

Machine Learning-Based Tagging of Vector Boson Polarization in the Leptonic Channel

Master-Arbeit
zur Erlangung des Hochschulgrades
Master of Science
im Master-Studiengang Physik

vorgelegt von

Lena Alshut
geboren am 11.10.2000 in Dresden

Institut für Kern- und Teilchenphysik
Fakultät Physik
Bereich Mathematik und Naturwissenschaften
Technische Universität Dresden
2026

Eingereicht am 2. Mai 2026

1. Gutachter: Prof. Dr. Frank Siegert
2. Gutachter: Prof. Dr. Michael Kobel

Abstract

The polarization of massive vector bosons is highly sensitive to the electroweak symmetry breaking and the Standard Model gauge symmetry structure. The comparison of polarization measurements with theoretical predictions provides a valuable insight into the Standard Model as well as Beyond the Standard Model physics.

Due to their short lifetimes, massive vector bosons cannot be observed directly but appear as intermediate particles. Their polarization is obtained indirectly from measurable, unpolarized observables using theoretically constructed polarization templates.

In this thesis, a machine learning-based approach for event-by-event vector boson polarization tagging is studied. Polarized samples of vector boson production processes are generated using the Monte Carlo event generator Sherpa where polarization fractions, defined as the ratio of polarized to unpolarized cross sections, are determined for each event. Neural networks are trained on the simulated samples to predict the event-wise polarization fractions of intermediate massive vector bosons based on experimentally accessible final-state kinematics. With this approach, the NN output can be directly compared to the true polarization distribution and the NN performance can be evaluated quantitatively.

The performance of two neural network architectures, a multilayer perceptron and a more advanced transformer model, is systematically compared. The study is conducted for $Z+jets$ and $W+jet$ production processes, as well as for the electroweak production of a W^+Z boson pair in association with two jets at fixed leading order and matched to the parton shower. Additionally, the effect of higher-order corrections in the strong coupling constant and the influence of different phase spaces are examined, with $Z+jets$ serving as the test bed.

Zusammenfassung

Die Polarisierung massiver Vektorbosonen ist hoch sensitiv gegenüber der elektroschwachen Symmetriebrechung und der Struktur der Eichsymmetrie des Standardmodells. Der Vergleich von Polarisationsmessungen mit theoretischen Vorhersagen liefert wertvolle Einblicke sowohl in das Standardmodell als auch in die Physik jenseits des Standardmodells.

Aufgrund ihrer kurzen Lebensdauer können massive Vektorbosonen nicht direkt beobachtet werden, sondern treten als intermediäre Teilchen auf. Folglich wird ihre Polarisierung indirekt aus messbaren, unpolarisierten Observablen unter Verwendung theoretisch konstruierter Polarisationsvorlagen ermittelt.

In dieser Arbeit wird ein auf maschinellem Lernen basierender Ansatz zur Polarisationsbestimmung von massiven Vektorbosonen untersucht. Mit dem Monte-Carlo-Ereignisgenerator Sherpa werden polarisierte Stichproben von Vektorbosonen-Produktionsprozessen erzeugt, wobei für jedes Ereignis die Polarisationsanteile, definiert als das Verhältnis von polarisierten zu unpolarisierten Wirkungsquerschnitten, bestimmt werden. Die neuronalen Netze werden anhand der simulierten Datensätzen trainiert, um die Polarisationsanteile massiver Vektorbosonen im Zwischenzustand auf der Grundlage experimentell zugänglicher Endzustands-Kinematik vorherzusagen. Mit diesem Ansatz ermöglicht einen direkten Vergleich der Vorhersage des neuronalen Netzes mit der tatsächlichen Polarisationsverteilung und eine quantitative Bewertung des neuronalen Netzes.

Die Leistungsfähigkeit zweier neuronaler Netzwerkarchitekturen, eines Multi-Layer-Perzeptrons und eines komplexeren Transformer-Modells, wird systematisch verglichen. Die Studie wird für Z +jets- und W +jet-Produktionsprozesse sowie für die elektroschwache Produktion eines W^+Z Bosonpaares in Verbindung mit zwei Jets in führender Ordnung durchgeführt und mit der Parton Shower gematcht. Zusätzlich werden der Effekt von Korrekturen höherer Ordnung in der starken Kopplungskonstante und der Einfluss verschiedener Phasenräume untersucht, wobei Z +jets als Testfall dient.

Contents

1	Introduction	1
2	Theoretical Background	3
2.1	Standard Model of particle physics	3
2.1.1	Electroweak Theory	4
2.1.2	Electroweak Symmetry Breaking	6
2.2	Polarization	8
2.2.1	Polarization of massive vector bosons	8
2.2.2	Definition of polarization fractions	9
3	Machine Learning methods for polarization extraction	13
3.1	Multilayer Perceptron	13
3.2	Transformer	15
3.3	Hyperparameter optimization with OPTIMA	18
3.4	Neural networks in vector boson polarization tagging	20
4	Simulation	23
4.1	Simulation of proton-proton collisions	23
4.2	Simulation of polarized cross sections in Sherpa	26
4.3	Simulated samples	26
4.3.1	Investigated processes	27
4.3.2	Settings of the event simulation	29
5	Results	31
5.1	Evaluation of the NN performance	31
5.2	$Z+jets$	36
5.2.1	Leading order results	36
5.2.2	Parton shower effects	41
5.2.3	Higher order effects	45
5.2.4	Phase space effects	64
5.3	$W+jets$	74
5.4	$WZjj$ -EW	84

6	Summary and Outlook	91
A	Mathematical Definitions	93
B	Samples and Sherpa runcards	95
	List of Figures	99
	List of Tables	101
	Bibliography	103

1 Introduction

Particle physics aims to understand the basic building blocks of matter and the forces through which they interact. Here, elementary particles as the smallest known constituents of the universe are studied. The most successful theoretical framework describing these particles, their properties and interactions is the Standard Model of particle physics. Testing its predictions requires extremely high energies, which are achieved in modern particle accelerators such as the Large Hadron Collider. Predicted more than five decades ago, experiments at the Large Hadron Collider observed a scalar particle consistent with the long-predicted Higgs boson relatively recently in 2012 [1][2]. This discovery provided strong evidence for the existence of the Higgs field and the mechanism of electroweak symmetry breaking. Up to now, the Standard Model predictions showed to be very accurate when compared to most experimental measurements. Nevertheless, the theory does not account for some observations such as a non-zero neutrino mass, gravitation as the fourth fundamental force or the existence of dark matter. Because of this, numerous extensions beyond the Standard Model have been proposed.

The longitudinal polarization of massive vector bosons arises directly from the Higgs mechanism. As a result, processes involving polarized massive vector boson production offer a powerful way to probe electroweak symmetry breaking. In the absence of the Higgs boson, the scattering cross section of longitudinally polarized massive vector bosons would increase with the square of the center-of-mass energy and violate unitarity. This highlights the strong connection between massive vector boson polarization and the underlying symmetry breaking mechanism. In addition, many massive vector boson production processes involve triple and quartic gauge couplings, making them sensitive to the fundamental gauge structure of the Standard Model. Massive vector boson polarization measurements provide an important tool for investigating new physics scenarios. In particular, various theories beyond the Standard Model predict modifications in massive vector boson scattering processes.

Due to their extremely short lifetime, massive vector bosons only appear as intermediate particles and cannot be measured directly. Therefore, their polarization is not directly accessible but can only be reconstructed from their stable decay products. In order to develop analyses for measurements of massive vector boson polarization, it is necessary to rely on polarized cross section templates generated by Monte Carlo event generators. These templates make it possible to extract polarization information from experimentally observable, unpolarized data. The simulation of polarized cross sections of massive vector bosons is possible with event

generators. For massive vector boson production processes, the polarization fraction of the intermediate boson as the ratio of polarized and unpolarized cross sections can be computed at phase-space-point level. This polarization fraction of each event strongly depends on the kinematics and parton types of the initial and final states. Experimentally, only the four-momenta of the final state particles can be reconstructed. As proposed by [3], artificial neural networks can be used to approximate event-wise polarization information from the experimentally available information. The network learns the relation between the final state kinematics and the polarization fraction of the intermediate massive vector boson at phase-space-point level. After a training on simulation data, the model can then predict the polarization fraction of the massive vector boson in measured events allowing a new approach for polarization tagging.

In this thesis, the approach of event-wise polarization fraction extraction by artificial neural networks is further examined. It presents detailed investigations of the predictive network performance and its behaviour for an increasingly complex event simulation while providing a characterization and explanation of the networks' shortcomings.

This thesis is structured as follows. In chapter 2, the Standard Model of particle physics is explained focussing on the electroweak theory and the electroweak symmetry breaking. Additionally, the polarization of massive vector bosons and their polarization fraction are introduced. Chapter 3 explains the concept of neural networks with a focus on the two network architectures used in this work: the Multilayer Perceptron and the Transformer architecture. A short description of the OPTIMA framework is given, which is used for the optimization of the neural network hyperparameters. In chapter 4, the general event simulation of proton-proton collisions as well as the simulation of polarized cross section with the Monte Carlo event generator Sherpa is described. The single boson and diboson production processes investigated in this thesis are motivated and the settings of the simulated samples are stated. Chapter 5 presents the results of the polarization extraction using neural networks. The first part of the chapter describes the approach on the network evaluation which is applied subsequently for all results. The $Z+jets$ production process serves as a test bed to investigate general predictive power of the neural networks. It is investigated in great detail for leading order in the strong coupling constant followed by parton shower, higher order and phase space effects. Furthermore, results of the $W+jets$ and $WZjj$ -EW production process are discussed. Finally, chapter 6 provides a summary of this thesis together with an outlook to possible future investigations and applications.

2 Theoretical Background

This chapter gives a general overview of the Standard Model of particle physics and an important property of the described particles, the polarization, following [4] and [5].

2.1 Standard Model of particle physics

The Standard Model (SM) of particle physics reflects today's understanding of the elementary particles and their fundamental interactions.

The SM is built on the principle that all matter is composed of elementary particles. According to their spin S , these elementary particles are characterized as fermions, which constitute matter, and bosons, which mediate interactions. Fermions ($S = 1/2$) include quarks and leptons arranged in three generations of increasing mass. Quarks combine to form hadrons such as protons and neutrons. Leptons include the electron and its heavier counterparts, along with their associated neutrinos. Each elementary particle has its corresponding antiparticle with same mass but opposite charge. The fundamental interactions among the fermions are described by the exchange of bosons with integer spin $S = 1$. The electromagnetic interaction is mediated by photons γ , the weak interaction by the $W^\pm$ and Z bosons and the strong interaction by gluons g . Gravity as the fourth known fundamental interaction is not described by the SM. The only scalar boson with $S = 0$ in the SM is the Higgs boson. It provides the weak bosons and the fermions with mass through the electroweak symmetry breaking described in subsection 2.1.2. An overview of all elementary particles and their fundamental properties is given in Table 2.1 and Table 2.2.

The SM is a relativistic Quantum Field Theory (QFT). Mathematically, it is formulated as a gauge theory with a local $U(1)_Y \otimes SU(2)_L \otimes SU(3)_C$ gauge symmetry. This means that transformations of this symmetry group leave the SM-Lagrangian invariant. The subgroups correspond to the three fundamental interactions described by the SM. Quantum Chromodynamics (QCD), based on the $SU(3)_C$ gauge symmetry, describes the dynamics of quarks and gluons, while the gauge group $U(1)_Y \otimes SU(2)_L$ describes both electromagnetic and weak interactions in the unified electroweak theory.

Table 2.1: Properties of the bosons ($S = 0, 1$) in the SM with mass m , width Γ , electric charge Q , weak isospin I_3^W and spin S . Data taken from [6].

Bosons		m [GeV]	Γ [GeV]	Q [e]	I_3^W	S
Gauge bosons	γ	0	0	0	0	1
	g	0	0	0	0	1
	Z	91.188	2.49	0	0	1
	$W^\pm$	80.369	2.14	± 1	± 1	1
Higgs boson	H	125.09	0.004	0	$1/2$	0

Table 2.2: Properties of the fermions ($S = 1/2$) in the SM with electric charge Q , weak isospin I_3^W and spin S .

Fermions	Generation			Q	I_3^W
	I	II	I		
Quarks	u	c	t	$2/3$	$1/2$
	d	s	b	$-1/3$	$-1/2$
Leptons	ν_e	ν_μ	ν_τ	0	$1/2$
	e	μ	τ	-1	$-1/2$

2.1.1 Electroweak Theory

The weak interaction is parity-violating as shown by C. Wu in her famous experiment [7]. In the electroweak theory, left- and right-handed components of the Dirac fermion field ψ are introduced. A Dirac fermion field ψ is composed of a left-handed field component ψ_L and a right-handed field component ψ_R , such that $\psi = \psi_L + \psi_R$. The components can be extracted by applying the projection operator $P_{L/R}$ with

$$\psi_{L/R} = P_{L/R} \psi = \frac{1}{2} \left(1 \mp \frac{\gamma^5}{2} \right) \psi. \quad (2.1)$$

Here, the γ^5 matrix is defined as in chapter A.

The weak interaction only couples to left-handed chiral states which form the weak isospin doublets

$$\begin{pmatrix} e \\ \nu_e \end{pmatrix}, \begin{pmatrix} \mu \\ \nu_\mu \end{pmatrix}, \begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix}, \begin{pmatrix} u \\ d \end{pmatrix}, \begin{pmatrix} c \\ s \end{pmatrix} \text{ and } \begin{pmatrix} t \\ b \end{pmatrix}. \quad (2.2)$$

The right-handed states behave like singlets under the weak interaction. This can be described as a rotation in the $SU(2)_L$ weak isospin space

$$\psi_L \rightarrow \exp \left(\frac{-iq\sigma\theta(x)}{\hbar c} \right) \psi_L, \quad (2.3)$$

with the coupling constant q , the local phase transformation $\theta(x)$ and the Pauli-matrices σ as defined in chapter A. Three new massless fields $W_\mu = (W_\mu^1, W_\mu^2, W_\mu^3)$ are introduced to

conserve invariance under this local gauge transformation. The transformation of these new fields compensates the local phase transformation in the Lagrangian of the weakly interacting particles.

The formulation of the electroweak theory as the unification of electromagnetic and weak interaction was proposed in [8], [9] and [9] because of their comparable coupling strengths at high energies. The electroweak interaction is described by a $U(1)_Y \otimes SU(2)_L$ gauge symmetry group. Here, Y is the hypercharge of the $U(1)_Y$ symmetry group. With $Y = 2(Q - I_3^W)$, it is derived from the electric charge Q (the conserved quantity of the electromagnetic interaction) and the third component of the weak isospin I_3^W (the conserved quantity of the weak interaction). As a mediating particle of the $U(1)_Y$ group, a neutral and massless gauge field B_μ is introduced.

The electroweak Lagrangian $\mathcal{L}_{\text{EW}}$ can be written as

$$\mathcal{L}_{\text{EW}} = \sum_{f_L, f_R} i\bar{\psi}_f(x)\gamma^\mu D_\mu \psi_f(x) - \frac{1}{4}W_{\mu\nu}^a W_a^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}. \quad (2.4)$$

The covariant derivative D_μ is defined as

$$D_\mu \psi_f(x) = [\partial_\mu + ig_W T^a W_\mu^a + ig' Y B_\mu] \psi_f(x), \quad (2.5)$$

with the coupling constants g' and g_W of the $U(1)_Y$ and $SU(2)_L$ gauge fields. The generators T^a of the $SU(2)_L$ group can be expressed as $T^a = \sigma^a/2$ with σ_a being the Pauli matrices defined in Appendix A.

The last two terms in Equation 2.4, which include the field strength tensors

$$W_{\mu\nu}^a = \partial_\mu W_\nu^a - \partial_\nu W_\mu^a + g_W \epsilon_{abc} W_{b,\mu} W_{c,\nu} \quad (2.6)$$

and

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu, \quad (2.7)$$

describe the gauge fields dynamics and produce terms of order three and four in the gauge fields. These terms appear as a consequence of the non-Abelian nature of the $SU(2)_L$ and can be interpreted as self-interactions among the gauge bosons (triple and quartic gauge couplings). The electroweak Lagrangian $\mathcal{L}_{\text{EW}}$ is invariant under the simultaneous transformation of the fermion field $\psi(x)$, the gauge fields $W_\mu^a(x)$ of $SU(2)_L$ and $B_\mu(x)$ of $U(1)_Y$:

$$\psi(x) \rightarrow \exp(\alpha(x)Y + i\beta(x)T^a)\psi \quad (2.8)$$

$$W_\mu^a(x) \rightarrow W_\mu^a(x) - \frac{1}{g_W} \partial_\mu \beta^a(x) - \epsilon_{abc} \beta^b(x) \beta^c(x) \quad (2.9)$$

$$B_\mu(x) \rightarrow B_\mu(x) - \frac{1}{g} \partial_\mu \alpha(x) \quad (2.10)$$

2.1.2 Electroweak Symmetry Breaking

The Lagrangian $\mathcal{L}_{\text{EW}}$ of the electroweak interaction in Equation 2.4 cannot describe what is measured in physical experiments. Gluons as the interaction particles of the strong interaction are observed to be massless and the symmetry of the strong gauge symmetry group $SU(3)_C$ is not broken. In contrast, fermions as well as $W^\pm$ and Z bosons are massive in nature and the symmetry of the $U(1)_Y \otimes SU(2)_L$ group is spontaneously broken. The introduction of mass terms quadratic in the gauge fields, like $\mathcal{L}_f = m\bar{\psi}\psi$ for the fermion mass and $\mathcal{L}_{W/Z} = -M_V^2 V_\mu V^\mu/2$ for the boson mass, would destroy the local gauge invariance of $\mathcal{L}_{\text{EW}}$. The electroweak symmetry breaking (EWSB) is described in the SM by a mechanism proposed by Brout and Englert [10] as well as Higgs [11]. Here, a complex scalar $SU(2)_L$ doublet $\Phi = (\Phi^+(x), \Phi^0(x))^T$ with the gauge transformation

$$\Phi(x) \rightarrow \exp(i\alpha(x)Y + i\beta_a(x)T_a)\Phi(x), \quad (2.11)$$

and $Y = \frac{1}{2}$ is proposed. A scalar potential with continuous symmetry (even order in Φ) and a ground state with non-zero vacuum expectation value is chosen, such that

$$V(\Phi) = -\mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 \quad (2.12)$$

This potential leads to a new term in the Lagrangian

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \Phi)(D_\mu \Phi)^\dagger - V(\Phi). \quad (2.13)$$

Here, λ and μ are scalar free parameters with $\lambda, \mu > 0$. The vacuum ground state Φ_0 is the state with the lowest energy but it is also degenerate due to the potential's rotational symmetry. Choosing $\Phi^+ = 0$, the potential minimum is located at $\Phi^0 = \sqrt{\frac{\mu^2}{2\lambda}}$. With a vacuum expectation value $v = \sqrt{-\frac{\mu^2}{\lambda}}$, the ground state Φ_0 of the vacuum is chosen to be

$$\Phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \quad (2.14)$$

With this, the $U(1)_Y \otimes SU(2)_L$ symmetry is broken. The ground state is invariant under the transformation

$$\Phi_0 \rightarrow \exp(i\theta(x)Q)\Phi_0 \quad \text{with} \quad Q = T_3 + Y = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}. \quad (2.15)$$

The symmetry described by the $U(1)_Q$ subgroup of the electromagnetic interaction is not broken by this mechanism. Consequently, photons as the gauge bosons of the QED retain massless as required and all interactions conserve the electric charge Q .

Expanding the complex doublet around the ground state leads to

$$\Phi \approx \begin{pmatrix} \Phi_1(x) + i\Phi_2(x) \\ \frac{1}{\sqrt{2}}(v + H(x)) + i\Phi_3 \end{pmatrix}, \quad (2.16)$$

with the real scalar fields Φ_1, Φ_2, Φ_3 and H . The Nambu-Goldstone theorem [12][13] predicts the existence of a Nambu-Goldstone boson for each generator of a spontaneously broken continuous symmetry. Φ_1, Φ_2 and Φ_3 are the massless Goldstone fields that can be eliminated by a gauge transformation so that

$$\Phi \approx \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}. \quad (2.17)$$

The term $(D_\mu \Phi)(D_\mu \Phi)^\dagger$ in Equation 2.18 then has the form

$$\mathcal{L}_{\text{Higgs I}} = \frac{1}{2}(\partial_\mu H)^2 + \frac{1}{8}g_W^2(v + H)^2|W_\mu^1 - W_\mu^2|^2 + \frac{1}{8}(v + H)^2|g_W W_\mu^3 - g' B_\mu|^2. \quad (2.18)$$

This includes the kinetic term for the Higgs boson as well as mass and interaction terms for the gauge bosons $W_\mu^1, W_\mu^2, W_\mu^3$ and B_μ . The known mass eigenstates of the electroweak interaction are then obtained by the diagonalization of the mass matrix. The fields B_μ and W_μ^3 mix to create the photon A_μ and the Z boson whereas the field components W_μ^1 and W_μ^2 mix to create the two charged $W^\pm$ bosons so that

$$A_\mu = \frac{g' B_\mu + g_W W_\mu^3}{\sqrt{g'^2 + g_W^2}}, \quad (2.19)$$

$$Z_\mu = \frac{-g' B_\mu + g_W W_\mu^3}{\sqrt{g'^2 + g_W^2}} \quad (2.20)$$

and

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2), \quad (2.21)$$

with the coupling constants g' and g_W of the $U(1)_Y$ and $SU(2)_L$ gauge fields.

Hence, Equation 2.18 can be expressed as

$$\mathcal{L}_{\text{Higgs I}} = \frac{1}{2}(\partial_\mu H)^2 + (v + H)^2 \left(\frac{g_W^2}{4} W_\mu^+ W^{-\mu} + \frac{g'^2 + g_W^2}{8} Z_\mu Z^\mu \right), \quad (2.22)$$

which now includes the mass terms of the gauge bosons as $m_W = g_W v/2$, $m_Z = v\sqrt{g'^2 + g_W^2}/2$ and $m_A = 0$. The other terms describe the coupling of one and two Higgs bosons to a gauge

boson pair. The term in Equation 2.13 including the Higgs potential itself can be written as

$$\mathcal{L}_{\text{Higgs II}} = -\lambda v^2 H^2 - \lambda v H^3 - \frac{1}{4}\lambda H^4. \quad (2.23)$$

In 2012, the Higgs boson was finally discovered at the LHC with the ATLAS detector [1] and the CMS detector [2]. The Higgs mass was measured to be $m_H = 125$ GeV and the vacuum expectation value was determined to be $v = 246$ GeV. From this, the free parameter λ of the potential and the Higgs self-couplings can be determined.

2.2 Polarization

2.2.1 Polarization of massive vector bosons

The polarization of a particle is defined as the alignment of its spin along a given axis. Commonly, the helicity h is used which is defined as the projection of a particle's spin $\vec{S}$ on its momentum $\vec{p}$ with

$$h = \vec{S} \cdot \frac{\vec{p}}{|\vec{p}|}. \quad (2.24)$$

For massive particles with $S = 1$, the helicity has three eigenvalues. The eigenvalues $h = 1$ and $h = -1$ correspond to the parallel and anti-parallel alignment of spin and momentum. They are called the *transverse polarization states* of the particle (right-handed for $h = 1$ and left-handed for $h = -1$). The eigenvalue $h = 0$ describes the orthogonal alignment of spin and momentum. It is referred to as the *longitudinal polarization state* and does not exist for massless particles like photons.

The helicity is only Lorentz invariant for massless particles which travel with speed of light. For massive particles, the helicity depends on the measurement's frame of reference. For these particles, it is possible to choose a second reference frame which moves faster with respect to the particle's current reference frame. In this second frame, the particle moves in opposite direction but the spin is unchanged and the particle's helicity value flips. Lorentz boosts can therefore change a particle's helicity. Reference frames commonly employed for single and multi-boson processes are the laboratory frame, the center-of-mass system of all hard-decaying particles or the center-of-mass system of the interacting partons. Since the helicity commutes with the Hamiltonian operator, it is a conserved quantity and a good quantum number to classify particles.

The $W^\pm$ and Z boson as interaction particles of the weak interaction acquire their mass by the EWSB mechanism in the SM as described in subsection 2.1.2. The existence of longitudinal polarization states of the massive vector bosons (MVB) is can therefore be explained by the EWSB mechanism and it is sensitive to the SM gauge symmetry structure.

Information on the polarization of the intermediate gauge boson can be derived from the

detected final state (FS) particles. The parity violating weak interaction leads to several possibly sensitive observables, e.g. the angular distribution of the FS leptons. The polarization of MVB affects the decay as well as the production process of the bosons. The $W^\pm$ couples exclusively to left-handed particles and right-handed antiparticles. Thus, for example, an intermediate W^+ boson decays into a right-handed positron whose small mass compared to the boson mass can be neglected. The positron is in an almost purely right-handed helicity state and its momentum and spin direction are aligned. Angular momentum conservation causes the alignment of the positron spin with the boson spin so that a right-handed polarized boson emits the positron in its direction of flight. The same considerations can be applied to the production of the $W^\pm$ boson when it only couples to left-handed quarks and right-handed anti-quarks. The spin of the boson is perpendicular to the quark momentum. Orthogonally emitted bosons will therefore be left-/right-handed while forward or backward emission leads to a longitudinally polarized MVB.

2.2.2 Definition of polarization fractions

In this section, polarized matrix elements and polarization fractions of intermediate massive vector bosons are introduced closely following [14] and [3]. First, this is explained for the case of single boson production.

The matrix element for the MVB production and its decay into a fermion pair is given by

$$\mathcal{M} = \mathcal{M}_\mu^{\mathcal{P}} \frac{i}{k^2 - m_V^2 + i\Gamma_V m_V} \left(-g^{\mu\nu} + \frac{k^\mu k^\nu}{m_V^2} \right) \mathcal{M}_\nu^{\mathcal{D}}. \quad (2.25)$$

Here, $\mathcal{M}_\mu^{\mathcal{P}}$ and $\mathcal{M}_\mu^{\mathcal{D}}$ describe the matrix elements of production and decay vertex, m_V and Γ_V are the mass and width of the MVB and k^μ describes the MVB momentum.

The tensor part of the propagator in Equation 2.25 can be expressed by

$$-g^{\mu\nu} + \frac{k^\mu k^\nu}{M^2} = \sum_{\lambda=1}^4 \epsilon_\lambda^\mu(k) \epsilon_\lambda^{\nu*}(k), \quad (2.26)$$

called the completeness relation with polarization vectors $\epsilon_\lambda^\mu(k)$ of the four polarizations λ . Three of the polarizations are physical, with two transverse states ($\lambda = +, -$) and one longitudinal state ($\lambda = L$). The fourth polarization state is unphysical and does not contribute if the MVB is on-shell or if it decays into massless leptons [15].

With this, the (unpolarized) matrix element in Equation 2.25 can now be expressed as a sum over the polarization states λ of the intermediate MVB, such that

$$\mathcal{M} = \sum_{\lambda=1}^4 \mathcal{M}_\mu^{\mathcal{P}} \epsilon_\lambda^\mu(k) \frac{i}{k^2 - M_V^2 + i\Gamma_V M_V} \epsilon_\lambda^{\nu*}(k) \mathcal{M}_\nu^{\mathcal{D}} = \sum_{\lambda} \mathcal{M}_\lambda. \quad (2.27)$$

Here, $\mathcal{M}_\lambda := \mathcal{M}_\lambda^{\mathcal{P}} \mathcal{M}_\lambda^{\mathcal{D}}$ defines the matrix element of the MVB as the product of production and decay amplitude with polarization λ . This factorization is only possible for on-shell MVB.

Approximations like the narrow-width approximation (NWA) are necessary so only the on-shell contributions of the intermediate MVBs remain and only physical polarization states contribute. In the NWA, the squared denominator of the MVB propagator is replaced by the delta function $\frac{\pi}{m_V \Gamma_V} \delta(k^2 - m_V^2)$. Processes with two intermediate MVBs include double-resonant but also non- or single-resonant diagrams. Referring to the number of resonant MVBs, double-resonant diagrams of diboson production processes contain only FS lepton lines that are connected to one of the MVBs. The polarization of MVBs can only be defined in the case of fully-resonant diagrams. Yet, not fully-resonant diagrams have to be considered to guarantee electroweak gauge invariance. With the NWA, the factorization of the amplitude into production and decay amplitude of on-shell MVBs holds for diboson processes and only diagrams of well-defined polarization contribute.

The absolute square of the unpolarized matrix element $\mathcal{M}$ in Equation 2.27 contains the incoherent sum over the polarized absolute squared matrix elements and a sum over all interference terms

$$|\mathcal{M}|^2 = \sum_{\lambda} |\mathcal{M}_\lambda|^2 + \sum_{\lambda \neq \lambda'} \mathcal{M}_\lambda^* \mathcal{M}_{\lambda'}. \quad (2.28)$$

The interference $\mathcal{M}_\lambda^* \mathcal{M}_{\lambda'}$ between different polarization states ($\lambda \neq \lambda'$) of the MVB vanishes when integrating over all events only if no lepton selection criteria are applied.

A transversely polarized amplitude can be defined as the coherent sum of left- and right-handed contributions

$$|\mathcal{M}_T|^2 = |\mathcal{M}_+|^2 + |\mathcal{M}_-|^2 + 2 \operatorname{Re}(\mathcal{M}_+^* \mathcal{M}_-). \quad (2.29)$$

The unpolarized amplitude therefore consists of the amplitudes of the longitudinal MVB polarization, the transverse polarization and interference terms as

$$|\mathcal{M}|^2 = |\mathcal{M}_L|^2 + |\mathcal{M}_T|^2 + 2 \operatorname{Re}(\mathcal{M}_L^* \mathcal{M}_+) + 2 \operatorname{Re}(\mathcal{M}_L^* \mathcal{M}_-). \quad (2.30)$$

For each phase-space point, it is now possible to calculate polarization fractions as the ratio of the polarized cross section over the unpolarized cross section of the event. As the fully differential cross section is directly proportional to the absolute squared matrix element, the polarization fractions of the MVB for a single boson production process are defined as

$$r_L = \frac{|\mathcal{M}_L|^2}{|\mathcal{M}|^2}, \quad r_T = \frac{|\mathcal{M}_T|^2}{|\mathcal{M}|^2}, \quad \text{and} \quad r_{int} = \frac{2 \operatorname{Re}(\mathcal{M}_L^* \mathcal{M}_T)}{|\mathcal{M}|^2} = 1 - r_L - r_T. \quad (2.31)$$

Analogously, it is possible to define polarized matrix elements and polarization fractions at phase-space-point level for diboson production processes. The unpolarized amplitude takes

the form

$$|\mathcal{M}|^2 = \sum_{\lambda, \lambda'} |\mathcal{M}_{\lambda\lambda'}|^2 + \sum_{\lambda\lambda' \neq \lambda''\lambda'''} \mathcal{M}_{\lambda\lambda'}^* \mathcal{M}_{\lambda''\lambda'''}. \quad (2.32)$$

In the polarized amplitude $|\mathcal{M}_{\lambda\lambda'}|^2$, the first index corresponds to the polarization λ of the first intermediate MVB, whereas the second index λ' corresponds to the polarization of the second intermediate MVB.

With

$$\begin{aligned} |\mathcal{M}_{TT}|^2 = & |\mathcal{M}_{++}|^2 + |\mathcal{M}_{+-}|^2 + |\mathcal{M}_{-+}|^2 + |\mathcal{M}_{--}|^2 \\ & + 2 \operatorname{Re}(\mathcal{M}_{++}^* \mathcal{M}_{--}) + 2 \operatorname{Re}(\mathcal{M}_{++}^* \mathcal{M}_{-+}) + 2 \operatorname{Re}(\mathcal{M}_{++}^* \mathcal{M}_{+-}) \\ & + 2 \operatorname{Re}(\mathcal{M}_{--}^* \mathcal{M}_{-+}) + 2 \operatorname{Re}(\mathcal{M}_{--}^* \mathcal{M}_{+-}) + 2 \operatorname{Re}(\mathcal{M}_{-+}^* \mathcal{M}_{+-}), \end{aligned} \quad (2.33)$$

$$|\mathcal{M}_{LT}|^2 = |\mathcal{M}_{L+}|^2 + |\mathcal{M}_{L-}|^2 + 2 \operatorname{Re}(\mathcal{M}_{L+}^* \mathcal{M}_{L-}), \quad (2.34)$$

and

$$|\mathcal{M}_{TL}|^2 = |\mathcal{M}_{+L}|^2 + |\mathcal{M}_{-L}|^2 + 2 \operatorname{Re}(\mathcal{M}_{+L}^* \mathcal{M}_{-L}), \quad (2.35)$$

the unpolarized amplitude can then be written as

$$\begin{aligned} |\mathcal{M}|^2 = & |\mathcal{M}_{LL}|^2 + |\mathcal{M}_{LT}|^2 + |\mathcal{M}_{TL}|^2 + |\mathcal{M}_{TT}|^2 \\ & + 2 \operatorname{Re}(\mathcal{M}_{L+}^* \mathcal{M}_{+L}) + 2 \operatorname{Re}(\mathcal{M}_{L+}^* \mathcal{M}_{-L}) + 2 \operatorname{Re}(\mathcal{M}_{L-}^* \mathcal{M}_{+L}) + 2 \operatorname{Re}(\mathcal{M}_{L-}^* \mathcal{M}_{-L}) \\ & + 2 \operatorname{Re}(\mathcal{M}_{LL}^* \mathcal{M}_{+L}) + 2 \operatorname{Re}(\mathcal{M}_{LL}^* \mathcal{M}_{-L}) + 2 \operatorname{Re}(\mathcal{M}_{LL}^* \mathcal{M}_{L+}) + 2 \operatorname{Re}(\mathcal{M}_{LL}^* \mathcal{M}_{L-}) \\ & + \sum_{\lambda\lambda' \neq \lambda''\lambda'''} 2 \operatorname{Re}(\mathcal{M}_{\lambda\lambda'}^* \mathcal{M}_{\lambda''\lambda'''}) \end{aligned} \quad (2.36)$$

with $\lambda\lambda' \in (++, +-, -+, --)$ and $\lambda''\lambda''' \in (+L, -L, L+, L-, LL)$.

In this case, the polarization fractions are defined as

$$r_{LL} = \frac{|\mathcal{M}_{LL}|^2}{|\mathcal{M}|^2}, \quad r_{LT} = \frac{|\mathcal{M}_{LT}|^2}{|\mathcal{M}|^2}, \quad r_{TL} = \frac{|\mathcal{M}_{TL}|^2}{|\mathcal{M}|^2}, \quad r_{TT} = \frac{|\mathcal{M}_{TT}|^2}{|\mathcal{M}|^2} \quad (2.37)$$

and

$$r_{int} = 1 - r_{LL} - r_{LT} - r_{TL} - r_{TT}. \quad (2.38)$$

3 Machine Learning methods for polarization extraction

The extraction of event-wise vector boson polarization fractions from experimentally accessible final state kinematics via neural networks was first proposed by [3]. Since this approach is further investigated, this chapter introduces the two neural network architectures Multilayer Perceptron and Transformer as well as the hyperparameter optimization framework OPTIMA used in this thesis.

3.1 Multilayer Perceptron

In Machine Learning, neural networks (NNs) are computational models used to approximate nonlinear mathematical functions and are inspired by the structure and functions of biological neural networks. They consist of connected artificial neurons typically arranged in layers. Many architectures arrange the layers in a chain. Information is passed from the first layer, the input layer, to the last layer, the output layer. Between these, one or multiple intermediate layers, the hidden layers, are situated. Deep neural networks are NNs with at least two hidden layers. The weights of the connections between the neurons determine the signal strength that is passed to the neurons in the next layer. During training, the model learns to adjust these weights via mechanisms like loss minimization and backpropagation which will be described in the following based on [16], [17] and [18].

