for anomaly detection*,
Phys. Lett. B **840** (2023) 137836, arXiv: 2210.02226 [hep-ph] (cit. on p. 31).

[96] J. A. Raine, S. Klein, D. Sengupta and T. Golling, *CURTAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals*,
Front. Big Data **6** (2023) 899345, arXiv: 2203.09470 [hep-ph] (cit. on p. 31).

[97] G. Kasieczka et al., *Anomaly detection under coordinate transformations*,
Phys. Rev. D **107** (2023) 015009, arXiv: 2209.06225 [hep-ph] (cit. on p. 31).

[98] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih and M. Sommerhalder,
Resonant anomaly detection without background sculpting, Phys. Rev. D **107** (2023) 114012,
arXiv: 2210.14924 [hep-ph] (cit. on pp. 31, 52).

[99] M. F. Chen, B. Nachman and F. Sala,
Resonant Anomaly Detection with Multiple Reference Datasets, (2022),
arXiv: 2212.10579 [hep-ph] (cit. on p. 31).

[100] T. Golling, S. Klein, R. Mastandrea and B. Nachman,
Flow-enhanced transportation for anomaly detection, Phys. Rev. D **107** (2023) 096025,
arXiv: 2212.11285 [hep-ph] (cit. on p. 31).

[101] T. Golling et al.,
The Interplay of Machine Learning–based Resonant Anomaly Detection Methods, (2023),
arXiv: 2307.11157 [hep-ph] (cit. on p. 31).

[102] D. Sengupta, S. Klein, J. A. Raine and T. Golling, *CURTAINs Flows For Flows: Constructing Unobserved Regions with Maximum Likelihood Estimation*, (2023),
arXiv: 2305.04646 [hep-ph] (cit. on p. 31).

[103] J. Neyman and E. S. Pearson,
"On the Problem of the Most Efficient Tests of Statistical Hypotheses",
Breakthroughs in Statistics: Foundations and Basic Theory,
New York, NY: Springer New York, 1992 73, ISBN: 978-1-4612-0919-5,
URL: https://doi.org/10.1007/978-1-4612-0919-5_6 (cit. on p. 32).

[104] E. M. Metodiev, B. Nachman and J. Thaler,
Classification without labels: Learning from mixed samples in high energy physics,
JHEP **10** (2017) 174, arXiv: 1708.02949 [hep-ph] (cit. on p. 32).

[105] G. Aad et al., *Dijet Resonance Search with Weak Supervision Using* $\sqrt{s} = 13$ TeV pp *Collisions in the ATLAS Detector*, Phys. Rev. Lett. **125** (13 2020) 131801,
URL: https://link.aps.org/doi/10.1103/PhysRevLett.125.131801 (cit. on p. 32).

[106] A. Sirunyan et al., *Measurement of the ttbb production cross section in the all-jet final state in pp collisions at s=13 TeV*, Physics Letters B **803** (2020) 135285, ISSN: 0370-2693, URL:
https://www.sciencedirect.com/science/article/pii/S0370269320300897
(cit. on p. 32).

[107] O. Amram and C. M. Suarez,
Tag N' Train: a technique to train improved classifiers on unlabeled data,
Journal of High Energy Physics **2021** (2021),
URL: https://doi.org/10.1007%2Fjhep01%282021%29153 (cit. on p. 33).

[108] T. Heimel, G. Kasieczka, T. Plehn and J. Thompson, *QCD or what?*, SciPost Physics **6** (2019),
ISSN: 2542-4653, URL: http://dx.doi.org/10.21468/SciPostPhys.6.3.030
(cit. on p. 33).

[109] M. Farina, Y. Nakai and D. Shih, *Searching for new physics with deep autoencoders*,
Physical Review D **101** (2020), ISSN: 2470-0029,
URL: http://dx.doi.org/10.1103/PhysRevD.101.075021 (cit. on p. 33).

[110] A. Blance, M. Spannowsky and P. Waite,
Adversarially-trained autoencoders for robust unsupervised new physics searches,
Journal of High Energy Physics **2019** (2019), ISSN: 1029-8479,
URL: http://dx.doi.org/10.1007/JHEP10(2019)047 (cit. on p. 33).

