GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Fakultät für
Physik

Master's Thesis

# Improving Event Reconstruction for Dileptonic Decays of Top-Quark Pairs Using Machine Learning

prepared by

**Siemen Henning Aulich**

from Emden

at the DESY Hamburg

| | |
|---|---|
| **Course:** | M.Phy.1610: Development and Realization of Scientific Projects in Theoretical Physics |
| **Course period:** | 7th April 2025 until 4th December 2026 |
| **First referee:** | Prof. Dr. Stan Lai |
| **Second referee:** | Dr. Katharina Behr |

# Abstract

Many of the recent highlights in particle physics research are related to top-quark physics. These include both the stringent tests of spin correlations and quantum effects in pairs of top quarks ($t\bar{t}$), and the observation of a possible quasi-bound state resonance in the $t\bar{t}$ invariant mass spectrum. The latter presents a very exciting opportunity to extend our current understanding of non-relativistic Quantum Chromodynamics. Both effects are predominantly studied in the dilepton decay channel of the top quark in a mass range close to the $t\bar{t}$ production threshold.

The dilepton final state contains two neutrinos, besides two charged leptons and two $b$-quark jets. Probing this system requires a precise reconstruction of the top quarks, which is complicated by the presence of the two neutrinos. While analytical regression strategies primarily focus on inferring the neutrino momenta, the problem of correctly assigning $b$-jets to their parent top quarks remains largely understudied. However, many of the sensitive variables used in $t\bar{t}$ precision measurements depend critically on the correct assignment of the jets.

Inspired by the success of machine learning architectures in tackling the assignment challenge for hadronic decay channels, this work investigates using a transformer model for the dilepton channel. An architecture specifically tailored to this channel's topology is shown to outperform all existing methods in assigning reconstructed particles to their parent partons, improving the resolution of relevant physics observables. Furthermore, it is investigated how these efforts can be combined with neutrino regression methods to offer a full reconstruction pipeline. Applying these methods to existing and upcoming analyses promises further enhancements in the sensitivity and precision of searches and measurements alike.

**Keywords:** Top Quark, Dileptonic Decay, Event Reconstruction, Machine Learning, Transformer Networks

# Contents

**1. Introduction 1**

**2. Theoretical Background 3**

2.1. The Standard Model of Particle Physics . . . . . . . . . . . . . . . . . . . . . 3

2.1.1. Electromagnetic Interaction . . . . . . . . . . . . . . . . . . . . 4

2.1.2. Weak Interaction . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.1.3. Strong Interaction . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.2. Top Quark Physics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.2.1. Spin Correlations in Top-Quark Pairs . . . . . . . . . . . . . . . 7

2.2.2. Top-Quark Pair Bound State Effects . . . . . . . . . . . . . . . . 8

2.2.3. Dileptonic Decay Channel . . . . . . . . . . . . . . . . . . . . . 8

2.2.4. Spin Sensitive Angular Variables . . . . . . . . . . . . . . . . . . 9

**3. Machine Learning in Particle Physics 10**

3.1. Supervised Learning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

3.1.1. Monte Carlo Training Data . . . . . . . . . . . . . . . . . . . . . 10

3.1.2. Training . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

3.1.3. Validation and Testing . . . . . . . . . . . . . . . . . . . . . . . 11

3.2. Neural Networks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

3.2.1. Dense Neural Network . . . . . . . . . . . . . . . . . . . . . . . 12

3.2.2. Recurrent Neural Networks . . . . . . . . . . . . . . . . . . . . . 13

3.2.3. Transformer . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

3.2.4. Training Neural Networks . . . . . . . . . . . . . . . . . . . . . . 15

**4. Analysis and Reconstruction 16**

4.1. Data and Simulation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

4.1.1. Simulation of Dileptonic Top-Quark Pair Events . . . . . . . . . . . 16

4.1.2. Object Definition . . . . . . . . . . . . . . . . . . . . . . . . . . 16

4.1.3. Event Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

4.2. Reconstruction of Top-Quark Pairs . . . . . . . . . . . . . . . . . . . . . 17

4.2.1. Neutrino Regression Methods . . . . . . . . . . . . . . . . . . . . 17

4.2.2. Jet Assignment Methods . . . . . . . . . . . . . . . . . . . . . . 18

**5. Machine Learning Based Jet Assignment 20**

5.1. Neural Network Architectures . . . . . . . . . . . . . . . . . . . . . . . . 20

5.1.1. FeatureConcatTransformer . . . . . . . . . . . . . . . . . . . . . 20

5.1.2. CrossAttentionTransformer . . . . . . . . . . . . . . . . . . . . . 22

5.2. Training Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

5.2.1. Metrics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

5.3. Hyperparameter Optimization . . . 23
5.4. Performance Comparison . . . 25
5.4.1. High Level Input Features . . . 26
5.4.2. Feature Evaluation . . . 27
5.4.3. Summary . . . 29

**6. Method Evaluation 30**
6.1. Accuracy Metrics . . . 30
6.1.1. Variable Dependent Performance . . . 31
6.2. Impact on Physics Observables . . . 33
6.3. Near-Threshold Optimized Model . . . 36

**7. Conclusion 38**
7.1. Summary . . . 38
7.2. Outlook . . . 39

**A. Additional Plots 40**
A.1. Model and Hyperparameter Studies . . . 40
A.2. Feature Distributions . . . 43
A.3. Performance Evaluation Plots . . . 45

# 1. Introduction

The top quark, being the heaviest known elementary particle, plays a crucial role in the Standard Model of particle physics. Due to its high mass, it stands out in two ways; for one, as having the highest Yukawa-coupling and secondly, as being the shortest lived elementary particle. While, the Yukawa-coupling makes it very interesting for Higgs physics, the short lifetime of the top quark presents it as the only window allowing us to look at the bare coupling of quarks. The top quark decays almost exclusively into a $W$ boson and a $b$-quark. These properties make processes involving top quarks as some of the most studied at the Large Hadron Collider (Lhc) at Cern.

Top quarks are predominantly produced in pairs ($t\bar{t}$) via the strong interaction in proton-proton collisions at the Lhc. Each $W$ boson coming from the top decay can decay either leptonically (into a charged lepton and neutrino) or hadronically (into a pair of quarks). The dileptonic decay channel, where both $W$ bosons decay leptonically, results in a final state with two charged leptons, two $b$-jets, and two neutrinos. This channel, while having a lower branching ratio compared to other decay modes, offers a cleaner experimental signature due to the presence of leptons and reduced hadronic activity. Leptons in particular can be reconstructed with efficiency in a hadron collider environment. However, the presence of two neutrinos in the final state poses significant challenges for event reconstruction, as they evade detection, leading to missing transverse energy ($\not{E}_{\mathrm{T}}$) in the event.

Due to its clean signature and sensitivity to various top-quark properties, the dileptonic $t\bar{t}$ channel is widely used for precision measurements of the top properties and spin correlations in particular. Two recent highlights in top-quark physics are the tests of spin correlations and quantum entanglement in pairs of top quarks [1], and the observation of a possible quasi-bound state resonance in the $t\bar{t}$ invariant mass spectrum [2, 3]. Both effects are predominantly studied in the dilepton decay channel of the top quark in a mass range close to the on-shell production threshold of two top quarks $m(t\bar{t}) \gtrsim 2 \cdot m_t \approx 350\,\mathrm{GeV}$. Compared to hadronic decays, the leptons allow a more accurate probing of the top quarks spin by preserving a larger fraction of the spin information [4, 5].