The Multilayer Perceptron (MLP) is a fully connected NN, which means that each neuron is connected to all neurons of the previous layer. Hence, the hidden layer neurons receive numerical inputs from all neurons in the preceding layer. A schematic depiction of a MLP, exemplary with five hidden layers, is shown in Figure 3.1. For each neuron, a weighted sum of the input values is calculated before adding a fixed bias term. This value is then passed through the activation function of the neuron. The output z^i of a neuron i can be calculated as

$$z_i = \sigma\left(\sum_j w_{ij}x_j + b_i\right) \quad (3.1)$$

with the activation function σ , the input x_j and weight w_{ij} from the neuron j in the previous layer and the bias term b_i . The weights and biases are trainable parameters, which means that

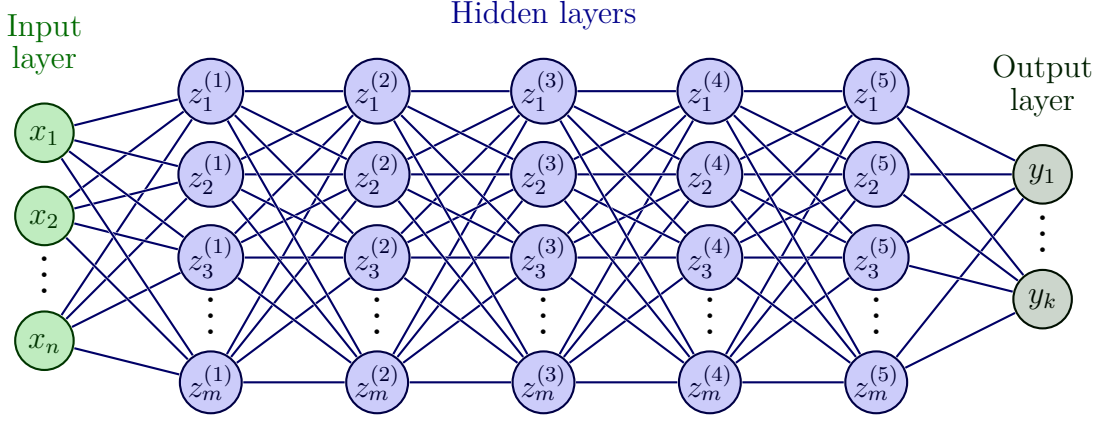


Figure 3.1: General architecture of a Multilayer Perceptron with input layer (green) of n input variables x_n , $h = 5$ hidden layers (purple) of equal neuron number m and a output layer (grey) of k output variables y_k . The output of the hidden layer neurons is denoted by z_m^h and a weight w_{ij}^h is assigned to each connections between the neurons.

they are adjusted during the training process. The output vector $\vec{z}$ of an entire layer can be expressed in terms of a input vector $\vec{v}$, the bias vector $\vec{b}$ and the weight matrix $\mathbf{W}$ as

$$\vec{z} = \sigma(\mathbf{W}\vec{x} + \vec{b}), \quad (3.2)$$

where the activation function σ is applied element-wise. If σ is non-linear, the NN is able to model non-linear relationships between the input and output variables [19]. The choice of the best activation function is highly dependent on the task and network architecture [20], which will further be discussed in section 3.3.

In the case of supervised learning, the model f is trained on a dataset of tuples $\langle \vec{x}_i, y_i \rangle$ where $\vec{x}_i$ is the input vector of the data point i and y_i its target variable. It learns to predict values $\hat{y}_i$ from the input $\vec{x}_i$ that are as similar to the target value y_i as possible. A loss function $\mathcal{L}(y_i, \hat{y}_i)$ is introduced that measures the distance of prediction and target values. The model parameters, the NN weights $\mathbf{W}$ and biases $\vec{b}$ are adjusted, such that the loss function $\mathcal{L}$ is optimized and the distance of predictions and targets is minimal. The weight update is achieved using the gradient descent algorithm where the gradient is calculated by backpropagation. Here, the model parameters ϑ are adapted according to the weight update rule

$$\vartheta_{s+1} = \vartheta_s - \eta \nabla_{\vartheta} \mathcal{L}(y_i, \hat{y}_i). \quad (3.3)$$

Here, the model parameters ϑ_{s+1} are derived from the parameters ϑ_s of the previous training steps by subtracting the gradient $\nabla_{\vartheta} \mathcal{L}(y_i, \hat{y}_i)$ of the loss function with respect to the model parameter scaled by a factor η . This factor η is called the learning rate and determines the step size of the weight update.

The optimization landscape of the problem is strongly defined by the loss function. The

most common loss function for classification tasks (discrete output variables) is the cross-entropy. For regression tasks with continuous output variables, the Means Squared Error MSE is typically used so that

$$\mathcal{L} = MSE(y_i, \hat{y}_i) = \frac{1}{n} \sum_i^N (y_i - \hat{y}_i)^2. \quad (3.4)$$

Here, y_i is the NN prediction for the data point i , $\hat{y}_i$ the corresponding target and n is the number of data points in the dataset considered in one backpropagation step.

The optimal weight and bias values are approached iteratively by running backpropagation on the entire training dataset or only subsets. The latter case is called Stochastic Gradient Descent (SGD) and includes the random splitting of the dataset into mini batches. The model parameters are updated according to the loss gradient of these subsets. Completing one batch is then referred to as a step while completing the entire dataset is referred to as an epoch. The batches can be randomized again after each epoch.

With multiple parameter updates per epoch, the convergence of the loss optimization is faster and can be more stable. Averaging over a subsample helps to avoid convergence at local minima of the loss landscape. The smaller the batches, the higher is the probability of gradient variance and weight update fluctuations which potentially prevents convergence. This makes the batch size an essential parameter to select before the training.

For the NN training in this thesis, another gradient descent algorithm is chosen. The Adaptive Moment Estimation optimizer (ADAM) is more complex and improves the convergence behaviour compared to the use of SGD. For the parameter update, it takes into account the gradients of past epochs (less probable convergence at local minima or saddle points) and scales the learning rate dynamically (higher step size for regions of lower gradient fluctuation that lead to a reduced convergence time) [21].

3.2 Transformer

The Transformer was introduced for sequential data in the field of Natural Language Processing (NLP). Former architectures like recurrent neural networks were not able to capture all information in data with longer sequences. This problem could be solved by proposing a self-attention mechanism which will be described in the following. The section is based on [22], where the Transformer model was first proposed. The Transformer architecture with an encoder-decoder structure did not only show better performance in translation tasks but was also more parallelizable and faster in training. Because of this, the Transformer gained more and more popularity in the past years, in NLP but also computer vision and audio processing [23]. At high-energy experiments like the ATLAS experiment at the LHC, transformer architectures are explored in their ability to capture extremely complex relations in measured or

simulated data (see e.g. [24] or [25]).

The proposed attention mechanism uses the scaled dot-product attention function defined as

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (3.5)$$

with query matrix Q , key matrix K and value matrix V . It is a weighted sum of the values with weights calculated from the dot products of the query with all keys. The matrix QK^T contains the dependencies of each sequence element with all others. It is scaled with the *softmax* function to obtain weight values between 0 and 1. This weighted sum finally determines how much attention is required by the different elements in the input sequence. Learnable parameters are introduced by the weight matrices $W_i^Q \in \mathbb{R}^{d_m \times d_k}$, $W_i^K \in \mathbb{R}^{d_m \times d_k}$ and $W_i^V \in \mathbb{R}^{d_m \times d_v}$, such that Equation 3.5 is calculated for $\text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ of each head i . Here, d_m denotes the model dimension, whereas d_k and d_v are the dimensions of the key and value matrix.

The multi-head attention mechanism now uses multiple attention heads $i = 1, \dots, h$ of the scaled dot-product attention. Complex input data often requires to attend to multiple aspects of the sequence elements. The query, key and value matrices are split and the obtained heads calculate multiple attention vectors simultaneously, which are then concatenated and multiplied by a trainable output weight matrix $W^O \in \mathbb{R}^{h \cdot d_v \times d_m}$ for a final output scaling. This allows attention to multiple representation subspaces at once.

The original Transformer in [22] consists of a encoder-decoder structure. The task of the encoder is to map the the input to an abstract continuous representation. The N stacked layers of the encoder consist of two sub-layers each. The first sub-layer uses the introduced multi-head self-attention mechanism whereas a fully connected feed-forward network, as described in the previous section, forms the subsequent sub-layer. Each sub-layer is equipped with a layer normalization and a residual connection. The decoder in the original Transformer uses the output of the encoder to generate sequential output. In this thesis, the Transformer architecture is not used to generate sequential output but applied on vectors of physical quantities as inputs to predict single numerical values of a continuous output variable. Therefore, the decoder is not necessary and only the encoder structure is used as shown in Figure 3.2.

Generally, models with higher capacity and more model parameter are more likely to capture complex relationships between input and output data. In [26], MLP-based architectures and Transformer-based architectures were compared regarding their expressive power in simultaneously modelling dependencies between multiple input variables. An exponential gap between the MLP-based mechanisms and the attention mechanism was shown. Thus, more parameters are needed for the MLP than for the Transformer to model highly complex relations in data. Furthermore, depth-width trade-offs of Transformer architectures showed significant parameter savings for certain tasks in [27]. Therefore, the Transformer encoder architecture

is employed and compared to the predictive performance of the MLP.

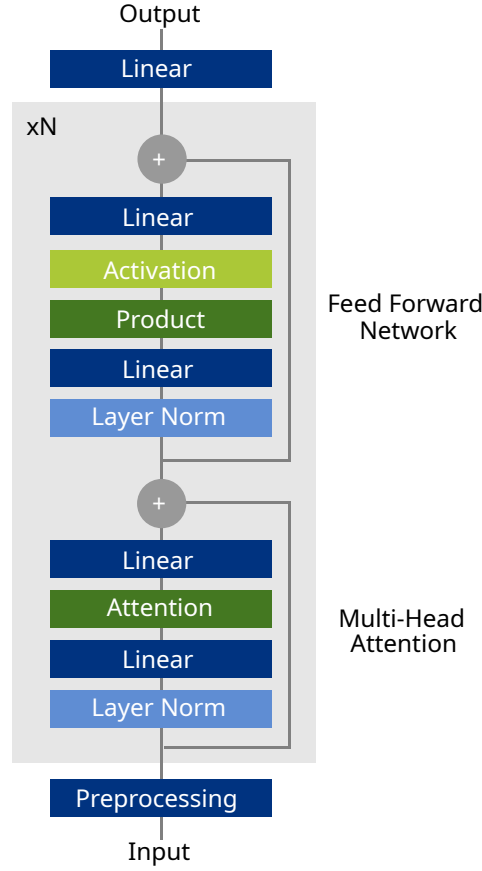


Figure 3.2: General Transformer encoder architecture as used in this thesis. The input is first preprocessed before it is propagated through N stacked layers, each consisting of a multi-head self-attention sub-layer and a fully connected feed-forward layer. Each of the sub-layers is equipped with a residual connection and a layer normalization. The illustration is based on [22] and [24].

In [28], a weight decay regularization is deployed by decoupling the weight decay from the optimization steps and shown to improve the generalization performance of models trained with ADAM. Therefore, the ADAM algorithm with decoupled weight decay (AdamW) is used for the Transformer opposed to the regularization of the MLP. In the past, the effect of layer normalization was shown to substantially affect the Transformer training. Originally, the Transformer used post-norm residual units. Here, the layer normalization was performed after each sub-layer and residual addition. But [29] demonstrated that pre-norm residual units with layer normalization before each sub-layer improved the model performance substantially. This benefits the backpropagation and training of deep architectures with many layers [30]. Pre-norm residual connections with small weight initialization values allow a training without warm-up and a larger learning rate [31] and is therefore employed for the Transformer.

3.3 Hyperparameter optimization with OPTIMA

The generalization performance of NNs describes how well the model generalizes to unseen data. It depends strongly on the applied regularization strength and the model capacity which are defined by hyperparameters. These are model parameters that are set manually before and are not adjusted during training. A careful selection of a model's hyperparameters prevents over- and underfitting and is therefore crucial. The optimization of the hyperparameter values aims to achieve the best generalization error. The manual test of different values is common but time-consuming, especially for complex models with a high number of hyperparameters. Here, the hyperparameters are adjusted by interpreting the validation error in relation to the training error. This search for a optimal set of hyperparameters can be automated. The hyperparameter optimization framework OPTIMA [32], which is used in this thesis, was developed within the work of and components relevant for this thesis are described in the following based on [33].

The input and output variables are first loaded and the input is preprocessed for the training. The search space of hyperparameter ranges is provided to the framework. The hyperparameter optimization uses Bayesian optimization, which, unlike grid search, does not include the training of each combination of hyperparameter values with defined variation steps. It is therefore more suitable for complex NNs with many hyperparameters. It uses surrogate functions as a probabilistic model of the NN performance as a function of the hyperparameters based on previous observations. During the optimization, the hyperparameters are adjusted according to the surrogate function that predicts regions of the search space in which the hyperparameter values should be tested next. The surrogate function is updated according to the results of the tested set of hyperparameters on the model performance. Each optimization ends with an evaluation to extract the best hyperparameters. The generalization performance of the resulting model is then estimated via k -fold cross-validation. Here, k models are trained on $k - 1$ subsets of the dataset and then validated on the remaining subset. The subsets used for training and validation are permuted for each fold, such that the model is tested on independent validation datasets. From this, k performance values are obtained that allow an estimate of the performance (calculated as the average value) and its uncertainties (calculated as the standard deviation). In this thesis, $k = 5$ folds are applied.

In the following, the hyperparameters of the NN architectures used in this thesis are introduced. The update of the model weight during training is governed by the `learning_rate` and the `batch_size` the average gradient is calculated on. Using the ADAM algorithm, it is also controlled by the decay rates `Adam_beta_1` and `one_minus_Adam_beta_2` of past gradients and the parameter `Adam_epsilon` for numerical stability. The regularization strength of the MLP is determined by the `dropout` rate which gives the percentage of randomly dropped neurons of each hidden layer to avoid extreme dependence on single neurons. Specific to the MLP are the number of hidden layers (`num_layers`) and the number of layer neurons (`units`).

The Transformer model is defined by the number of heads of the multihead attention layer (`n_heads`), the dimension of these attention heads (`attention_head_dim`), the dimension of the feedforward network layer (`FF_dim`) and the overall number of transformer blocks (`num_transformer_blocks`). The dropout is set via the dropout value of the encoder layers (`dropout_encoder`) and of the final feedforward network (`dropout_pred`). The `aggregation` parameter determines wheather the encoder output is aggregated as the arithmetic mean or maximal value of all the values before it is propagated to the output layer.

An additional hyperparameter of NNs is the activation function of the neurons. The first activation function in NNs was the sigmoid function $\sigma(x) = \frac{e^x}{1+e^x}$, which models the neuronal activation in the brain. With increasing size and depth of the NNs, the training of the models was impaired by vanishing gradients and other activation functions were investigated. The Rectified Linear Unit (ReLU) function defined as $\text{ReLU}(x) = \max(0, x)$ became popular due to its simplicity and computational efficiency but, with it, the problem of "dying neurons" occurred [34]. Neurons with consistently negative input can become inactive over the course of the training because the updating of their weights stops. Many alternative activation functions have been proposed but most showed inconsistent performance gains across models and tasks [20]. One proposed alternative is the Gaussian Error Linear Unit (GeLU) which gained popularity with models like BERT and GPT. It is defined as $\text{GeLU}(x) = x \cdot \Phi(x)$ with $\Phi(x)$ being the standard Gaussian cumulative distribution. It was demonstrated to outperform ReLU in Transformer models by better optimization and gradient flow [35]. Another proposed activation function is the Swish function defined as

$$\text{Swish}(x) = \frac{xe^x}{1 + e^x}. \quad (3.6)$$

In [36], it was shown to outperform ReLU on deeper networks across a number of different tasks. Compared to GeLU, Swish profits from the similarity to ReLU with its computational efficiency and gains popularity. The Transformer uses the GeLU activation function which was shown to allow better training in Transformer models across multiple tasks [35]. The Swish activation function is chosen for the activation function of the neurons in the MLP with the weight and bias initialization schemes suggested in [37] due to the computational efficiency. Table 3.1 and Table 3.2 show the search spaces of the hyperparameter optimization. In [33], it was shown that the impact of the optimization varies for the different model hyperparameters. Therefore, some hyperparameters are not optimized but set to default value for all optimizations in this thesis. They are `learning_rate` = 0.001, `Adam_beta_1` = 0.6, `one_minus_Adam_beta_2` = 0.001 and `Adam_epsilon` = $1e^{-8}$.

Table 3.1: Hyperparameter optimization search space of the MLP.

Hyperparameter	Search space
num_layers	$[2, 8] \in \mathbb{Z}$
units	$[16, 256] \in \mathbb{Z}$
dropout	$[0.0, 0.4] \in \mathbb{R}$
batch_size	$[64, 2048] \in \mathbb{Z}$

Table 3.2: Hyperparameter optimization search space of the Transformer.

Hyperparameter	Search space
attention_head_dim	$[4, 16] \in \mathbb{Z}$
n_heads	$[2, 8] \in \mathbb{Z}$
FF_dim	$[8, 128] \in \mathbb{Z}$
dropout_encoder	$[0.0, 0.1] \in \mathbb{R}$
num_transformer_blocks	$[2, 8] \in \mathbb{Z}$
aggregation	['mean', 'max']
dropout_pred	$[0.0, 0.3] \in \mathbb{R}$
batch_size	$[256, 2048] \in \mathbb{Z}$

3.4 Neutral networks in vector boson polarization tagging

As discussed in chapter 2, the EWSB explained by the Higgs mechanism can be investigated by the polarization of MVB at high-energy experiments like the LHC. W and Z bosons of the electroweak interaction acquire mass by coupling to the Higgs boson and therefore obtain a longitudinal polarization state. Observing a difference between measured and predicted production of polarized MVB can give insights into BSM mechanisms realized in our universe. Interest in the polarization of MVB has grown steadily, as it is part of more and more analyses of the LHC. Here, the MVB polarization is investigated in single boson processes (see e.g. [38] and [39]) as well as diboson production processes (see e.g. [40] and [41]). With the Large Hadron Collider reaching higher luminosities, Run-3 and the High-Luminosity LHC are expected to deliver enough statistics to study polarized signals in progressively rarer processes. To extract the polarization of MVB in the LHC, multiple approaches were proposed in the past. For Run-1 data, polarization fractions were obtained from angular coefficients of the decay angles of the MVB [42][43]. Since the necessary analytic decomposition of the decay angle distributions only holds in the decay-inclusive phase space, this method comes with practical limitations. A second method was proposed, which uses polarized templates obtained from Monte Carlo event simulation. For each polarization state, a polarized template is calculated separately and the independent templates are fitted simultaneously to the unpolarized signal [15][44]. This method is restricted to observables that discriminate most between the polarization modes. As a result, the full polarization information contained in the data may not be

exploited completely. Machine Learning can be used to extract the polarization fractions of the intermediate MVB. Typically, an event-by-event classification into transverse or longitudinal polarized events is performed based on kinematic observables [45][33]. In [3], an regression on the event-wise polarization fractions of Z bosons was proposed first. A MLP was trained to predict the polarization fractions on an event-by-event basis which were derived from MC simulations at leading order with and without the presence of parton shower effects. The computation of the MVB polarization fraction, defined at the phase-spacepoint level as the ratio of polarized over unpolarized squared matrix element, requires the kinematic knowledge of all initial and final state particles. The momenta and type of the initial state particles is not available experimentally and, thus, only the momenta of the final state particles as the available information is passed to a NN. From this, the NN is trained on MC simulation to predict the polarization fraction. The model can later be applied to data to extract the polarization of MVB which cannot be measured directly. This method allows to combine kinematic information of LHC events and theoretical considerations for each event separately. It is not restricted to variables most sensitive to discriminate the polarization states. Also, there are no phase space requirements.

The described ML approach of [3] will be developed further in this thesis. Mainly, five aspects will be evolved:

- Formerly, two polarized predictions are generated by a reweighting of the unpolarized sample. One is produced by reweighting according to the polarization fraction predicted by the network. The other one is obtained by reweighting according to the true polarization fraction which is calculated by the MC simulation and used to train the network. The kinematics of the final state particles in both predictions are then compared. In this thesis, the distributions of the raw events weights are presented and the predicted and true polarization fractions will be compared directly instead of relying on reweighted distributions.
- So far, only one NN architecture, a MLP, is investigated. In comparison with Transformer-based architectures, MLPs show an exponential gap in expressive power in terms of their ability to model dependencies among multiple input variables simultaneously [26]. Thus, MLPs require far more parameters than Transformers to represent equally complex relations in data. Therefore, this thesis compares the performance of a MLP and a Transformer.
- The work of [3] includes no optimization of the NN hyperparameters. The parameters not adjusted during training were set beforehand to fixed values. To find a set of hyperparameters, which are optimal for the specific task at hand, the framework OPTIMA is used to optimize the hyperparameters of the MLP and Transformer in this thesis. A lack in predictive performance of the network is therefore not a consequence of a specific

(at worst arbitrary) hyperparameter choice but shows a systematic inability or difficulty of the network in the polarization extraction.

- Up to now, the MLP was only trained to predict the longitudinal polarization fraction r_L as output variable. The transverse polarization r_T of the MVB is not included. Since some BSM theories also predict differences in the transverse polarization [46], the extraction of all polarization fractions provides a wider and more sensitive probe of potential deviations from SM predictions. In this thesis, both longitudinal and transverse MVB polarization fractions will be considered.
- In the former work, the proposed method is only applied to $Z+jet$ production at leading order with and without parton shower effects. This thesis also investigates the effect of higher orders of the strong coupling constant and phase space effects in $Z + jets$ as well as single W^+ boson and $WZjj$ -EW diboson production processes, which were not explored before.

4 Simulation

The polarization of the MVBs cannot be measured directly because MVBs have a very short lifetime and appear as unstable, intermediate particles in experiments. In the thesis, NNs are trained on simulated events of the particle collision, during which MVBs are first produced and decay afterwards. In such simulations, the full information on all particles including the polarization of the intermediate vector boson(s) is available. To investigate the NN behaviour, simulations of polarized vector boson processes are necessary for training and a quantitative prediction vs. truth comparison. Simulating the proton-proton collisions requires extensive theoretical work due to the colour confinement of the QCD at long distances and considered detector phase spaces. The processes cannot be derived analytically and Monte Carlo (MC) event generators are used in good approximation. There are multiple general-purpose event generators, the most popular are Sherpa [47], Pythia8 [48] and Herwig [49]. The samples used in this thesis are generated with Sherpa (Simulation of High-Energy Reactions of PArticles). The first section of this chapter introduces fundamental concepts of the simulation of pp collisions with Sherpa. Section 4.2 briefly discusses the recently implemented option to simulate polarized cross sections in Sherpa [14]. The last section of this chapter examines the simulated samples of vector boson production processes and their settings.

4.1 Simulation of proton-proton collisions

The complex nature of processes occurring in proton-proton collisions makes analytical calculation unfeasible. For this reason, simulations based on Monte Carlo (MC) methods were developed to investigate the processes. This section presents such MC simulations of proton-proton collisions in Sherpa based on [4] and [50].

QCD describes a big part of the collision and shows a strongly energy-dependent behaviour. Because of this, the collision simulation is factorized into phases of different energy scales. The hard process describes the inelastic scattering of two partons from the colliding protons and occurs at the highest event energies in the event simulation. At the energy scale of proton-proton collisions in high energy experiments like the LHC, the partons in the protons behave as quasi-free particles. In QCD, higher energies or smaller distances of the partons decrease the coupling between the partons (asymptotic freedom). Since the strong coupling constant α_s is small at high energies, the hard process can be calculated by perturbation theory. The matrix

element of the hard process is calculated up to a given order from the Feynman diagrams connecting IS and FS particles. The phase space integration is performed together with the simulation of the events. The event generation is completed for fixed-order calculations but not in cases where a more experimentally realistic event generation is desired.

The parton content and structure of the protons is parametrized by the Parton Distribution Function (PDF). It describes the probability to find a parton q with the proton's longitudinal momentum fraction x_q within the proton. The proton consists of two up-quarks and one down-quark, called valence quarks, while pairs of quarks and anti-quarks are repeatedly generated and annihilated by small fluctuations of the proton energy (sea quarks). These sea quarks have low momentum or exist on a short time scale. Larger energies of the colliding protons increase the probability of heavy quarks within the proton and so the PDF depends on the energy Q of the proton. The inelastic scattering partons can radiate particles or originate from the splitting of other particles. The PDF of the valence quarks peaks for $x_q \approx 1/3$, while the gluon PDF is large for low momentum transfer. Since it is very unlikely for one parton to carry the entire proton energy, the PDF decreases for high x_q . In the generation, unstable particles produced in the hard process (top quarks, Higgs, $W^\pm$ or Z boson) decay into stable FS particles.

The FS particles as well as the IS particles of the hard process can radiate photons, quarks and gluons of lower energies. These softer emissions are simulated with the help of the Parton Shower (PS) algorithm in perturbation theory. The PS simulates how the high-energetic quarks and gluons radiate additional partons as they evolve from a hard process down to lower energy scales by approximating QCD in the soft and collinear limits. It uses a probabilistic, stepwise branching algorithm to generate successive emissions ordered in a chosen evolution variable to produce a realistic final state. Different mechanisms are implemented in the PS model to avoid double counting of the emissions from the hard process and the PS at higher orders. While the process can be described properly by perturbation theory at leading order, higher orders of the process appear at lower energies with a higher coupling constant and phenomenological models have to be introduced.

In nature, only colour-neutral states are observed. Therefore, the final state of the hard process and the PS form colour-neutral states in a process called hadronization. At low energy scales, the potential energy between quark - anti-quark pairs increases with increasing separation. At sufficient potential energy, new pairs of quarks and anti-quarks are created (QCD confinement). When their kinetic energy is low enough, these pairs form colour-neutral bound states with longer lifetime which are observable by the detector whereas unstable hadrons decay before they are detected. Since the colliding protons do not only consist of the scattering partons, the other partons can themselves take part in collisions of lower energy. These interaction processes not associated to the hard final state are called the underlying event. Remaining parts of the scattered partons after the hard process are described by beam remnants and

the interaction of multiple partons are included by multi-parton interactions. With the MC event generator Sherpa, events of particle collisions can be simulated up to NLO QCD and approximative NLO (nLO) EW [47]. The generation of realistic events is achieved by the simulation of parton showers, QED radiation, hadronization and multi-parton interactions. The different stages of the event generation in Sherpa are illustrated in Figure 4.1.

The origin of the coordinate system used in the Sherpa simulation lies in the collision point. From this collision point, the z-axis is defined by the incoming particle beams, whereas the x-axis and y-axis are oriented perpendicular, such that the x-y plane describes the plane transverse to the beam direction. The three axes form a Cartesian right-handed coordinate system. A spherical coordinate system is introduced where the azimuthal angle ϕ describes the angle in the x-y plane and the polar angle θ is defined as the angle with respect to the beam line.

The samples generated in this thesis are on particle level meaning that momenta, energies and origins of all particles are known. In a more realistic setting, it would further be necessary to simulate the pile-up of other proton-proton collisions during the bunch crossing and the detector response which considers geometry and material of the detector.

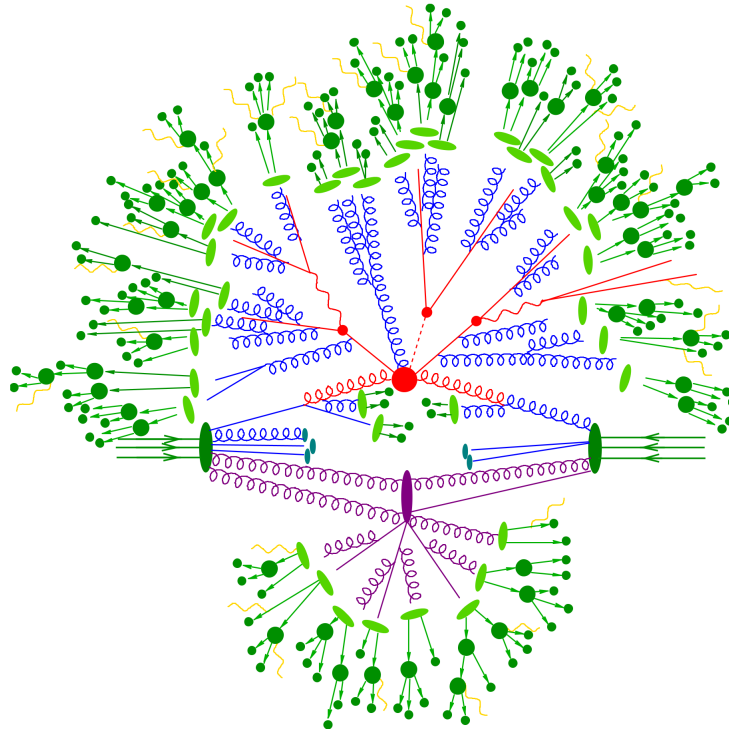


Figure 4.1: The MC event generation steps in Sherpa: hard scattering process (red), initial and final state parton shower (blue), hadronization (light green), hadronic decays (dark green), radiation of photons (yellow) and underlying event (violet) [47].

4.2 Simulation of polarized cross sections in Sherpa

Within the work of [51], the simulation of polarized cross sections of intermediate MVBs was implemented in Sherpa. It is described in the following based on [14] and [52]. The polarized cross sections can be calculated at fixed LO, LO+PS and in approximation with NLO QCD corrections on the MVB production.

Sherpa simulates the cross sections of all possible polarizations in a single simulation run during the event generation. The polarized cross sections are provided together with the interference as additional weights for each generated event. Two definitions of transversely polarized cross sections are provided. In the incoherent definition, only the left- and right-handed contributions are added, whereas in the coherent definition used in this thesis, the left-right interference terms, not included in the transverse polarized weight, are also taken into account.

The polarization for intermediate MVBs can only be defined for processes which factorize into production and decay part of the MVBs. The narrow-width approximation described in subsection 2.2.2 is applied to only consider fully resonant diagrams. This means, all FS particles of the decay originate from the resonant intermediate MVB which is considered on-shell. All spin correlations are preserved during the production and decay of the MVB.

The polarization vectors of the MVBs are expressed in the helicity basis so that they are eigenvectors of the helicity operator. Since the helicity is only Lorentz invariant for massless particles, a reference system must be chosen for a unambiguous definition of the MVB polarization (see subsection 2.2.1). This reference system can be the laboratory frame, the center-of-mass system of all particles decaying in the hard process or the center-of-mass system of the two interacting partons. For single boson production processes, the laboratory frame ("Lab") is used as the reference frame for the definition of the polarization which is the common reference frame for single boson polarization investigations. For the diboson samples, the center-of-mass system of the produced W boson and Z boson from the hard process ("COM") is applied since this is most common for diboson polarization studies.

4.3 Simulated samples

This section first provides an overview of the single boson and diboson production processes on which the polarization tagger is applied. Then, the simulation settings and the phase space definitions used for the sample generation are described. In this thesis, only the leptonic decay channels of MVBs are considered.

4.3.1 Investigated processes

***Z+jets* and *W+jets* production**

First, a single boson in association with a defined number of jets is produced in the hard process. At leading order, the $Z+1j$ and the W^++1j production of coupling order $\mathcal{O}(\alpha_S\alpha_{EW}^2)$ are investigated. These are one of the simplest processes of MVB production and, due to their high cross section, they play an important role as background in many of the investigated SM processes at high-energy particle colliders. The Feynman diagrams contributing at LO are depicted in Figure 4.2a. The radiation of additional gluons from IS and FS particles through the QCD Parton Shower algorithm is shown in Figure 4.2b, exemplary for one of the hard processes at LO. Virtual corrections contributing to the calculations at NLO QCD are presented in Figure 4.2c. In the subsequent leptonic decay, the Z boson decays in a pair $\mu^+\mu^-$ of two muons with opposite electric charge. The W^+ boson decays into a pair $\mu^+\nu_\mu$ of a charged muon and its corresponding neutrino.

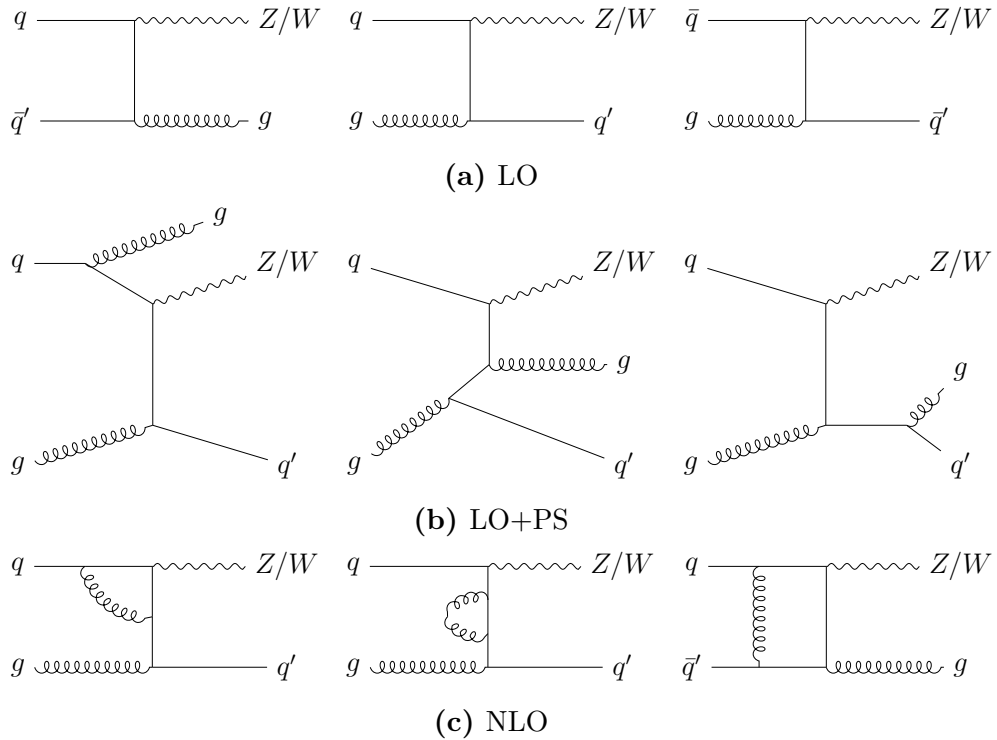


Figure 4.2: Feynman diagrams of the hard scattering process in $Z+jet$ and $W+jet$ at leading order (LO), leading order with IS and FS parton shower (LO+PS) and virtual corrections at next-to-leading-order (NLO) in QCD.

***WZjj*-EW production**

Vector Boson Scattering (VBS) describes processes in which two MVBs as initial particles scatter during the hard interaction. The SM predicts a very small cross section for the VBS process. With increasing luminosity at the LHC, it became possible to measure these rare

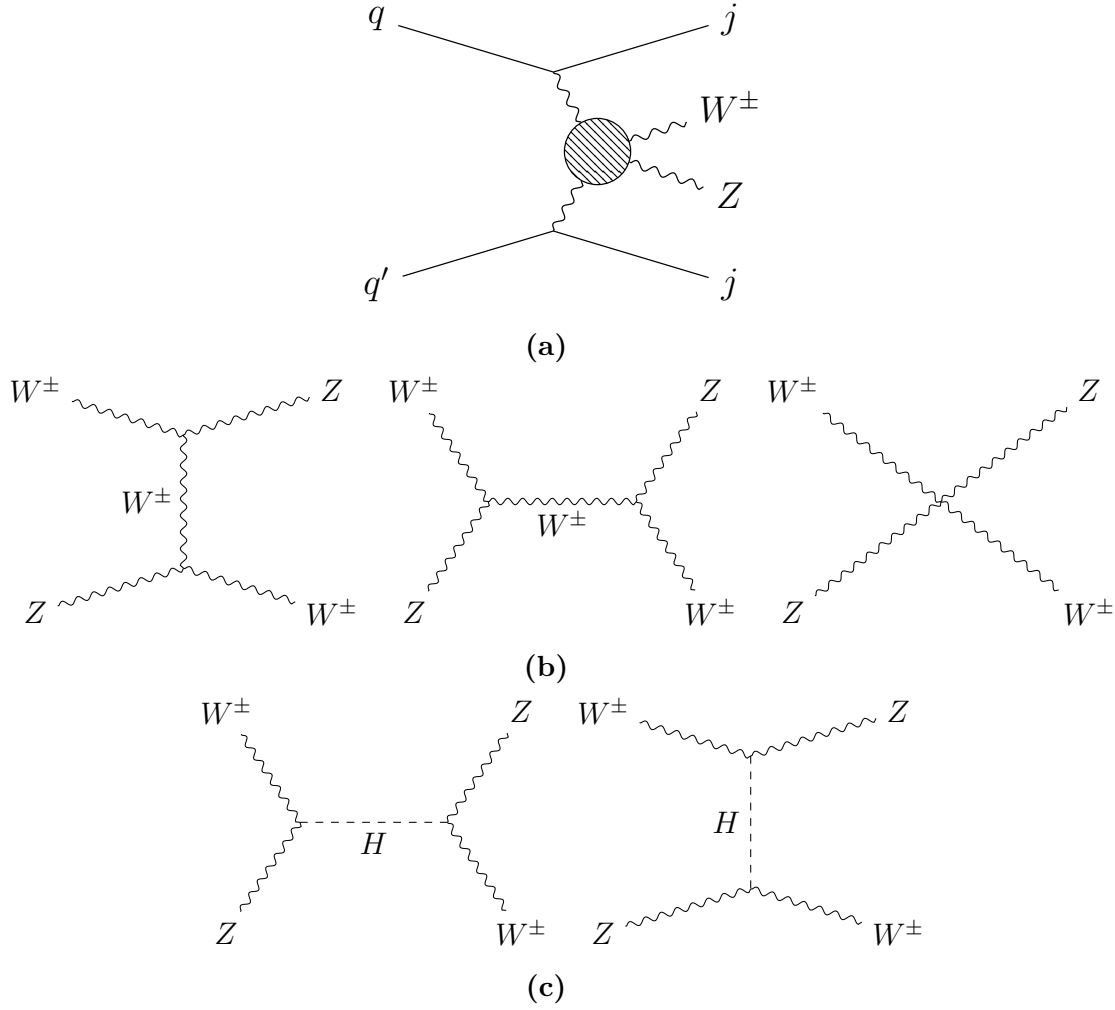


Figure 4.3: Double-resonant Feynman diagram of the electroweak $W^\pm Z$ boson pair production in association with two jets at leading order (a) and the corresponding vector boson scattering contributions of triple- and quartic MVB self coupling (b) as well as the MVB with Higgs coupling (c).

processes experimentally in the last years (see e.g. [40],[41] and [53]). This process is especially interesting due to its relevance for testing the SM structure and possible new physics. The Lagrangian of the electroweak theory contains triple and quartic gauge boson self interactions as described in subsection 2.1.2. The VBS of two gauge bosons is highly sensitive to the realization of the Higgs mechanism. Without the contributions from the doubly longitudinal polarized vector boson pair, the VBS cross section would grow continuously with increasing IS center-of-mass energy [4]. At sufficiently high energies, the scattering probability would excel 1, which would violate unitarity. Thus, the VBS is of high interest because the large unitarity-violating contributions from gauge and Higgs boson diagrams cancel for the scattering of longitudinal MVBs in the SM [5]. This cancellation could be affected by BSM physics where the VBS rate could possibly be increased. Polarization measurements in VBS processes provide insights to the SM gauge symmetry structure as well as the realization of the EWSB.