[111] T. S. Roy and A. H. Vijay, *A robust anomaly finder based on autoencoders*, 2020,
arXiv: 1903.02032 [hep-ph] (cit. on p. 33).

[112] B. Nachman and D. Shih, *Anomaly detection with density estimation*, Phys. Rev. D **101** (7 2020) 075042, URL: https://link.aps.org/doi/10.1103/PhysRevD.101.075042 (cit. on p. 33).

[113] G. Kasieczka et al., *The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics*, Rept. Prog. Phys. **84** (2021) 124201, arXiv: 2101.08320 [hep-ph] (cit. on p. 35).

[114] A. Mullin et al., *Does SUSY have friends? A new approach for LHC event analysis*, JHEP **02** (2021) 160, arXiv: 1912.10625 [hep-ph] (cit. on p. 35).

[115] *Search for supersymmetry in final states with disappearing tracks in proton-proton collisions at* 13 TeV, tech. rep., CERN, 2023, URL: http://cds.cern.ch/record/2859611 (cit. on p. 35).

[116] G. Aad et al., *Search for supersymmetry in final states with missing transverse momentum and three or more b-jets in 139 fb^{-1} of proton–proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Eur. Phys. J. C **83** (2023) 561, arXiv: 2211.08028 [hep-ex] (cit. on p. 35).

[117] A. Tumasyan et al., *Search for top squarks in the four-body decay mode with single lepton final states in proton-proton collisions at $\sqrt{s}$ = 13 TeV*, JHEP **06** (2023) 060, arXiv: 2301.08096 [hep-ex] (cit. on p. 35).

[118] A. Choudhury, A. Mondal, S. Mondal and S. Sarkar, *Improving sensitivity of trilinear RPV SUSY searches using machine learning at the LHC*, (2023), arXiv: 2308.02697 [hep-ph] (cit. on p. 35).

[119] A. M. Sirunyan et al., *Search for supersymmetry in proton-proton collisions at $\sqrt{s}$ = 13 TeV in events with high-momentum Z bosons and missing transverse momentum*, JHEP **09** (2020) 149, arXiv: 2008.04422 [hep-ex], URL: https://doi.org/10.1007%2Fjhep09%282020%29149 (cit. on pp. 35, 38, 44, 46, 50, 51, 83, 85).

[120] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, Journal of High Energy Physics **2014** (2014), eprint: 1405.0301, URL: https://doi.org/10.1007%2Fjhep07%282014%29079 (cit. on pp. 37, 62, 76).

[121] R. D. Ball et al., *Parton distributions from high-precision collider data*, Eur. Phys. J. C **77** (2017) 663, arXiv: 1706.00428 [hep-ph] (cit. on p. 37).

[122] P. Artoisenet, R. Frederix, O. Mattelaer and R. Rietkerk, *Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations*, JHEP **03** (2013) 015, arXiv: 1212.3460 [hep-ph] (cit. on p. 37).

[123] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, (2022), arXiv: 2203.11601 [hep-ph] (cit. on p. 37).

[124] A. M. Sirunyan et al., *Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements*, Eur. Phys. J. C **80** (2020) 4, arXiv: 1903.12179 [hep-ex] (cit. on p. 37).

[125] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv: 1405.0301 [hep-ph] (cit. on p. 37).

[126] J. de Favereau et al., *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, JHEP **02** (2014) 057, arXiv: 1307.6346 [hep-ex] (cit. on p. 37).

[127] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, JHEP **03** (2011) 015, arXiv: 1011.2268 [hep-ph] (cit. on p. 40).

[128] P. N. Bhattiprolu, S. P. Martin and J. D. Wells, *Criteria for projected discovery and exclusion sensitivities of counting experiments*, The European Physical Journal C **81** (2021), URL: https://doi.org/10.1140%2Fepjc%2Fs10052-020-08819-6 (cit. on p. 46).

[129] A. Tumasyan et al., *Search for higgsinos decaying to two Higgs bosons and missing transverse momentum in proton-proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **05** (2022) 014, arXiv: 2201.04206 [hep-ex] (cit. on pp. 47, 50, 51, 56).