Both analyses rely on measurements of angular distributions and correlations between the decay products of the top quarks. Probing this system requires a precise reconstruction of the top quarks from their decay products, which is complicated by the presence of the two neutrinos. While analytical neutrino regression strategies have been studied extensively, the problem of correctly assigning $b$-jets to their parent top quarks remains largely unstudied. In channels with hadronically decaying top quarks, various methods for jet assignment have been developed and employed successfully [6]. However, in the dileptonic channel, the jet assignment problem was often simplified or overlooked due to a focus on neutrino reconstruction. However, many of the sensitive variables used in measurements of spin correlations depend critically on the correct assignment of the jets.

Misassignments degrade the resolution of reconstructed observables, ultimately limiting the precision of measurements.
Inspired by the success of machine-learning architectures in tackling the assignment challenge for hadronic decay channels, this work investigates the use of a transformer model for the dilepton channel. It introduces the necessary theoretical background in Chapter 2 before detailing the use of machine learning in particle physics in Chapter 3. The reconstruction methods are outlined in Chapter 4, followed by the development of the machine learning-based jet assignment in Chapter 5. For this, different transformer architectures are explored and optimized. The performance of the developed methods is evaluated in Chapter 6, where the impact on relevant physics observables is assessed. Finally, Chapter 7 summarizes the findings and discusses potential future directions for further improving dileptonic $t\bar{t}$ event reconstruction.

# 2. Theoretical Background

## 2.1. The Standard Model of Particle Physics

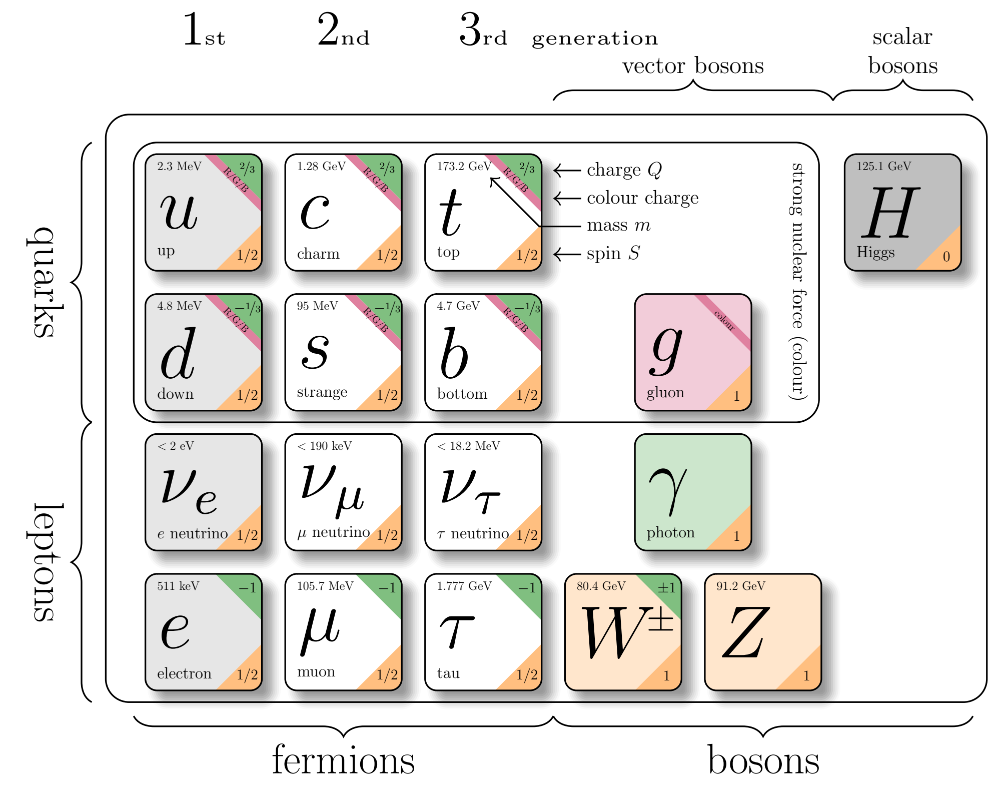


**Figure 2.1.:** The particles of the Standard Model. Particle properties taken from Ref. [7].

The Standard Model (SM) of particle physics is a quantum field theory describing three of the four fundamental interactions of nature—electromagnetic, weak and strong—and the elementary particles on which they act. It is formulated as a renormalizable, Lorentz-invariant gauge theory with gauge group

$$SU(3)_C \times SU(2)_L \times U(1)_Y,$$

and its structure has been established in a series of seminal works [8–13]. A summary of the particle content is shown in Figure 2.1.

The elementary particles fall into two categories, fermions and bosons. Fermions are spin-$\frac{1}{2}$ fields that constitute matter, organised into three generations of quarks and leptons. Quarks carry colour charge and participate in the strong interaction; leptons do not.

Bosons are integer-spin gauge fields mediating the fundamental forces: the photon for electromagnetism, the $W^{\pm}$ and $Z$ bosons for the weak interaction, and eight gluons for the strong interaction. The Higgs boson completes the SM and is responsible for electroweak symmetry breaking (EWSB), giving mass to the weak gauge bosons and, through Yukawa couplings, to the fermions. The SM Lagrangian can be written schematically as

$$\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{fermion}} + \mathcal{L}_{\text{Yukawa}} + \mathcal{L}_{\text{Higgs}},$$

where the first term contains the gauge kinetic terms and self-interactions, the second describes the fermions and gauge covariant derivatives, the third encodes the Yukawa couplings to the Higgs field, and the last term contains the Higgs potential and EWSB mechanism.

Below, the three interaction sectors are briefly summarised, with emphasis on the aspects relevant for top-quark physics and LHC phenomenology.

### 2.1.1. Electromagnetic Interaction

The electromagnetic interaction is described by quantum electrodynamics (QED), the $U(1)_{\text{EM}}$ gauge theory that remains after EWSB. Its massless gauge boson, the photon, couples to electric charge and does not self-interact. Because QED is an Abelian theory, its coupling runs only mildly with energy and remains perturbative at all experimentally accessible scales.

### 2.1.2. Weak Interaction

The weak interaction forms, together with electromagnetism, the electroweak theory based on the gauge group $SU(2)_L \times U(1)_Y$. The vacuum expectation value of the Higgs field spontaneously breaks the symmetry to $U(1)_{\text{EM}}$, giving masses to the weak gauge bosons through the Higgs mechanism [14–16]. The resulting massive $W^{\pm}$ and $Z$ bosons mediate a short-range force.