In this thesis, the electroweak production of a W^+Z boson pair in association with two jets is investigated at the coupling order of $\mathcal{O}(\alpha_{EW}^6)$. The contributing Feynman diagrams include triple gauge boson self interaction, quartic gauge boson self interaction as well as gauge boson with Higgs interaction. These are shown in Figure 4.3.

The VBS process is difficult to study in experiments since it is extremely rare: the gauge bosons cannot be directly produced and they decay after a very short time. In proton-proton collisions at high energies, the IS quarks in the colliding protons couple to the gauge bosons before they are detected as jets. The IS quarks only interact electroweakly and continue travelling almost in beam direction. The two jets therefore have high invariant mass ($m_{jj} > 500$ GeV) and a significant rapidity difference ($\eta_{j1} \cdot \eta_{j2} < 0$). This characteristic signature allows to increase the contributions of VBS diagrams in the sample in form of VBS selection criteria, so that the extremely rare process can be studied [51]. Instead of Z bosons, these diagrams can also contain photons. Forcing the lepton-anti-lepton invariant mass close to the Z mass allows to reduce these photon contributions. Contributions from resonant top quarks not containing VBS are suppressed by a veto on bottom quarks in the simulation. After the W^+Zjj production, the Z boson decays in a pair l^+l^- of two charged leptons with same flavour but opposite electric charge. The W boson decays into a pair $l'\nu_{l'}$ of a charged lepton and its corresponding neutrino.

4.3.2 Settings of the event simulation

The events are generated with Sherpa 3.0 [47] and analysed with Rivet 4.0 (see [54] and [55]), which is an analysis tool for MC event generators. The Rivet analysis allows to apply further selection criteria, e.g. on the intermediate MVB reconstructed by these final state particles. Sherpa runcards of the single boson and diboson production are presented in chapter B exemplarily at LO+PS.

The setting of the simulation parameters applied for the event generation are described in the following. The matrix element generators COMIX [56] and AMEGIC [57] are augmented by OpenLoops [58] for virtual matrix element calculations at NLO. For the identification of the QCD jets, the anti- k_T algorithm [59] is employed with a resolution parameter of $R = 0.4$. The energy of both incoming proton beams is set to $E = 6500$ GeV, accounting for a center-of-mass energy of the collision at $\sqrt{s} = 13$ GeV. The PDF set NNPDF31_nlo_as_0118 from the LHAPDF library [60] is used. The electroweak scheme applies the G_μ scheme where $G_\mu = 1.16639 \cdot 10^{-5}$ GeV⁻². The QCD renormalization and factorization scales are chosen dynamically with $\mu_R = \mu_F = m_Z^2$ for the $Z+jets$ samples, $\mu_R = \mu_F = m_W^2$ for the $W+jets$ samples and $\mu_R = \mu_F = m_{WZ}/2$ for the $WZjj$ -EW samples. The MVBs are set to have a mass of $m_W = 80.399$ GeV and $m_Z = 91.1876$ and a width of $\Gamma_W = 2.085$ GeV and $\Gamma_Z = 2.4952$. The transverse mass of the W boson, which is restricted for the $WZjj$ -EW samples is defined according to [61] as

$$m_T^W = \sqrt{2p_{miss}^T p_{l'}^T [1 - \cos \Delta\phi(p_{miss}^T, p_{l'}^T)]}, \quad (4.1)$$

with $\Delta\phi(p_{miss}^T, p_{l'}^T)$ being the angle between the transverse momentum $p_{l'}^T$ of the W boson decay lepton and the missing transverse momentum p_{miss}^T .

For the analysis of the simulated events, modified versions of the Rivet standard analyses MC_ZJETS for the $Z + jets$ samples and MC_WJETS for the $W + jets$ samples are used. For the $WZjj$ -EW samples, the analysis ATLAS_2018_I1711223 from [61] is applied. For each sample, 10,000,000 events are generated.

The phase space definitions of the generated samples are presented in Table 4.1 and Table 4.2. Almost all samples in this thesis are generated in the generator phase space with selection criteria on the leading jet of the hard process. In case of Z boson production, a mass window for the Z boson is applied. For the W boson, this is not possible: it cannot be reconstructed from the decay leptons since only the transverse momentum and not the full four-momentum of the neutrino can be estimated in experiments. Only one of the $Z + jets$ samples is generated in the fiducial phase space with additional selection criteria on the charged decay leptons.

Table 4.1: Selection criteria of the $Z+jets$ and $W+jets$ samples generated with Sherpa for the generator and the fiducial phase space according to [3].

Phase space	Selection criteria
Generator	$p_{T,j} > 10 \text{ GeV}$ $ \eta_j < 5$ $76 \text{ GeV} < m_Z < 106 \text{ GeV}$ (only for $Z+jets$)
Fiducial	$p_{T,j} > 20 \text{ GeV}$ $ \eta_j < 4$ $p_{T,l} > 20 \text{ GeV}$ $ \eta_l < 2.7$ $81 \text{ GeV} < m_Z < 101 \text{ GeV}$ (only for $Z+jets$)

Table 4.2: Selection criteria of the $WZjj$ -EW datasets generated with Sherpa in the generator phase space according to [61].

Selection criteria
perfect b-veto
$p_{T,l^+} > 15 \text{ GeV}$
$p_{T,l^-} > 15 \text{ GeV}$
$p_{T,\nu} > 20 \text{ GeV}$
$ \eta_l < 2.5$
$81 \text{ GeV} < m_Z < 101 \text{ GeV}$
$m_{T,W} > 30 \text{ GeV}$
$p_{T,j} > 40 \text{ GeV}$
$ \eta_j < 4.5$
$\eta_{j_1} \cdot \eta_{j_2} < 0$
$m_{jj} > 500 \text{ GeV}$

5 Results

This section presents the results on the ML-based polarization tagging of vector bosons in the leptonic decay channel. Neural networks are trained to predict the event-wise polarization fraction of the intermediate MVB from the four momenta of FS particles which are also accessible in experiments. Two NN architectures are applied and compared: a MLP, described in section 3.1 and a Transformer, described in section 3.2. The MVB polarization is obtained from the simulation of proton-proton collisions with the MC generator Sherpa.

Events generated for this thesis are observed to have up to 18 jets when the PS is included in the event generation. The MLP can only be provided with a fixed input size which allows a training on a fixed number of FS particles, exclusively. In contrast to the MLP, the Transformer accepts a varying input length of the events. The Transformer is trained in two different ways: once on the minimal number of FS particles (as the MLP) and once on all FS particles which additionally include all jets resulting from the PS. Latter is only possible for samples that include parton showering.

The optimization of the model hyperparameters is performed with the framework OPTIMA (described in section 3.3). This allows to find the best possible NN for the sample it is optimized on. The observed differences in predictive performance can then be related to the systematic ability to extract the polarization fraction of the event from the kinematics of the FS particles. Thus, the specific NN architecture of the tagger differs for each sample..

To investigate the polarization tagger features systematically, the tagger is applied to several samples of $Z+jets$, $W+jets$ and $WZjj$ -EW production processes. Each of these samples is assigned a label which is used throughout this thesis. An overview of all generated samples with their labels and characteristics is presented in Table B.1 together with Sherpa runcards of the event generation exemplary for each process at LO+PS in chapter B.

5.1 Evaluation of the NN performance

This section presents how the performance of the neural network-based polarization tagger is evaluated. During the training, the NN parameters are adjusted so that the predicted polarization fraction ("prediction") for each event is as close as possible to the true polarization fraction calculated by Sherpa ("truth"). This is achieved by minimizing the loss on the training dataset in terms of the mean squared error MSE (see section 3.1). With increasing model

capacity (number of model parameters), the risk of overfitting increases. In this case, the model starts to memorize the true polarization fraction of an event. To have an unbiased estimate of the generalization error, only the NN predictions on the validation dataset are used to assess the model performance. Since the events of the validation dataset are not included in the training dataset, they provide an estimation how well the model generalizes to new data.

To now quantify the performance of the neural network, the deviation of NN prediction and true fraction is calculated on the events of the validation dataset. For the single boson production with two polarization fractions, MSE_{val} of r_L and r_T can be compared with the $MSE_{val,L}$, where only prediction and truth of r_L are included. With the number of events in the validation dataset N_{val} , the predicted polarization fraction value $r_{i,L}$ and $r_{i,T}$ as well as the true value polarization fraction $\hat{r}_{i,L}$ and $\hat{r}_{i,T}$ for the event i in the validation dataset, they are defined as

$$MSE_{val} = \frac{1}{2N_{val}} \sum_{i=1}^{N_{val}} ((r_{i,L} - \hat{r}_{i,L}) + (r_{i,T} - \hat{r}_{i,T}))^2 \quad (5.1)$$

and

$$MSE_{val,L} = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} (r_{i,L} - \hat{r}_{i,L})^2. \quad (5.2)$$

For the diboson production with four polarization fractions, the MSE_{val} is calculated for r_{LL} , r_{LT} , r_{TL} and r_{TT} whereas $MSE_{val,LL}$ only includes the double-longitudinal polarization fraction r_{LL} , such that

$$MSE_{val} = \frac{1}{4N_{val}} \sum_{i=1}^{N_{val}} ((r_{i,LL} - \hat{r}_{i,LL}) + (r_{i,LT} - \hat{r}_{i,LT}) + (r_{i,TL} - \hat{r}_{i,TL}) + (r_{i,TT} - \hat{r}_{i,TT}))^2 \quad (5.3)$$

and

$$MSE_{val,LL} = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} (r_{i,LL} - \hat{r}_{i,LL})^2. \quad (5.4)$$

Besides the quantitative evaluation of the predictive NN performance, two different visualizations are chosen to compare the NN predictions on the validation dataset with the true polarization fractions provided by Sherpa. In the following, both visualizations will be described in detail to facilitate the understanding of the subsequent sections. Exemplary, the results of the MLP trained and applied on $pp \rightarrow Z + j + X$ @ LO+PS are presented.

The first visualization is the comparison of the predicted and true distribution of the polarization fractions, as shown in Figure 5.1. Here, the number of events is plotted as a function of the polarization fraction, once for the true value calculated by the MC generator and once for the predicted value of the NN. In this case, each depicted distribution includes all events since a predicted and true value for r_L as well as r_T is assigned to each event. The ratio plots compare the predicted vs. true number of events in each of the polarization fraction value bins. By this, the difference in the number of events for a certain range of the polarization fraction

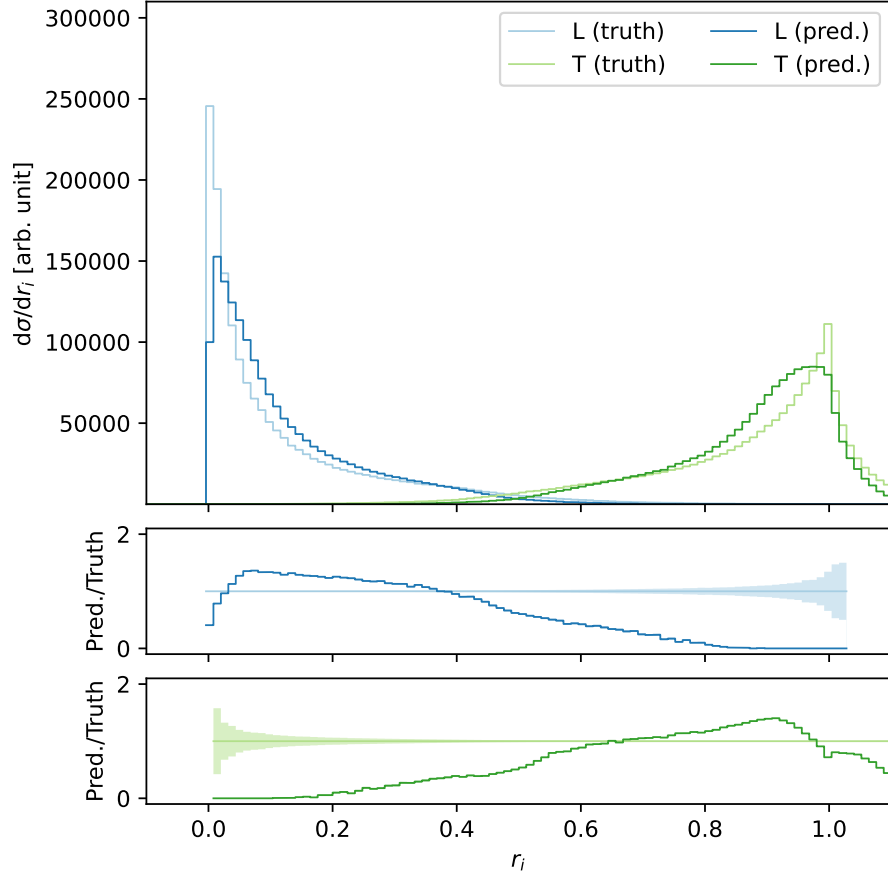


Figure 5.1: Distributions of the polarization fractions r_L (blue) and r_T (green) for NN predictions (pred.) and MC true value (truth) with absolute number of events in the validation dataset (top), relative difference of predicted and true distribution for r_L (middle) and relative difference of predicted and true distribution r_T (bottom).

value can be examined. It is possible to deduct in Figure 5.1 that the NN overestimates the number of events at $r_L = 0.1$ and underestimates the number of events at $r_T = 0.9$. From this, it is not possible to draw conclusion on the over- or underestimation of the polarization fraction value itself, e.g. on the predicted value of events with $r_{L,truth} = 0.3$.

The latter can be investigated in the second visualization which is used throughout this thesis. Here, the relative difference between the predicted and true value of the polarization fraction is calculated for each event. The distribution of this relative prediction error $RPE = \frac{r_{pred.}}{r_{truth}}$ is then plotted as a function of the true polarization fraction value. This results in a two-dimensional histogram as shown in Figure 5.2a, where the number of events with RPE is depicted depending on $r_{L,truth}$, the true value of the longitudinal polarization fraction of the validation dataset. From the distribution of the polarization fractions shown in Figure 5.1, it is already evident that the number of events decreases drastically for higher values of $r_{L,truth}$. Because of this, the RPE distribution is hardly observable in the unnormalized 2D-histogram in Figure 5.2a for higher $r_{L,truth}$. Therefore, a normalization step is performed where the number of events in each $r_{L,truth} \times RSE$ bin is divided by the overall number of events in

the corresponding $r_{L,truth}$ bin. The result is shown in Figure 5.2b. The RPE distribution is observable also for higher $r_{L,truth}$ values but information on the total number of events between different $r_{L,truth}$ bins is lost. For different vertical $r_{L,truth}$ bins, the same colour does not correspond to the same number of events in these histograms. An ideal model with $r_{L,pred} = r_{L,truth}$ would show $RPE = 1$ for all values of $r_{L,truth}$. Therefore, a smaller vertical distance between dark green regions (maximal number of events) and the horizontal black line at $RPE = 1$ corresponds to a better predictive performance of the polarization tagger. For the diboson production process, the RPE of the double-longitudinal polarization fraction r_{LL} in the validation dataset is depicted.

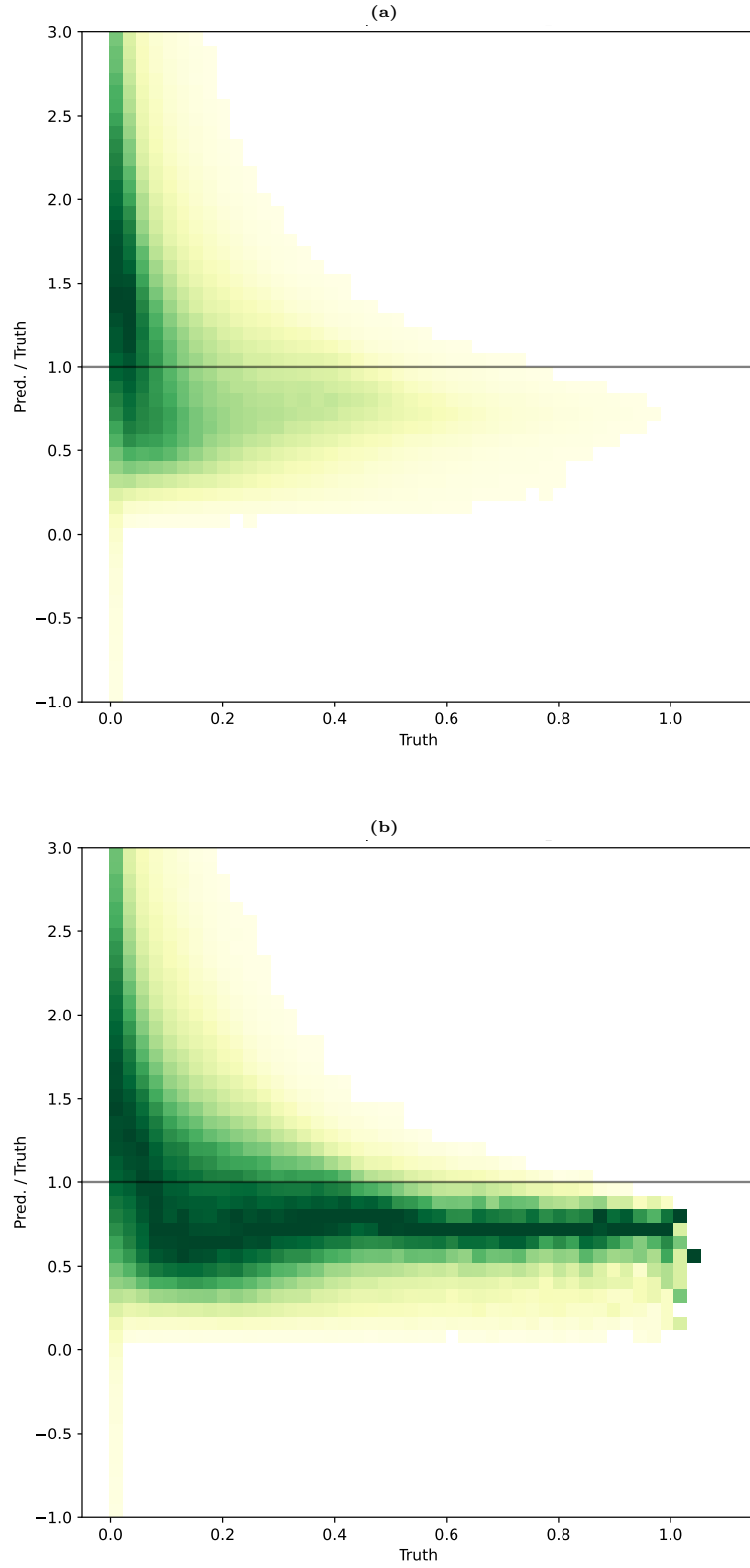


Figure 5.2: Distribution of the relative prediction error $RPE = \frac{r_{L,pred.}}{r_{L,truth}}$ (y-axis) of NN predictions $r_{L,pred.}$ and MC truth $r_{L,truth}$ of the longitudinal polarization fraction r_L for the validation dataset in dependence of the MC true value $r_{L,truth}$ (x-axis) (a). This is then plotted normalized by the number of events in the $r_{L,truth}$ bin (b).

5.2 $Z+jets$

In this section, the polarization extraction for the single Z boson production in association with at least one jet is investigated at fixed leading-order (LO), leading-order matched to parton shower (LO+PS) and next-to-leading order in QCD matched to parton-shower (NLO+PS). The Z boson decays leptonically into two charged muons μ^+ and μ^- . The MLP and the Transformer are trained to predict the polarization fractions r_L and r_T of the intermediate Z boson from the FS kinematics. The tagger is applied to samples with increasing complexity of the included events. In this thesis, $Z+jets$ production serves as the first test bed to identify general features of the polarization tagger that are independent of the observed process. Therefore, more samples are investigated for $Z+jets$ than for the following processes.

The results of the $Z+jets$ samples at LO, LO+PS and NLO+PS for both MLP and Transformer are shown in Figure 5.3-5.18. The computed loss on the validation dataset in terms of the MSE_{val} and $MSE_{val,L}$ is shown in Table 5.1 and Table 5.2 for the MLP and the Transformer.

5.2.1 Leading order results

Starting with fixed-order LO samples allows to construct cases in which there are no ambiguities in the polarization extraction, meaning where the matrix element depends only on the final-state momenta that can be unambiguously assigned to the particles. As presented in the following, the network is, in principle, capable of learning the correct relationship between FS kinematics and MVB polarization fraction.

a) Zj_LO_noPDF

The Zj_LO_noPDF sample contains a prediction for $dg \rightarrow Zd$ @ LO with fixed initial states and a beam energy of $E_{IS} = 75$ MeV. Thus, the IS particles and their kinematics the same for all events and the MVB polarization is independent of the IS.

The polarization fraction distributions of the MLP in Figure 5.3a and of the Transformer in Figure 5.4a show a high overlap and deviate only in regions of r_{truth} with an extremely low number of events. The number of events decreases drastically to higher values of r_L and lower values of r_T . The polarization fractions of the events show values of $r_L < 0.5$ and $r_T > 0.5$ in this sample, which is briefly discussed in the following section.

The RPE of the MLP in Figure 5.11a and of the Transformer in Figure 5.12a show a desired ratio of predicted and true r_L values close to $RPE = 1$ for $r_{L,truth} > 0.1$. For $r_{L,truth} < 0.1$, the MLP exhibits an under- and overestimation whereas the Transformer exhibits only an underestimation of the fraction. The scale here is chosen for better comparability with the following results.

The computed loss on the validation dataset in Table 5.1 is very low and mainly governed by statistics since mean value and error of the cross validation folds are of (almost) the same

order. The Transformer shows a slightly better MSE than the MLP but both are comparable. The results show that the MLP and the Transformer perform very well. The high ability of the tagger to extract the polarization fractions is explained by the unambiguous polarized matrix elements for fixed IS from which the polarization fractions are computed.

b) Zj_L0_s

The Zj_L0_s sample contains a prediction for the partonic channel $qg \rightarrow Zq$ @ LO. Since the PDF of the colliding protons is included here, the events now vary in the IS particles and their kinematics. The tagger only receives information on the FS particles and can only reconstruct the IS up to a commutation of the IS momenta.

The polarization fraction distributions of the MLP in Figure 5.3b and of the Transformer in Figure 5.4b show some difficulties of the tagger to predict the polarization fractions. Events are observable with $r_L > 0.5$ and $r_T < 0.5$, that did not exist in the Zj_L0_noPDF sample. The wider value spread of the polarization fraction is therefore related to the varying energy of the colliding partons and resulting variations on the IS in the events when including the PDF. With lower number of events, the statistical fluctuations increase drastically. Events with polarization fractions of $r_L \gg 0.5$ and $r_T \ll 0.5$ are much rarer, less of them are included in the training dataset and the prediction of these events is therefore harder for the tagger. MLP and Transformer show a systematic underestimation of the number of events as a clear constant negative off-set in the ratio plots for $r_L > 0.5$ and $r_T < 0.5$. The region of $r_L < 0.5$ and $r_T > 0.5$ shows a slight overestimation of the predicted number of events. Including the PDF leads to a sharpening of the peaks in the predicted and true distribution at $r_L \approx 0.0$ and $r_T \approx 1.0$. Especially the peak region $r_T \approx 1.0$ of the transverse fraction depicts a difficulty for the MLP as well as for the Transformer to predict the number of events correctly.

The RPE of the MLP in Figure 5.11b and of the Transformer in Figure 5.12b show that the spread of the RPE is much higher compared to the Zj_L0_noPDF sample. The majority of events have a $RPE < 1$ so that, for most events, the predicted value is lower than the true value for r_L , especially in the region of $0.1 < r_L < 0.7$. It is important to notice, that the described overestimation of the number of events for $r_L < 0.5$ (in Figure 5.3b and Figure 5.4b) does not contradict the underestimation of the polarization fractions for $0.1 < r_L < 0.7$ in the RPE . Events with $r_{L,pred.} < r_{L,truth}$ will be counted in bins with lower values of r_L than in the true distribution, which leads to an overestimation of the number of events there. Again, the statistical fluctuations increase for $r_L \gg 0.5$ due to the lower number of events in this region that are included in the training.

The loss on the validation dataset in Table 5.1 increased drastically compared to Zj_L0_noPDF and no difference in the predicted performance of MLP and Transformer is observed.

In summary, the varying IS kinematics introduce an ambiguity, such that the reconstruction of the MVB polarization and the prediction of polarization fractions is impaired.

c) Zj_L0_s_qposz

The phase-space used for Zj_L0_s does not contain restrictions on the initial state momenta, thus, it includes events with $p_{z,q}>0$ and $p_{z,q}<0$. An ambiguity under the commutation of the IS kinematics remains and the MVB polarization cannot be reconstructed completely. To resolve this ambiguity, a restriction on the phase space is introduced, such that only events with the IS quark in positive and the gluon in negative beam direction are considered. The Zj_L0_s_qposz sample contains a prediction for $qg \rightarrow Zq$ @ LO with $p_{z,q}>0$.

The polarization fraction distributions of the MLP in Figure 5.3c and of the Transformer in Figure 5.4c show a very good overlap. The under- and overestimation in the number of events across wide regions observed for the Zj_L0_s sample vanishes when excluding events with $p_{z,q}<0$. The statistical fluctuations in the prediction in regions with a low number of events still exist. The predicted polarization fractions in regions with high number of events show great improvement, especially in the peak region $r_T \approx 1.0$.

The RPE of the MLP in Figure 5.11c and of the Transformer in Figure 5.12c also exhibits an improvement of the predictions compared to the Zj_L0_s sample. The spread of the RPE is comparable with Zj_L0_noPDF where all IS and FS are fixed. The Transformer shows an even smaller spread of the predictions, which is also measurable quantitatively in the loss on the validation dataset in Table 5.1. The high predictive performance of MLP and Transformer lead to a MSE that is mainly governed by statistical fluctuations and are of the same order as the Zj_L0_noPDF results.

The restriction on the kinematics of the IS particles leads to an severe improvement of the tagger predictions. The observed deviation between NN prediction and true polarization fraction can be entirely attributed to the ambiguity caused by the IS kinematics.

d) Zj_L0

The Zj_L0 sample contains a prediction for $pp \rightarrow Zj$ @ LO, i.e. it includes predictions for all three partonic channels: $q\bar{q} \rightarrow Zg$, $qg \rightarrow Zq$ and $\bar{q}g \rightarrow Z\bar{q}$. The corresponding Feynman diagrams are shown in Figure 4.2a. Thus, in addition to the IS momenta, the missing knowledge of the partonic channel constitutes a new ambiguity.

The distributions of the predicted and true polarization fractions of the MLP in Figure 5.3d and of the Transformer in Figure 5.4d show substantial deviations of the predicted and true event number. The region of $r_L > 0.5$ and $r_T < 0.5$ exhibits a strong underestimation in the predicted distribution with a nearly constant off-set for $0.6 < r_L < 0.8$ and $0.2 < r_T < 0.4$ in the ratio plot. For $r_L < 0.5$ and $r_T > 0.5$, the overestimation of the number of events is observed with peaks at $r_L \approx 0.05$, $r_T \approx 0.9$ and $r_T \approx 1.0$.

The RPE of the MLP in Figure 5.11d and of the Transformer in Figure 5.12d show the majority of events with $RPE < 1$. Notice the change in the depicted range of the RPE . The RPE is much higher than for the samples with one partonic channel.

Quantitatively, the loss on the validation dataset in Table 5.1 is approximately twice as high as for the Zj_LO_s with only one channel and two to three magnitudes higher than for the Zj_LO_noPDF or $Zj_LO_s_qposz$ sample.

To explain the observed behaviour, one has to note that the polarization fractions of an event depend on the kinematics and types of the IS and FS particles. Including events from multiple partonic channels means that events can have similar FS kinematics but different initial and final state particle types and therefore different polarization fractions. In this case, the network receives similar input values but totally different corresponding target output values. Predicting polarization fractions closer to the mean value can then be encouraged during training. This is the case when the ambiguity in the events prevents the network from being able to reconstruct the true polarization fractions from the information available. If the network is not provided with the information necessary to distinguish between these events, it is convenient to predict closer to the mean of the output variable, which provides a training loss minimum.

This explains the observed over- and underestimation of the number of events in the polarization fractions distribution in Figure 5.3d and Figure 5.4d. Events with $r_L > 0.5$ and $r_T < 0.5$ are rare and $r_L \approx 0.0$ and $r_T \approx 1.0$ count the highest number of events. The mean polarization fraction value of r_L and r_T is dominated by values with a high number of events. The longitudinal polarization fraction is predicted closer to the average longitudinal fraction. Thus, the predicted number of events is underestimated for high r_L , since these events are now counted at lower values of r_L , and overestimated around $r_{L,mean} \approx 0.05$. The same holds true for the transverse polarization fraction, where the predicted number of events is underestimated for low r_T , since they are counted for higher values of r_T , and overestimated around $r_{T,mean} \approx 0.9$.

This explanation is supported by the results of the RPE in Figure 5.11d and Figure 5.12d. The underestimated value of the polarization fraction in the majority of events for $r > 0.05$ leads to an overestimation of the number of events for lower values. The substantial number of overestimated values ($RPE < 1$) for $0 < r_L < 0.05$ contributes to the peak of the number of events at $r_L \approx 0.05$ in the ratio plots of the polarization fraction distributions in Figure 5.3d and Figure 5.4d.

Including all partonic channels of $pp \rightarrow Zj$ @ LO causes drastic difficulties in predicting the polarization fractions compared to samples that included only one $Z + j$ production channel. Events with the same FS kinematics might have different polarization fractions due to varying IS and FS particle types. The ambiguity in the events affects the reconstruction of the polarized matrix element negatively as long as no additional information on the IS and FS particles is provided to the network.

e) Zj_L0 with channel information

Following the thoughts on the Zj_L0 results, the ability of the network to extract the polarization fractions should increase when it is provided with additional information on the hard process of each event. One important information is the partonic channel, i.e. the initial and final state particle types. Two of the three FS particles of each event are the charged leptons μ^+ and μ^- that result from the leptonic decay of the intermediate Z boson. The events vary in the third FS particle, the jet type, which can be a quark, an anti-quark or a gluon. This is encoded in the particle identification (PID) in event generators. Knowledge about the jet type allows conclusions about the partonic channel of the event.

The new approach is to provide this information on the jet type ((anti-)quark or gluon) as additional input into the network (besides p^μ of μ^+ , μ^- and j). With this, the network is able to distinguish between events from different channels that are ambiguous otherwise. The network is trained on the same Zj_L0 sample as in the previous section that includes the full $pp \rightarrow Z + jet$ @ LO process. It is now also provided with the information whether the jet type in the event is an (anti-)quark or a gluon.

The polarization fraction distributions of the MLP in Figure 5.3e and of the Transformer in Figure 5.4e show an improvement on the predicted number of events compared to the Zj_L0 sample. The off-set of underestimation for $0.6 < r_L < 0.8$ and $0.2 < r_T < 0.4$ in the ratio plot is mitigated and the overestimation of number of events in $r_L < 0.5$ and $r_T < 0.5$ is observably reduced, especially for the former peak regions at $r_L \approx 0.05$, $r_T \approx 0.9$ and $r_T \approx 1.0$.

The RPE of the MLP in Figure 5.11e and of the Transformer in Figure 5.12e show a drastic improvement in the r_L predictions. The differences between predicted and true fractions are much smaller for most events and distributed much more symmetrically around $RPE = 1$, compared to the Zj_L0 sample without additional information on the jet. This improvement is also quantitatively measurable in the lower computed loss on the validation dataset in Table 5.1. The MSE is now closer to the loss on the Zj_L0_s sample which included only events with a quark as the jet type.

Improvements in the prediction of the polarization fractions can be explained by the resolved ambiguity in the events. The information on the jet particle allows the network to distinguish between events originating from different processes. This information is necessary to reconstruct the polarization fraction from the FS kinematics. The reduced over- and underestimation in the described regions show that the network relies less on the average polarization fraction. A general under- and overestimation in the described regions is nevertheless still observable. The network is provided with information whether the jet type is an (anti-)quark or gluon, but there is no distinction between anti-quark and quark, as this information is not experimentally available. The ambiguity in these events is therefore not resolved and the network cannot perform as good as for the Zj_L0_s sample which only includes events with a quark as jet type.

From the LO results of the $Z+jets$ production process presented so far, it can be concluded that deviations from perfect performance can be explained entirely by missing information. In the absence of missing information, the neural networks can learn the relationship perfectly, they are complex enough and can be trained stable to obtain the best possible solution. It is therefore reasonable to assume that, even in the presence of missing information, the neural network result represents the best possible solution. The greatest influence comes from uncertainties regarding the parton channel. Uncertainty regarding the IS momentum (reversal) is smaller but still relevant. Uncertainty regarding the parton channel can be resolved by additional jet type information, which would also be experimentally accessible (at least approximately) via quark-gluon taggers.

5.2.2 Parton shower effects

In the following, the effect of the parton shower is investigated. All samples studied so far only include the hard scattering of the two incoming jets, which leads to the radiation of the Z boson and the leading jet particle. Now, the PS approximation is applied to the event generation. Gluons (QCD shower) and photons (QED shower) can be radiated from the IS (initial state radiation, ISR) and FS (final state radiation, FSR) of the hard process. This radiation affects the kinematics of the IS and FS particles. If not stated otherwise, both QCD and QED shower evolution are included for initial and final state.

The MLP and the Transformer are trained on the four-momenta of the two charged decay leptons and the leading jet. As explained in the beginning of this chapter, it is now possible to also train the Transformer on the four-momenta of the two charged decay leptons and all jets in the events, including the hard process jet and all jets produced by the PS. In this case, the Transformer is referred to as "trained on all jets".

a) Zj_LOPS_s

The Zj_LOPS_s sample contains predictions for $qg \rightarrow Z + q + X$ @ LO+PS. Here, only one partonic channel is included. The polarization fraction distributions of the MLP in Figure 5.5a, the Transformer in Figure 5.6a and the Transformer trained on all jets in Figure 5.7a show a strong increase in the systematic deviations between true and predicted distribution. Now, the ability of the network to reconstruct the fraction from the available kinematics of the FS particles becomes limited. The network shows a much stronger underestimation behaviour compared to the Zj_LO_s sample without PS. This underestimation of the number of events is present in a larger range of $r_L > 0.4$ and $r_T < 0.6$. Together with the observable overestimation in the regions $0.01 < r_L < 0.4$ and $0.6 < r_T < 1.0$, the network is hardly able to reconstruct the polarization fraction distribution in any region. Also, the network performs poor in the peak regions at $r_L = 0$ and $r_T = 1$, despite the high number of events included in the training. Since

this is also observed for significantly smaller sample sizes, the observed strong deviations must be related to a fundamental inability of the tagger and is not related to the model training.