[130] A. M. Sirunyan et al., *Search for Physics Beyond the Standard Model in Events with High-Momentum Higgs Bosons and Missing Transverse Momentum in Proton-Proton Collisions at 13 TeV*, Phys. Rev. Lett. **120** (2018) 241801, arXiv: 1712.08501 [hep-ex] (cit. on pp. 48, 50, 51).

[131] T. Finke et al., *Back To The Roots: Tree-Based Algorithms for Weakly Supervised Anomaly Detection*, 2023, arXiv: 2309.13111 [hep-ph] (cit. on p. 52).

[132] *Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV*, tech. rep., CERN, 2024, URL: https://cds.cern.ch/record/2892677 (cit. on p. 53).

[133] A. Collaboration, *The quest to discover supersymmetry at the ATLAS experiment*, 2024, arXiv: 2403.02455 [hep-ex] (cit. on p. 55).

[134] R. Barbier et al., *R−Parity-violating supersymmetry*, Physics Reports **420** (2005) 1, ISSN: 0370-1573, eprint: hep-ph/0406039, URL: http://dx.doi.org/10.1016/j.physrep.2005.08.006 (cit. on p. 56).

[135] A. Redelbach, *Searches for Prompt R−Parity-Violating Supersymmetry at the LHC*, 2015, arXiv: 1512.05956 [hep-ex] (cit. on p. 56).

[136] J. L. Feng, K. T. Matchev and T. Moroi, *Multi - TeV scalars are natural in minimal supergravity*, Phys. Rev. Lett. **84** (2000) 2322, arXiv: hep-ph/9908309 (cit. on p. 56).

[137] R. Kitano and Y. Nomura, *Supersymmetry, naturalness, and signatures at the LHC*, Phys. Rev. D **73** (2006) 095004, arXiv: hep-ph/0602096 (cit. on p. 56).

[138] M. Papucci, J. T. Ruderman and A. Weiler, *Natural SUSY Endures*, JHEP **09** (2012) 035, arXiv: 1110.6926 [hep-ph] (cit. on p. 56).

[139] G. G. Ross, K. Schmidt-Hoberg and F. Staub, *Revisiting fine-tuning in the MSSM*, JHEP **03** (2017) 021, arXiv: 1701.03480 [hep-ph] (cit. on p. 56).

[140] H. Baer, V. Barger, S. Salam, D. Sengupta and X. Tata, *The LHC higgsino discovery plane for present and future SUSY searches*, Phys. Lett. B **810** (2020) 135777, arXiv: 2007.09252 [hep-ph] (cit. on p. 56).

[141] G. Aad et al., *Search for direct production of winos and higgsinos in events with two same-charge leptons or three leptons in pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **11** (2023) 150, arXiv: 2305.09322 [hep-ex] (cit. on p. 56).

[142] G. Aad et al., *Search for nearly mass-degenerate higgsinos using low-momentum mildly-displaced tracks in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, (2024), arXiv: 2401.14046 [hep-ex] (cit. on p. 56).

[143] G. Aad et al., *Search for pair production of higgsinos in events with two Higgs bosons and missing transverse momentum in $\sqrt{s} = 13$ TeV pp collisions at the ATLAS experiment*, (2024), arXiv: 2401.14922 [hep-ex] (cit. on p. 56).

[144] A. Hayrapetyan et al., *Combined search for electroweak production of winos, binos, higgsinos, and sleptons in proton-proton collisions at $\sqrt{s} = 13$ TeV*, (2024), arXiv: 2402.01888 [hep-ex] (cit. on p. 56).

[145] J. A. Evans and Y. Kats, *LHC coverage of RPV MSSM with light stops*, Journal of High Energy Physics **2013** (2013), ISSN: 1029-8479, eprint: 1209.0764, URL: http://dx.doi.org/10.1007/JHEP04(2013)028 (cit. on p. 56).

[146] J. M. Butterworth, J. R. Ellis, A. R. Raklev and G. P. Salam, *Discovering Baryon-Number Violating Neutralino Decays at the LHC*, Physical Review Letters **103** (2009), ISSN: 1079-7114, eprint: 0906.0728, URL: http://dx.doi.org/10.1103/PhysRevLett.103.241803 (cit. on p. 57).