A defining property of the weak interaction is its *chiral* structure: charged-current interactions couple only to left-handed fermions and right-handed antifermions. This feature, predicted in [17] and confirmed in the Wu experiment [18], leads to maximal parity violation and characteristic angular distributions of weak decay products. The vertex factor of the charged-current interaction is given by

$$\frac{-ig}{2\sqrt{2}}\gamma^{\mu}(1-\gamma^{5}), \tag{2.1}$$

where $g$ is the weak coupling constant, $\gamma^{\mu}$ are the Dirac matrices. The projection operator $(1-\gamma^{5})$ leads to a coupling exclusively to the left-handed components. Based on this, the left-handed particles are arranged in $SU(2)_L$ doublets, while the right-handed particles are singlets.

For quarks, the weak eigenstates $d', s', b'$ are not identical to their mass eigenstates $d, s, b$. This misalignment is described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix [19, 20], which enables flavour-changing coupling in charged-current interactions.

### 2.1.3. Strong Interaction

The strong interaction is described by quantum chromodynamics (QCD), a non-Abelian gauge theory with gauge group $SU(3)_C$. Quarks carry colour charge, while gluons carry a colour and anticolour index, giving rise to self-interactions through three- and four-gluon vertices. These features lead to two fundamental properties:

**Asymptotic freedom.** The QCD coupling decreases at high momentum transfer [11, 21], allowing perturbative calculations of hard processes such as top-quark pair production. In the high-energy regime of the LHC, quarks and gluons effectively behave as quasi-free particles.

**Confinement.** At low energies, the QCD coupling becomes large, preventing coloured states from existing in isolation. Instead, they form colour-neutral bound states—hadrons. The transition from short-distance partons to long-distance hadrons involves nonperturbative dynamics encapsulated through hadronisation models.
A crucial consequence for collider physics is that quarks and gluons produced in hard interactions undergo a cascade of QCD emissions before hadronising into jets. This process is typically described in terms of:

- **Parton distribution functions (PDFs)**, characterising the proton structure and entering via the QCD factorisation theorem [22].
- **Initial- and final-state radiation (ISR/FSR)**, arising from soft and collinear gluon emission.
- **Parton showers**, which resum logarithmically enhanced emissions and are implemented in Monte Carlo generators such as PYTHIA, first developed in [23, 24] and extended in modern formulations [25].
- **Hadronisation models**, such as the Lund string model [26, 27], which describe the confinement-driven formation of hadrons.

These aspects of QCD are essential for the reconstruction of hadronic final states and for Monte Carlo simulations of LHC physics.

## 2.2. Top Quark Physics

The top quark is the heaviest known elementary particle. The prediction of its existence arose due the already observed CP-Violation in the kaon system, which required a third generation of quarks to be explained within the SM framework [20]. Furthermore, the discovery of the bottom quark in 1977 [28] implied the existence of its weak isospin partner, the top quark. The top quark was eventually discovered in 1995 by the CDF and DØ collaborations at the TEVATRON collider [29, 30]. Currently, the world average of the top-quark mass is measured to be [7]

$$m_t = 172.56 \pm 0.31\,\text{GeV}. \tag{2.2}$$

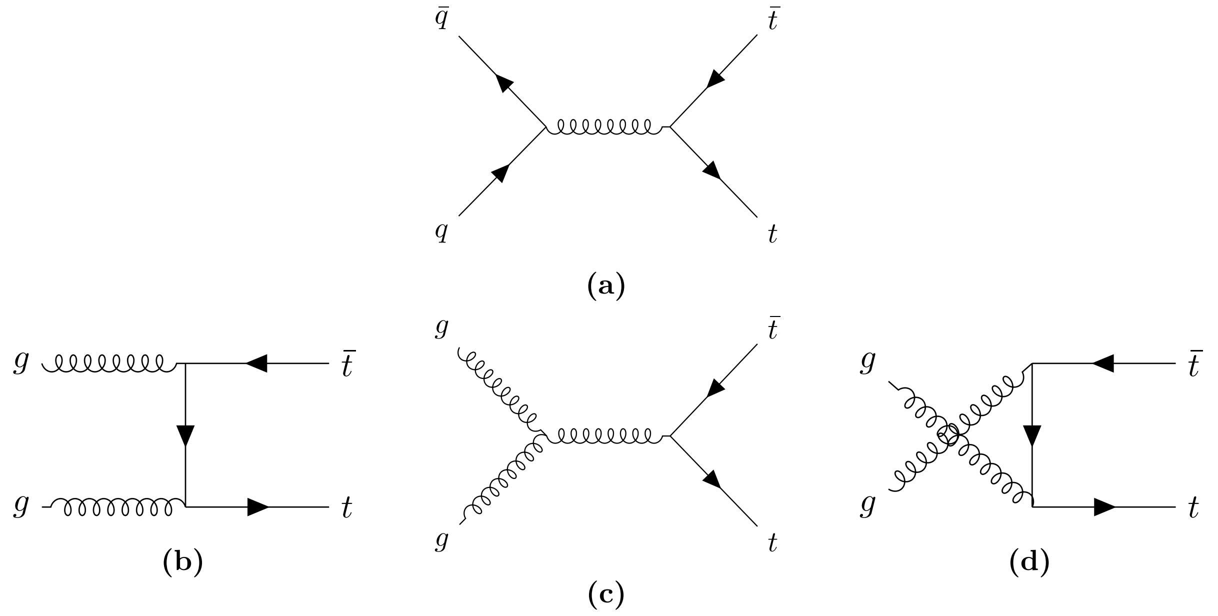


**Figure 2.2.:** Leading order Feynman diagrams for top-quark pair production in (a) quark-antiquark annihilation and (b), (c) & (d) gluon fusion (ggF).

Along with the highest mass, the top quark also has the strongest Yukawa coupling to the Higgs boson, making it interesting for studies of the Higgs mechanism and electroweak symmetry breaking.

Due to its large mass, the top quark also has the shortest lifetime of all quarks, about $5 \times 10^{-25}$ s [7], which is shorter than the typical hadronisation timescale of $3 \times 10^{-24}$ s [31]. Therefore, the top is the only quark that decays before it hadronises, allowing for direct studies of its properties through its decay products.

Notably, the top quark also decays before its spin can decorrelate, preserving spin information in the angular distributions of its decay products [32]. This unique feature enables precise measurements of spin correlations and polarisation of top quarks.

At hadron colliders like the LHC, top quarks are predominantly produced in pairs via the strong interaction. The leading order (LO) Feynman diagrams for top-quark pair production are shown in Figure 2.2. At the LHC, the dominant production mechanism is gluon fusion (ggF). This is due to the high gluon density in the proton at the relevant Bjorken-$x$ values for top-quark production.

The CKM matrix element $V_{tb}$ is close to unity [7], leading to a nearly exclusive decay of the top quark into a $W$ boson and a $b$ quark. The $W$ can subsequently decay either leptonically into a charged lepton and a neutrino, or hadronically into a pair of quarks, with branching ratios of about $\mathrm{BR}(W \to \ell\nu) = 1/3$ and $\mathrm{BR}(W \to q\bar{q}) = 2/3$, respectively [7].