The RPE of the MLP in Figure 5.13a, the Transformer in Figure 5.14a and the Transformer trained on all jets in Figure 5.15a show a strong underestimation of the predicted values of r_L since the majority of events exhibit $RPE < 1$. This underestimation shows an almost constant distribution for a wide range of $r_L > 0.1$ in all networks. The loss on the validation dataset in Table 5.1 and Table 5.2 shows that, quantitatively, the Transformer trained on all jets performs slightly better, but the MSE of the networks is one order of magnitude larger compared to for the equivalent Zj_LO_s sample without PS.

The IS and FS radiation of gluons and photons has a huge effect on the particles kinematics in the hard process. Unlike before, the IS momenta no longer need to be parallel to the beam direction, since these are the momenta after the ISR, and only the IS momenta prior to ISR are parallel to the beam. That is, for both IS partons, only the hard process momenta plus the momenta of all particles emitted from that parton. The network would therefore have to correctly identify the one hard process jet, correctly assign all remaining jets to the initial state partons and final state jets. Only in this case, it would have the same degree of ambiguity as in the fixed-order case. The major new uncertainty is therefore which jet was emitted where in the process. If that is not clear, the IS momenta cannot be determined, as they now have more degrees of freedom than before. In that case, the network must average over all possible configurations for the given FS momenta, and predicts closer to the average polarisation fraction.

b) Zj_LOPS

The Zj_LOPS sample includes events with $pp \rightarrow Z + j + X @ LO+PS$. Thus, all partonic channels are included here.

The polarization fraction distributions of the MLP in Figure 5.5b, the Transformer in Figure 5.6b and the Transformer trained on all jets in Figure 5.7b show a very similar effect of the PS as described in the previous section. Compared to the Zj_LO results, the underestimation in the number of events increases drastically for $r_L > 0.4$ and $r_T < 0.6$, as well as the overestimation in $r_L < 0.4$ and $r_T > 0.6$. Also, the mismodelling of the peak region at $r_L = 0$ and $r_T = 1$ is stronger. For the Transformer trained on all jets, the described effects are slightly less strong.

The RPE of the MLP in Figure 5.13b, the Transformer in Figure 5.14b and the Transformer trained on all jets in Figure 5.15b show a much wider spread in the region of $r_L < 0.5$ compared to the Zj_LOPS_s sample with only one channel. Thus, a stronger under- and overestimation in the r_L value is observable in this region. The loss on the validation dataset in Table 5.1 and Table 5.2 is slightly smaller for the Transformer trained on all jets. MLP and Transformer trained on the leading jet show the same predictive performance. It is surprising that the MSE

is slightly lower compared to the result of the `Zj_LOPS_s` sample. It would be easier for the network to predict polarization fractions of all partonic channels compared to the `Zj_LOPS_s` sample only if the predictive performance varies significantly between the channels. Based on the investigation of all partonic channels separately, this was not found to be the case. Therefore, it is concluded that the effect of the PS overshadows the expected loss increase from `Zj_LOPS_s` to `Zj_LOPS` as observed in `Zj_LO_s` and `Zj_LO` without PS.

Including the PS in the event generation leads to the radiation of interaction particles from the IS and FS and, thus, a variation in the IS particle energies. Since the IS kinematics of the event are unknown experimentally, they cannot be provided to the network. This ambiguity affects the estimation of the polarized matrix element. The ability of the network to reconstruct the fraction from the available kinematics of the FS particles becomes limited.

The different aspects of the parton showering are investigated in the following. By this, it is possible to further understand the impact of the PS in the event generation and the reasons for the NN difficulties in the polarization fraction extraction.

c) `Zj_LOPS_noQED`

The `Zj_LOPS_noQED` sample includes events with $pp \rightarrow Z + j + X$ @ LO+PS(QCD). Here, only the QCD radiation is included, whereas the QED shower is excluded. By this, the effect of QCD radiation can be compared to the effect of the full shower in the `Zj_LOPS` sample. The kinematic of the leptons is crucial for reconstructing the polarization fraction. If photons are emitted by the leptons, this could alter their kinematics. Although lepton dressing is used to cluster photons close to the leptons back onto the leptons and thus reconstruct the original momentum, it is unclear whether this is sufficient.

The polarization fraction distributions of the MLP in Figure 5.5c, the Transformer in Figure 5.6c and the Transformer trained on all jets in Figure 5.7c show no major improvements in the predicted number of events compared to the `Zj_LOPS` sample, where the QED shower is included. The same holds true for the *RPE* of the MLP in Figure 5.13c, the Transformer in Figure 5.14c and the Transformer trained on all jets as shown in Figure 5.15c. The loss on the validation dataset in Table 5.1 and Table 5.2 measures only a very small improvement.

Since the networks show very comparable results for the `Zj_LOPS` and the `Zj_LOPS_noQED` sample, the effect of the QED shower is negligible. The only difference between the `Zj_LO` and the `Zj_LOPS_noQED` sample is the QCD shower that allows the additional radiation of gluons by the IS and FS particles. Therefore, it can be concluded that the QCD shower causes the major difficulty for predicting the polarization fractions from the FS kinematics in samples with PS. Thus, either the QED showering does not significantly affect the kinematics of leptons, or lepton dressing is effective. From this, the question arises, how the IS and FS radiation affects the tagger results.

d) Zj_LOPS_noFSR

The Zj_LOPS_noFSR sample includes events with $pp \rightarrow Z + j + X$ @ LO+ISR. Gluons and photons are only radiated from the IS particles.

The polarization fraction distributions of the MLP in Figure 5.5d, the Transformer in Figure 5.6d and the Transformer trained on all jets in Figure 5.7d only show a slight improvement of the predicted number of events in the peak region of $r_T = 1$. The difference between the Zj_LOPS_noFSR and the Zj_LOPS sample are rather small.

The RPE of the MLP in Figure 5.13d, the Transformer in Figure 5.14d and the Transformer trained on all jets in Figure 5.15d show no observable difference compared to the results of Zj_LOPS where the FSR is included. The loss on the validation dataset in Table 5.1 and Table 5.2 measures a small improvement on the MSE .

This shows that the radiation originating from a hard-process jet does not introduce significant additional uncertainties (at least not significantly relative to ISR). Furthermore, the uncertainty as to whether a jet was emitted from the FS or the IS is apparently irrelevant. The conclusion is, that only the uncertainty regarding the IS momentum, arising from the lack of knowledge about which jet was emitted from which initial state parton, can be decisive.

e) Zj_LOPS_noISR

The Zj_LOPS_noISR sample includes events with $pp \rightarrow Z + j + X$ @ LO+FSR, such that gluons and photons are only radiated from the FS particles. This means that the IS particles are once again parallel to the beam, which simplifying the reconstruction of the IS up to a commutation of the IS (provided all FS particles can be used).

The polarization fraction distributions of the MLP in Figure 5.5e, the Transformer in Figure 5.6e and the Transformer trained on all jets in Figure 5.7e show a strong improvement compared to the Zj_LOPS results. The distributions are indeed very similar to the Zj_LO results without parton showering. The remaining discrepancy is therefore dominated by the uncertainty regarding the partonic channel and the IS momentum flip, not by the change in the jet momentum. This is interesting, as the MLP only receives a single jet as input, but it is yet able to reconstruct the hard process jet momentum accurately.

The RPE of the MLP in Figure 5.13e, the Transformer in Figure 5.14e and the Transformer trained on all jets in Figure 5.15e also show a much smaller distribution of RPE which is comparable to the Zj_LO results. Quantitatively, the loss on the validation dataset in Table 5.1 and Table 5.2 measures a small effect of the FS radiation and is one order of magnitude smaller than the measured effect of the IS radiation.

The results show that the exclusion of IS radiation in the PS leads to a drastic increase in the predictive power of the model. This can be explained by the fact that the the matrix element and the polarization fraction of the intermediate Z boson highly depend on the IS kinematics. The IS kinematics are affected strongly by the IS showering, where the IS particles radiate

gluons and quarks of the boson in the hard process. The uncertainty regarding the direction of the partons, and thus regarding the emission angle of the Z boson, might be the reason for the observed great difficulties for the polarization extraction.

f) Zj_LOPS_s_qposz_noISR

According to the results presented in the previous sections, the tagger has the highest predictive performance when (I) only one partonic channel is considered, (II) a restriction on the kinematics of the IS particles is applied and (III) only FS radiation of additional gluons is allowed. Thus, the NN performance on the Zj_LOPS_s_qposz_noISR sample is investigated, which contains a prediction for $qg \rightarrow Z + q + X$ @ LO+FSR with $p_{z,q} > 0$.

The polarization fraction distributions of the MLP in Figure 5.5f, the Transformer in Figure 5.6f and the Transformer trained on all jets in Figure 5.7 show good agreement of predicted and true distribution for r_L and r_T , which is comparable to the results of Zj_LO_s_qposz without PS. With the selection $p_{z,q} > 0$, the deviations between predicted and true distribution decrease compared to the Zj_LOPS_s sample. The underestimation of events for $r_L > 0.5$ and $r_T < 0.5$ is resolved. Now, stronger statistical fluctuations are observable in the regions $r_L < 0.2$ and $r_T > 0.8$ with a very low number of events

The RPE of the MLP in Figure 5.13f, the Transformer in Figure 5.14f and the Transformer trained on all jets in Figure 5.15f shows well the non-vanishing effect of the FSR, since the RPE exhibits a wider distribution than for the Zj_LO_s_qposz with no showering. The Transformer trained on all jets performs much better. This is to be expected, as it can fully reconstruct the hard process jet, whereas the other two models can only use one of the jets. The loss on the validation dataset in Table 5.1 and Table 5.2 is one order of magnitude smaller than the Zj_LOPS sample with all channels, no restriction on IS kinematics and included IS radiation. The FS radiation increases the MSE about one order of magnitude compared to the Zj_LO_s_qposz without PS.

5.2.3 Higher order effects

This section studies the effects of higher orders in QCD on the polarization fraction extraction. For that, two predictions with a higher perturbative accuracy were produced and are discussed in the following.

a) Zj_NLO_PS

The Zj_NLO_PS sample contains a prediction for $pp \rightarrow Z + j + X$ @ NLO(QCD)+PS. Due to real corrections, up to two jets are produced in the hard process, with the remaining jets originating from the PS.

The polarization fraction distributions of the MLP in Figure 5.8a, the Transformer in Figure 5.9a and the Transformer trained on all jets in Figure 5.10a show a change in the distribution of the true polarization fraction compared to the LO(+PS) samples. Including additional diagrams by real and virtual corrections seems to have a high impact in the calculation of the polarized matrix element and the polarization fraction. The number of events with higher values of r_L and lower values of r_T increases drastically. Even a plateau is observable for $0.2 < r_L < 0.4$ and $0.6 < r_T < 0.8$. The ratio of the true and predicted distribution increases and shows new characteristics compared to the Zj_LOPS sample. The peak in the event number overestimation shifts towards the region of $0.2 < r_L < 0.3$ and $0.7 < r_T < 0.8$, where a plateau is now observable. The peak regions of $r_L \approx 0$ and $r_T \approx 1$ show much worse results than in the samples without higher order corrections.

The RPE of the MLP in Figure 5.16a, the Transformer in Figure 5.17a and the Transformer trained on all jets in Figure 5.18a shows a wider distribution than in the Zj_LOPS sample. A high number of events with a true transverse polarization fraction $r_T > 1.0$ is observed. The loss on the validation dataset in Table 5.1 and Table 5.2 underlines the drastic decrease in predictive performance. The MSE is more than twice as high as for the Zj_LOPS sample. The Transformer trained on all jets performs slightly better than the MLP and the Transformer trained on the leading jet. Due to the calculation of the leading jet at NLO QCD, the additional jets included in the training carry information relevant for reconstructing the polarized matrix element which leads to greater ambiguity. One possible reason could be that the two hard process jets are more difficult to distinguish from the PS jets. This makes it harder, on the one hand, to decide which of the FS particle momenta is directly relevant, and, on the other hand, it complicates the reconstruction of the IS momenta from the remaining jets. This could be investigated further with additional intermediate steps, e.g. without ISR or FSR. Since information on the hard process is also encoded in the additional jets, the Transformer trained on all information available shows better results in predicting the polarization fraction of the intermediate Z boson. Now, it is not sufficient anymore to provide the network with only one jet from the hard process.

b) Zj_NLO_2j_LO_2j_PS

The Zj_NLO_2j_LO_2j_PS sample includes $pp \rightarrow Z+1-2j@NLO(QCD)+3-4j@LO+PS$. In the hard process, up to two jets are included at NLO in QCD and up to four jets at LO in the matrix element calculation. Additional jets are generated by the parton shower.

The polarization fraction distributions of the MLP in Figure 5.5, the Transformer in Figure 5.6 and the Transformer trained on all jets in Figure 5.7 show an even higher relative number of events with a high value of $r_{L,truth}$, as well as a low value of $r_{T,truth}$. Together with the Zj_NLO_PS sample, this behaviour could be related to the increasing complexity of the hard process. All three NNs show slightly smaller deviations between predicted and true distributions in the

region of overestimation, which is unexpected.

The *RPE* of the MLP in Figure 5.13, the Transformer in Figure 5.14 and the Transformer trained on all jets in Figure 5.15 are distributed very similar to the Zj_NLO_PS sample.

The loss on the validation dataset in Table 5.1 and Table 5.2 shows only small deviations from the Zj_NLO_PS sample. Here, the advantage of the Transformer, which can be trained on a varying number of FS particles, is even more evident. This is not surprising since important information on the matrix element of the process is included in all jets from the hard process. Thus, the second hard process jet leads to significant additional ambiguities, whereas further hard process jets do not introduce any significant additional uncertainties.

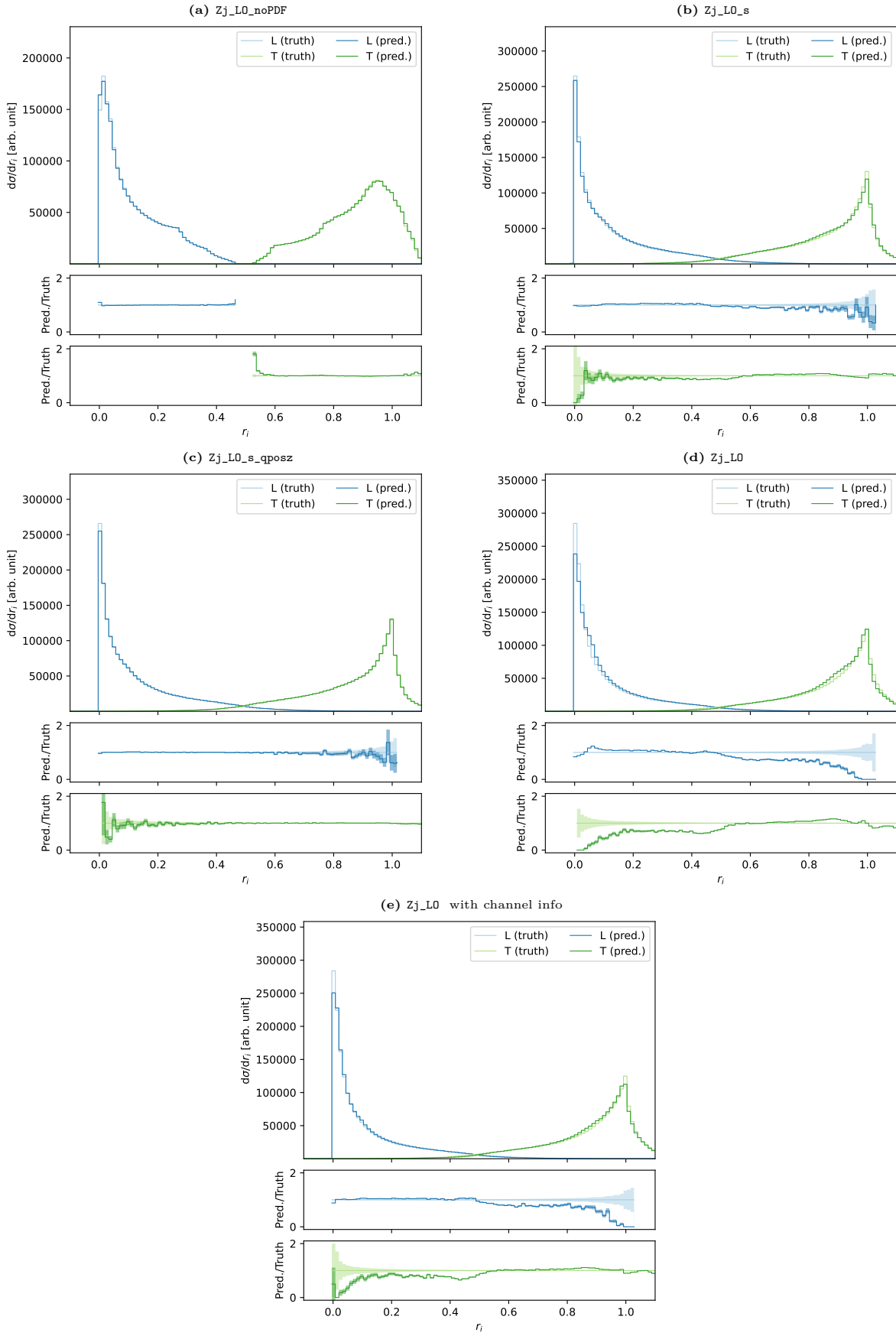


Figure 5.3: Predicted and true distribution of r_L and r_T of the MLP for $Z+jets$ @ LO.

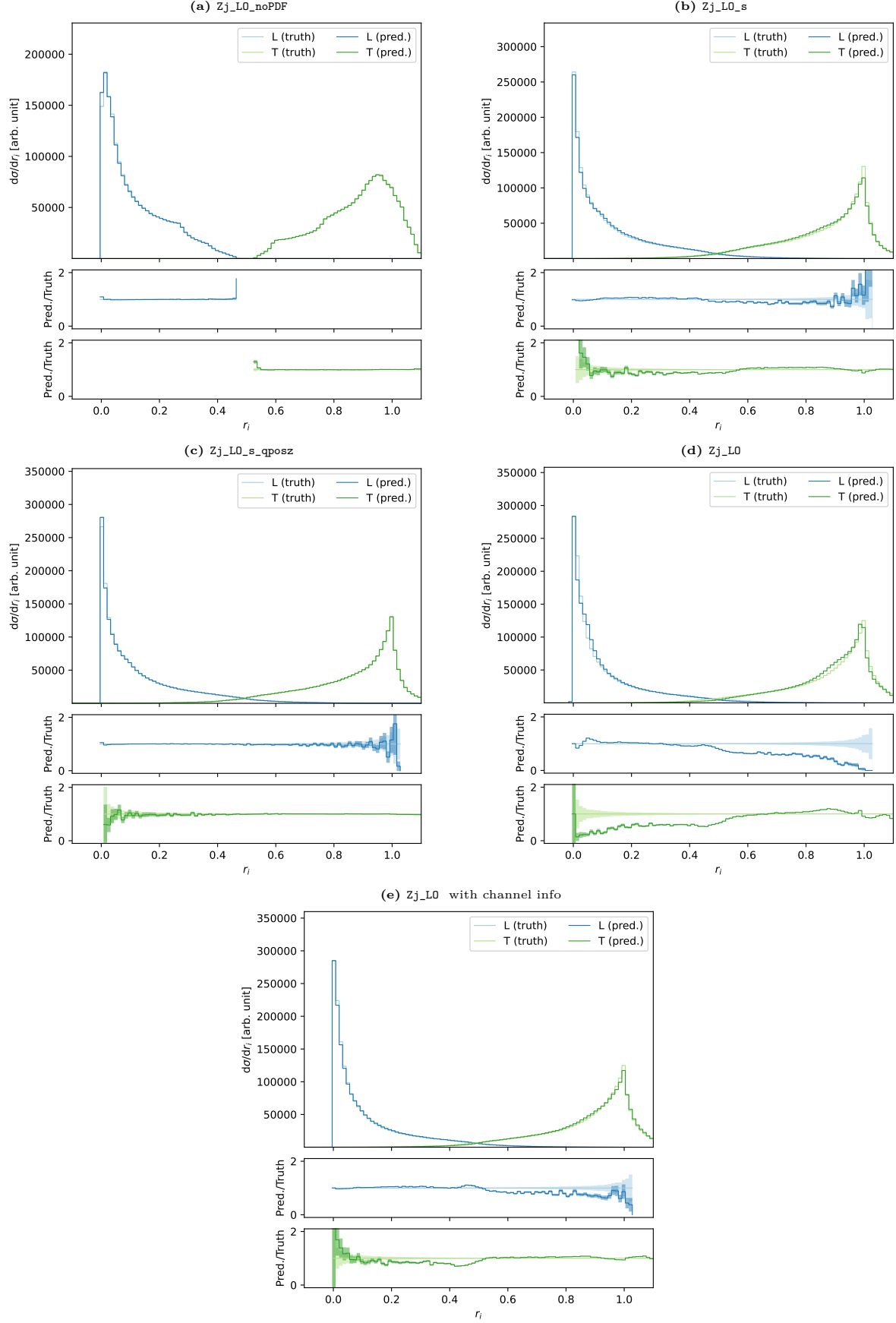


Figure 5.4: Predicted and true distribution of r_L and r_T of the Transformer for $Z+jets$ @ LO.

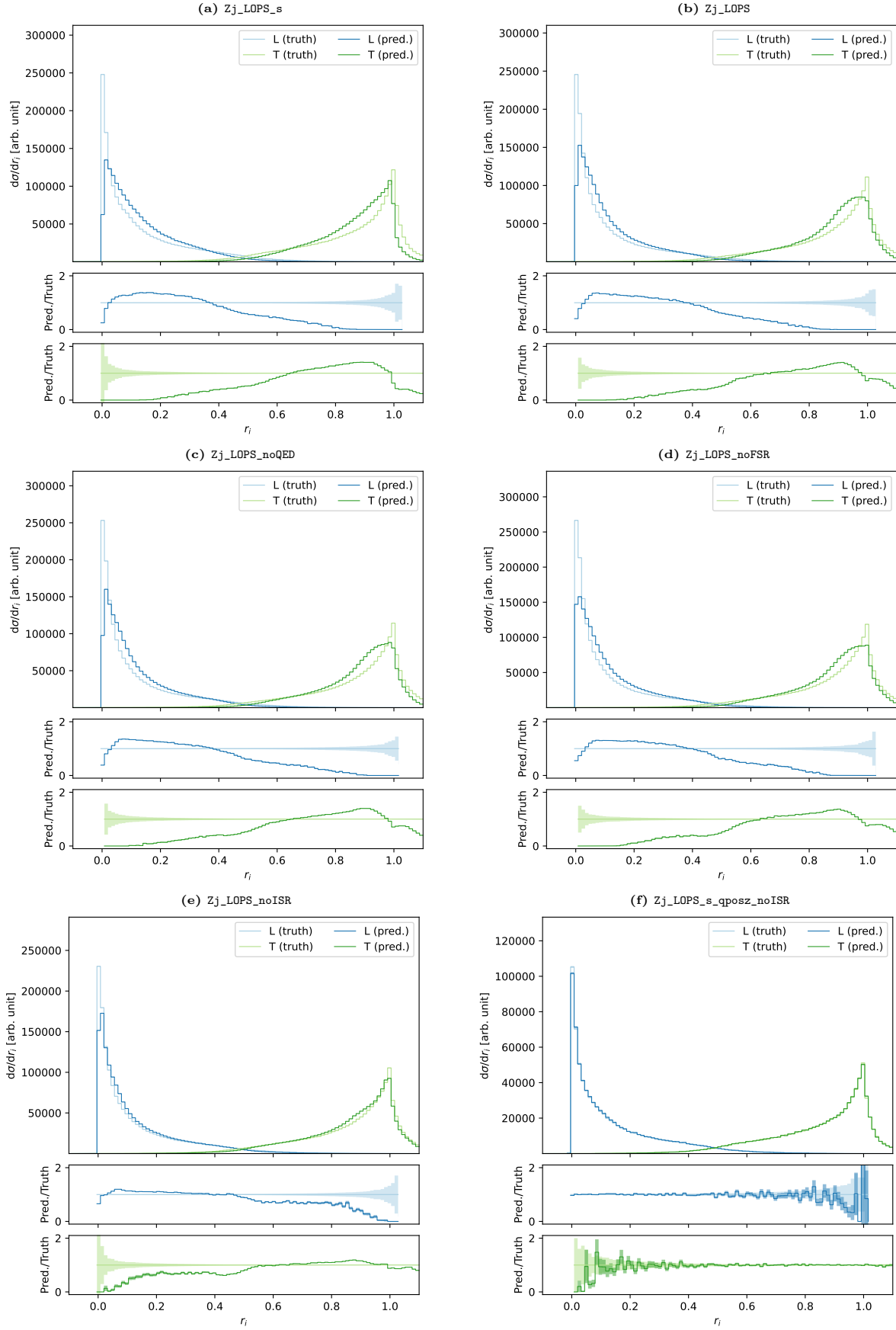


Figure 5.5: Predicted and true distribution of r_L and r_T of the MLP trained $Z+jets$ @ LO+PS.

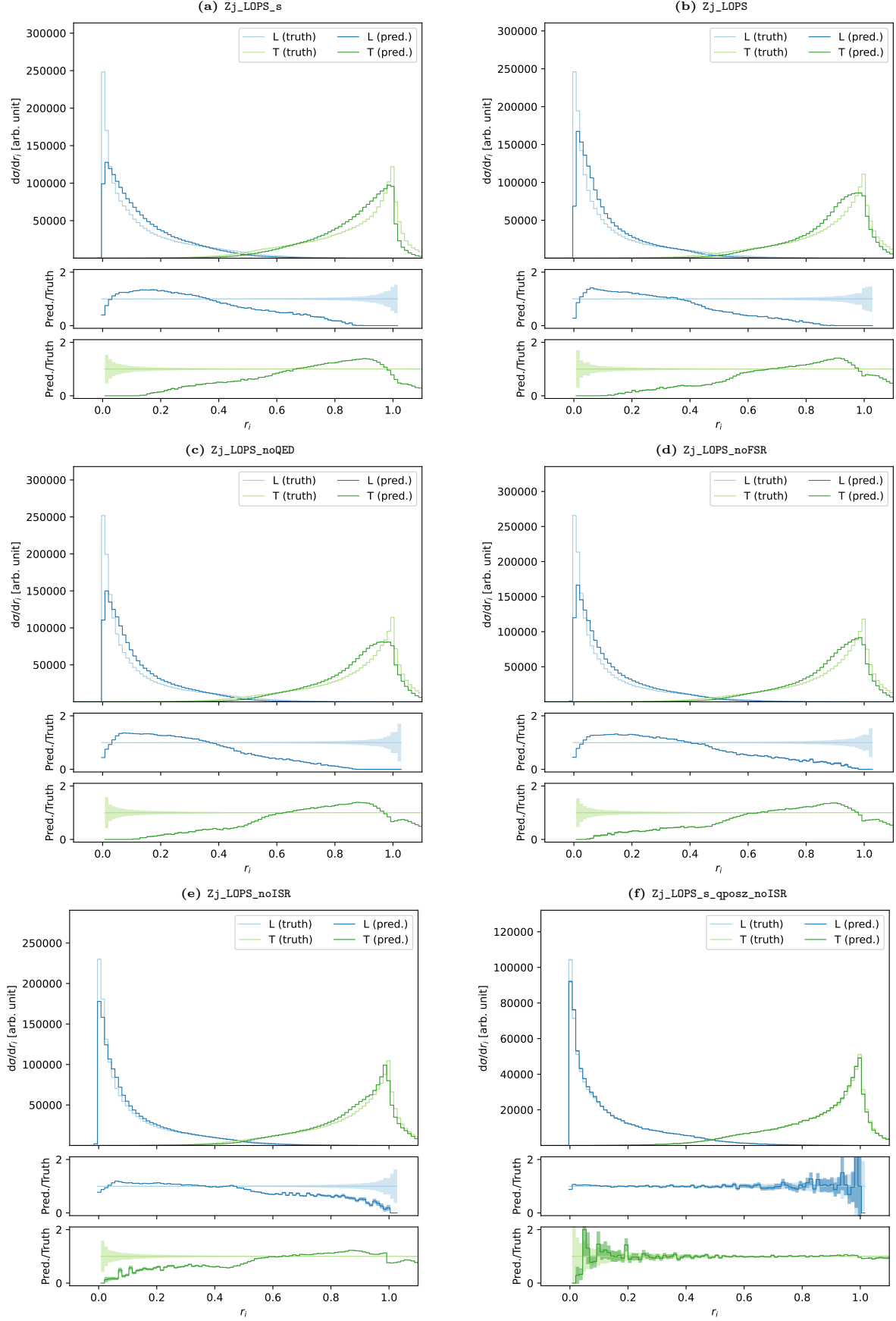


Figure 5.6: Predicted and true distribution of r_L and r_T of the Transformer for $Z+jets$ @ LO+PS.

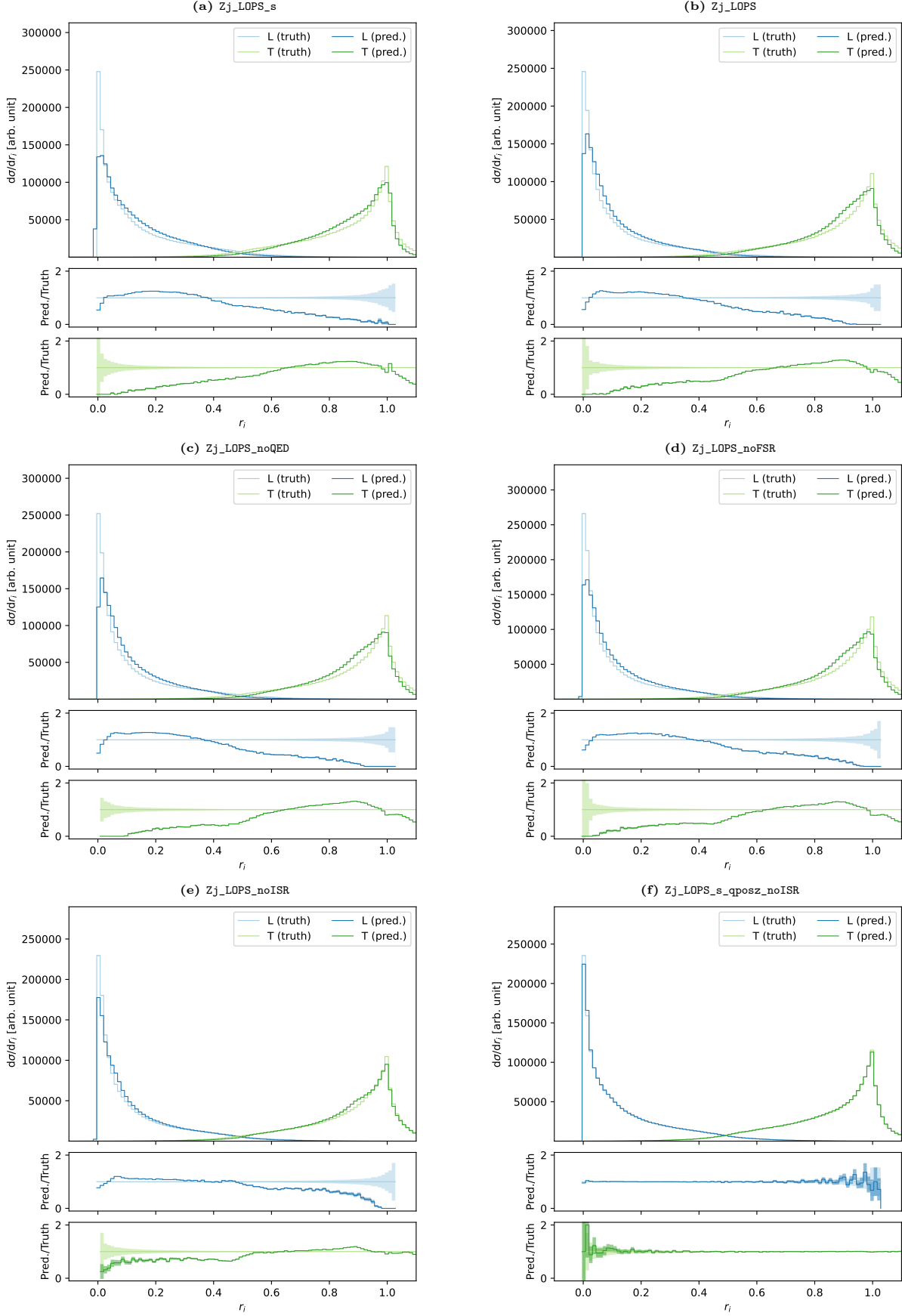


Figure 5.7: Predicted and true distribution of r_L and r_T of the Transformer trained all jets for $Z+jets$ @ LO+PS.

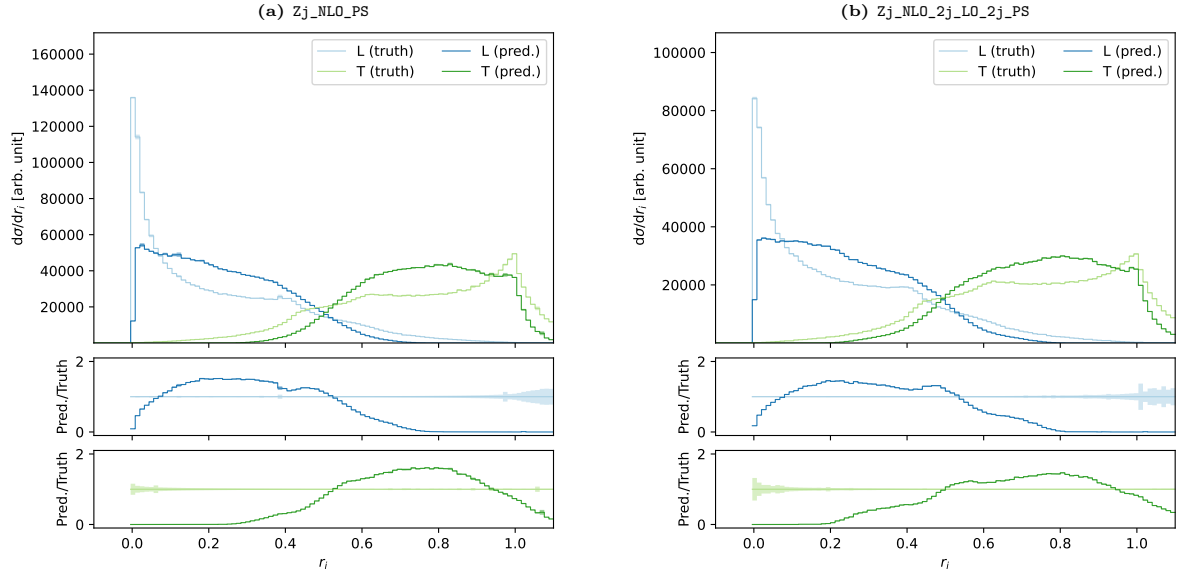


Figure 5.8: Predicted and true distribution of r_L and r_T of the MLP $Z+jets$ @ NLO.