[147] J. Pearkes, W. Fedorko, A. Lister and C. Gay,
Jet Constituents for Deep Neural Network Based Top Quark Tagging, 2017,
arXiv: 1704.02124 [hep-ex] (cit. on p. 57).

[148] M. Aaboud et al.,
Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC,
The European Physical Journal C **79** (2019), ISSN: 1434-6052,
URL: http://dx.doi.org/10.1140/epjc/s10052-019-6847-8 (cit. on p. 57).

[149] G. Aad et al.,
Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1,
The European Physical Journal C **77** (2017), ISSN: 1434-6052,
URL: http://dx.doi.org/10.1140/epjc/s10052-017-5004-5 (cit. on p. 57).

[150] S. Macaluso and D. Shih, *Pulling out all the tops with computer vision and deep learning*,
Journal of High Energy Physics **2018** (2018), ISSN: 1029-8479, eprint: 1803.00107,
URL: http://dx.doi.org/10.1007/JHEP10(2018)121 (cit. on pp. 57, 65).

[151] T. Plehn, A. Butter, B. Dillon and C. Krause, *Modern Machine Learning for LHC Physicists*,
2022, arXiv: 2211.01421 [hep-ph] (cit. on p. 57).

[152] P. T. Komiske, E. M. Metodiev and M. D. Schwartz,
Deep learning in color: towards automated quark/gluon jet discrimination,
Journal of High Energy Physics **2017** (2017), ISSN: 1029-8479, eprint: 1612.01551,
URL: http://dx.doi.org/10.1007/JHEP01(2017)110 (cit. on p. 57).

[153] G. Kasieczka, T. Plehn, M. Russell and T. Schell,
Deep-learning top taggers or the end of QCD?, Journal of High Energy Physics **2017** (2017),
eprint: 1701.08784, URL: https://doi.org/10.1007%2Fjhep05%282017%29006
(cit. on pp. 57, 63).

[154] H. Lv, D. Wang and L. Wu,
Deep learning jet images as a probe of light Higgsino dark matter at the LHC,
Phys. Rev. D **106** (5 2022) 055008, eprint: 2203.14569,
URL: https://link.aps.org/doi/10.1103/PhysRevD.106.055008 (cit. on p. 57).

[155] J. Guo, J. Li, T. Li, F. Xu and W. Zhang,
Deep learning for R−parity violating supersymmetry searches at the LHC,
Phys. Rev. D **98** (7 2018) 076017, eprint: 1805.10730,
URL: https://link.aps.org/doi/10.1103/PhysRevD.98.076017 (cit. on p. 57).

[156] J. S. H. Lee, I. Park, I. J. Watson and S. Yang,
Quark-Gluon Jet Discrimination Using Convolutional Neural Networks,
Journal of the Korean Physical Society **74** (2019) 219, ISSN: 1976-8524, eprint: 2012.02531,
URL: http://dx.doi.org/10.3938/jkps.74.219 (cit. on p. 57).

[157] J. Filipek, S.-C. Hsu, J. Kruper, K. Mohan and B. Nachman,
Identifying the Quantum Properties of Hadronic Resonances using Machine Learning, 2021,
arXiv: 2105.04582 [hep-ph] (cit. on p. 57).

[158] T. Han, I. M. Lewis, H. Liu, Z. Liu and X. Wang,
A Guide to Diagnosing Colored Resonances at Hadron Colliders, 2023,
arXiv: 2306.00079 [hep-ph] (cit. on p. 57).

[159] M. e. a. Aaboud,
Electron Identification with a Convolutional Neural Network in the ATLAS Experiment,
tech. rep., CERN, 2023, URL: https://cds.cern.ch/record/2850666 (cit. on p. 57).

[160] M. e. a. Aaboud, *Quark versus Gluon Jet Tagging Using Jet Images with the ATLAS Detector*,
tech. rep., CERN, 2017, URL: https://cds.cern.ch/record/2275641 (cit. on p. 57).

[161] A. e. a. Sirunyan, *Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques*, Journal of Instrumentation **15** (2020) P06005, ISSN: 1748-0221,
eprint: 2004.08262, URL: http://dx.doi.org/10.1088/1748-0221/15/06/P06005
(cit. on p. 57).