Based on these $W$ decay modes, top-quark pair events are categorised into three channels. In the **dileptonic channel**, both $W$ bosons decay leptonically ($W^+ \to \ell^+\nu_\ell$, $W^- \to \ell^-\bar{\nu}_\ell$), resulting in two charged leptons, two neutrinos and two $b$ quarks in the final state. This channel has a branching ratio of approximately 1/9. The **semileptonic channel** occurs when one $W$ boson decays leptonically and the other hadronically ($W \to q\bar{q}'$), yielding one charged lepton, one neutrino, four quarks (including two $b$ quarks) in the final state. This is the most common channel, with a branching ratio of about 4/9. Finally, in the **all-hadronic channel**, both $W$ bosons decay hadronically, producing six quarks (including two $b$ quarks) in the final state. This channel has the highest branching ratio of approximately 4/9 but suffers from large multijet background contamination in the busy environment of hadron colliders like the LHC.

### 2.2.1. Spin Correlations in Top-Quark Pairs

As elaborated before, the top quark decays before it can hadronise or its spin decorrelates. Therefore, the spin information of the top quark is directly transferred to its decay products. In top-quark pair production, the spins of the top and antitop quarks are correlated due to the production mechanism. Due to the chiral nature of the weak interaction, these spin correlations manifest in the angular distributions of the decay products, particularly the charged leptons in the dileptonic decay channel.

Recent studies at the LHC began to explore potential effects of quantum entanglement in top-quark pairs [1]. Entanglement generally describes a quantum mechanical phenomenon where the quantum states of two or more particles become correlated in such a way that the state of one particle cannot be described independently of the state of the other(s), even when the particles are separated by large distances [33]. This property is expressed as the system having a non-seperable density matrix. While a general description of this phenomenon is beyond the scope of this thesis, entanglement can be tested using specific criterions. One such criterion is the *Peres-Horodecki criterion* [34, 35], which states that if the partial transpose of the density matrix of a bipartite system has at least one negative eigenvalue, then the system is entangled. Considering the top-quark pair's spin as a bipartite system of two spin-$\frac{1}{2}$ particles (qubits), this criterion can be applied to test for entanglement in their spin states.

In Ref. [36] it was proposed to measure entanglement in top-quark pairs produced at the LHC by analysing the angular distributions of their decay products, specifically the charged leptons in the dileptonic decay channel. The study showed that by measuring the opening angle between the two leptons, one can construct an observable sensitive to the entanglement of the spin system of the top quarks. Due to the kinematics of the top-quark pair production, the entanglement is expected to be more pronounced in certain regions of phase space, one such region is at low invariant masses of the top-quark pair system. One particular region of interest is the so-called *threshold region*, where the top-quark pair is produced with low relative velocity. In this region, the top quark is produced as a spin singlet state in about $\sim 80\%$ of the cases, leading to a strong entanglement signature [37]. The ATLAS collaboration has recently reported the first experimental evidence of quantum correlations consistent with entangled spin quantum states in top-quark pairs [1]. Measurements of this variable, are particularly sensitive in the dilepton channel, because, the lepton has the highest spin-analysing power [4, 5]. The spin-analysing power measure, to which extent the spin-information of the parent is imprinted in the angular distributions of the decay products. Due to QCD effects, this is lower for light jets compared to leptons.

### 2.2.2. Top-Quark Pair Bound State Effects

Even though the top quark decays before it can hadronise, near the production threshold of top-quark pairs, the strong interaction between the top and top antiquarks can lead to the formation of transient bound states, coined as toponium [37–39]. These bound state effects can influence the production cross section and kinematic distributions of top-quark pairs near threshold. The CMS collaboration has recently observed an excess in the invariant mass distribution of top-quark pairs near threshold, consistent with the presence of toponium bound state effects [3]. The modelling of these effects is based on non-relativistic QCD (NRQCD) calculations, which account for the strong interaction dynamics between the top and top aniquarks in the near-threshold region [40, 41]. Understanding and accurately modelling these bound state effects is crucial for precision measurements of top-quark properties and for searches for new physics in top-quark pair production.

This study is also investigating the threshold region of top-quark pair production and makes use of angular correlations between the decay products as a probe of the underlying dynamics, including potential bound state effects.

### 2.2.3. Dileptonic Decay Channel

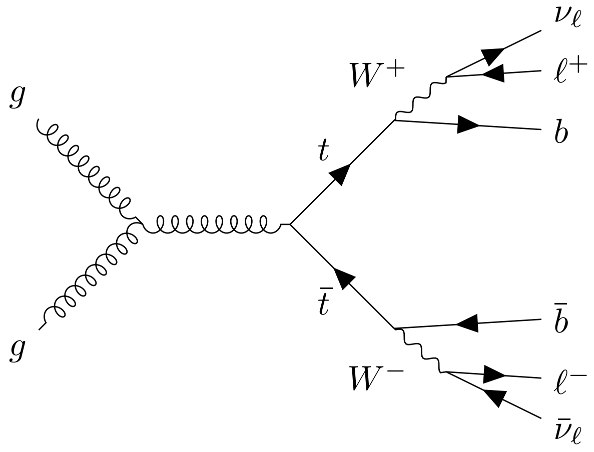


**Figure 2.3.:** Leading order Feynman diagram for top-quark pair production with dileptonic decay.

The dileptonic decay channel, where both $W$ bosons decay leptonically, has a branching ratio of about 1/9 and results in a final state with two charged leptons, two neutrinos and two $b$ quarks, as illustrated in Figure 2.3. Due to the lower branching ratio, the dileptonic channel has a smaller event yield compared to the other channels. However, it offers a cleaner signature with reduced background contamination, making it particularly suitable for precision measurements of top-quark properties. The presence of two neutrinos in the final state leads to missing transverse energy ($\not{E}_\mathrm{T}$) in the event, complicating the full reconstruction of the top quark's kinematics. Advanced techniques, such as kinematic fitting and multivariate analyses, are often employed to address these challenges and extract maximum information from the dileptonic events.

### Kinematic Constraints

Assuming both $W$ bosons and top quarks are on-shell, the following kinematic constraints can be applied to the dileptonic decay channel,

$$(p_{\ell_1} + p_{\nu_1})^2 = m_W^2, \tag{2.3}$$
$$(p_{\ell_2} + p_{\nu_2})^2 = m_W^2, \tag{2.4}$$
$$(p_{b_1} + p_{\ell_1} + p_{\nu_1})^2 = m_t^2, \tag{2.5}$$
$$(p_{b_2} + p_{\ell_2} + p_{\nu_2})^2 = m_t^2, \tag{2.6}$$
$$(\vec{p}_{\nu_1} + \vec{p}_{\nu_2})_x = \not{E}_{\mathrm{T}x}, \tag{2.7}$$
$$(\vec{p}_{\nu_1} + \vec{p}_{\nu_2})_y = \not{E}_{\mathrm{T}y}. \tag{2.8}$$

If one assumes the neutrinos to be massless, these six equations provide constraints on the six unknown components of the two neutrino momenta. Therefore, additional assumptions or techniques are required to fully reconstruct the event kinematics in the dileptonic channel. Due to the algebraic nature of the constraints, multiple solutions can exist for a given event, leading to ambiguities in the reconstruction. Due to the effects of detector resolution and off-shell particles, this system of equation can also have no real-valued solution [42, 43]. Various methods have been developed to address these challenges, including likelihood-based approaches [44], matrix element methods, and machine learning techniques [45], which aim to optimally utilise the available information and improve the accuracy of the top quark kinematic reconstruction in dileptonic events.