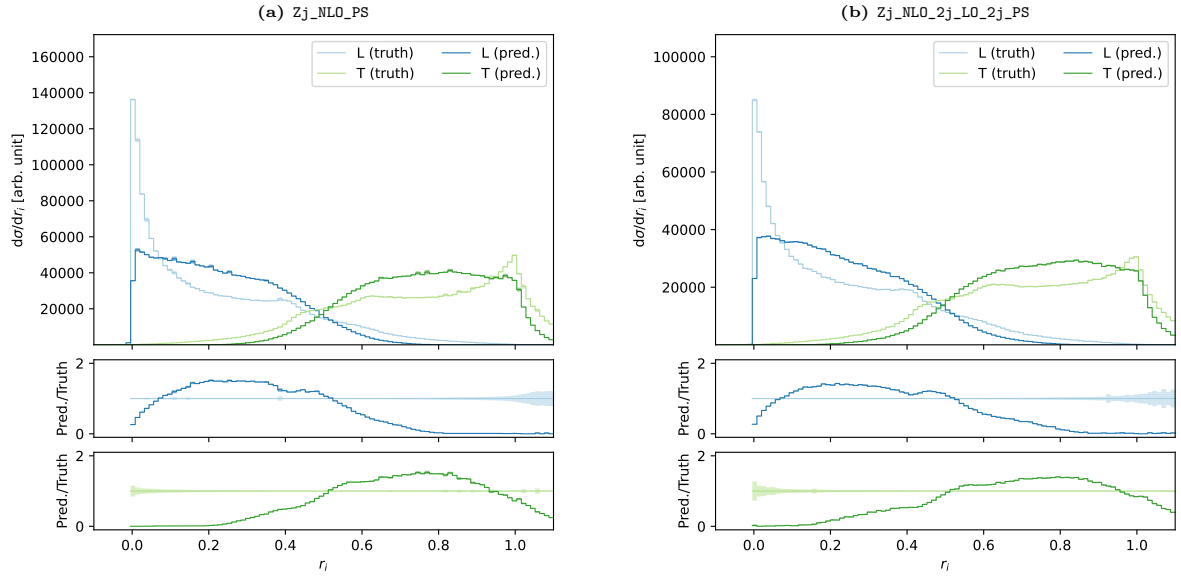


Figure 5.9: Predicted and true distribution of r_L and r_T of the Transformer for $Z+jets$ @ NLO.

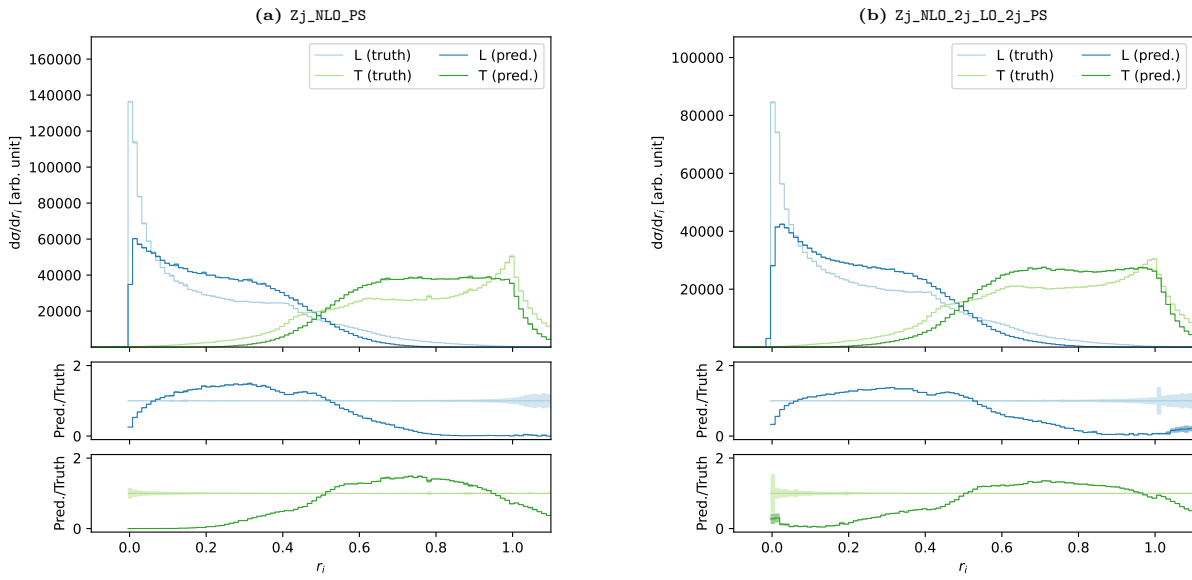


Figure 5.10: Predicted and true distribution of r_L and r_T of the Transformer trained on all jets for $Z+jets$ @ NLO.

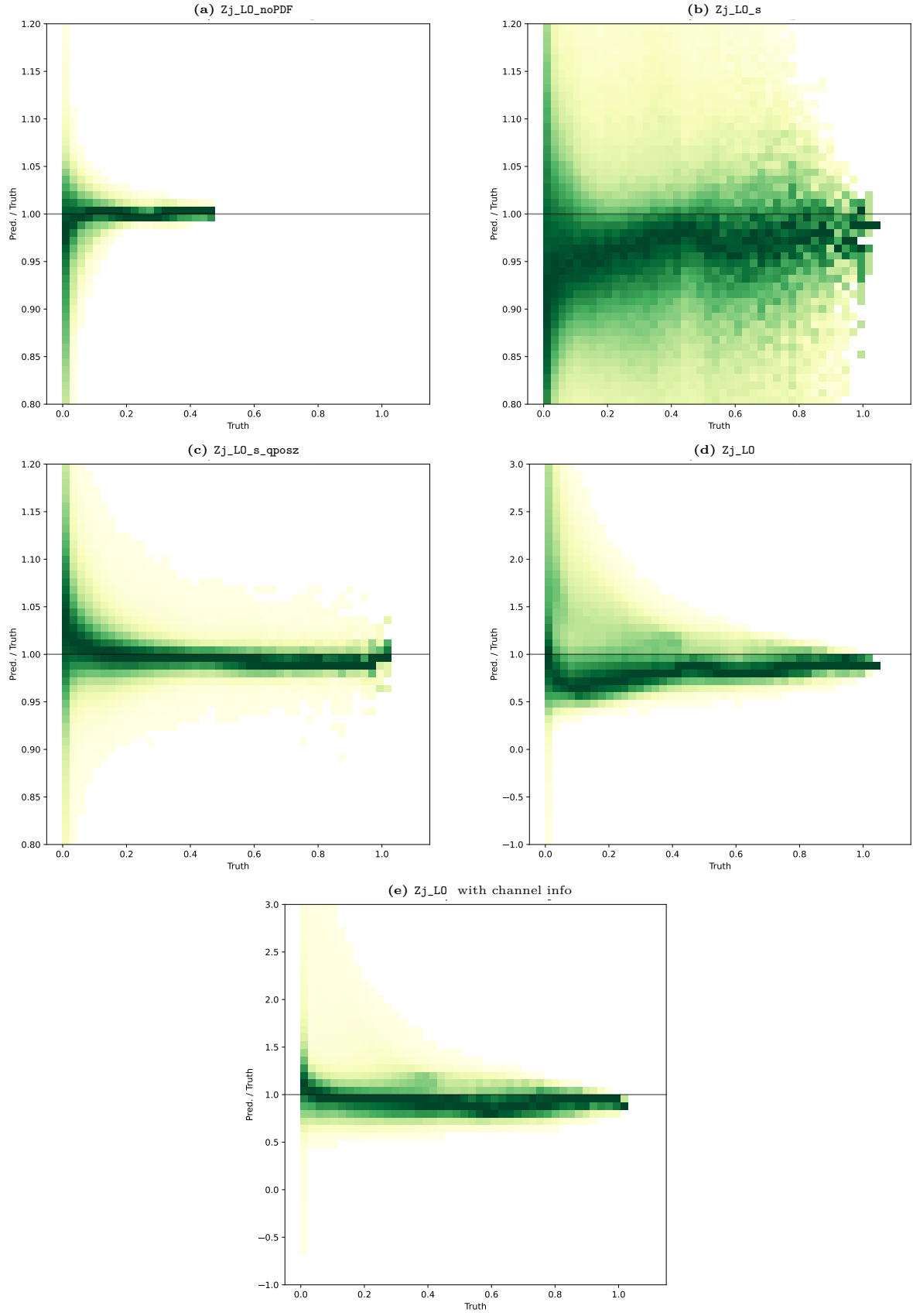


Figure 5.11: Relative prediction error of $r_{L,val}$ of the MLP for $Z+jets$ @ LO.

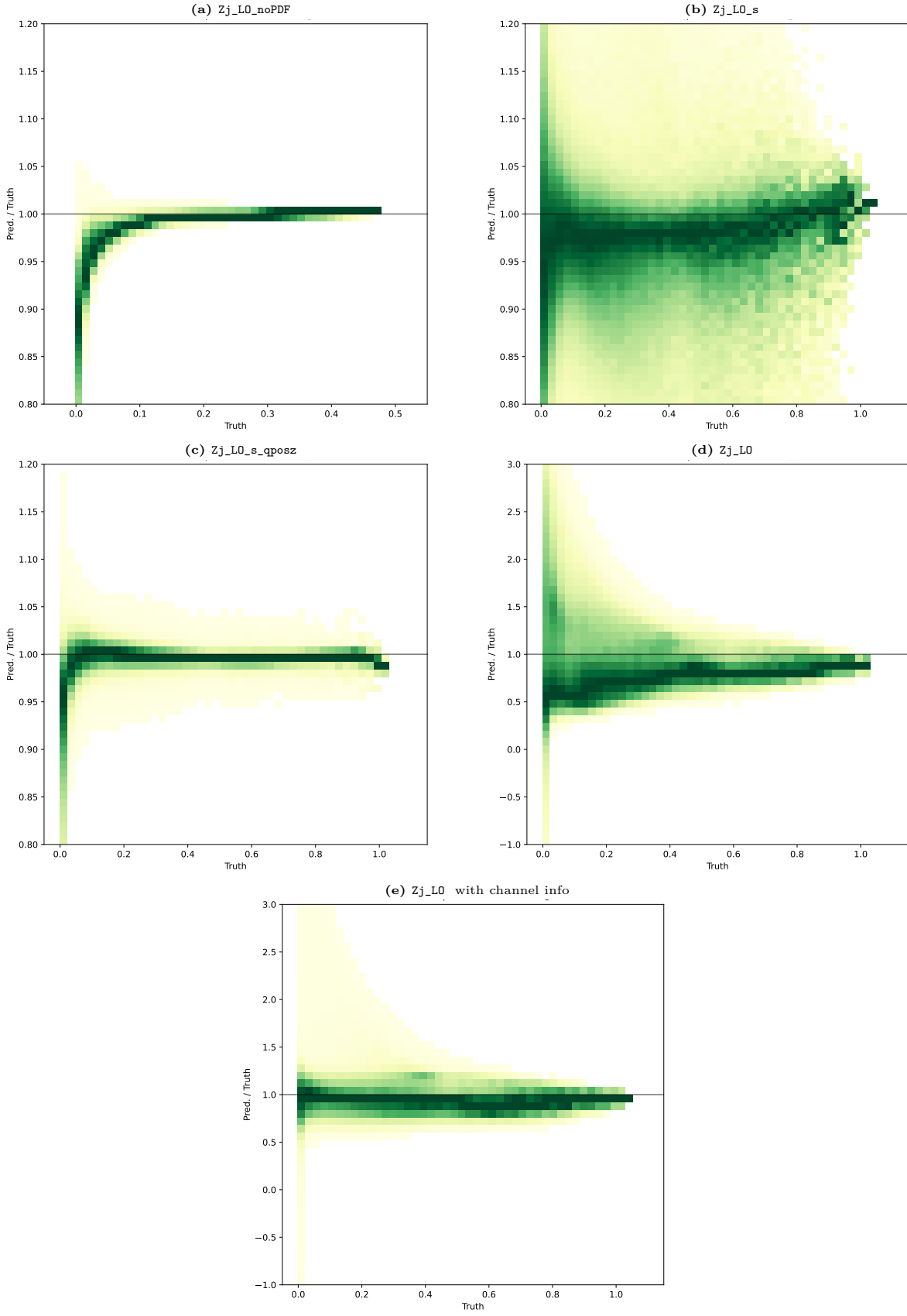


Figure 5.12: Relative prediction error of $r_{L,val}$ of the Transformer for $Z+jets$ @ LO.

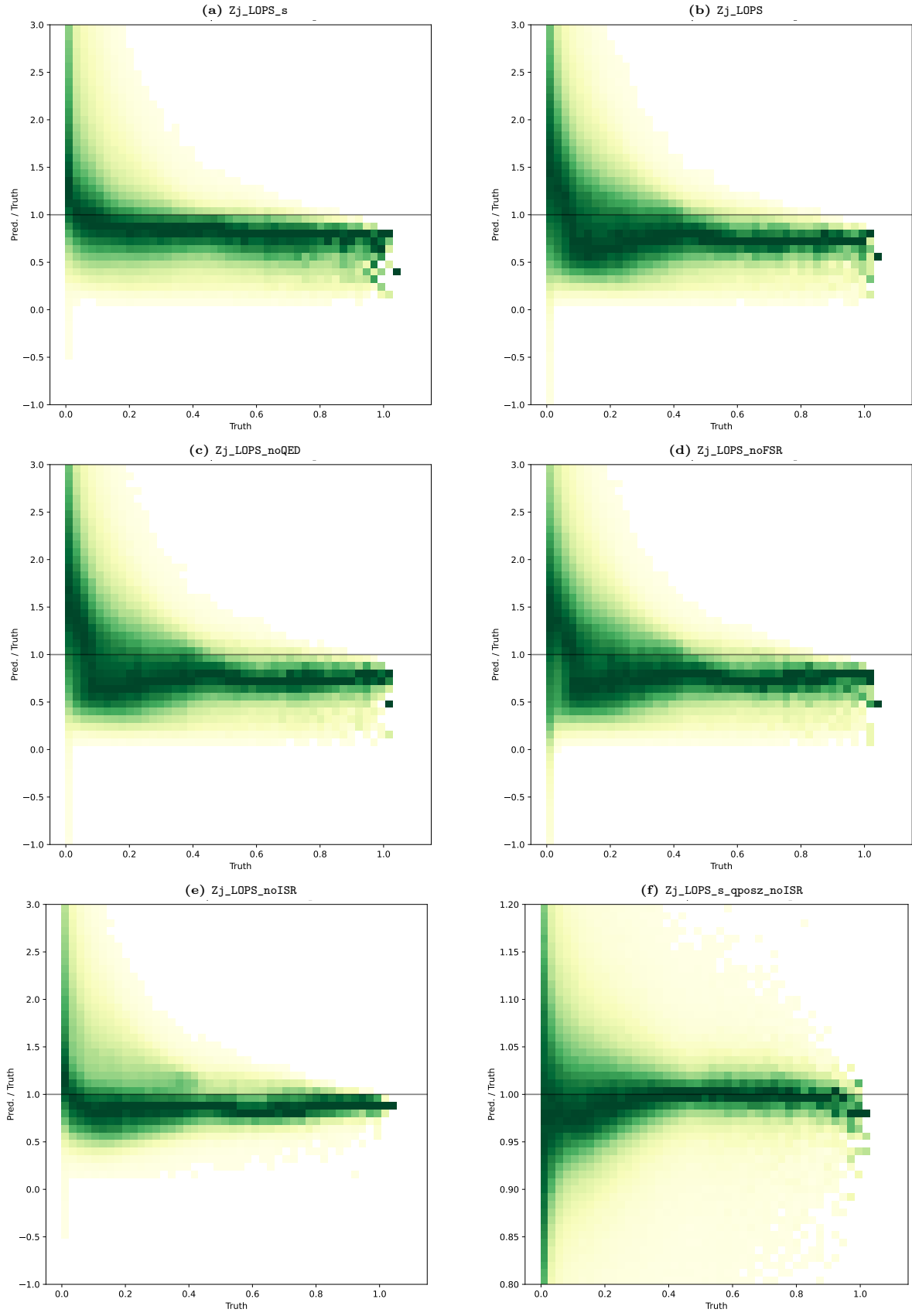


Figure 5.13: Relative prediction error of $r_{L,val}$ of the MLP trained $Z+jets$ @ LO+PS.

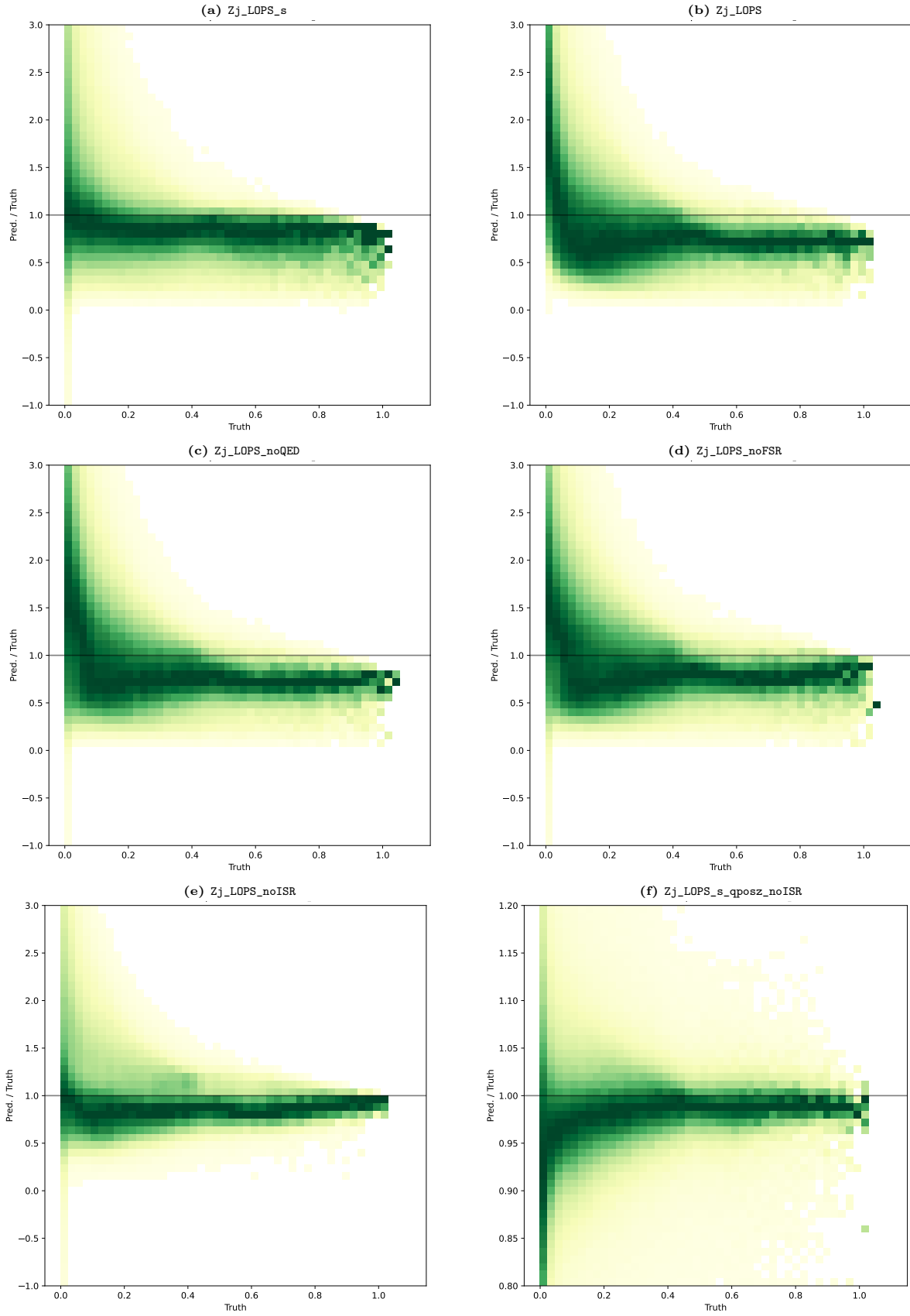


Figure 5.14: Relative prediction error of $r_{L,val}$ of the Transformer for $Z+jets$ @ LO+PS.

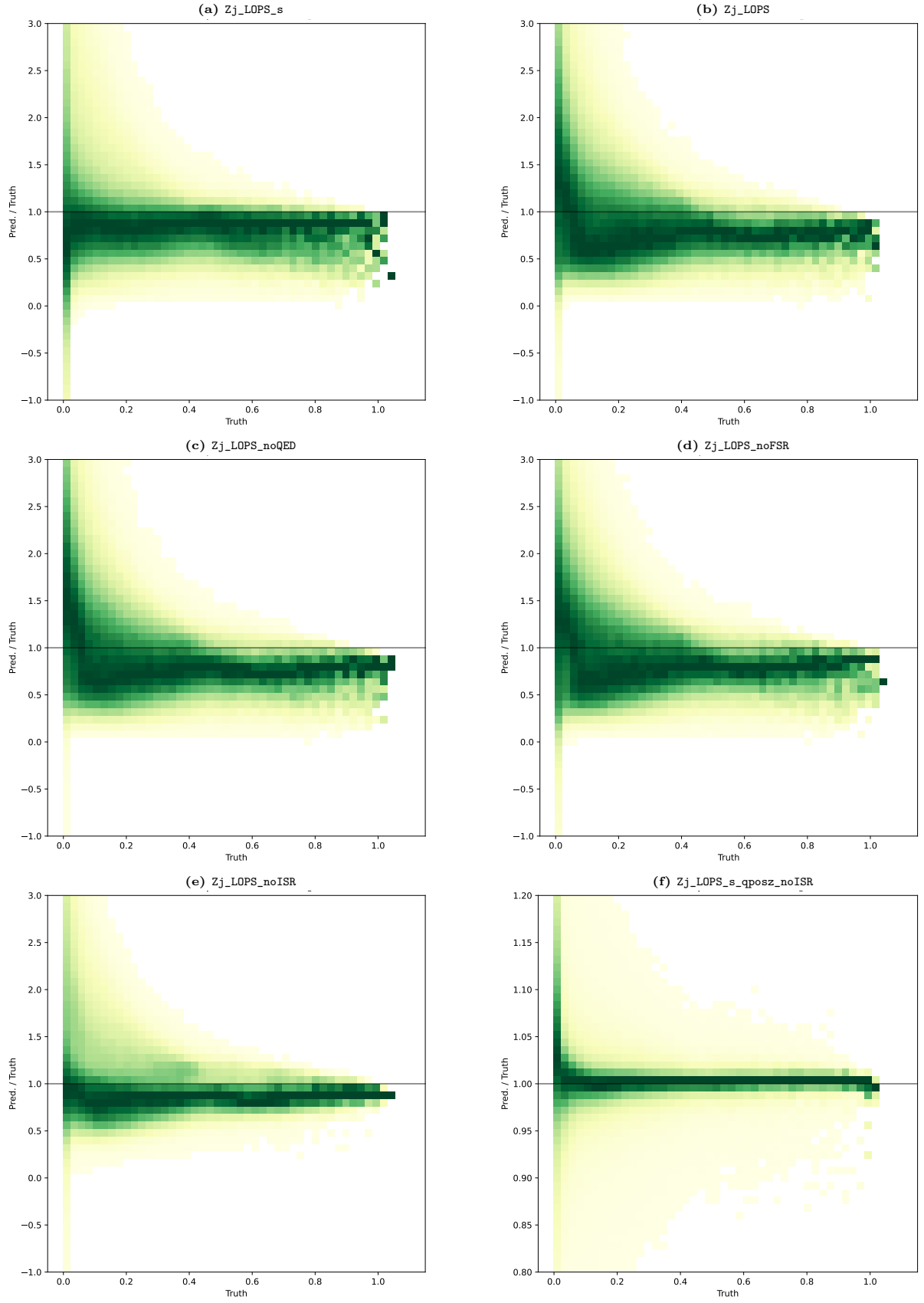


Figure 5.15: Relative prediction error of $r_{L,val}$ of the Transformer trained all jets for $Z+jets$ @ LO+PS.

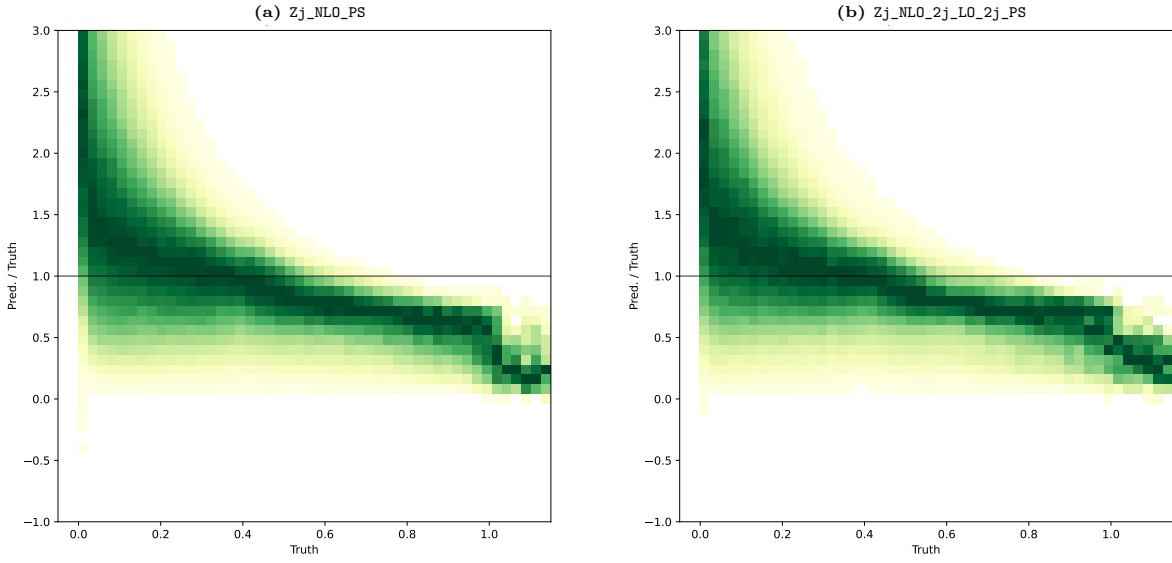


Figure 5.16: Relative prediction error of $r_{L,val}$ of the MLP $Z+jets$ @ NLO.

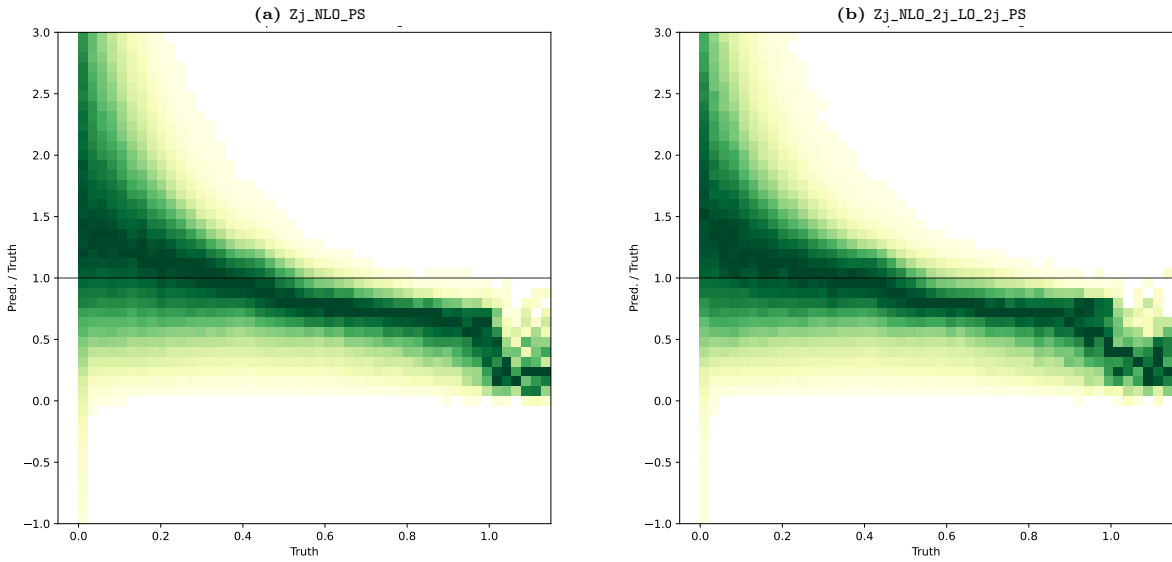


Figure 5.17: Relative prediction error of $r_{L,val}$ of the Transformer for $Z+jets$ @ NLO.

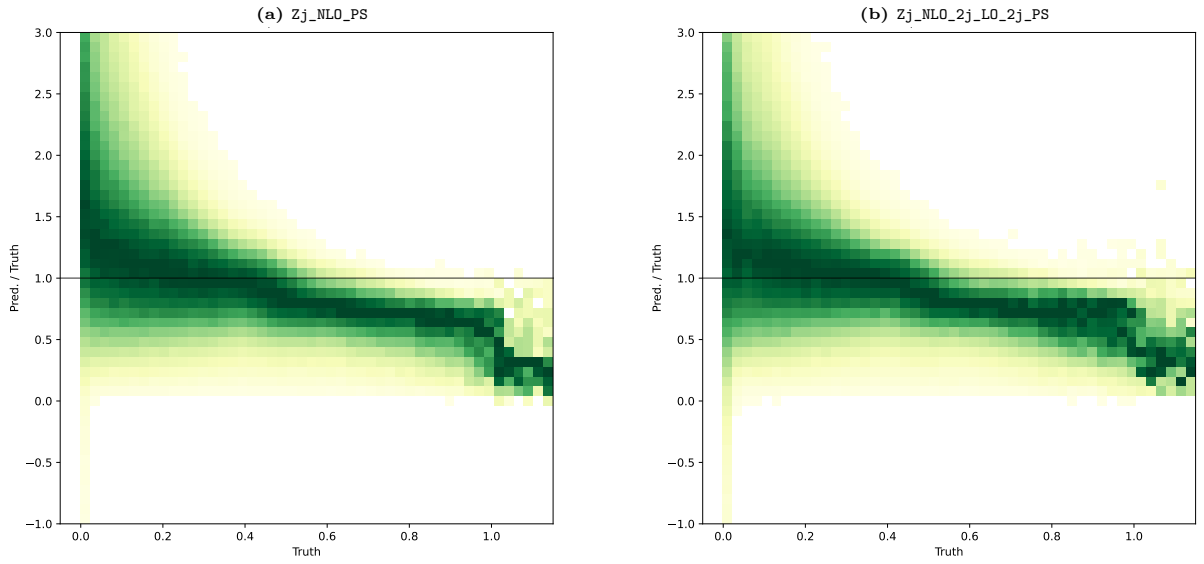


Figure 5.18: Relative prediction error of $r_{L,val}$ of the Transformer trained on all jets for $Z+jets$ @ NLO.

Table 5.1: Mean Squared Error of predicted and true polarization fractions of the validation dataset for the MLP and the Transformer for $Z+jets$, calculated for both longitudinal and transverse polarization fraction r_L and r_T (MSE_{val}) as well as for only the longitudinal polarization fraction r_L ($MSE_{val,L}$).

Sample	MLP		Transformer	
	MSE_{val}	$MSE_{val,L}$	MSE_{val}	$MSE_{val,L}$
Zj_LO_noPDF	0.000 002 8 $\pm 0.000 002 0$	0.000 001 75 $\pm 0.000 000 54$	0.000 001 03 $\pm 0.000 000 67$	0.000 000 58 $\pm 0.000 000 32$
Zj_LO_s	0.001 902 $\pm 0.000 022$	0.001 450 $\pm 0.000 023$	0.001 896 $\pm 0.000 017$	0.001 446 $\pm 0.000 022$
Zj_LO_s_qposz	0.000 038 $\pm 0.000 011$	0.000 032 9 $\pm 0.000 009 5$	0.000 006 4 $\pm 0.000 001 9$	0.000 005 7 $\pm 0.000 001 8$
Zj_LO	0.003 565 $\pm 0.000 055$	0.002 640 $\pm 0.000 045$	0.003 587 $\pm 0.000 047$	0.002 672 $\pm 0.000 045$
Zj_LO with channel info	0.002 164 6 $\pm 0.000 008 8$	0.001 457 5 $\pm 0.000 008 5$	0.002 082 $\pm 0.000 016$	0.001 389 $\pm 0.000 013$
Zj_LOPS	0.009 620 $\pm 0.000 048$	0.007 822 $\pm 0.000 047$	0.009 678 $\pm 0.000 036$	0.007 874 $\pm 0.000 027$
Zj_LOPS_noQED	0.009 510 $\pm 0.000 016$	0.007 726 $\pm 0.000 019$	0.009 574 $\pm 0.000 021$	0.007 771 $\pm 0.000 030$
Zj_LOPS_s	0.010 379 $\pm 0.000 015$	0.008 786 $\pm 0.000 011$	0.010 380 $\pm 0.000 020$	0.008 791 $\pm 0.000 020$
Zj_LOPS_noFSR	0.009 074 $\pm 0.000 022$	0.007 340 $\pm 0.000 012$	0.009 084 $\pm 0.000 033$	0.007 359 $\pm 0.000 037$
Zj_LOPS_noISR	0.004 183 3 $\pm 0.000 003 6$	0.003 199 1 $\pm 0.000 003 3$	0.004 227 $\pm 0.000 022$	0.003 223 $\pm 0.000 016$
Zj_LOPS_s_qposz_noISR	0.000 386 $\pm 0.000 025$	0.000 332 $\pm 0.000 021$	0.000 367 $\pm 0.000 013$	0.000 314 9 $\pm 0.000 004 4$
Zj_NLO_PS	0.023 754 $\pm 0.000 011$	0.020 129 $\pm 0.000 089$	0.023 692 $\pm 0.000 050$	0.020 092 $\pm 0.000 062$
Zj_NLO_2j_LO_2j_PS	0.023 900 $\pm 0.000 093$	0.019 939 $\pm 0.000 079$	0.023 94 $\pm 0.000 13$	0.019 961 $\pm 0.000 069$

Table 5.2: Mean Squared Error of predicted and true polarization fractions of the validation dataset for the Transformer trained on all jets for $Z+jets$.

Sample	Transformer on all jets	
	MSE_{val}	$MSE_{val,L}$
Zj_LOPS_s	0.008 908 5 $\pm 0.000\ 007\ 9$	0.007 189 3 $\pm 0.000\ 008\ 4$
Zj_LOPS	0.009 437 $\pm 0.000\ 029$	0.007 952 $\pm 0.000\ 035$
Zj_LOPS_noQED	0.008 807 $\pm 0.000\ 023$	0.007 102 $\pm 0.000\ 023$
Zj_LOPS_noFSR	0.008 275 3 $\pm 0.000\ 009\ 6$	0.006 636 $\pm 0.000\ 010$
Zj_LOPS_noISR	0.004 265 $\pm 0.000\ 027$	0.003 268 $\pm 0.000\ 038$
Zj_LOPS_s_qposz_noISR	0.000 161 4 $\pm 0.000\ 003\ 7$	0.000 135 9 $\pm 0.000\ 003\ 4$
Zj_NLO_PS	0.022 642 $\pm 0.000\ 086$	0.019 134 $\pm 0.000\ 087$
Zj_NLO_2j_LO_2j_PS	0.021 615 $\pm 0.000\ 055$	0.017 960 $\pm 0.000\ 027$

5.2.4 Phase space effects

This section studies the effects of the phase space, in which the polarization tagger is trained and applied, on the extraction of the MVB polarization fractions of the events. MLPs are often trained in phase spaces larger than the final phase space in which they are ultimately applied, for example when used as discriminants in measurements. Even out-of-phase-space events can provide useful information to the network and improve the NN performance generally compared to an exclusive training in the fiducial phase space. In this thesis, the MLP is always provided with input vectors of the same size and the choice of phase space merely represents an event selection. For the Transformer applied to all jets, the phase space applies a minimum p_T cut to all jets in the event which does not correspond to a pure event selection, but rather determines the jets provided to the Transformer as input. Consequently, the structure of the input changes depending on the phase space, which could potentially lead to incorrect predictions. Furthermore, it is also interesting to see to what extent the networks are capable of extrapolating to unknown phase space regions.

The question arises, how strongly the NN predictions depend on the phase space of the events that the network is trained on, especially in cases where training and prediction are performed in different phase spaces. In all other sections, the tagger is trained and validated on events in the generator phase space. In this section, the tagger is trained on events in the generator phase space or the fiducial phase space before it is applied to events in the generator phase space or the fiducial phase space. The definition of the phase spaces used in this work is presented in subsection 4.3.2. The corresponding subsamples are obtained from the Zj_LOPS sample that includes events of $pp \rightarrow Z + j + X @ LO+PS$.

The polarization fraction distributions are shown for the MLP in Figure 5.19, the Transformer in Figure 5.20 and the Transformer trained on all jets in Figure 5.21. The RPE is shown for the MLP in Figure 5.22, the Transformer in Figure 5.23 and the Transformer trained on all jets in Figure 5.24. The loss on the validation dataset is presented in Table 5.3 and Table 5.4.

a) Narrowing of the phase space

In the first step, the tagger is trained on events in the generator phase space and applied on events in the fiducial phase space. Thus, the phase space of the polarization fraction predictions is more narrow compared to the phase space of the training. To evaluate the effect, the result is compared to a network trained and applied on fiducial phase space. All the results show almost no difference in whether the tagger is trained on the generator or fiducial phase space before it is applied to events in the fiducial phase space ("train: generator - pred.: fiducial" vs. "train: fiducial - pred.: fiducial"). This is the case for the polarization fraction distributions of the MLP in Figure 5.19a and 5.19b, the Transformer in Figure 5.20a and 5.20b, as well as the Transformer trained on all jets in Figure 5.21a and 5.21b.