[162] J. Liu et al., *Deep-Learning-Based Kinematic Reconstruction for DUNE*, 2020,
arXiv: 2012.06181 [physics.ins-det], URL: https://arxiv.org/abs/2012.06181
(cit. on p. 57).

[163] P. T. Komiske, E. M. Metodiev and J. Thaler, *Energy flow networks: deep sets for particle jets*,
Journal of High Energy Physics **2019** (2019), ISSN: 1029-8479,
URL: http://dx.doi.org/10.1007/JHEP01(2019)121 (cit. on p. 58).

[164] H. Qu and L. Gouskos, *Jet tagging via particle clouds*, Physical Review D **101** (2020),
ISSN: 2470-0029, URL: http://dx.doi.org/10.1103/PhysRevD.101.056019
(cit. on p. 58).

[165] Y. Wang et al., *Dynamic Graph CNN for Learning on Point Clouds*, 2019,
arXiv: 1801.07829 [cs.CV] (cit. on p. 58).

[166] A. Tumasyan et al., *Search for Nonresonant Pair Production of Highly Energetic Higgs Bosons Decaying to Bottom Quarks*, Physical Review Letters **131** (2023), ISSN: 1079-7114,
URL: http://dx.doi.org/10.1103/PhysRevLett.131.041803 (cit. on p. 58).

[167] A. Tumasyan et al., *Search for Higgs Boson Decay to a Charm Quark-Antiquark Pair in Proton-Proton Collisions at $\sqrt{s}$=13TeV*, Physical Review Letters **131** (2023), ISSN: 1079-7114,
URL: http://dx.doi.org/10.1103/PhysRevLett.131.061801 (cit. on p. 58).

[168] S. Gong et al., *An efficient Lorentz equivariant graph neural network for jet tagging*,
Journal of High Energy Physics **2022** (2022), ISSN: 1029-8479,
URL: http://dx.doi.org/10.1007/JHEP07(2022)030 (cit. on p. 58).

[169] H. Qu, C. Li and S. Qian, *Particle Transformer for Jet Tagging*, 2024, arXiv: 2202.03772 [hep-ph] (cit. on pp. 59, 60).

[170] *Constituent-Based Quark Gluon Tagging using Transformers with the ATLAS detector*, (2023) (cit. on p. 60).

[171] V. Mikuni and F. Canelli, *Point cloud transformers applied to collider physics*, Machine Learning: Science and Technology **2** (2021) 035027, ISSN: 2632-2153, eprint: 2102.05073, URL: http://dx.doi.org/10.1088/2632-2153/ac07f6 (cit. on p. 60).

[172] F. A. Di Bello et al., *Reconstructing particles in jets using set transformer and hypergraph prediction networks*, The European Physical Journal C **83** (2023), ISSN: 1434-6052, eprint: 2212.01328, URL: http://dx.doi.org/10.1140/epjc/s10052-023-11677-7 (cit. on p. 60).

[173] L. Builtjes et al., *Attention to the strengths of physical interactions: Transformer and graph-based event classification for particle physics experiments*, 2024, arXiv: 2211.05143 [hep-ph] (cit. on p. 60).

[174] A. Hammad, S. Moretti and M. Nojiri, *Multi-scale cross-attention transformer encoder for event classification*, 2024, arXiv: 2401.00452 [hep-ph] (cit. on p. 60).

[175] T. Finke, M. Krämer, A. Mück and J. Tönshoff, *Learning the language of QCD jets with transformers*, Journal of High Energy Physics **2023** (2023), ISSN: 1029-8479, eprint: 2303.07364, URL: http://dx.doi.org/10.1007/JHEP06(2023)184 (cit. on p. 60).

[176] M. He and D. Wang, *Quark/Gluon Discrimination and Top Tagging with Dual Attention Transformer*, 2023, arXiv: 2307.04723 [hep-ph] (cit. on p. 60).

[177] A. Hammad and M. M. Nojiri, *Streamlined jet tagging network assisted by jet prong structure*, 2024, arXiv: 2404.14677 [hep-ph] (cit. on p. 60).