## 2.2.4. Spin Sensitive Angular Variables

To study spin correlations and entanglement in top-quark pairs, a spin-sensitive angular variable is used. For the measurements of the bound-state effects, an additional angular variable is employed. Both variables are defined using the charged leptons from the dileptonic decay of the top-quark pair.

To define the variables, the momenta of the leptons are first boosted into their respective parent top-quark rest frames $\hat{\ell}_t^+$ and $\hat{\ell}_t^-$. The inner product of these unit vectors then defines the variable

$$c_{\mathrm{hel}} = \hat{\ell}_t^+ \cdot \hat{\ell}_t^- \,. \tag{2.9}$$

Analogously, one can define the variable $c_{\mathrm{han}}$, where either of the lepton momenta components parallel to the parent top-quark momentum in the $t\bar{t}$ rest frame is flipped before taking the inner product [46, 47]. This variable adds important context by enabling isolation of scalar states.

# 3. Machine Learning in Particle Physics

Since this work focuses on improving the event reconstruction of dileptonic $t\bar{t}$ decays using machine learning techniques, this chapter aims to provide an overview over the fundamental concepts and methodologies employed in machine learning.
From a computational perspective, machine learning can be viewed as the optimization of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ that maps input data points $\mathbf{x} \in \mathbb{R}^n$ to output predictions $\mathbf{y} \in \mathbb{R}^m$. The function $f$ is parameterized by a set of learnable parameters $\theta$, which are adjusted during the training process to minimize a predefined *loss function* $L(\mathbf{y}, f(\mathbf{x}; \theta))$. The loss function quantifies the discrepancy between the predicted outputs and the true target values, guiding the optimization process.

## 3.1. Supervised Learning

Machine learning can be broadly categorized into supervised and unsupervised learning. In supervised learning, the model is trained on a labelled dataset, where each input data point is associated with a corresponding target output. The goal of the model is to learn a mapping from inputs to outputs, enabling it to make accurate predictions on unseen data.

### 3.1.1. Monte Carlo Training Data

In high energy physics, supervised learning is commonly performed by training models on simulated datasets. The way the Monte Carlo Event modelling is performed in particle physics, makes it naturally suited for supervised learning tasks.
Events are usually simulated using a cascade of different programs, each simulating a different state of the event. First the hard scattering process is simulated using matrix element generators. These programs calculate the probabilities of different particle interactions based on the underlying physics theories, such as the Standard Model. Using these probabilities, they generate events that represent the initial state of the particles after the collision. The possible kinematic phase space is sampled according to these probabilities, resulting in a set of particles with specific momenta and energies. This type of event variables is often referred to as *parton-level*.
Next, the parton showering and hadronization processes are simulated. Parton showering models the emission of additional particles from the initial partons, while hadronization simulates the formation of hadrons from quarks and gluons. These processes are crucial for accurately modelling the final state particles observed in detectors.

Finally, detector simulation programs are used to simulate the interaction of particles with the detector material, producing realistic detector responses. This includes simulating the energy deposits in calorimeters, hits in tracking detectors, and other relevant signals.
While the actual detector response is simulated using complex detector simulation software, for many machine learning applications, a simplified representation of the detector response is sufficient. This can involve smearing the particle momenta and energies according to the detector resolution, applying efficiency corrections, and simulating the effects of pile-up.
This is because, for the reconstruction of the physics objects, such as jets, leptons, and missing transverse energy, highly sophisticated algorithms are used that already take into account the detector effects (Note that these algorithms may also be machine learning based). Therefore, the input features for machine learning models can often be derived directly from these reconstructed objects, rather than relying on the raw detector signals. The reconstructed object event variables are called *reco-level*.

### 3.1.2. Training

During the training phase, the model is presented with a set of input features derived from the reco-level event variables, along with their corresponding target outputs, which are typically derived from the parton-level event variables. The model learns to map the input features to the target outputs by minimizing a loss function that quantifies the difference between the predicted and true values. Common loss functions include mean squared error for regression tasks and cross-entropy loss [48] for classification tasks. The training process involves iteratively (each step is called *epoch*) updating the model's parameters using an optimization algorithm to minimize the loss function.

### 3.1.3. Validation and Testing

To evaluate the performance of the trained model, it is essential to validate and test it on independent datasets that were not used during training. This helps to assess the model's generalization capabilities and ensures that it can make accurate predictions on unseen data. A common practice is to split the available data into training, validation, and test sets. The validation set is used to tune hyperparameters and monitor the model's performance during training, while the test set is reserved for the final evaluation of the model's performance.
Since in particle physics, simulated data is often used for training and background modelling, one typically employs a k-folding strategy, where the data is divided into k subsets. The model is trained k times, each time using a different subset as the validation set and the remaining subsets for training. This approach helps to use the available data more efficiently.

## 3.2. Neural Networks

As mentioned earlier, machine learning models can be viewed as parameterized functions that map input data to output predictions. Neural networks are a class of machine learning

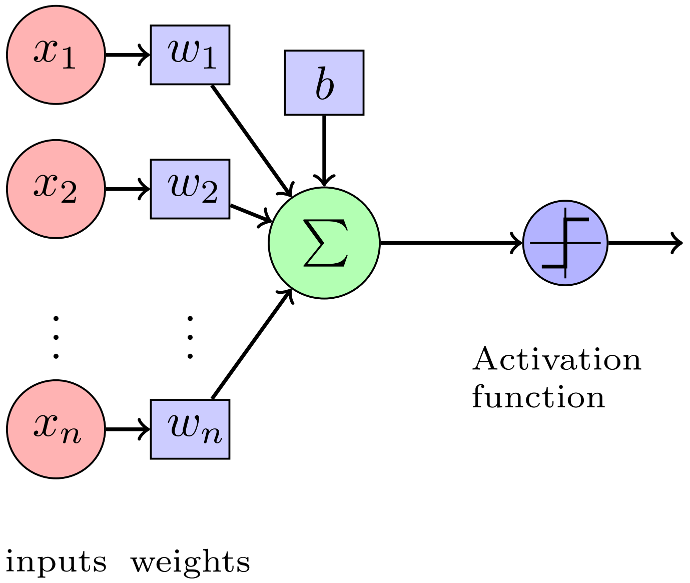


**Figure 3.1.:** Schematic of a single artificial neuron (perceptron). The neuron receives multiple input signals, each associated with a weight, and computes a weighted sum of these inputs, optionally a so-called bias is added. An activation function is then applied to this sum to produce the neuron's output.

models that are inspired by the structure and function of biological neural networks in the human brain. They consist of interconnected artificial neurons that process and transform the input data to produce the desired output. The simplest unit of a neural network is the artificial neuron, also known as a perceptron [49]. A perceptron takes multiple input signals, each associated with a weight, and computes a weighted sum of these inputs. An activation function is then applied to this sum to produce the neuron's output. This structure is illustrated in Figure 3.1. The general structure of a perceptron is inspired by biological neurons. The activation function introduces non-linearity into the model, allowing it to learn complex patterns in the data. Common activation functions include the sigmoid function, hyperbolic tangent (tanh), and rectified linear unit (ReLU) [50].