Same holds true for the RPE of the MLP in Figure 5.22a and 5.22b, the Transformer in

Figure 5.23a and 5.23b, as well as the Transformer trained on all jets in Figure 5.24a and 5.24b. The loss on the validation dataset in Table 5.3 and Table 5.4 shows a better *MSE* of the Transformer trained on all jets compared to the MLP and the Transformer trained on the leading jet for both phase space combinations. Thus, the Transformer trained on all jets shows no problem with the different input structure. In fact, it is significantly better at utilising knowledge from the larger phase space than a MLP or a Transformer trained on a single jet, which is both surprising and reassuring.

The information included in the PS jets is observed to be beneficial for the polarization extraction in this case. No phase space effect is observed for the MLP and Transformer trained on a single jet, as the performance for training in the generator and fiducial phase space are very similar. Since the fiducial phase space is included in the generator phase space, this is not surprising.

The true polarization fractions show a slightly different distribution for events in the fiducial phase space compared to events in the generator phase space. The additional phase space cuts cause a less extreme peak for $r_L = 0$ and $r_T = 1$, and the number of events with a higher r_L and lower r_T is much higher relative to the entire sample. The different shape of the predicted number of events is caused by the different underlying distribution of the true polarization fractions, where mainly the peaks are excluded, but not a different prediction behaviour itself. This assumption is supported by the fact that the distribution of the *RPE* is very similar for all investigated cases. Thus, the tagger learns the same relations between fraction and FS particles, but is then only applied on events with more restrictions on these FS kinematics.

b) Widening of the phase space

In the second step, the tagger is trained on events in the fiducial phase space and applied on events in the generator phase space. Thus, the phase space of the polarization fraction prediction is wider than the phase space of the training events. To evaluate the effect, the result is compared to a network which is trained and applied on the generator phase space.

All the results show almost no difference in whether the tagger is trained on the generator or fiducial phase space before it is applied to events in the generator phase space ("train: generator - pred.: generator" vs. "train: fiducial - pred.: generator"). This is the case for the polarization fraction distributions of the MLP in Figure 5.19c and 5.19d, the Transformer in Figure 5.20c and 5.20d, as well as the Transformer trained on all jets in Figure 5.21c and 5.21d, despite slight differences in the sharpness of the predicted peak. Nevertheless, the Transformer performs significantly worse at extrapolating across all jets. This is not surprising, given that its input structure with significantly more jets is completely different.

This can also be observed for the *RPE* of the MLP in Figure 5.22c and 5.22d, the Transformer in Figure 5.23c and 5.23d, as well as the Transformer trained on all jets in Figure 5.24a and 5.24b. The loss on the validation dataset in Table 5.3 and Table 5.4 shows a slight phase space

effect, since the MSE for the tagger trained on fiducial phase space is slightly higher than when trained on the generator phase space. Nevertheless, the taggers ability to extrapolate events in regions of the generator phase space is remarkable.

As already discussed for the narrowing of the phase space, the different shape of the predicted number of events is not caused by a different prediction behaviour of the NN itself but by the different underlying distribution of the true polarization fractions. This leads to the assumption that the tagger learns the same fundamental relations between fraction and FS particles independent of the training phase space. This also explains why the networks are so good at extrapolating into the larger phase space.

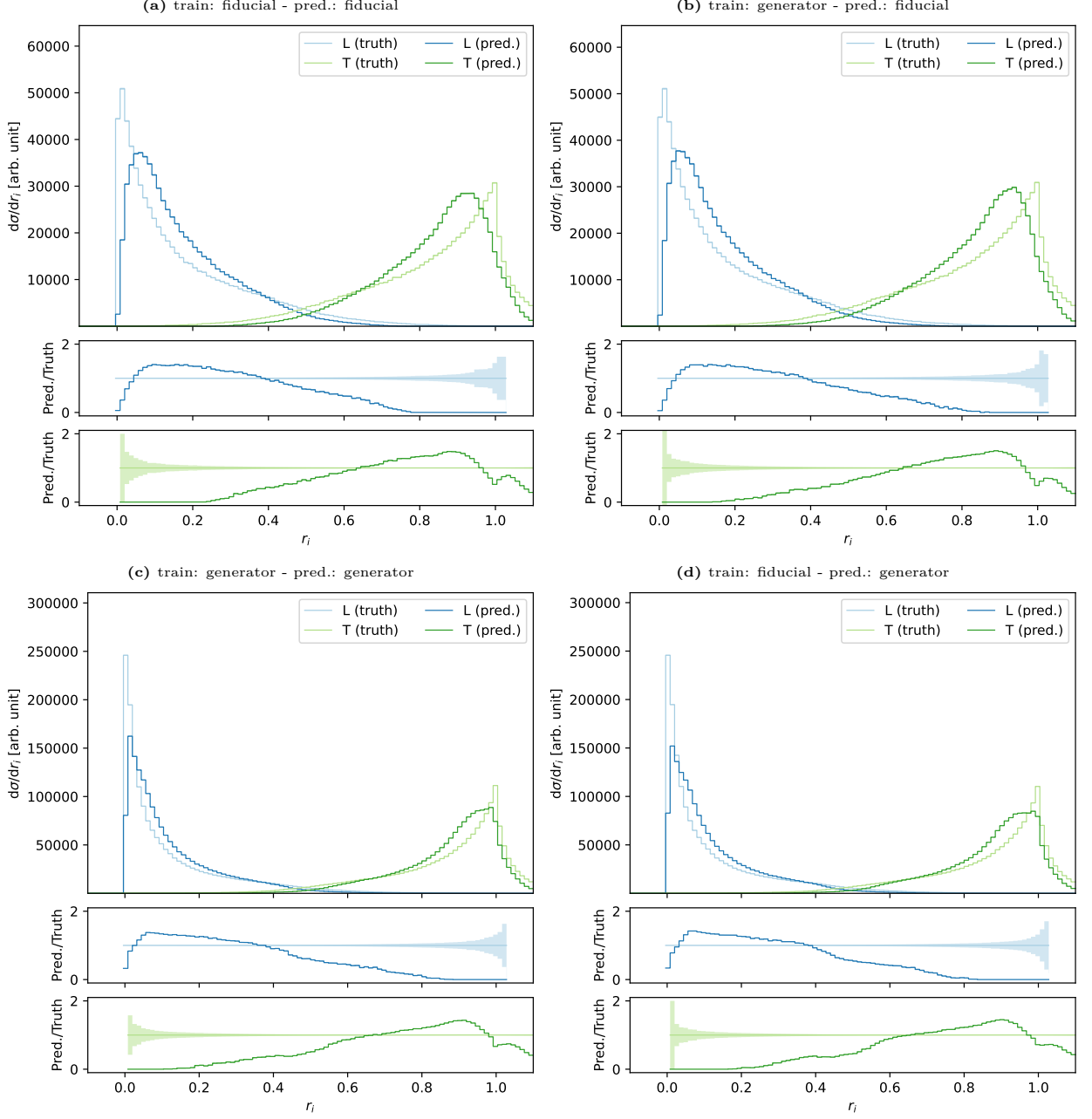


Figure 5.19: Predicted and true distribution of r_L and r_T of the MLP for $pp \rightarrow Z + j + X$ @ LO+PS. The NN is trained in the generator phase space (train: generator) or in the fiducial phase space (train: fiducial) and afterwards validated in the generator phase space (pred.: generator) or the fiducial phase space (pred.: fiducial).

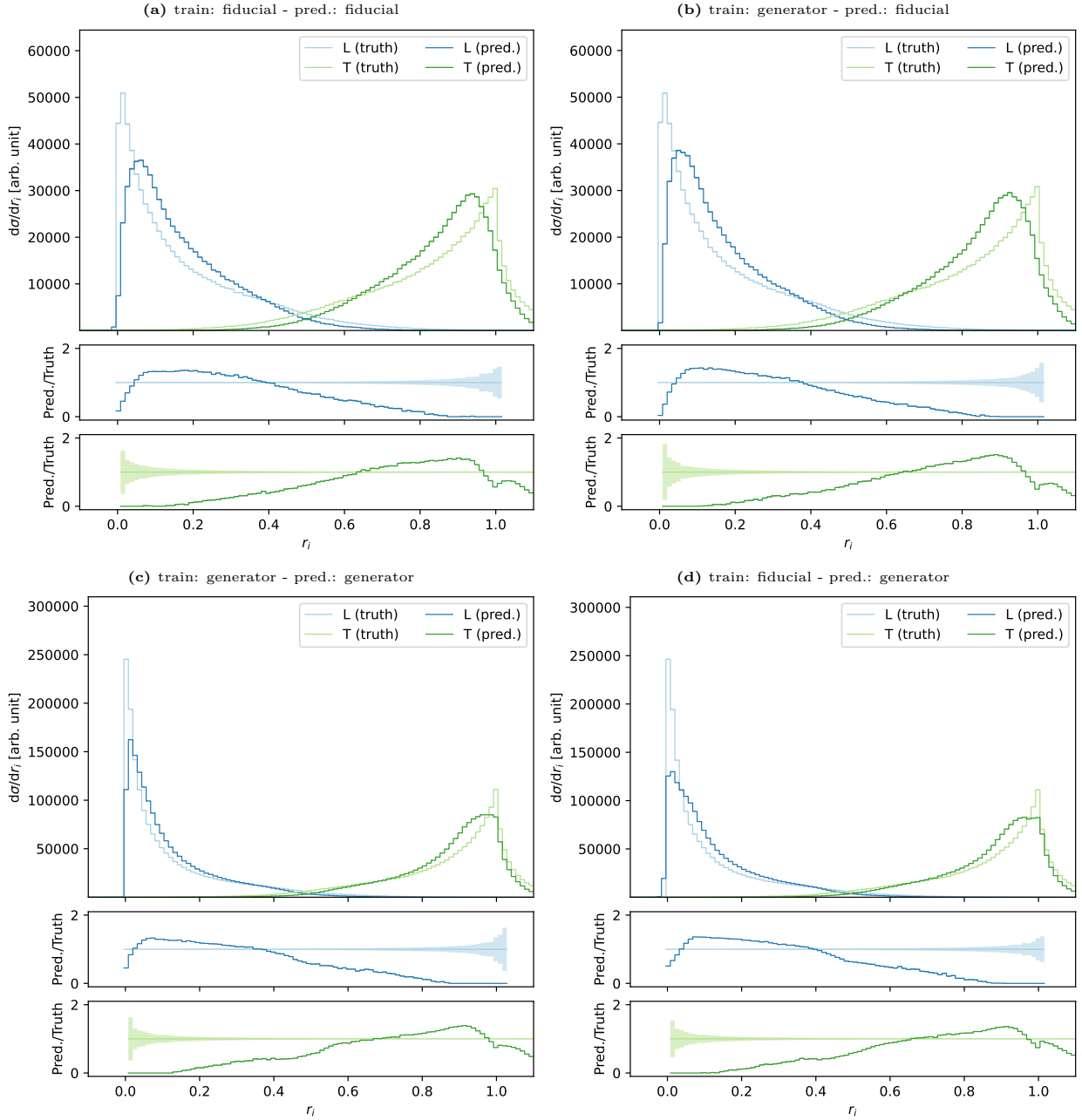


Figure 5.20: Predicted and true distribution of r_L and r_T of the Transformer for $pp \rightarrow Z + j + X$ @ LO+PS.

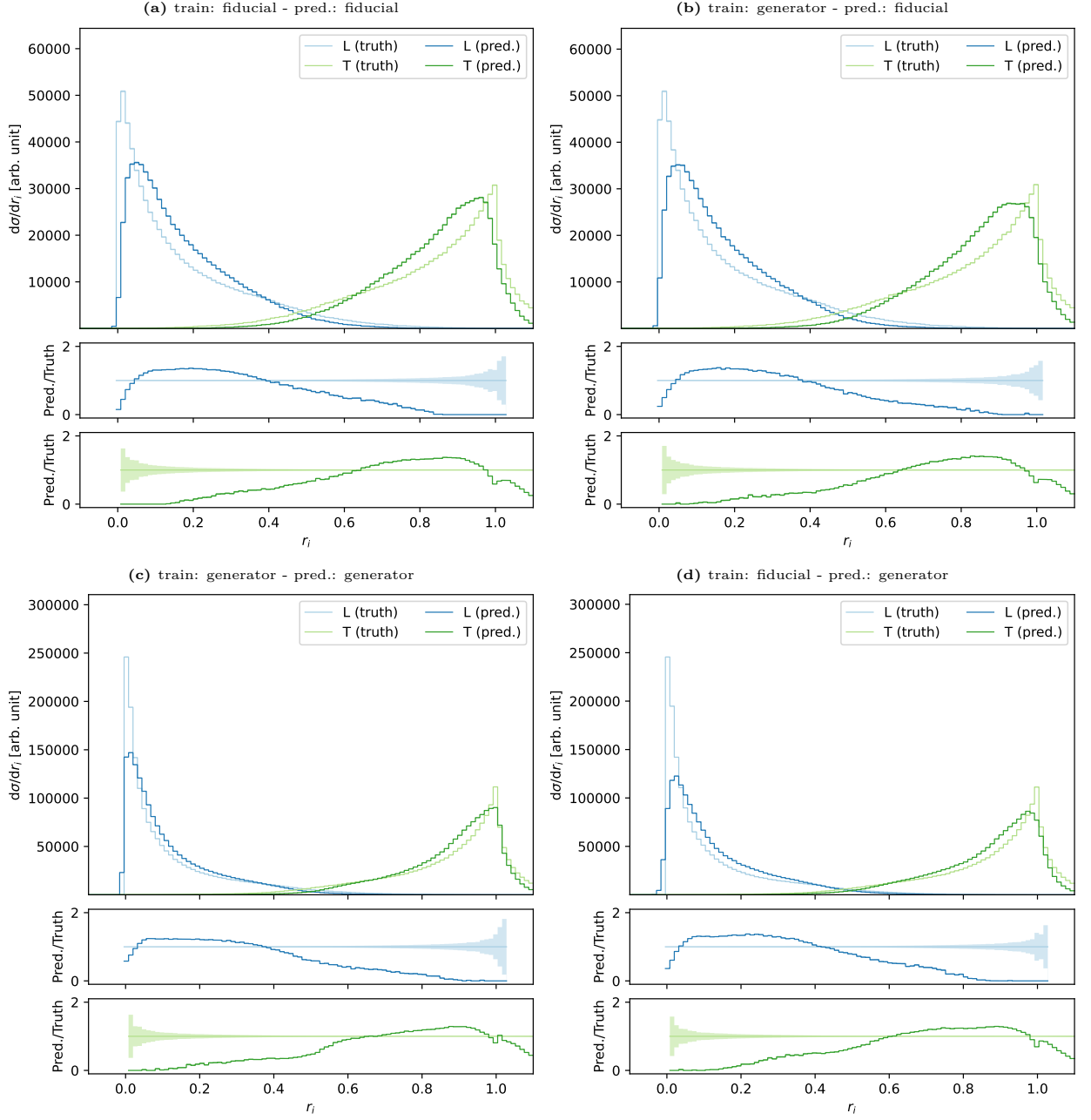


Figure 5.21: Predicted and true distribution of r_L and r_T of the Transformer trained on all jets for $pp \rightarrow Z + j + X$ @ LO+PS.

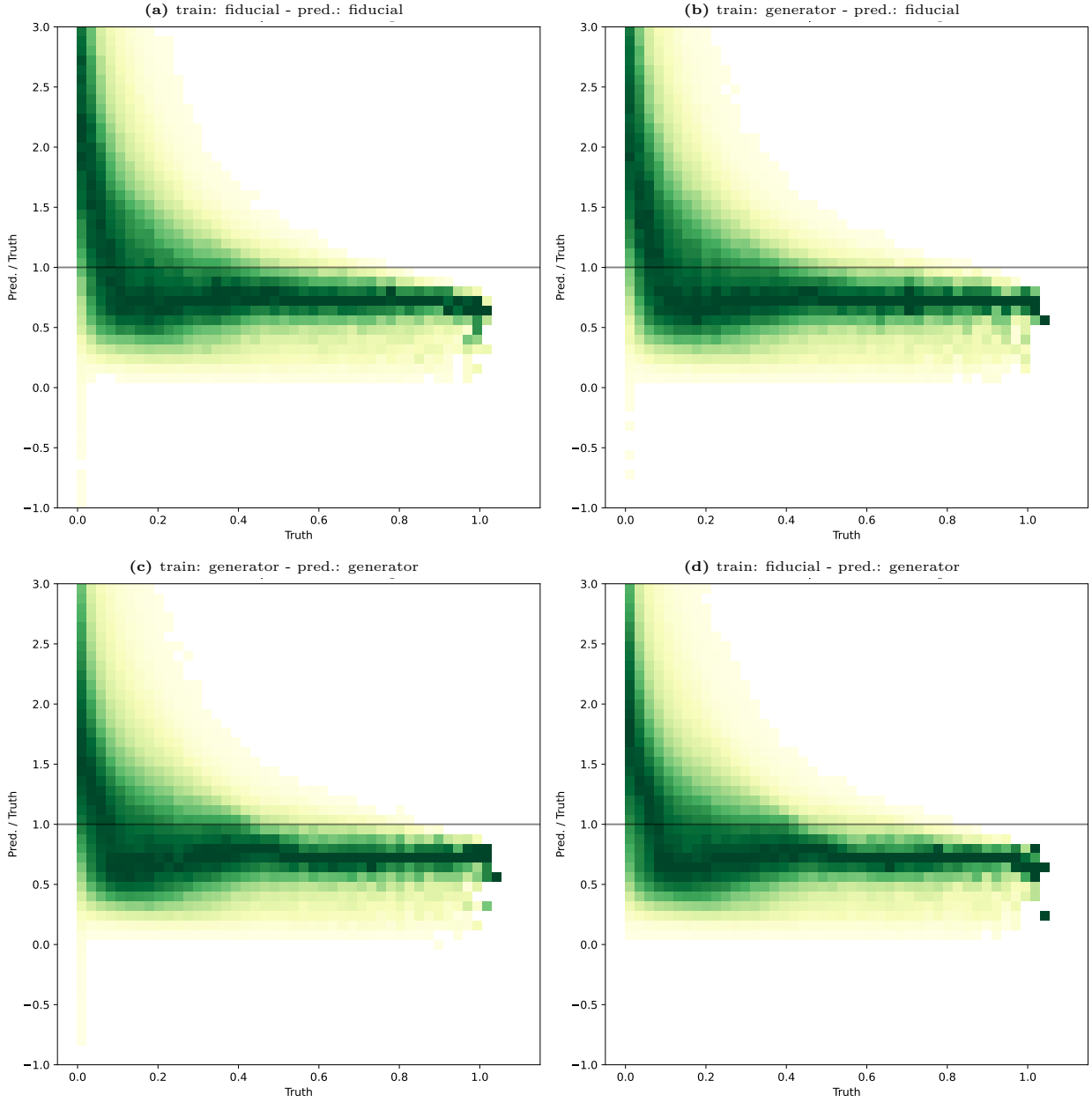


Figure 5.22: Relative prediction error of $r_{L,val}$ of the MLP for $pp \rightarrow Z + j + X$ @ LO+PS. The NN is trained in the generator phase space (train: generator) or in the fiducial phase space (train: fiducial) and afterwards validated in the generator phase space (pred.: generator) or the fiducial phase space (pred.: fiducial).

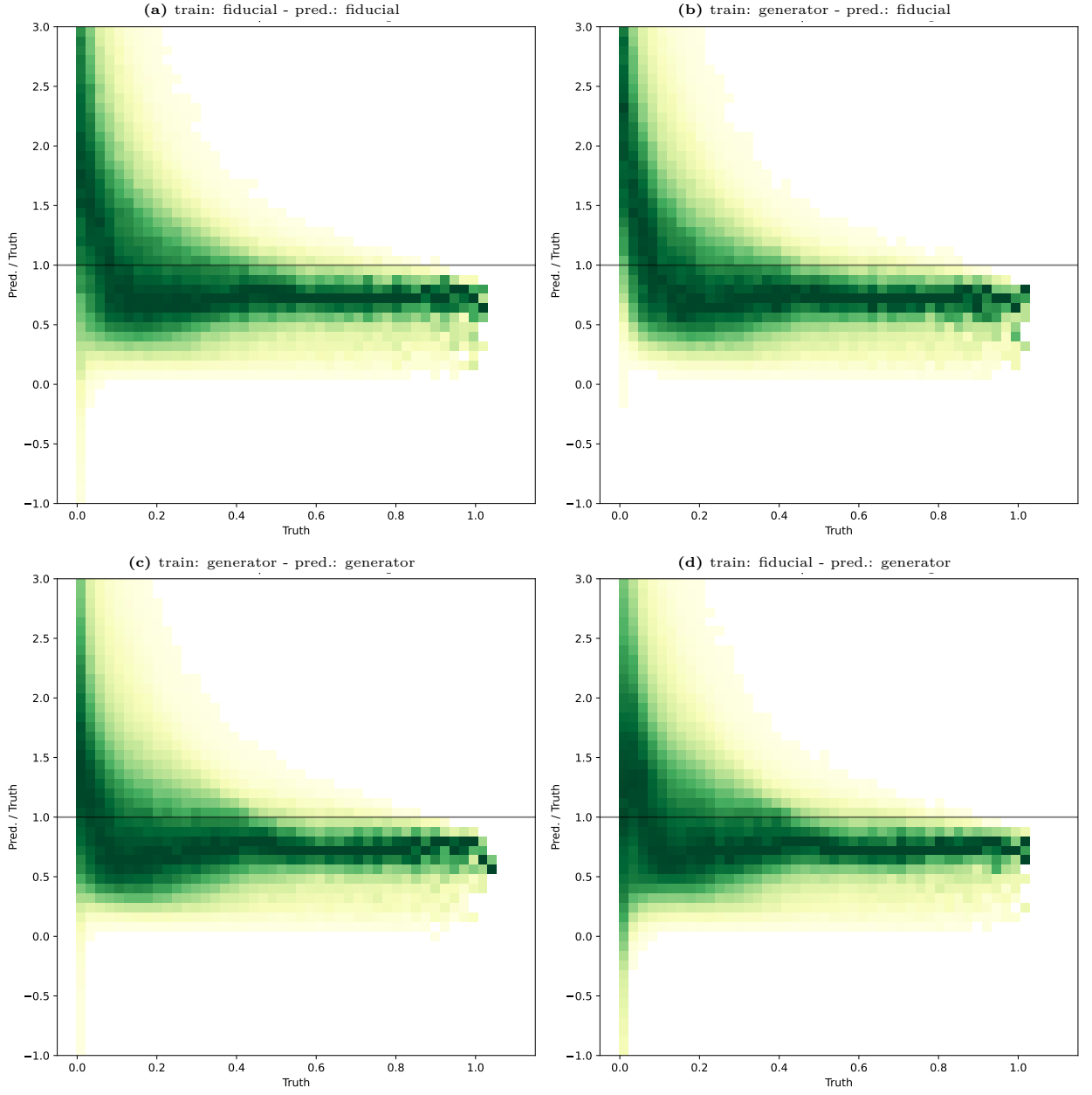


Figure 5.23: Relative prediction error of $r_{L,val}$ of the Transformer for $pp \rightarrow Z + j + X$ @ LO+PS.

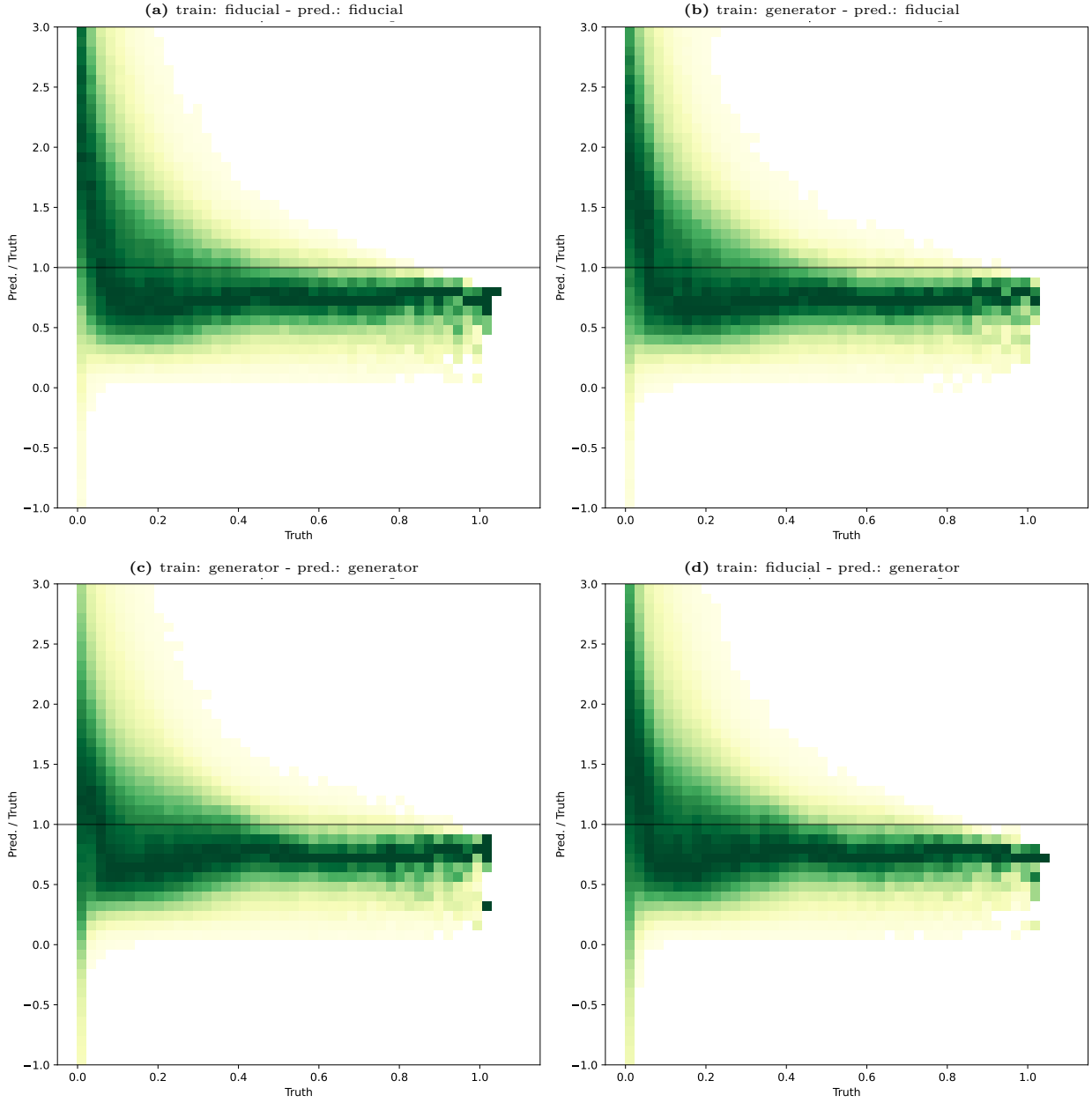


Figure 5.24: Relative prediction error of $r_{L,val}$ of the Transformer trained on all jets for $pp \rightarrow Z + j + X$ @ LO+PS.

Table 5.3: Mean Squared Error of predicted and true polarization fractions of the validation dataset for the MLP and the Transformer for $pp \rightarrow Zj$ @ LO. MLP and Transformer are trained on generator (train: generator) or fiducial (train: fiducial) phase space and validated by predicting r_L and r_T on events in generator (pred.: generator) or fiducial (pred.: fiducial) phase space.

Sample	MLP		Transformer	
	MSE_{val}	$MSE_{val,L}$	MSE_{val}	$MSE_{val,L}$
train: generator	0.009 659	0.007 850	0.009 691	0.007 892
pred.: generator	$\pm 0.000\,019$	$\pm 0.000\,018$	$\pm 0.000\,041$	$\pm 0.000\,043$
train: fiducial	0.009 980	0.007 869	0.010 003	0.008 208
pred.: generator	$\pm 0.000\,026$	$\pm 0.000\,020$	$\pm 0.000\,061$	$\pm 0.000\,073$
train: fiducial	0.014 635	0.012 444	0.014 615	0.012 443
pred.: fiducial	$\pm 0.000\,020$	$\pm 0.000\,025$	$\pm 0.000\,044$	$\pm 0.000\,038$
train: generator	0.014 615	0.012 433	0.014 517	0.012 364
pred.: fiducial	$\pm 0.000\,031$	$\pm 0.000\,021$	$\pm 0.000\,061$	$\pm 0.000\,061$

Table 5.4: Mean Squared Error of predicted and true polarization fractions of the validation dataset for the Transformer trained on all jets for $pp \rightarrow Zj$ @ LO.

Sample	MLP	
	MSE_{val}	$MSE_{val,L}$
train: generator	0.009 112	0.007 367
pred.: generator	$\pm 0.000\,036$	$\pm 0.000\,037$
train: fiducial	0.010 695	0.008 859
pred.: generator	$\pm 0.000\,043$	$\pm 0.000\,042$
train: fiducial	0.013 843	0.011 692
pred.: fiducial	$\pm 0.000\,021$	$\pm 0.000\,028$
train: generator	0.013 757	0.011 361
pred.: fiducial	$\pm 0.000\,055$	$\pm 0.000\,048$

5.3 $W+jets$

In this section, single W^+ boson production in association with at least one jet is investigated at LO and LO+PS. The W^+ boson decays leptonically into a positively charged lepton and its corresponding neutrino. This neutrino can not be measured experimentally and information on the prompt neutrino momentum is only available in the simulation. In experiments, the transverse momentum of the neutrino is estimated by the missing transverse momentum p_{miss}^T which is also available in the simulation. This is based on the assumption that the incoming protons carry no transverse momentum, such that the total transverse momentum p_{total} of the outgoing particles must vanish. The missing transverse momentum $p_{miss}^T = -p_{total}$ must then originate from the undetected neutrino. With the $W+jets$ production, the effect of the incomplete knowledge of final state momenta compared to the knowledge of the prompt neutrino can be investigated.

Generally, much higher values of the true polarization fractions are observed for the W^+ boson than for the Z boson presented in the previous section. This is related to the exclusive coupling of the W^+ boson to left-handed fermions and right-handed anti-fermions as a consequence of its maximal parity violation. Similar high polarization fraction values for the Z boson can artificially be achieved by turning of the right-handed coupling component of the $Z \rightarrow f\bar{f}$ vertex in the simulation. The distributions of the true polarization fractions exhibits a broadening for r_L as well as r_T , which affects the training stability of the NN. This can be avoided by a logarithmic pre-scaling of the polarization fractions in form of $\log(r_i + 1)$.

The results for $W+jets$ at LO and LO+PS of both the MLP and the Transformer are shown in Figure 5.25 and 5.28. The computed loss on the validation dataset in terms of MSE_{val} and $MSE_{val,L}$ is shown in Table 5.5.

a) Wj_LO_noPDF

The Wj_LO_noPDF sample contains a prediction for $ug \rightarrow W^+d$ @ LO with fixed initial states and a beam energy of $E_{IS} = 75$ MeV. The IS particles and their kinematics are fixed for all events, comparable to the Zj_LO_noPDF sample. Here, the polarized matrix element and the polarization fraction are independent of the IS. The full kinematic information on the prompt neutrino is provided to the network.

The polarization fraction distributions of the MLP in Figure 5.25a and of the Transformer in Figure 5.26a show, first of all, slight differences in the true polarization fractions compared to the Z boson production process (see e.g. Figure 5.3a). In the Wj_LO_noPDF sample, the range of the polarization fraction values is much wider and events exhibit values of $r_L > 0.5$ and $0.3 < r_T < 0.5$. Nevertheless, the majority of events has a fraction of $r_L \approx 0$ and $r_T \approx 1$ as for the Z boson production process. The predicted and true distribution are in good agreement for a wide range of r_L and r_T . For higher values of r_L , the number of events decreases and

statistical fluctuations increasingly dominate the ratio of true and predicted events for MLP and Transformer. These fluctuations are not observed in the `Zj_LO_noPDF` results, since no events exist for $r_L < 0.5$ in this process.

The *RPE* of the MLP in Figure 5.27a and of the Transformer in Figure 5.28a show that the majority of events exhibit a ratio of predicted and true r_L value close to $RPE = 1$. The MLP shows a small underestimation of the longitudinal polarization fraction for $r_{L,truth} < 0.2$, which increases to lower values of $r_{L,truth}$. The loss on the validation dataset in Table 5.5 are comparable for MLP and Transformer and mainly governed by statistics.

Overall, the networks perform very well in predicting the polarization fraction of the `Wj_LO_noPDF` sample. This is expected since the events do not differ in the IS of the hard process, which allows to reconstruct the polarized matrix element and the polarization fraction of the intermediate W^+ boson from the FS kinematics.

b) `Wj_LO_s_qposz`

The `Wj_LO_s_qposz` sample contains a prediction for the partonic channel $qg \rightarrow W^+q'$ @ LO with $p_{z,q} > 0$. Again, the momentum of the neutrino is provided. The events vary in the IS particles type and kinematics, but a restriction is imposed on the IS such that the IS quark always moves in positive and the gluon in negative beam direction. This approach leads to a good predictive NN performance for Z boson production (see `Zj_LO_s_qposz` results).

The polarization fraction distributions of the MLP in Figure 5.25b and of the Transformer in Figure 5.26b show a slightly different shape and range of the true polarization fractions when the IS momenta and particle type is not fixed for all events compared to the `Wj_LO_noPDF` sample. The distribution of r_T exhibits a sharper peak of the number of events at $r_T \approx 1$ and events with $0 < r_T < 0.3$ exist, which are not observed for the case of fixed IS kinematics in the `Wj_LO_noPDF` results. Both architectures show a good agreement between predicted and true distribution. For increasing r_L , a higher statistical fluctuation in the ratio between predicted and true distribution is observed similar to the `Wj_LO_noPDF` results. Strong deviations are only observed for $r_T < 0.05$. Nevertheless, a small constant underestimation of the number of events for $0.5 < r_T < 0.25$, as well as a systematic small fluctuation at $r_T = 1.0$ are evident in the ratio plot of both the MLP and the Transformer.

The *RPE* of the MLP in Figure 5.27b and of the Transformer in Figure 5.28b show a clear increase in the ratio of predicted and true r_L values compared to `Wj_LO_noPDF` with fixed IS kinematics. For $r_L < 0.2$, the Transformer exhibits overestimation in the majority of events, whereas the *RPE* is distributed more symmetrically around $RPE = 1$ for the MLP. The high statistical fluctuation for $r_L > 1$ in MLP and Transformer is caused by the extreme low number of events there. The loss on the validation dataset in Table 5.5 exhibits a *MSE* which is, again, mainly governed by statistical fluctuations. It shows an increase compared to the *MSE* of the `Wj_LO_noPDF` sample, but both are still in the same order of magnitude.

Compared to `Wj_LO_noPDF` sample, the kinematics of the IS particles are not fixed. The tagger only receives the FS kinematics and is not able to exactly reconstruct the IS determining the polarized matrix element of the process. The restriction on the IS, which only allows one of the two possible configurations of the incoming direction of the colliding particles, resolves the IS ambiguity in the events and supports the polarization fraction prediction.

c) `Wj_LO`

The `Wj_LO` sample contains a prediction for $pp \rightarrow W^+j$ @ LO. Thus, it includes predictions for all partonic channels $q\bar{q}' \rightarrow W^+g$, $qg \rightarrow W^+q'$ and $\bar{q}g \rightarrow W^+\bar{q}'$, whose corresponding Feynman diagrams are shown in Figure 4.2a. Now, the missing knowledge of the IS kinematics and the partonic channel constitutes an ambiguity. The network receives the four-momentum of the prompt neutrino.

The polarization fraction distributions of the MLP in Figure 5.25c and of the Transformer in Figure 5.26c show dramatic deviations of MC truth and NN predictions. The longitudinal fraction exhibits a strong and non-constant overestimation of the number of events for $r_L < 0.4$ and an underestimation increasing to higher values for $r_L > 0.4$. The transverse fraction shows a steadily increasing ratio of predicted over true events for $0 < r_T < 1$ to higher values. Therefore, an underestimation in the number of events is detected for $0 < r_L < 0.6$, which leads to an overestimation for $0.6 < r_L < 1$. Peaks of overestimation are situated at $r_L \approx 0.1$, $r_L \approx 0.3$, $r_T \approx 0.85$ and $r_T = 1$ for both architectures. The tagger shows no good agreement between the true and predicted distribution in any region of the polarization fractions.