[178] A. Hammad, P. Ko, C.-T. Lu and M. Park, *Exploring Exotic Decays of the Higgs Boson to Multi-Photons at the LHC via Multimodal Learning Approaches*, 2024, arXiv: 2405.18834 [hep-ph] (cit. on p. 60).

[179] V. K. et al., *Search for pair-produced resonances decaying to jet pairs in proton–proton collisions at s = 8 TeV*, Physics Letters B **747** (2015) 98, ISSN: 0370-2693, eprint: 1412.7706, URL: https://www.sciencedirect.com/science/article/pii/S0370269315002889 (cit. on p. 60).

[180] G. e. a. Aad, *A search for top squarks with R−parity-violating decays to all-hadronic final states with the ATLAS detector in* $\sqrt{s} = 8$ *TeV proton-proton collisions*, Journal of High Energy Physics **2016** (2016), ISSN: 1029-8479, eprint: 1601.07453, URL: http://dx.doi.org/10.1007/JHEP06(2016)067 (cit. on p. 60).

[181] M. e. a. Aaboud, *A search for pair-produced resonances in four-jet final states at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, Eur. Phys. J. C **78** (2018) 250, arXiv: 1710.07171 [hep-ex] (cit. on p. 60).

[182] A. M. e. a. Sirunyan, *Search for pair-produced resonances decaying to quark pairs in proton-proton collisions at* $\sqrt{s} = 13$ TeV, Phys. Rev. D **98** (11 2018) 112014, eprint: 1808.03124, URL: https://link.aps.org/doi/10.1103/PhysRevD.98.112014 (cit. on p. 60).

[183] A. M. e. a. Sirunyan, *Search for top squarks in final states with two top quarks and several light-flavor jets in proton-proton collisions at* $\sqrt{s} = 13$ TeV, Phys. Rev. D **104** (3 2021) 032006, eprint: 2102.06976, URL: https://link.aps.org/doi/10.1103/PhysRevD.104.032006 (cit. on pp. 61, 62, 74, 76, 78, 79).

[184] R. D. Ball et al., *Parton distributions from high-precision collider data*, The European Physical Journal C **77** (2017), eprint: 1706.00428, URL: https://doi.org/10.1140%2Fepjc%2Fs10052-017-5199-5 (cit. on p. 62).

[185] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, 2022, arXiv: 2203.11601 [hep-ph] (cit. on p. 62).

[186] A. M. e. a. Sirunyan, *Extraction and validation of a new set of CMS Pythia8 tunes from underlying-event measurements*, The European Physical Journal C **80** (2020), ISSN: 1434-6052, eprint: 1903.12179, URL: http://dx.doi.org/10.1140/epjc/s10052-019-7499-4 (cit. on p. 62).

[187] M. L. Mangano, M. Moretti, F. Piccinini and M. Treccani, *Matching matrix elements and shower evolution for top-pair production in hadronic collisions*, Journal of High Energy Physics **2007** (2007) 013, ISSN: 1029-8479, eprint: hep-ph/0611129, URL: http://dx.doi.org/10.1088/1126-6708/2007/01/013 (cit. on p. 62).

[188] J. de Favereau et al., *DELPHES 3: a modular framework for fast simulation of a generic collider experiment*, Journal of High Energy Physics **2014** (2014), ISSN: 1029-8479, eprint: 1307.6346, URL: http://dx.doi.org/10.1007/JHEP02(2014)057 (cit. on p. 62).

[189] A. Mertens, *New features in Delphes 3*, Journal of Physics: Conference Series **608** (2015) 012045, URL: https://dx.doi.org/10.1088/1742-6596/608/1/012045 (cit. on p. 62).

[190] M. Cacciari, G. P. Salam and G. Soyez, *FastJet user manual*, The European Physical Journal C **72** (2012), eprint: 1111.6097, URL: https://doi.org/10.1140%2Fepjc%2Fs10052-012-1896-2 (cit. on p. 62).

[191] A. Chakraborty et al., *Revisiting jet clustering algorithms for new Higgs Boson searches in hadronic final states*, The European Physical Journal C **82** (2022), eprint: 2008.02499, URL: https://doi.org/10.1140/epjc/s10052-022-10314-z (cit. on p. 63).