### 3.2.1. Dense Neural Network

A dense neural network (DNN), also known as a fully connected neural network [51], is a type of artificial neural network where each neuron in one layer is connected to every neuron in the subsequent layer. This architecture allows for the learning of complex relationships between input features and output targets. A schematic of a dense neural network is shown in Figure 3.2.

A DNN consists of an input layer, one or more hidden layers, and an output layer. The input layer receives the input features, while the hidden layers perform a series of transformations on the data using the weights and biases associated with each neuron. The output layer produces the final predictions. Each layer applies an activation function to the weighted sum of inputs from the previous layer, enabling the network to learn non-

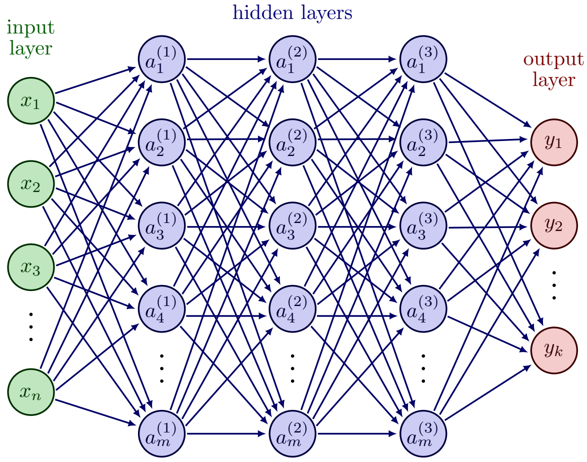


**Figure 3.2.:** Schematic of a dense neural network with multiple layers. Each layer consists of multiple neurons, and each neuron in one layer is connected to every neuron in the subsequent layer.

linear mappings. The depth (number of layers) and width (number of neurons per layer) of the network can be adjusted.

### 3.2.2. Recurrent Neural Networks

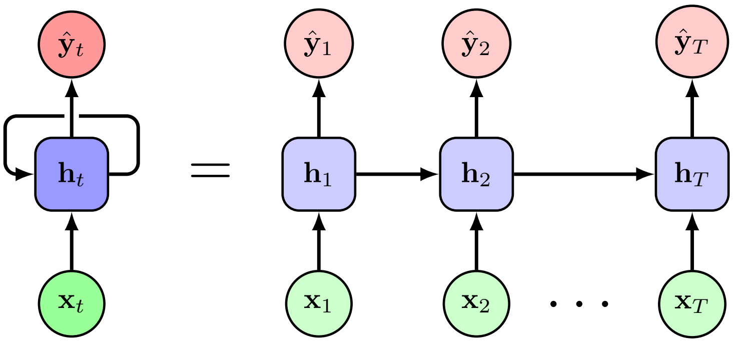
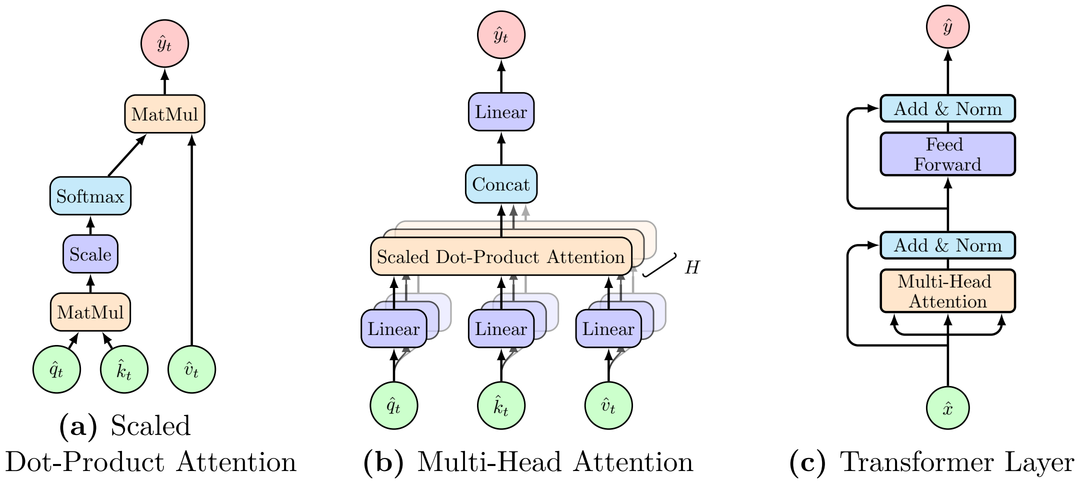


**Figure 3.3.:** Schematic of a recurrent neural network (RNN). The RNN processes sequences of data by maintaining a hidden state that captures information from previous time steps. The unfolded representation illustrates how the RNN operates over multiple time steps.

Dense neural networks require their inputs to have a fixed size. However, in many applications, including particle physics, the input data can vary in size. For example, the number of particles detected in an event can differ from one event to another. To handle such sequential data, recurrent neural networks (RNNs) [52] are employed. RNNs are designed to process sequences of data by maintaining a hidden state that captures information from previous time steps. A schematic of an RNN is shown in Figure 3.3.

RNNs consist of a series of interconnected neurons that process input data sequentially. At each time step, the RNN takes an input vector and combines it with the hidden state from the previous time step to produce a new hidden state. This hidden state is then used to generate the output for the current time step. The recurrent connections allow the RNN to retain information from previous inputs, enabling it to learn temporal dependencies in the data. One common variant of RNNs is the Long Short-Term Memory (LSTM) network [53], which addresses the vanishing gradient problem and allows for learning long-term dependencies more effectively. Another caveat of RNNs is that they are inherently sequential, making them less suited for data without inherent order, such as sets of particles in an event.

### 3.2.3. Transformer



**(a)** Scaled Dot-Product Attention **(b)** Multi-Head Attention **(c)** Transformer Layer

**Figure 3.4.:** Different components of the transformer architecture from smallest to largest unit: (a) Scaled Dot-Product Attention, (b) Multi-Head Attention, and (c) Transformer Layer.

The transformer architecture deals with sequential data, similar to RNNs, but it does so using an attention mechanism rather than recurrent connections. This allows transformers to process information in parallel, making them more efficient for training on large datasets. Additionally, transformer can capture long-range dependencies in the data more effectively than RNNs, as they relate all elements of the input sequence to each other through attention mechanisms.