The *RPE* of the MLP in Figure 5.27c and of the Transformer in Figure 5.28c also show a drastic decline in the tagger's ability to predict the longitudinal polarization fraction compared to the `Wj_LO_s_qposz` results. The number of events with higher values of *RPE* is tremendously increased. For $r_L < 0.3$, an nearly exclusive underestimation of the longitudinal polarization fraction is detected and a notable amount of events with $RPE > 1$ is observed around $r_L \approx 0.2$. This overestimation of r_L for a considerable number of event leads to the overestimation for slightly higher r_L , which is observed in the peak of $r_L \approx 0.3$ in the ratio plots of the polarization fraction distributions in Figure 5.25c and Figure 5.26c. Nevertheless, the distributions of the MLP and the Transformer are comparable in most regions. The loss on the validation dataset in Table 5.5 is increased about 2-3 orders of magnitude compared to the `Wj_LO_s_qposz` result with fixed IS and FS particles. According to the *MSE*, the MLP performs slightly better than the Transformer averaged over all cross-validation folds.

The drastic decline in predictive performance from a sample with varying IS and FS of the hard process compared to a sample with fixed IS and FS is already observed, described and explained for the case of the Z boson production (see `Zj_LO_s_qposz` vs. `Zj_LO` results). The polarized matrix element and the polarization fraction of the intermediate W^+ boson highly depend on the IS and FS of the hard process. Here, the network derives r_L and r_T only

from the FS kinematics without further information on IS and FS particles. This provides an ambiguity of the events, which leads the network to predict the mean value of the polarization fraction for given FS kinematics. Other than the exact reconstruction of the matrix element, which is not possible from the information available, a prediction closer to the mean of the output variable provides a minimum of training loss. This leads to an overestimation peak in the number of events situated around the mean value of the polarization fraction $r_L \approx 0.1$ and $r_T \approx 0.85$, observable in Figure 5.25c and Figure 5.26c.

d) Wj_L0 with p_{miss}^T

To investigate the effect of the missing transverse momentum p_{miss}^T as an estimate of the neutrino momentum, the tagger is trained on the Wj_L0 sample, whereas the experimentally available p_{miss}^T is provided instead of the momentum of the truth neutrino from the simulation.

The polarization fraction distributions of the MLP in Figure 5.25d and of the Transformer in Figure 5.26d show an overall higher disagreement between the predicted and true number of events compared to Wj_L0 with prompt neutrino information. Main peaks of overestimation at $r_L \approx 0.1$ and $r_T \approx 0.9$ in the ratio plots now clearly dominate the distribution whereas the peak at $r_L \approx 0.3$ is less predominant. The ratio of predicted and true number of events shows a strongly developing underestimation for higher values of $r_L > 0.4$ and lower values of $r_T < 0.6$. The same is observed with an overall smaller underestimation when the kinematics of the prompt neutrino is provided.

The *RPE* of the MLP in Figure 5.27d and of the Transformer in Figure 5.28d show an increase in the number of events with higher *RPE* but with similar overall distribution compared to Wj_L0 with a prompt neutrino. The loss on the validation dataset in Table 5.5 increases about 20-30% in the *MSE*.

Overall, providing the missing transverse momentum p_{miss}^T as an estimate of the neutrino momentum leads to a significant decrease of the predictive performance of the tagger. The intermediate W^+ boson is reconstructed from its two decay leptons and, thus, a loss of information on the neutrino kinematics affects the reconstruction of the W^+ boson, as well as its polarization state. This information loss could possibly be reduced partially by the approach of auxiliary task learning. Here, the network is not only trained on the main task but also a second, auxiliary, task that helps the performance in the main task. The network could be trained to predict the truth neutrino momentum, additionally to the polarization fractions, from the experimentally available information. By this, the network learns the relation between missing transverse momentum and prompt neutrino kinematics. Since knowledge about the prompt neutrino is demonstrated to be advantageous for the polarization fraction prediction, this auxiliary task could then improve the performance in the main task.

e) + f) Wj_LOPS

The Wj_LOPS sample contains a prediction for $pp \rightarrow W^+ + j + X$ @ LO+PS. The tagger again receives the information on the missing transverse momentum instead of the four-momentum of the prompt neutrino. The QCD and QED parton showering as ISR and FSR is included, such that gluons and photons are radiated by the IS and FS of the hard process. As already introduced for $Z+jets$, the Transformer can now be trained not only on the four-momenta of both decay leptons and the leading jet of the hard process but also on the jets resulting from the parton shower. Since the number of PS jets and therefore the size of the input vector varies between the events, this is not possible for the MLP, as it requires a constant input size. The PS effect is already investigated and discussed to great extend for the single Z boson production (see subsection 5.2.2). There is no reason to assume that the general NN behaviour in the intermediate steps differs from the observations for $Z+jets$. Thus, only the effect of the full PS on the polarization fraction prediction is investigated for the single W^+ boson production process.

The polarization fraction distributions of the MLP in Figure 5.25e, the Transformer in Figure 5.26e and the Transformer trained on all jets in Figure 5.26f show an overall increased deviation between predicted and true distributions for r_L and r_T compared to the Wj_LO results with p_{miss}^T . The tagger performs much worse in modelling the peak regions of the r_L and r_T distribution. The over- and underestimation of the number of events, which is observed for the Wj_LO sample, increases even more when including the parton shower. The peaks of overestimation at $r_L \approx 0.15$ and $r_T \approx 0.85$ indicate that regions exist where the predicted number of events is up to twice as high as the true number, which is not observed in any of the other samples in this thesis. No major improvement can be observed for the Transformer training on all jets compared to only the leading jet of the hard process.

The RPE of the MLP in Figure 5.27e, the Transformer in Figure 5.28e and the Transformer trained on all jets in Figure 5.28f also show an increase in the number of events with $RPE \ll 1$ and a broader distribution of RPE values compared to the Wj_LO results with p_{miss}^T without PS. The majority of events exhibit a strong underestimation of the longitudinal polarization fraction, which is more extreme compared to the one for the Z boson production (see Zj_LOPS results). No major differences are observed between the distributions of the MLP and the two Transformer results. The loss on the validation dataset in Table 5.1 is increased about 15-20% compared to the Wj_LO results with p_{miss}^T . No improvement is measured when including all jets in the Transformer training according to the MSE .

The decrease in the predictive performance of the polarization tagger is expected and also observed in the case of single Z boson production. The parton shower leads to the radiation of additional gluons and photons from IS and FS, which leads to a strong variation in the IS particle kinematics of the events not provided to the network. This ambiguity in the events negatively affects the reconstruction of the intermediate W^+ boson.

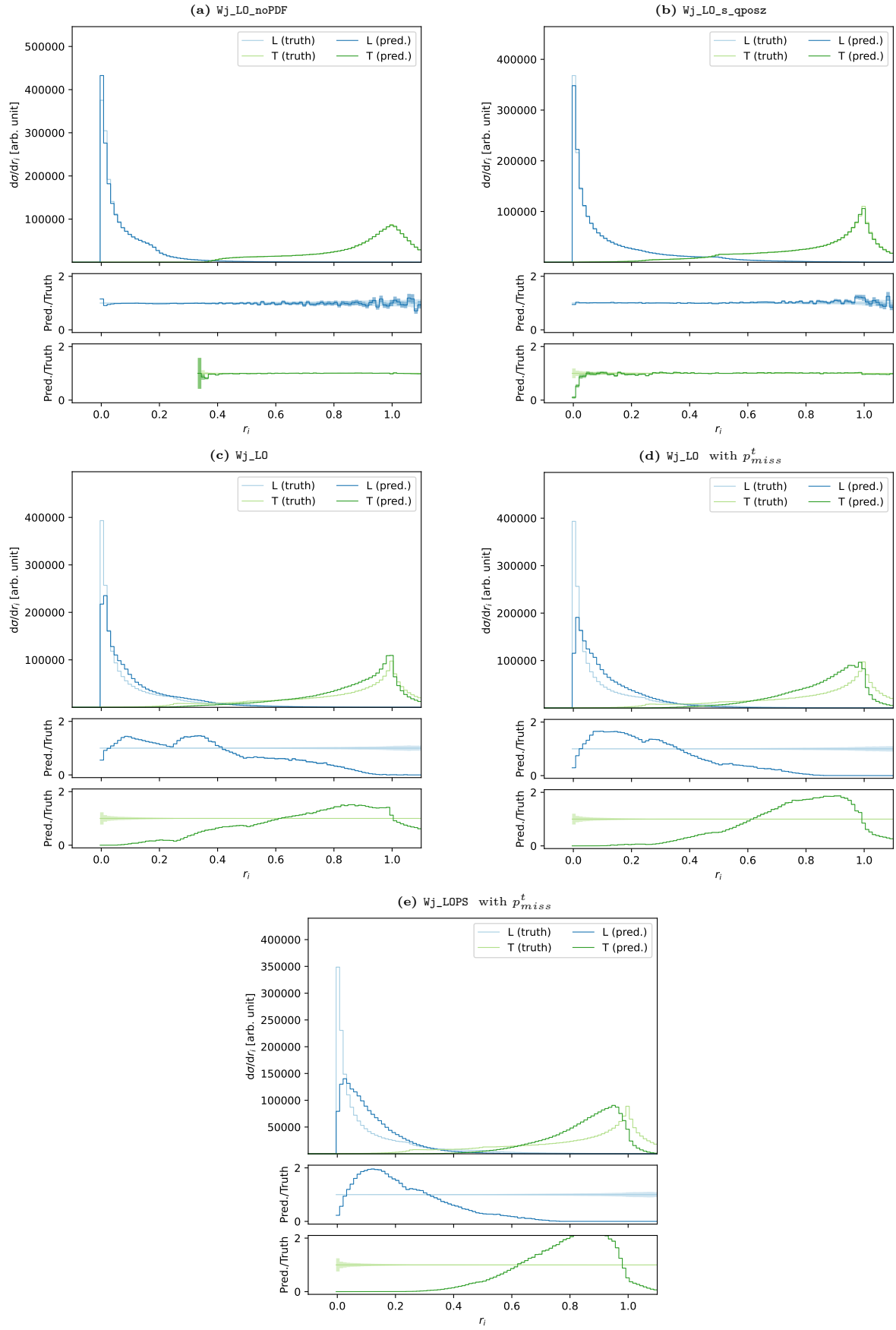


Figure 5.25: Predicted and true distribution of r_L and r_T of the MLP for $W+jets$.

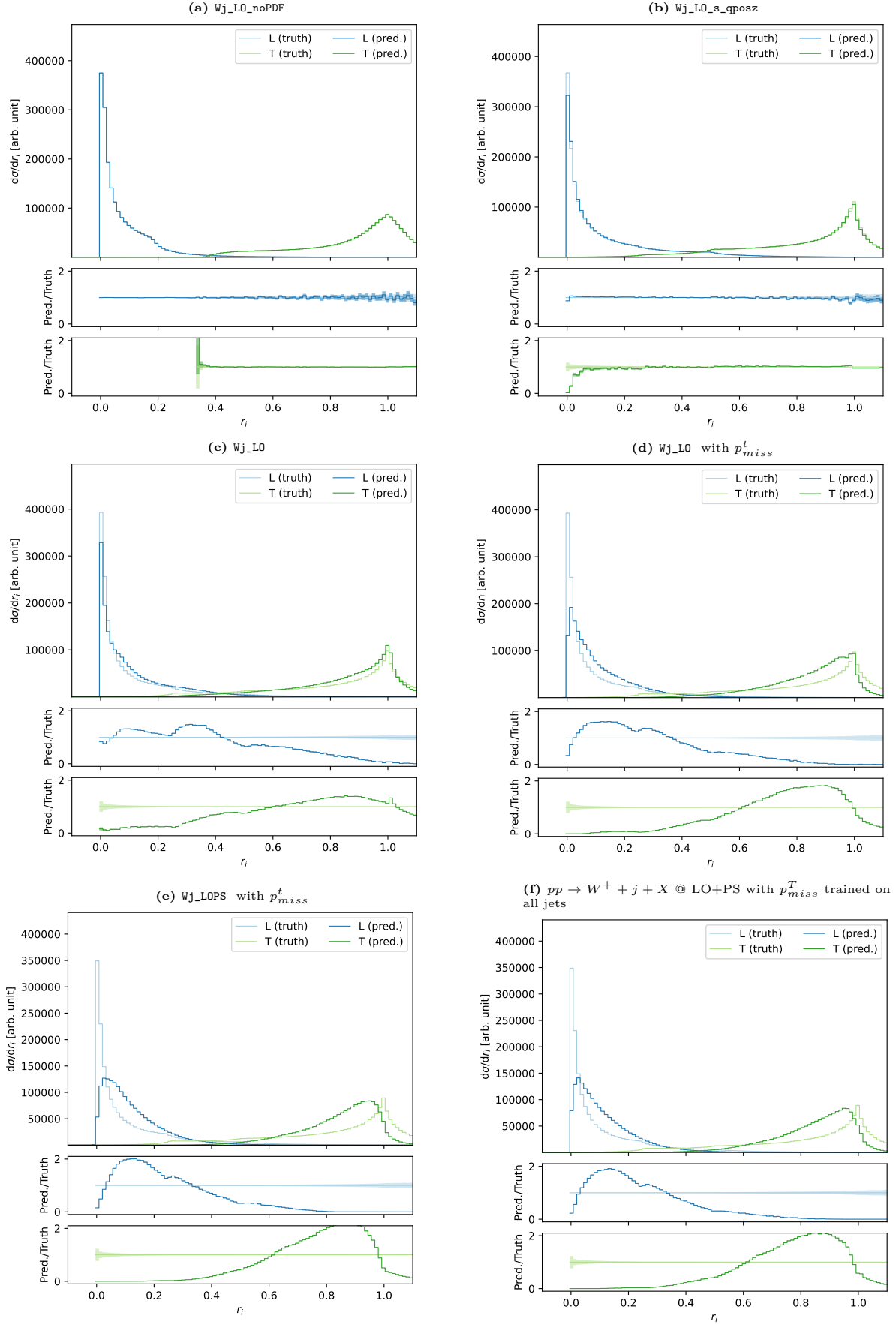


Figure 5.26: Predicted and true distribution of r_L and r_T of the Transformer for $W+jets$.

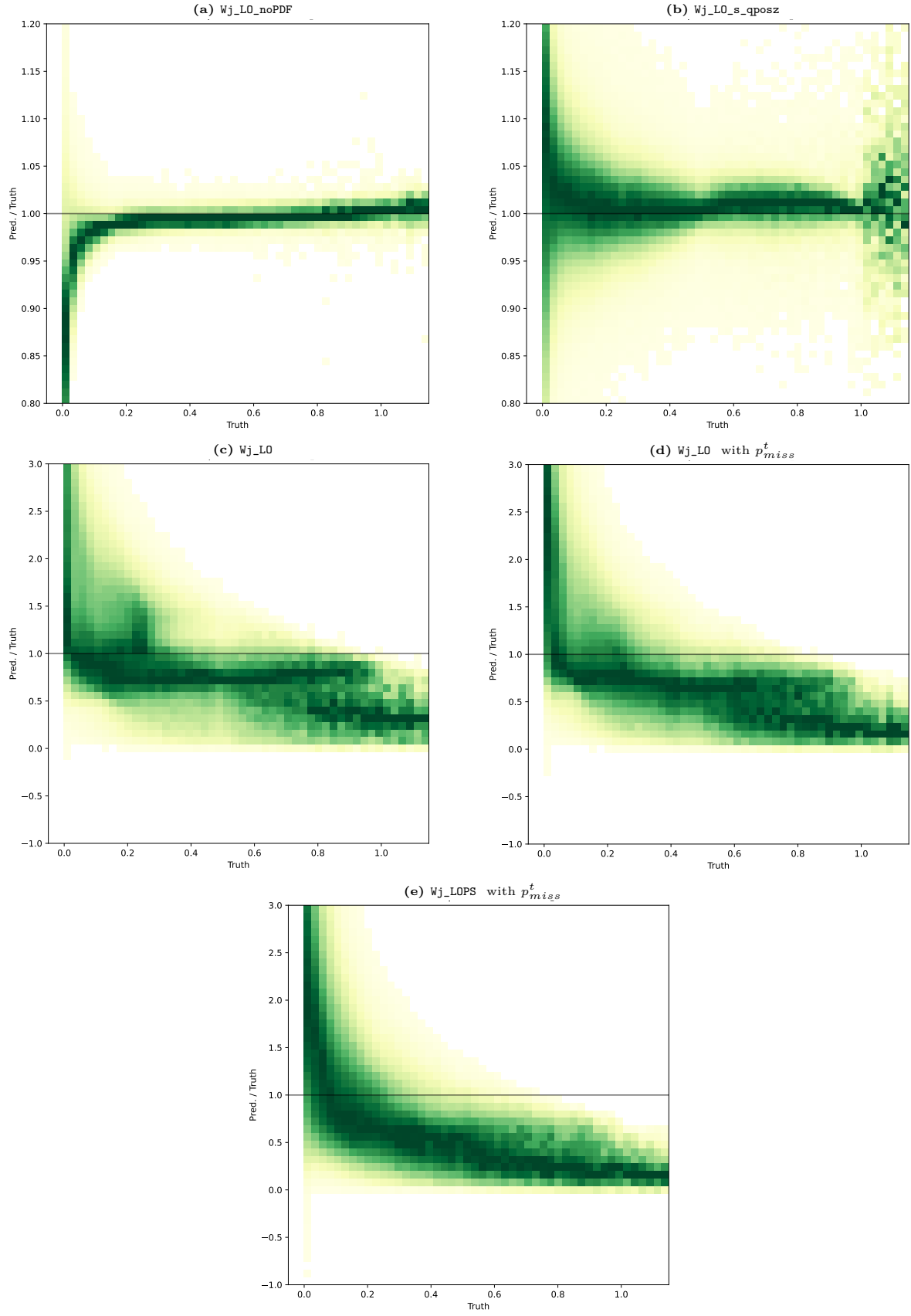


Figure 5.27: Relative prediction error of $r_{L, val}$ of the MLP for $W+jets$.

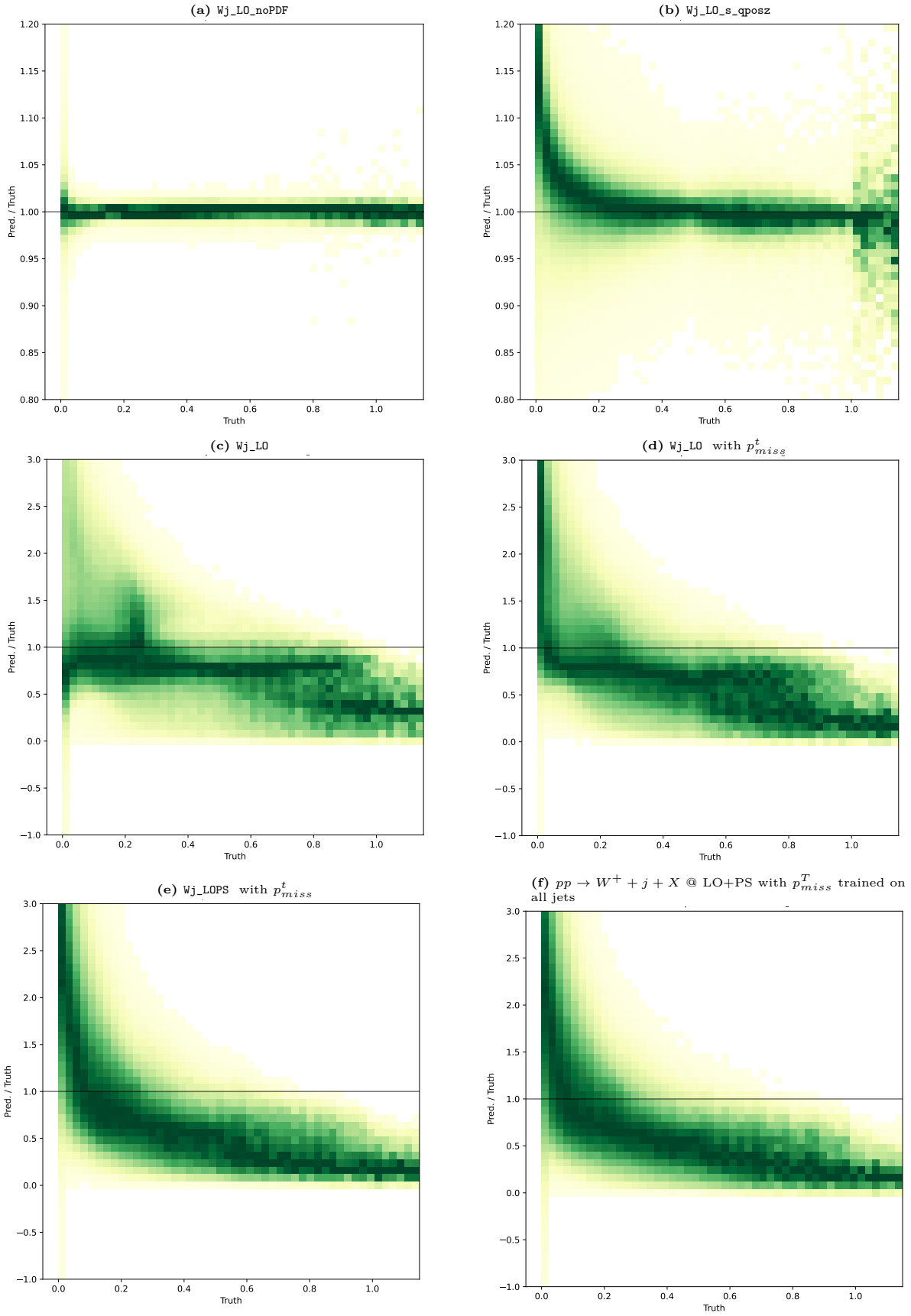


Figure 5.28: Relative prediction error of $r_{L,val}$ of the Transformer for $W+jets$.

Table 5.5: Mean Squared Error of predicted and true polarization fractions of the validation dataset for the MLP and the Transformer for $W+jets$.

Sample	MLP		Transformer	
	MSE_{val}	$MSE_{val,L}$	MSE_{val}	$MSE_{val,L}$
Wj_LO_noPDF	0.000 017 0 $\pm 0.000 005 1$	0.000 012 6 $\pm 0.000 004 4$	0.000 011 4 $\pm 0.000 007 3$	0.000 010 1 $\pm 0.000 008 1$
Wj_LO_s_qposz	0.000 099 $\pm 0.000 014$	0.000 092 $\pm 0.000 013$	0.000 097 $\pm 0.000 012$	0.000 089 $\pm 0.000 012$
Wj_LO	0.013 168 $\pm 0.000 082$	0.009 918 $\pm 0.000 067$	0.013 478 $\pm 0.000 073$	0.010 187 $\pm 0.000 078$
Wj_LO with p_{miss}^T	0.016 350 $\pm 0.000 063$	0.012 760 $\pm 0.000 060$	0.016 299 $\pm 0.000 075$	0.012 708 $\pm 0.000 092$
Wj_LOPS with p_{miss}^T	0.019 010 $\pm 0.000 052$	0.015 599 $\pm 0.000 056$	0.019 029 $\pm 0.000 096$	0.015 62 $\pm 0.000 13$
Wj_LOPS with p_{miss}^T trained on all jets			0.019 142 $\pm 0.000 088$	0.015 75 $\pm 0.000 12$

5.4 $WZjj$ -EW

The polarization tagging for events with two intermediate bosons is investigated with the electroweak production of a W^+Z boson pair in association with two jets ($WZjj$ -EW) at the coupling order of $\mathcal{O}(\alpha_{EW}^6)$ at LO and LO+PS.

As described in section 4.3.1, the investigation of vector boson scattering is of high interest due to the relevance for testing the SM structure and possible new physics. The contribution of VBS to the overall process can be increased by making use of the characteristic signature of a high invariant mass and a significant rapidity difference of the two jets.

For the case of diboson production, four polarization fractions r_{LL} , r_{LT} , r_{TL} and r_{TT} can be defined. The first index refers to the polarization of the W^+ boson, whereas the second index denotes the polarization of the Z boson.

Since the distribution of the true polarization fractions exhibits the same broadening as observed in single W^+ boson production, the training stability is again achieved by the pre-scaling of the polarization fraction by $\log(r + 1)$.

The results of the $WZjj$ -EW samples at LO and LO+PS for the MLP are shown in Figure 5.29 and 5.30, and the computed loss on the validation dataset in terms of the MSE_{val} and $MSE_{val,LL}$ is shown in Table 5.6.

a) $WZjj_EW_LO_fixed$

The $WZjj_EW_LO_fixed$ sample contains a prediction for $u\bar{d} \rightarrow W^+Zu\bar{u}$ @ LO with $p_{z,u} > 0$. Here, the kinematics of the prompt neutrino are provided to the network. Thus, all IS and FS particles are fixed and the IS kinematics is reconstructable from the FS kinematics which the network receives as input.

The polarization fraction distributions of the MLP in Figure 5.29a show an overall very good agreement between predicted and true distribution for almost the entire range of $r_i < 1$. An underestimation in the number of events is observable only for the double-transverse polarization fraction at $r_i > 0.95$. It is evident that all polarization fractions have a peak $r_i = 0$. This is different compared to the single boson production, where the transverse polarization fraction peaked at $r_T = 1$. The tagger exhibits some difficulty to model these peak regions at $r_i = 0$ for all polarization fractions.

The RPE of the MLP in Figure 5.30a also shows a narrow distribution near $RPE = 1$ with a broadening at $r_{LL} \approx 0$. The loss on the validation dataset for the MLP is presented in Table 5.6. It cannot be compared to previous results of other processes due to the different number of polarization fractions included in the calculations and the different distribution of these values.

The network is able to predict the polarization fraction of the intermediate W^+ and Z boson comparably good. This is expected since no ambiguity exists in the dataset since IS and FS

particles do not vary between events. Therefore, the network can reconstruct the polarized matrix element and the polarization fraction from the provided FS kinematics.

b) WZjj_EW_LO

The WZjj_EW_LO sample contains a prediction for $pp \rightarrow W^+ Z jj$ @ LO. Again, the kinematics of the prompt neutrino are provided to the network. Now, the IS and FS are not fixed.

The polarization fraction distributions of the MLP in Figure 5.29b shows stronger deviations compared to the WZjj_EW_LO_fixed results. For r_{LL} , r_{LT} and r_{TL} , an overestimation of the number of events is observed for $r_i < 0.5$, followed by an underestimation for $r_i > 0.5$. This behaviour is similar to the one observed for the longitudinal polarization fraction in $Z+jets$. The steady overestimation in the number of events for $0.1 < r_{TT} < 1.0$ was not observed before.

The *RPE* of the MLP in Figure 5.30b shows many events with higher deviation of true and predicted fractions and the loss on the validation dataset for the MLP in Table 5.6 is almost one order higher compared to the WZjj_EW_LO_fixed results.

Allowing all possible combinations of IS and FS particles in the event leads to a much more complex relation between the FS kinematics and the polarized matrix element. The network does not have sufficient information to learn this more complex relation, which is reflected in worse results compared to the WZjj_EW_LO_fixed sample.

c) WZjj_EW_LO with p_{miss}^T

Now, the tagger is trained on the WZjj_EW_LO sample, but only the missing transverse momentum p_{miss}^T instead of the prompt neutrino momentum is provided.

The polarization fraction distributions of the MLP in Figure 5.29c show increased deviations in the true and predicted number of events compared to the WZjj_EW_LO sample with prompt neutrino momentum. This is observable as an overall stronger trend of under- as well as overestimation in the described regions. This becomes even more evident in the *RPE* of the MLP in Figure 5.30c, where especially the number of events with under-predicted polarization fractions increases drastically for the entire range of r_{LL} . The loss on the validation dataset for the MLP in Table 5.6 increases by about 45% compared to WZjj_EW_LO with the prompt neutrino.

The estimation of the neutrino kinematics from the W^+ boson decay by the missing transverse momentum introduces major difficulties for the network compared to the full prompt neutrino information. The differences between both seem to have a great impact on the predictive power of the tagger, but this additional neutrino information is not available experimentally. One idea could be the introduction of an auxiliary task in which the network is trained to predict not only the polarization fraction but also the prompt neutrino kinematics from the experimentally available information. This could help the network to learn the relation between

missing transverse momentum and prompt neutrino kinematics, which could then facilitate the prediction of the polarization fraction.

d) WZjj_EW_LOPS with p_{miss}^T

The WZjj_EW_LOPS sample contains a prediction for $pp \rightarrow W^+ Z jj$ @ LO+PS. As an estimate of the neutrino momentum, the missing transverse momentum p_{miss}^T is provided to the network. The polarization fraction distributions of the MLP in Figure 5.29d show again an enhanced trend of under- and overestimation compared to WZjj_EW_LO, both with and without p_{miss}^T . The overestimation of the number of events for $r_i < 0.5$, as well as the underestimation for $r_i > 0.5$ increases for r_{LL} , r_{LT} and r_{TL} . Also, the steady overestimation in the number of events for $0.1 < r_{TT} < 1.0$ develops even stronger. The event number of stronger under-predicted values of r_{LL} , depicted in the RPE of the MLP in Figure 5.30d, is enhanced and the loss on the validation dataset for the MLP in Table 5.6 increases slightly compared to the WZjj_EW_LO results.

Including the parton shower leads to a more complex relation between the FS kinematics and the polarization fraction and the network exhibits great difficulties. This is reflected in the worse polarization fraction predictions compared to the WZjj_EW_LO sample without PS.

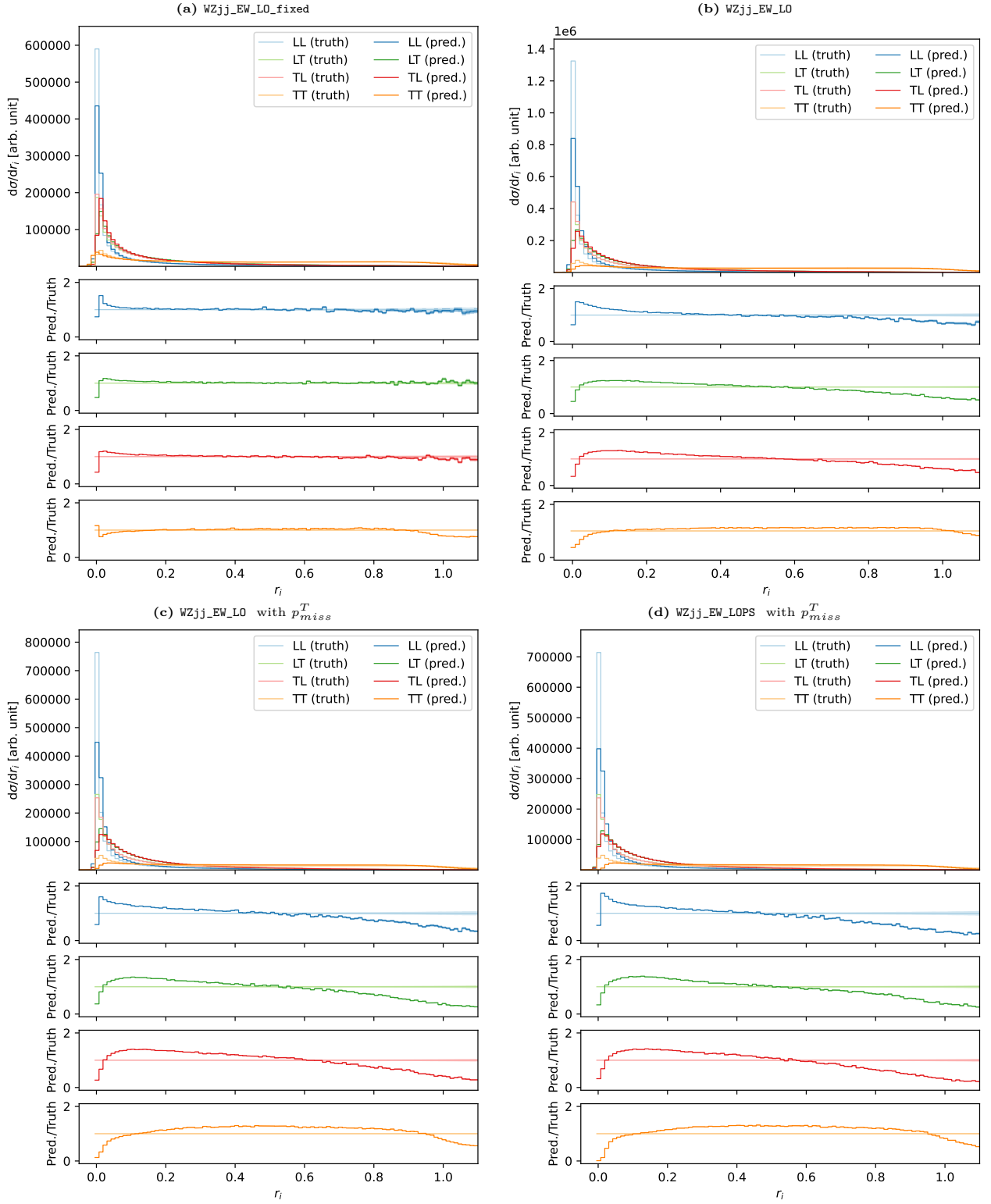


Figure 5.29: Predicted and true distribution of the polarization fractions r_{LL} , r_{TL} , r_{LT} and r_{TT} of the $WZjj$ -EW production samples for the MLP.

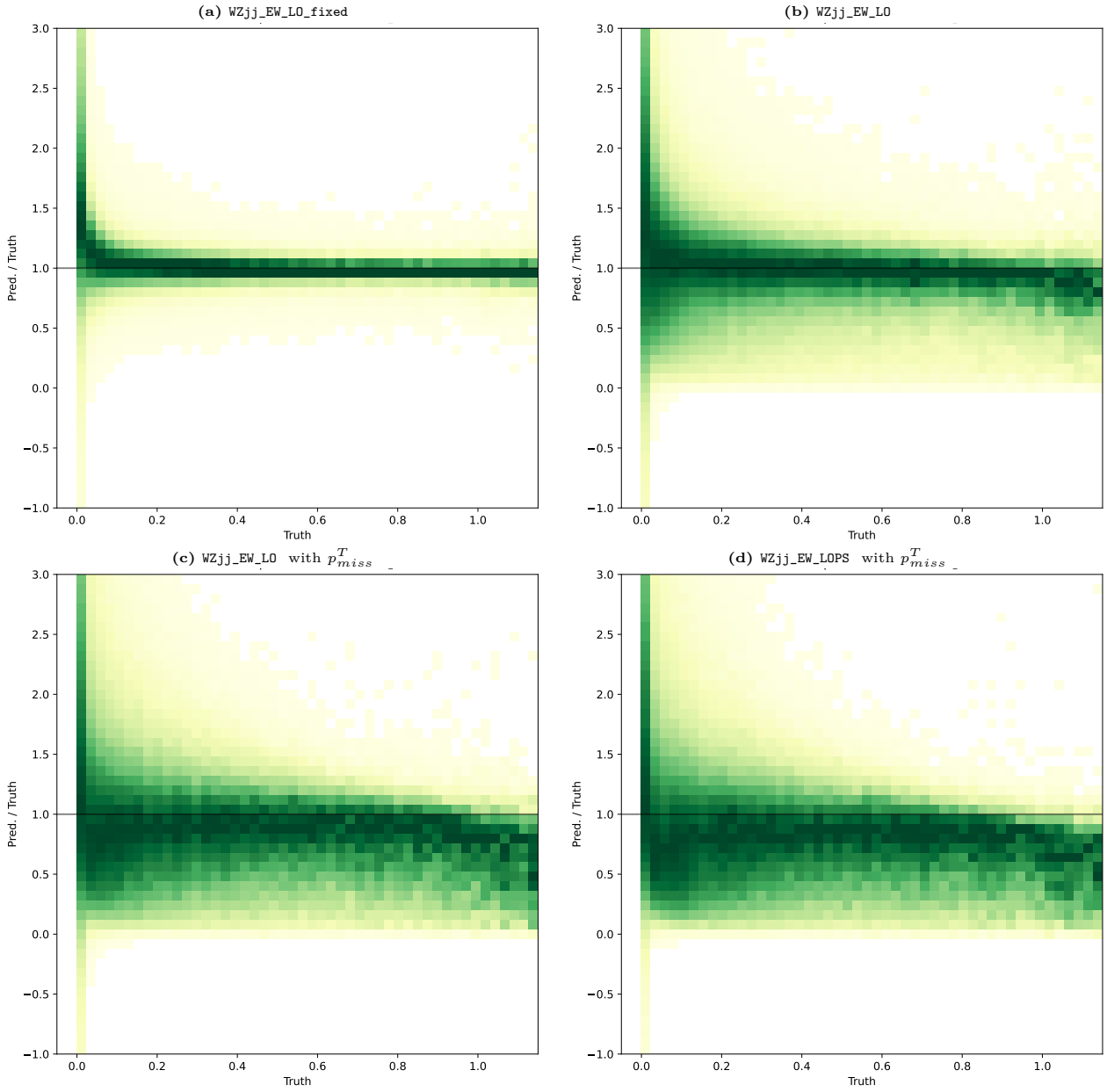


Figure 5.30: Relative prediction error of $r_{LL, val}$ of the $WZjj$ -EW production samples for the MLP.