[192] D. Krohn, J. Thaler and L.-T. Wang, *Jets with variable R*, Journal of High Energy Physics **2009** (2009) 059, ISSN: 1029-8479, eprint: 0903.0392, URL: http://dx.doi.org/10.1088/1126-6708/2009/06/059 (cit. on p. 63).

[193] A. Paszke et al., *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, 2019, arXiv: 1912.01703 [cs.LG] (cit. on p. 65).

[194] M. Usman et al., *Particle Multi-Axis Transformer for Jet Tagging*, 2024, arXiv: 2406.06638 [hep-ph] (cit. on p. 66).

[195] A. Hook, E. Izaguirre, M. Lisanti and J. G. Wacker, *High multiplicity searches at the LHC using jet masses*, Physical Review D **85** (2012), ISSN: 1550-2368, eprint: 1202.0558, URL: http://dx.doi.org/10.1103/PhysRevD.85.055029 (cit. on p. 74).

[196] C. Bernaciak, M. S. A. Buschmann, A. Butter and T. Plehn, *Fox-Wolfram moments in Higgs physics*, Physical Review D **87** (2013), ISSN: 1550-2368, eprint: 1212.4436, URL: http://dx.doi.org/10.1103/PhysRevD.87.073014 (cit. on p. 74).

[197] C. Borschensky et al., *Squark and gluino production cross sections in pp collisions at $\sqrt{s}$ = 13, 14, 33 and 100 TeV*, The European Physical Journal C **74** (2014), ISSN: 1434-6052, eprint: 1407.5066, URL: http://dx.doi.org/10.1140/epjc/s10052-014-3174-y (cit. on pp. 76, 78).

[198] N. Kumar and S. P. Martin, *Vectorlike leptons at the Large Hadron Collider*, Physical Review D **92** (2015), ISSN: 1550-2368, eprint: 1510.03456, URL: http://dx.doi.org/10.1103/PhysRevD.92.115018 (cit. on p. 76).

[199] O. Aberle et al., *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*, CERN Yellow Reports: Monographs, Geneva: CERN, 2020, URL: https://cds.cern.ch/record/2749422 (cit. on p. 81).

[200] R. B. et al., *root-project/root: v6.18/02*, version v6-18-02, 2019, URL: https://doi.org/10.5281/zenodo.3895860 (cit. on p. 84).

[201] C. Borschensky et al.,
Squark and gluino production cross sections in pp collisions at $\sqrt{s}$ = 13, 14, 33 and 100 TeV,
Eur. Phys. J. C **74** (2014) 3174, arXiv: 1407.5066 [hep-ph] (cit. on p. 84).

[202] A. L. Read, *Presentation of search results: the CLs technique*,
Journal of Physics G: Nuclear and Particle Physics **28** (2002) 2693,
URL: https://dx.doi.org/10.1088/0954-3899/28/10/313 (cit. on p. 84).

[203] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*,
Journal of Machine Learning Research **12** (2011) 2825 (cit. on p. 86).

List of Figures

List of Tables

Acronyms

ANODE Anomaly Detection with Density Estimation.

BSM Physics beyond the Standard Model.

CART Classification and Regression Trees.

CATHODE Classifying Anomalies THrough Outer Density Estimation.

CKM Cabibbo-Kobayashi-Maskawa.

CNN Convolutional Neural Networks.

CoAtNet Combination of Depthwise **Co**nvolution and self-**At**tention.

CWoLa Classification Without Labels.

ECAL Electromagnetic Calorimeter.

GBDT Gradient Boosted Decision Tree.

HCAL Hadron Calorimeter.

KDE Kernel Density Estimation.

LHC Large Hadron Collider.

LSP Lightest Supersymmetric Particle.

MADE Masked Autoencoder for Distribution Estimation.

MAF Masked Autoregressive Flow.

MaxViT Multi-AXis VIsion Transformer.

MLP Multilayer Perceptron.

MSSM Minimal Supersymmetric Standard Model.

NLSP next-to-lightest supersymmetric particle.

PDF Probability Density Function.

QCD Quantum Chromodynamics.

ROC Receiver Operating Characteristic.

RPV R-parity violating.

SB Side Band.

SM Standard Model of particle physics.

SR Signal Region.

SUSY Supersymmetry.

ViT Vision Transformer.