The core component of the transformer architecture is the attention mechanism, which allows the model to focus on different parts of the input sequence when making predictions. The most commonly used attention mechanism is the scaled dot-product attention [54], illustrated in Figure 3.4a. In this mechanism, three vectors are computed for each input element: the query ($q$), key ($k$), and value ($v$) vectors. The attention scores are calculated by taking the dot product of the query and key vectors, scaling them by the square root of the dimension of the key vectors, and applying a softmax function to obtain attention weights. These weights are then used to compute a weighted sum of the value vectors, producing the output of the attention mechanism. Multi-head attention, shown in Figure 3.4b, extends the scaled dot-product attention by allowing the model to attend

to different parts of the input sequence simultaneously. This is achieved by using multiple parallel attention heads, each with its own set of learned linear transformations for the query, key, and value vectors. The outputs of all attention heads are concatenated and linearly transformed to produce the final output.
The key, query and value vectors are typically obtained by applying learned linear transformations to the input embeddings. Depending on the type of relational information one wants to capture, one can choose to derive these vectors from different input sources. For example, in the *self-attention mechanism*, all three vectors are derived from the same input sequence ($\hat{q}_t = \hat{k}_t = \hat{v}_t = \hat{x}_t$), allowing the model to relate different elements within the same sequence. In *cross-attention*, the query vectors are derived from one sequence, while the key and value vectors are derived from another sequence ($\hat{q}_t = \hat{x}_{1,t}$, $\hat{k}_t = \hat{v}_t = \hat{x}_{2,t}$), enabling the model to relate information between two different sequences.

**Symmetries** In many applications, including particle physics, the input data exhibits certain symmetries that can be exploited to improve model performance.
The attention mechanism has inherent permutation equivariances, that can be exploited. Self-attention is permutation equivariant to the input sequence, meaning that permuting the input elements results in a corresponding permutation of the output elements. This property is particularly useful when dealing with sets or unordered data, where the order of the elements does not carry any inherent meaning. Cross-attention is permutation equivariant to permutations of the query inputs and separately to permutations of the key/value inputs. This allows the model to effectively relate information between two different sequences, regardless of their order.

### 3.2.4. Training Neural Networks

Training neural networks involves optimizing the model's parameters (weights and biases) to minimize the loss function. The number of parameters in a neural network can be quite large, especially in deep architectures, making the optimization process computationally intensive. The most commonly used optimization algorithm for training neural networks is stochastic gradient descent (SGD) [55], which iteratively updates the model's parameters based on the gradients of the loss function with respect to the parameters. Modern variants of SGD, such as Adam [56], incorporate adaptive learning rates and momentum to improve convergence. The backpropagation algorithm [51] is used to efficiently compute the gradients of the loss function with respect to the model's parameters. It involves propagating the error signal backward through the network, allowing for the calculation of gradients at each layer. These gradients are then used by the optimization algorithm to update the parameters. The algorithm relies on the chain rule and nested structure of the neural network to compute the gradients efficiently. Regularization techniques, such as dropout [57] and weight decay, are often employed during training to prevent overfitting and improve the model's generalization capabilities. Dropout randomly deactivates a fraction of the neurons during training, forcing the network to learn more robust features. Weight decay adds a penalty term to the loss function based on the magnitude of the weights, discouraging overly complex models and promoting generalization.

# 4. Analysis and Reconstruction

## 4.1. Data and Simulation

This section describes the data and simulation samples used in this analysis. At the current stage of this thesis, only simulation samples have been studied; the data samples will be described in the final version. The simulated samples are produced using Monte Carlo (MC) event generators as described in Section 3.1.1.

While a complete analysis requires detailed investigation of background processes contributing to the signal regions, this thesis focuses on developing and evaluating reconstruction methods. Therefore, only the signal process of $t\bar{t}$ production with dileptonic decays is described here. The background processes and their simulation will be discussed in the final thesis.

### 4.1.1. Simulation of Dileptonic Top-Quark Pair Events

The signal process of $t\bar{t}$ production with dileptonic decays is simulated using Powheg [58] at next-to-leading order (NLO) in perturbative quantum chromodynamics (QCD). The Powheg generator is interfaced with Pythia [25] for parton showering, hadronization, and underlying event modelling.

To link the reco-level particles to the parton-level objects, a parton matching is performed. The parton-level objects are defined as the top quarks and their decay products after final state radiation before hadronization. The matching is done by iteratively associating reconstructed jets to partons based on angular distance criteria. The decay products of the top quarks are matched within a cone of $\Delta R \leq 0.3$ and $\Delta R \leq 0.1$ for jets and leptons ($e$,$\mu$) respectively. If multiple reconstructed objects are found within this cone, the one with the smallest $\Delta R$ is chosen.

### 4.1.2. Object Definition

The reconstruction of physics objects—electrons, muons, jets, and missing transverse energy ($\not{E}_\mathrm{T}$)—is performed using the standard Atlas reconstruction algorithms [59].

**Electrons** are required to have transverse momentum $p_\mathrm{T} > 5$ GeV and pseudorapidity $|\eta| < 2.47$, excluding the transition region between the barrel and endcap calorimeters ($1.37 < |\eta| < 1.52$). Electrons must satisfy the Tight identification criteria and be isolated from other activity in the detector.

**Muons** are reconstructed using information from both the inner detector and the muon spectrometer. They are required to have $p_{\mathrm{T}} > 5$ GeV and $|\eta| < 2.5$. Muons must satisfy the Tight identification criteria and be isolated.

**Jets** are reconstructed using the anti-$k_t$ algorithm [60] with a radius parameter of $R = 0.4$. They are required to have $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$. $b$-tagging is performed using the GN2 algorithm [61] to identify jets originating from $b$-quarks.

**Missing Transverse Energy** ($\not{E}_{\mathrm{T}}$) is calculated from the negative vector sum of the transverse momenta of all reconstructed objects in the event, including a soft term to account for energy not associated with reconstructed objects.

### 4.1.3. Event Selection

Events are selected to contain exactly two oppositely charged leptons (electrons or muons). One of the leptons must have $p_{\mathrm{T}} > 25$ GeV, while the other must have $p_{\mathrm{T}} > 5$ GeV. At least two jets are required, with at least one being $b$-tagged at the 77% working point (WP). Additional selection criteria are applied to suppress background contributions, which will be detailed in the final thesis.

For the training and evaluation of the reconstruction methods, only events that pass the selection criteria and have a successful parton matching are considered.

## 4.2. Reconstruction of Top-Quark Pairs

Reconstructing the kinematics of dileptonic $t\bar{t}$ events requires an assignment of reconstructed objects to the parton-level decay products. This section describes the different reconstruction methods developed and evaluated in this thesis.

While the leptons are directly identified from the reconstructed objects, the assignment of jets to the $b$-quarks from the top quark decays is ambiguous due to the presence of multiple jets in the event. Additionally, the two neutrinos from the $W$ boson decays are not directly detected, leading to missing information in the event reconstruction. Based on this, the reconstruction breaks down into two main tasks:

**Neutrino Regression** Estimating the momenta of the two neutrinos using the measured $\not{E}_{\mathrm{T}}$ and other event kinematics.

**Jet Assignment** Assigning the reconstructed jets to the $b$-quarks from the top quark decays.

### 4.2.1. Neutrino Regression Methods

Due to the current stage of the thesis, only one baseline method for neutrino regression has been used. The method of choice for the final thesis will be determined after evaluating multiple approaches.