Table 5.6: Mean Squared Error of predicted and true polarization fractions of the validation dataset for the MLP and the Transformer for $WZjj$ -EW, calculated for all polarization fractions r_{LL} , r_{TL} , r_{LT} and r_{TT} (MSE_{val}) as well as for only the double-longitudinal polarization fraction r_{LL} ($MSE_{val,LL}$).

Sample	MLP	
	MSE_{val}	$MSE_{val,LL}$
WZjj_EW_LO_fixed	0.001 071 $\pm 0.000\ 073$	0.000 481 $\pm 0.000\ 053$
WZjj_EW_LO	0.009 161 $\pm 0.000\ 042$	0.003 468 $\pm 0.000\ 025$
WZjj_EW_LO with p_{miss}^T	0.013 254 $\pm 0.000\ 034$	0.006 250 $\pm 0.000\ 041$
WZjj_EW_LOPS with p_{miss}^T	0.014 45 $\pm 0.000\ 26$	0.006 88 $\pm 0.000\ 20$

6 Summary and Outlook

This thesis studies the polarization tagging of massive vector bosons with neural networks. A Multilayer Perceptron and a Transformer are trained on the kinematics of the final state particles to predict fully-differential polarization fractions of intermediate Z and W bosons in the leptonic decay channel. The NN is trained on simulation data generated with the Monte Carlo event generator Sherpa, which is able to compute the event-wise polarization fractions of the MVB during the event generation. The regression-based formulation of polarization tagging allows for a quantitative comparison of NN predictions and MC truth of the MVB polarization fractions. Therefore, it bears a more detailed understanding as well as a systematic improvement of the polarization tagging. This work extends the formerly proposed regression-based polarization tagging for single $Z + j$ production at LO and LO+PS by multiple aspects. The $Z+jets$ production process is successively investigated at LO, LO+PS and higher orders in QCD. The effect of the parton shower is investigated closely by comparing QED and QCD radiation as well as initial state vs. final state radiation. The gluon radiation before the hard process shows to hinder the correct NN prediction of the polarization fractions most, whereas the inclusion of QED shower and the FSR do not affect the predictions significantly.

The inclusion of higher orders in QCD leads to a substantial change in the distribution of the simulated fractions compared to the ones of the LO+PS samples. The strong effect of NLO corrections in QCD for the event-wise polarization fractions is not yet fully understood. For these processes, the training of the Transformer with all jets resulting from the hard process and the PS shows a strong improvement compared to the training on only the leading jet. Here, substantial information is included in the subsequent jets compared to the LO+PS case, that is necessary to reconstruct the hard process and its intermediate Z boson.

The effect of the phase space is investigated by a training and application of the tagger in the generator vs. fiducial phase space. The tagger shows a high generalizability since the prediction in the generator phase space are highly comparable for a training in the fiducial or generator phase space. This means that the tagger is able to predict almost as good in a phase space which was not included during training.

The $W+jets$ production process is explored at LO and LO+PS. Due to the leptonic decay of the intermediate W^+ boson into a charged lepton and a neutrino, the effect of the prompt neutrino vs. the experimentally available missing transverse momentum is investigated. As expected, the information on the prompt neutrino kinematics allows a better reconstruction of

the intermediate polarized W^+ boson leading to better predictions of the polarization fractions. The production of a W^+Z boson pair in association with two jets at the coupling order of $\mathcal{O}(\alpha_{EW}^6)$ is investigated at LO and LO+PS. The investigation of vector boson scattering is of high interest due to the relevance for testing the electroweak sector and possible new physics. The characteristic signature of VBS events (high invariant mass and a significant rapidity difference of the two jets) allows to increase their contribution in the measurable process. As expected, the experimentally available missing transverse momenta as a proxy of the neutrino kinematics and the inclusion of a parton shower worsen the prediction of the four polarization fractions in the diboson case.

The systematic investigation of all effects required for a unique mapping of phase-space points to polarisation fraction has revealed what constitutes the best possible approximation of the true polarization fraction distribution. All observed differences are fundamental and stem from a lack of physical information, rather than from sub-optimal training or network architecture.

For future studies, the investigation of the effect of higher orders on the polarization fractions should be continued. The strong effect of NLO+PS on the distribution of the MC truth fractions and the prediction performance should be investigated in more detail. The inclusion of higher orders is expected to be more similar to the real data. Thus, the training on higher order samples is beneficial for the application of the tagger later. The generalizability of the tagger to different phase spaces is a very promising result for its application, since the tagger is expected to be trained on simulation data in generator phase space but applied on real data in fiducial phase space. The investigation of the impact of the higher orders and the phase space should be extended to the $W+jets$ and $WZjj$ -EW production process to confirm the results. The effect of hadronization and detector effects like pile-up is not investigated so far, but should be included into future work since this is crucial for the application on real data.

This work only investigated the polarization of massive vector bosons in the leptonic decay channel. Since the simulation of polarized MVB cross sections in the hadronic decay channel is also possible in Sherpa, the studies should be extended to the polarization tagging of hadronically decaying massive vector bosons.

The question remains how exactly the tagger could be applied on real data. First investigations showed that this approach allows the best possible separation of the polarization states. So far, the polarized samples were produced by fitting theoretically derived templates to observables of the unpolarized sample. It is now possible to use the predicted longitudinal and transverse polarization fraction as additional event weights to ultimately obtain polarized samples. Since the tagger is trained on one specific process, the combination with further taggers trained on different processes as well as NN classifiers, which identify the process of the event, could be possible. In this case, each tagger determines the polarization fractions for one possible process, and the result of the most probable process according to the classifier is provided.

A Mathematical Definitions

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (\text{A.1})$$

$$\gamma^0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad \gamma^1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \quad (\text{A.2})$$

$$\gamma^2 = \begin{pmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & i & 0 & 0 \\ -i & 0 & 0 & 0 \end{pmatrix} \quad \gamma^3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$
$$\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad (\text{A.3})$$

B Samples and Sherpa runcards

This section presents an overview on the samples generated with Sherpa for the for the investigated $Z+jets$, $W+jets$ and $WZjj$ -EW production processes (see Table B.1). It also shows the runcards of the event generation, exemplary for the single boson and diboson production at LO+PS.

Table B.1: Description of the datasets generated with Sherpa. The hard process is calculated at leading order (LO) or next-to-leading order (NLO) in QCD. When matched with the parton shower algorithm (PS), events differ in whether gluons (QCD) or photons (QED) are radiated in the PS and whether the showering takes place before (initial state radiation, ISR) or after the hard process (final state radiation, FSR). The momentum of the incoming quark in the hard process can be restricted to positive beam direction ($p_{z,q} > 0$).

Label	Order	Process	Parton Shower				Further features
			QCD	QED	ISR	FSR	
Zj_LO_noPDF	LO	$dg \rightarrow Zd$					
Zj_LO_s	LO	$qg \rightarrow Zq$					
Zj_LO_s_qposz	LO	$qg \rightarrow Zq$					$p_{z,q} > 0$
Zj_LO	LO	$pp \rightarrow Zj$					
Zj_LOPS_s	LO+PS	$qg \rightarrow Zq$	✓	✓	✓	✓	
Zj_LOPS	LO+PS	$pp \rightarrow Zj$	✓	✓	✓	✓	
Zj_LOPS_noQED	LO+PS	$pp \rightarrow Zj$	✓	✗	✓	✓	
Zj_LOPS_noFSR	LO+PS	$pp \rightarrow Zj$	✓	✓	✓	✗	
Zj_LOPS_noISR	LO+PS	$pp \rightarrow Zj$	✓	✓	✗	✓	
Zj_LOPS_s_qposz_noISR	LO+PS	$qg \rightarrow Zq$	✓	✓	✗	✓	$p_{z,q} > 0$
Zj_NLO_PS	NLO+PS	$pp \rightarrow Zj$	✓	✓	✓	✓	
Zj_NLO_2j_LO_2j_PS	NLO+PS	$pp \rightarrow Zj(jj)$	✓	✓	✓	✓	
Wj_LO_noPDF	LO	$ug \rightarrow W^+d$					
Wj_LO_s_qposz	LO	$qg \rightarrow W^+q'$					$p_{z,q} > 0$
Wj_LO	LO	$pp \rightarrow W^+j$					
Wj_LOPS	LO+PS	$pp \rightarrow W^+j$	✓	✓	✓	✓	
WZjj_EW_LO_fixed	LO	$u\bar{d} \rightarrow W^+Zu\bar{u}$					$p_{z,q} > 0$
WZjj_EW_LO	LO	$pp \rightarrow W^+Zjj$					
WZjj_EW_LOPS	LO+PS	$pp \rightarrow W^+Zjj$	✓	✓	✓	✓	

```

BEAMS: 2212
BEAM_ENERGIES: 6500

ME_GENERATORS:
- Comix

FRAGMENTATION: None
MI_HANDLER: None
BEAM_REMNANTS: false

PDF_LIBRARY: LHAPDFSherpa
PDF_SET: NNPDF31_nlo_as_0118
ALPHAS: {USE_PDF: 1}

EW_SCHEME: Gmu
GF: 0.0000116639
WIDTH_SCHEME: Fixed

MEPS:
  CORE_SCALE: VAR{91.188*91.188}

COMIX_DEFAULT_GAUGE: 0

EVENTS: 200k

PARTICLE_DATA:
  23:
    Mass: 91.1876
    Width: 0
  24:
    Mass: 80.399
    Width: 0

HARD_DECAYS:
  Enabled: true
  Mass_Smearing: 1
  Channels:
    23,13,-13: {Status: 2}
  Pol_Cross_Section:
    Enabled: true
    Reference_System: Lab

PROCESSES:
- 93 93 -> 23 93:
  Order: {QCD: 1, EW: 1}

SELECTORS:
- FastjetFinder:
  Algorithm: antikt
  N: 1
  PTMin: 10.0
  YMax: 5.0

ANALYSIS: Rivet
RIVET:
  --analyses:
    - MC_ZJETS
  --ignore-beams: 1

```

Listing 1: Sherpa runcard for the $Z+jets$ @ $LO+PS$ sample

```

BEAMS: 2212
BEAM_ENERGIES: 6500

ME_GENERATORS:
- Comix

FRAGMENTATION: None
MI_HANDLER: None
BEAM_REMNANTS: false

PDF_LIBRARY: LHAPDFSherpa
PDF_SET: NNPDF31_nlo_as_0118
ALPHAS: {USE_PDF: 1}

EW_SCHEME: Gmu
GF: 0.0000116639
WIDTH_SCHEME: Fixed

MEPS:
  CORE_SCALE: VAR{80.419*80.419}

COMIX_DEFAULT_GAUGE: 0

EVENTS: 200k

PARTICLE_DATA:
  23:
    Mass: 91.1876
    Width: 0
  24:
    Mass: 80.399
    Width: 0

HARD_DECAYS:
  Enabled: true
  Mass_Smearing: 1
  Channels:
    24,14,-13: {Status: 2}
  Pol_Cross_Section:
    Enabled: true
    Reference_System: Lab

PROCESSES:
- 93 93 -> 24 93:
  Order: {QCD: 1, EW: 1}

SELECTORS:
- FastjetFinder:
  Algorithm: antikt
  N: 1
  PTMin: 10.0
  YMax: 5.0

ANALYSIS: Rivet
RIVET:
  --analyses:
    - MC_WJETS
  --ignore-beams: 1

```

Listing 2: Sherpa runcard for the W +jets @LO+PS sample

```

BEAMS: 2212
BEAM_ENERGIES: 6500.0

WIDTH_SCHEME: Fixed

SCALES: VBF

COMIX_DEFAULT_GAUGE: 0

PARTICLE_DATA:
  5:
    Mass: 4.95
    Width: 0.
    Massive: 1
  23:
    Mass: 91.1876
    Width: 2.4952
  24:
    Mass: 80.399
    Width: 2.085

EW_SCHEME: Gmu
GF: 1.166397e-5
PDF_LIBRARY: LHAPDFSherpa
USE_PDF_ALPHAS: 1
PDF_SET: PDF4LHC21_40_pdfas

PSI:
  ITMIN: 15000
  NOPT: 12

HARD_DECAYS:
  Enabled: true
  Channels:
    23,11,-11: {Status: 2}
    23,13,-13: {Status: 2}
    24,12,-11: {Status: 2}
    24,14,-13: {Status: 2}
  Pol_Cross_Section:
    Enabled: true
    Reference_System: COM

PROCESSES:
- 93 93 -> 23 24 93 93:
  Order: {QCD: 0, EW: 4}

SELECTORS:
- FastjetSelector:
  Expression: Mass(p[4]+p[5])>110
  Algorithm: antikt
  N: 2
  PTMin: 15.0
  DR: 0.4

ANALYSIS: Rivet
RIVET:
  --analyses:
    - MC_WZJETS
  --ignore-beams: 1

```

Listing 3: Sherpa runcard for the $WZjj$ @LO+PS sample

List of Figures

3.1	General architecture of a Multilayer Perceptron with input layer (green) of n input variables x_n , $h = 5$ hidden layers (purple) of equal neuron number m and a output layer (grey) of k output variables y_k . The output of the hidden layer neurons is denoted by z_m^h and a weight w_{ij}^h is assigned to each connections between the neurons.	14
3.2	General Transformer encoder architecture as used in this thesis. The input is first preprocessed before it is propagated through N stacked layers, each consisting of a multi-head self-attention sub-layer and a fully connected feed-forward layer. Each of the sub-layers is equipped with a residual connection and a layer normalization. The illustration is based on [22] and [24].	17
4.1	The MC event generation steps in Sherpa: hard scattering process (red), initial and final state parton shower (blue), hadronization (light green), hadronic decays (dark green), radiation of photons (yellow) and underlying event (violet) [47].	25
4.2	Feynman diagrams of the hard scattering process in $Z+jet$ and $W+jet$ at leading order (LO), leading order with IS and FS parton shower (LO+PS) and virtual corrections at next-to-leading-order (NLO) in QCD.	27
4.3	Double-resonant Feynman diagram of the electroweak $W^\pm Z$ boson pair production in association with two jets at leading order (a) and the corresponding vector boson scattering contributions of triple- and quartic MVB self coupling (b) as well as the MVB with Higgs coupling (c).	28
5.1	Distributions of the polarization fractions r_L (blue) and r_T (green) for NN predictions (pred.) and MC true value (truth) with absolute number of events in the validation dataset (top), relative difference of predicted and true distribution for r_L (middle) and relative difference of predicted and true distribution r_T (bottom).	33
5.2	Distribution of the relative prediction error $RPE = \frac{r_{L,pred.}}{r_{L,truth}}$ (y-axis) of NN predictions $r_{L,pred.}$ and MC truth $r_{L,truth}$ of the longitudinal polarization fraction r_L for the validation dataset in dependence of the MC true value $r_{L,truth}$ (x-axis) (a). This is then plotted normalized by the number of events in the $r_{L,truth}$ bin (b).	35
5.3	Predicted and true distribution of r_L and r_T of the MLP for $Z+jets$ @ LO.	48
5.4	Predicted and true distribution of r_L and r_T of the Transformer for $Z+jets$ @ LO.	49
5.5	Predicted and true distribution of r_L and r_T of the MLP trained $Z+jets$ @ LO+PS.	50

5.6	Predicted and true distribution of r_L and r_T of the Transformer for $Z+jets$ @ LO+PS.	51
5.7	Predicted and true distribution of r_L and r_T of the Transformer trained all jets for $Z+jets$ @ LO+PS.	52
5.8	Predicted and true distribution of r_L and r_T of the MLP $Z+jets$ @ NLO.	53
5.9	Predicted and true distribution of r_L and r_T of the Transformer for $Z+jets$ @ NLO.	53
5.10	Predicted and true distribution of r_L and r_T of the Transformer trained on all jets for $Z+jets$ @ NLO.	54
5.11	Relative prediction error of $r_{L,val}$ of the MLP for $Z+jets$ @ LO.	55
5.12	Relative prediction error of $r_{L,val}$ of the Transformer for $Z+jets$ @ LO.	56
5.13	Relative prediction error of $r_{L,val}$ of the MLP trained $Z+jets$ @ LO+PS.	57
5.14	Relative prediction error of $r_{L,val}$ of the Transformer for $Z+jets$ @ LO+PS.	58
5.15	Relative prediction error of $r_{L,val}$ of the Transformer trained all jets for $Z+jets$ @ LO+PS.	59
5.16	Relative prediction error of $r_{L,val}$ of the MLP $Z+jets$ @ NLO.	60
5.17	Relative prediction error of $r_{L,val}$ of the Transformer for $Z+jets$ @ NLO.	60
5.18	Relative prediction error of $r_{L,val}$ of the Transformer trained on all jets for $Z+jets$ @ NLO.	61
5.19	Predicted and true distribution of r_L and r_T of the MLP for $pp \rightarrow Z + j + X$ @ LO+PS. The NN is trained in the generator phase space (train: generator) or in the fiducial phase space (train: fiducial) and afterwards validated in the generator phase space (pred.: generator) or the fiducial phase space (pred.: fiducial).	67
5.20	Predicted and true distribution of r_L and r_T of the Transformer for $pp \rightarrow Z + j + X$ @ LO+PS.	68
5.21	Predicted and true distribution of r_L and r_T of the Transformer trained on all jets for $pp \rightarrow Z + j + X$ @ LO+PS.	69
5.22	Relative prediction error of $r_{L,val}$ of the MLP for $pp \rightarrow Z + j + X$ @ LO+PS. The NN is trained in the generator phase space (train: generator) or in the fiducial phase space (train: fiducial) and afterwards validated in the generator phase space (pred.: generator) or the fiducial phase space (pred.: fiducial).	70
5.23	Relative prediction error of $r_{L,val}$ of the Transformer for $pp \rightarrow Z + j + X$ @ LO+PS.	71
5.24	Relative prediction error of $r_{L,val}$ of the Transformer trained on all jets for $pp \rightarrow Z + j + X$ @ LO+PS.	72
5.25	Predicted and true distribution of r_L and r_T of the MLP for $W+jets$	79
5.26	Predicted and true distribution of r_L and r_T of the Transformer for $W+jets$	80
5.27	Relative prediction error of $r_{L,val}$ of the MLP for $W+jets$	81
5.28	Relative prediction error of $r_{L,val}$ of the Transformer for $W+jets$	82
5.29	Predicted and true distribution of the polarization fractions r_{LL} , r_{TL} , r_{LT} and r_{TT} of the $WZjj$ -EW production samples for the MLP.	87
5.30	Relative prediction error of $r_{LL,val}$ of the $WZjj$ -EW production samples for the MLP.	88

List of Tables

2.1	Properties of the bosons ($S = 0, 1$) in the SM with mass m , width Γ , electric charge Q , weak isospin I_3^W and spin S . Data taken from [6].	4
2.2	Properties of the fermions ($S = 1/2$) in the SM with electric charge Q , weak isospin I_3^W and spin S	4
3.1	Hyperparameter optimization search space of the MLP.	20
3.2	Hyperparameter optimization search space of the Transformer.	20
4.1	Selection criteria of the $Z+jets$ and $W+jets$ samples generated with Sherpa for the generator and the fiducial phase space according to [3].	30
4.2	Selection criteria of the $WZjj$ -EW datasets generated with Sherpa in the generator phase space according to [61].	30
5.1	Mean Squared Error of predicted and true polarization fractions of the validation dataset for the MLP and the Transformer for $Z+jets$, calculated for both longitudinal and transverse polarization fraction r_L and r_T (MSE_{val}) as well as for only the longitudinal polarization fraction r_L ($MSE_{val,L}$).	62
5.2	Mean Squared Error of predicted and true polarization fractions of the validation dataset for the Transformer trained on all jets for $Z+jets$	63
5.3	Mean Squared Error of predicted and true polarization fractions of the validation dataset for the MLP and the Transformer for $pp \rightarrow Zj$ @ LO. MLP and Transformer are trained on generator (train: generator) or fiducial (train: fiducial) phase space and validated by predicting r_L and r_T on events in generator (pred.: generator) or fiducial (pred.: fiducial) phase space.	73
5.4	Mean Squared Error of predicted and true polarization fractions of the validation dataset for the Transformer trained on all jets for $pp \rightarrow Zj$ @ LO.	73
5.5	Mean Squared Error of predicted and true polarization fractions of the validation dataset for the MLP and the Transformer for $W+jets$	83
5.6	Mean Squared Error of predicted and true polarization fractions of the validation dataset for the MLP and the Transformer for $WZjj$ -EW, calculated for all polarization fractions r_{LL} , r_{TL} , r_{LT} and r_{TT} (MSE_{val}) as well as for only the double-longitudinal polarization fraction r_{LL} ($MSE_{val,LL}$).	89

B.1 Description of the datasets generated with Sherpa. The hard process is calculated at leading order (LO) or next-to-leading order (NLO) in QCD. When matched with the parton shower algorithm (PS), events differ in whether gluons (QCD) or photons (QED) are radiated in the PS and whether the showering takes place before (initial state radiation, ISR) or after the hard process (final state radiation, FSR). The momentum of the incoming quark in the hard process can be restricted to positive beam direction ($p_{z,q} > 0$).	95
---	----

Bibliography

- [1] The ATLAS Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,” *Physics Letters B*, vol. 716, no. 1, pp. 1–29, Sep. 2012. DOI: 10.1016/j.physletb.2012.08.020
- [2] The CMS Collaboration, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,” *Physics Letters B*, vol. 716, no. 1, pp. 30–61, Sep. 2012. DOI: 10.1016/j.physletb.2012.08.021
- [3] M. Grossi, M. Incudini, M. Pellen, and G. Pelliccioli, “Amplitude-assisted tagging of longitudinally polarised bosons using wide neural networks,” *The European Physical Journal C*, vol. 83, no. 8, 2023. DOI: 10.1140/epjc/s10052-023-11931-y
- [4] J.-E. Nitschke, “Observation of Longitudinal-Longitudinal Interactions at High p_T^Z and Studies of the Radiation Amplitude Zero Effect in $W^\pm Z$ Events with the ATLAS Detector at the Large Hadron Collider,” Ph.D. dissertation, Technischen Universität Dresden, 2024. [Online]. Available: <https://repository.cern/records/eb2kp-nva28>
- [5] C. Bittrich, “Evidence for Scattering of Electroweak Gauge Bosons in the $W^\pm Z$ Channel with the ATLAS Detector at the Large Hadron Collider,” Ph.D. dissertation, Technische Universität Dresden, Jul. 2020. [Online]. Available: <https://repository.cern/records/za3hy-z5z39>
- [6] The Particle Data Group Collaboration, “Review of Particle Physics,” *Phys. Rev. D*, vol. 110, p. 030001, 3 Aug. 2024. DOI: 10.1103/PhysRevD.110.030001
- [7] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson, “Experimental Test of Parity Conservation in Beta Decay,” *Phys. Rev.*, vol. 105, pp. 1413–1415, 4 Feb. 1957. DOI: 10.1103/PhysRev.105.1413
- [8] S. Weinberg, “A Model of Leptons,” *Phys. Rev. Lett.*, vol. 19, pp. 1264–1266, 1967. DOI: 10.1103/PhysRevLett.19.1264
- [9] A. Salam, “Weak and Electromagnetic Interactions,” *Conf. Proc. C*, vol. 680519, pp. 367–377, 1968. DOI: 10.1142/9789812795915_0034
- [10] F. Englert and R. Brout, “Broken Symmetry and the Mass of Gauge Vector Mesons,” *Phys. Rev. Lett.*, vol. 13, pp. 321–323, 9 1964. DOI: 10.1103/PhysRevLett.13.321

- [11] P. W. Higgs, “Broken Symmetries and the Masses of Gauge Bosons,” *Phys. Rev. Lett.*, vol. 13, J. C. Taylor, Ed., pp. 508–509, 1964. DOI: 10.1103/PhysRevLett.13.508
- [12] J. Goldstone, A. Salam, and S. Weinberg, “Broken Symmetries,” *Phys. Rev.*, vol. 127, pp. 965–970, 3 Aug. 1962. DOI: 10.1103/PhysRev.127.965
- [13] Y. Nambu, “Quasi-Particles and Gauge Invariance in the Theory of Superconductivity,” *Phys. Rev.*, vol. 117, pp. 648–663, 3 Feb. 1960. DOI: 10.1103/PhysRev.117.648
- [14] M. Hoppe, M. Schönherr, and F. Siegert, “Polarised cross sections for vector boson production with SHERPA,” 2023. arXiv: 2310.14803 [hep-ph].
- [15] A. Ballestrero, E. Maina, and G. Pelliccioli, “ W boson polarization in vector boson scattering at the LHC,” *Journal of High Energy Physics*, vol. 2018, no. 3, Mar. 2018. DOI: 10.1007/jhep03(2018)170
- [16] P. Steinbach, *Lecture notes in Machine Learning in Natural Sciences*, Nov. 2023.
- [17] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction,” *Math. Intell.*, vol. 27, pp. 83–85, Nov. 2004. DOI: 10.1007/BF02985802
- [18] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning,” *Genet Program Evolvable Mach*, vol. 19, pp. 305–307, 2018. DOI: 10.1007/s10710-017-9314-z
- [19] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989, ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8
- [20] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, “Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark,” 2022. arXiv: 2109.14545 [cs.LG].
- [21] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” 2017. arXiv: 1412.6980 [cs.LG].
- [22] A. Vaswani et al., “Attention Is All You Need,” 2023. DOI: 10.48550/arXiv.1706.03762 arXiv: 1706.03762 [cs].
- [23] T. Lin, Y. Wang, X. Liu, and X. Qiu, “A Survey of Transformers,” 2021. arXiv: 2106.04554 [cs.LG].
- [24] J. Brehmer et al., “A Lorentz-Equivariant Transformer for All of the LHC,” *SciPost Phys.*, no. arXiv:2411.00446, Dec. 22, 2024. arXiv: 2411.00446 [hep-ph].
- [25] The ATLAS Collaboration, “Flavour tagging with graph neural networks with the ATLAS detector,” CERN, Geneva, Tech. Rep., 2023. arXiv: 2306.04415. [Online]. Available: <https://cds.cern.ch/record/2860610>,
- [26] D. Navon and A. M. Bronstein, “Transformer Vs. MLP-Mixer: Exponential Expressive Gap For NLP Problems,” 2022. arXiv: 2208.08191 [cs.CL].

-
- [27] L. Chen, B. Peng, and H. Wu, “Theoretical limitations of multi-layer Transformer,” 2024. arXiv: 2412.02975 [cs.LG].
 - [28] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” 2019. arXiv: 1711.05101 [cs.LG].
 - [29] M. X. Chen et al., “The best of both worlds: Combining recent advances in neural machine translation,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 76–86. DOI: 10.18653/v1/P18-1008
 - [30] Q. Wang et al., “Learning Deep Transformer Models for Machine Translation,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1810–1822. DOI: 10.18653/v1/P19-1176
 - [31] T. Q. Nguyen and J. Salazar, “Transformers without Tears: Improving the Normalization of Self-Attention,” 2019. DOI: 10.5281/ZENODO.3525484
 - [32] “OPTIMA: an Optimization Platform for Tuning Input Variables and Model Parameters,” Accessed: Apr. 20, 2026. [Online]. Available: <https://pypi.org/project/optima-ml/>
 - [33] E. Bachmann, “Polarization studies in same-sign W-boson pair production at the LHC with the ATLAS detector,” M.S. thesis, Technische Universität Dresden, 2023. [Online]. Available: <https://repository.cern/records/6gx3x-aqh11>
 - [34] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel: Omnipress, 2010, pp. 807–814. DOI: 10.5555/3104322.3104425
 - [35] D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs),” 2023. arXiv: 1606.08415 [cs.LG].
 - [36] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for Activation Functions,” 2017. arXiv: 1710.05941 [cs.NE].
 - [37] S. Hayou, A. Doucet, and J. Rousseau, “On the Selection of Initialization and Activation Function for Deep Neural Networks,” 2018. arXiv: 1805.08266 [stat.ML].
 - [38] The CMS Collaboration, “Measurement of the Polarization of W Bosons with Large Transverse Momenta in $W+jets$ Events at the LHC,” *Phys. Rev. Lett.*, vol. 107, p. 021 802, 2 Jul. 2011. DOI: 10.1103/PhysRevLett.107.021802

- [39] The ATLAS Collaboration, “Measurement of the polarisation of W bosons produced with large transverse momentum in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS experiment,” *The European Physical Journal C*, vol. 72, no. 5, May 2012. DOI: 10.1140/epjc/s10052-012-2001-6
- [40] The ATLAS Collaboration, “Measurement of $W^\pm Z$ production cross sections and gauge boson polarisation in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector,” *The European Physical Journal C*, vol. 79, no. 6, Jun. 2019. DOI: 10.1140/epjc/s10052-019-7027-6
- [41] The CMS Collaboration, “Measurement of the inclusive and differential WZ production cross sections, polarization angles, and triple gauge couplings in pp collisions at $\sqrt{s} = 13$ TeV,” *Journal of High Energy Physics*, vol. 2022, no. 7, Jul. 2022. DOI: 10.1007/jhep07(2022)032
- [42] C. Bittrich, “Study of Polarization Fractions in the Scattering of Massive Gauge Bosons $W^\pm Z \rightarrow W^\pm Z$ with the ATLAS Detector at the Large Hadron Collider,” M.S. thesis, Technische Universität Dresden, May 2015. [Online]. Available: <https://repository.cern/records/sxktp-sw975>
- [43] Z. Bern et al., “Left-handed W bosons at the LHC,” *Physical Review D*, vol. 84, no. 3, Aug. 2011. DOI: 10.1103/physrevd.84.034008
- [44] A. Ballestrero, E. Maina, and G. Pelliccioli, “Polarized vector boson scattering in the fully leptonic WZ and ZZ channels at the LHC,” *Journal of High Energy Physics*, vol. 2019, no. 9, Sep. 2019. DOI: 10.1007/jhep09(2019)087
- [45] A. S. Tegetmeier, “Multivariate Analysis of Polarized Vector Boson Scattering of a W and Z Boson with the ATLAS Detector,” M.S. thesis, Technische Universität Dresden, 2019. [Online]. Available: <https://repository.cern/records/hyhac-h0c28>
- [46] S. Brass, C. Fleper, W. Kilian, J. Reuter, and M. Sekulla, “Transversal modes and Higgs bosons in electroweak vector-boson scattering at the LHC,” *The European Physical Journal C*, vol. 78, no. 11, Nov. 2018. DOI: 10.1140/epjc/s10052-018-6398-4
- [47] E. Bothmann et al., “Event generation with Sherpa 3,” *Journal of High Energy Physics*, vol. 2024, Dec. 2024. DOI: 10.1007/JHEP12(2024)156
- [48] C. Bierlich et al., “A comprehensive guide to the physics and usage of PYTHIA 8.3,” *SciPost Physics Codebases*, Nov. 2022. DOI: 10.21468/SciPostPhysCodeb.8
- [49] M. Bähr et al., “Herwig++ Physics and Manual,” *European Physical Journal C*, vol. 58, pp. 639–707, Dec. 2008. DOI: 10.1140/epjc/s10052-008-0798-9
- [50] A. M. Burger, “Measurement of Vector Boson polarisation in WZ production at the LHC with the ATLAS detector,” Ph.D. dissertation, Université Grenoble Alpes, 2018. [Online]. Available: <https://theses.hal.science/tel-02132816v1>

-
- [51] M. Hoppe, “Implementation and phenomenology of polarized cross sections in the simulation of vector boson production with the SHERPA event generator,” M.S. thesis, Technische Universität Dresden, 2023. [Online]. Available: <https://repository.cern/records/smpmw-xyt24>
 - [52] The Sherpa Collaboration. “Sherpa User Manual,” Accessed: Apr. 1, 2026. [Online]. Available: <https://sherpa-team.gitlab.io/sherpa/master/manual/parameters/hard-decays>
 - [53] The ATLAS Collaboration, “Observation of gauge boson joint-polarisation states in $W^\pm Z$ production from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector,” *Physics Letters B*, vol. 843, p. 137 895, Aug. 2023. DOI: 10.1016/j.physletb.2023.137895
 - [54] C. Bierlich et al., “Robust Independent Validation of Experiment and Theory: Rivet version 4 release note,” 2024. arXiv: 2404.15984 [hep-ph].
 - [55] A. Buckley et al., “Rivet user manual,” *Computer Physics Communications*, vol. 184, no. 12, pp. 2803–2819, 2013. DOI: 10.1016/j.cpc.2013.05.021
 - [56] T. Gleisberg and S. Hoeche, “Comix, a new matrix element generator,” *JHEP*, vol. 12, p. 039, 2008. DOI: 10.1088/1126-6708/2008/12/039 arXiv: 0808.3674 [hep-ph].
 - [57] F. Krauss, R. Kuhn, and G. Soff, “AMEGIC++ 1.0: A Matrix element generator in C++,” *JHEP*, vol. 02, p. 044, 2002. DOI: 10.1088/1126-6708/2002/02/044 arXiv: hep-ph/0109036.
 - [58] F. Cascioli, P. Maierhofer, and S. Pozzorini, “Scattering Amplitudes with Open Loops,” *Phys. Rev. Lett.*, vol. 108, p. 111 601, 2012. DOI: 10.1103/PhysRevLett.108.111601 arXiv: 1111.5206 [hep-ph].
 - [59] M. Cacciari, G. P. Salam, and G. Soyez, “The anti- k_t jet clustering algorithm,” *Journal of High Energy Physics*, vol. 2008, no. 04, p. 063, Apr. 2008. DOI: 10.1088/1126-6708/2008/04/063
 - [60] A. Buckley et al., “LHAPDF6: parton density access in the LHC precision era,” *Eur. Phys. J. C*, vol. 75, p. 132, 2015. DOI: 10.1140/epjc/s10052-015-3318-8 arXiv: 1412.7420 [hep-ph].
 - [61] The ATLAS Collaboration, “Observation of electroweak $W^\pm Z$ boson pair production in association with two jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector,” *Physics Letters B*, vol. 793, pp. 469–492, Jun. 2019. DOI: 10.1016/j.physletb.2019.05.012

Danksagung

An dieser Stelle möchte ich die Gelegenheit nutzen, um Danke zu sagen:

- Prof. Frank Siegert für die Betreuer meiner Masterarbeit und die Leitung dieser großartigen Arbeitsgruppe voller Respekt, Unterstützung und Miteinander,
- Erik Bachmann und Mareen Hoppe für die Begleitung des Projektes und ihr großes Engagement, welches mir als starkes Vorbild dient,
- Joany Manjarres und Angela Burger für die Möglichkeit, die Arbeitsgruppe in Toulouse kennenzulernen und so herzlich aufgenommen zu werden,
- Lea Baumann und Florens Förster für die schönen und bereichernden gemeinsamen Momente im Büro,
- Antonia Bähr für die tolle Freundschaft, die stetige Unterstützung und den unumstößlichen Glauben an mich,
- Nick Maryshchak für die gesamte gemeinsame Studienzeit und dass du immer für mich da bist.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit im Rahmen der Betreuung am Institut für Kern- und Teilchenphysik ohne unzulässige Hilfe Dritter verfasst habe und alle verwendeten Quellen als solche gekennzeichnet habe.

KI-basierte Sprachmodelle habe ich ausschließlich zur Unterstützung bei der stilistischen Überarbeitung schriftlicher Texte verwendet. Alle wissenschaftlichen Inhalte, Interpretationen und Schlussfolgerungen stammen ausschließlich von mir.

Ort, Datum

Unterschrift