The baseline method used here is called $\nu^2$-Flows [45] and uses normalizing flows to model the conditional probability distribution of the neutrino momenta given the observed event kinematics. The model is trained on simulated dileptonic $t\bar{t}$ events, where the true neutrino momenta are known from the parton-level information. The trained model can then be used to sample possible neutrino momenta for new events, allowing for a probabilistic reconstruction of the event kinematics. For each event, multiple samples of neutrino momenta are drawn from the model, and the one with the highest probability density is selected as the reconstructed neutrino momenta. A detailed outline of the $\nu^2$-Flows method and its implementation can be found in Ref. [45]. For this thesis, this method serves to establish a baseline for neutrino regression to evaluate the impact of improved jet assignment methods. Note, however, that this method has not been used in Atlas analyses yet. However, due to matters of software compatibility, it was the only viable option at this stage of the thesis. Since $\nu^2$-Flows processes the full event information to make a prediction for the neutrino momenta, it is agnostic to the jet assignment. This is drastically different from the conventional methods, which were used in previous Atlas analyses [36]. These methods are interdependent with the jet assignment and mostly based on exploiting the kinematic constraints of the event, such as the invariant masses of the $W$ bosons and top quarks outlined in Section 2.2.3.

### 4.2.2. Jet Assignment Methods

This thesis explores several methods for jet assignment, ranging from traditional algorithms to machine learning approaches. From recent publications e.g. [36], two conventional methods have been implemented as baselines:

**Jet-Selection** Both methods start by selecting the two $b$-jet candidates in the event based on the $b$-tagging at the 77% WP. If more than two $b$-tagged jets are present, the two with the highest $p_{\mathrm{T}}$ are chosen. If only one $b$-tagged jet is found, the non-tagged jet with the highest $b$-tagging score is selected as the second candidate.

**$\chi^2$-Method**

For both possible assignments of the two $b$-jet candidates to the parton-level $b$-quarks, the invariant masses of the reconstructed top quarks are computed,

$$\chi^2 = (m_{b_1,\nu,\ell^+} - m_t)^2 + (m_{b_2,\bar{\nu},\ell^-} - m_t)^2 \tag{4.1}$$

where $m_t = 172.5$ GeV is the top quark mass. The assignment with the smaller $\chi^2$ value is chosen.

**$\Delta R$-Method**

This method assigns the $b$-jet candidates based on their angular distance to the leptons.

$$\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} \tag{4.2}$$

The jet-lepton-pair with the smaller $\Delta R$ value is assigned to each other. The remaining jet is assigned to the other lepton.

Note, that the $\chi^2$-Method relies on the reconstructed neutrino momenta to compute the invariant masses of the top quarks. Therefore, it is inherently linked to the neutrino

regression method used. In contrast, the $\Delta R$-Method is independent of the neutrino momenta. Further, it should be noted, that due to its reliance on the neutrino reconstruction and the use of $\nu^2$-Flows here, the $\chi^2$-method should also be considered a machine learning based approach.
The $\Delta R$-Method was used in previous ATLAS analyses [36] for events, that failed the kinematic reconstruction required for the $\chi^2$-Method. However, in this thesis, both methods are evaluated on the full dataset for comparison.

# 5. Machine Learning Based Jet Assignment

In addition to the conventional jet assignment methods described in Section 4.2.2, this thesis investigates machine learning-based approaches to improve the accuracy of jet assignment in dileptonic $t\bar{t}$ events. Machine learning techniques have shown promise in various aspects of high-energy physics, such as flavour tagging [62].
For the high permutation complexity of jet assignments in the all-jets channel, SPANet [6] has demonstrated significant improvements over traditional methods. These architectures leverage attention mechanisms to improve the accuracy. Based on these successes, this thesis explores the adaptation of similar architectures for the dileptonic $t\bar{t}$ channel, which presents unique challenges due to the presence of two neutrinos and fewer jets.
Albeit a variety of architectures were tested during the development (including RNNs), two manifested as the best performing ones.

## 5.1. Neural Network Architectures

### 5.1.1. FeatureConcatTransformer

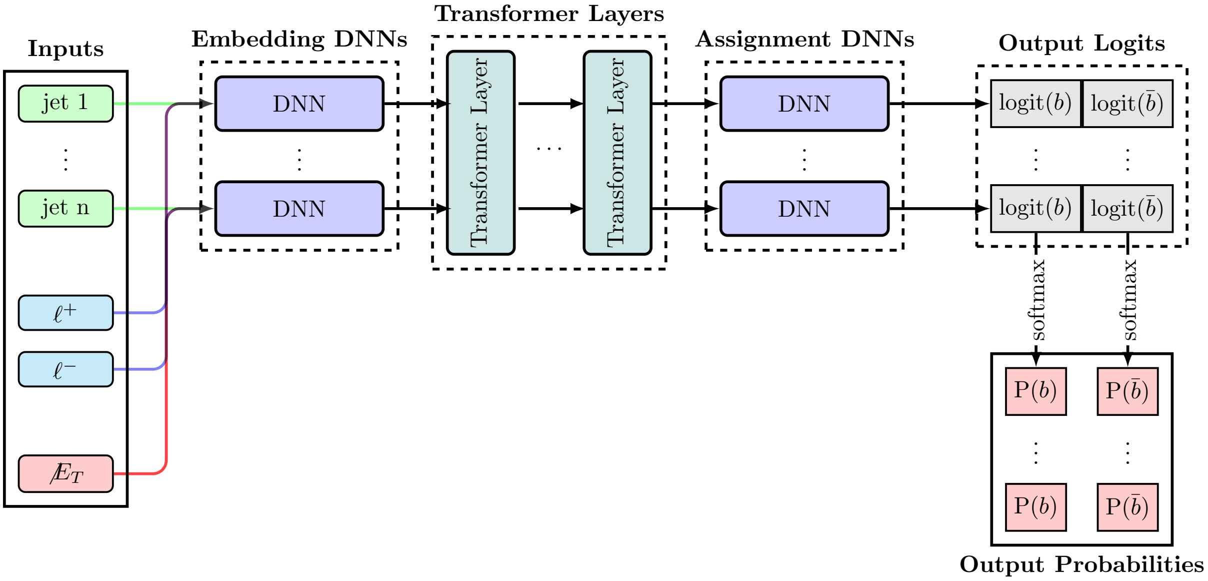


**Figure 5.1.:** Schematic of the **FeatureConcatTransformer** architecture.

The **FeatureConcatTransformer** architecture, illustrated in Figure 5.1 consists of the following components:

- **Input Features:** The model takes as input a set of features for each reconstructed object (jets and leptons), including kinematic variables ($E$, $p_{\mathrm{T}}$, $\eta$, $\phi$), $b$-tag score working point, and event-level features such as missing transverse energy ($\not{E}_{\mathrm{T}}$). The features of both leptons and the missing energy are concatenated to each jet's feature vector to provide context.

- **Embedding DNNs:** The concatenated feature sequence is processed through a shared embedding DNN to transform the raw features into a higher-dimensional representation. This size of the embedding is a hyperparameter of the model, called the *embedding dimension*. A default of 3 layers is used for the DNNs, where the dimension increases exponentially, to match the target dimension.

- **Transformer Layers:** The embedded features are passed through multiple transformer layers, which utilize self-attention mechanisms to capture the relationships between different reconstructed objects.

- **Assignment DNNs:** The output from the transformer layers is fed into assignment DNNs that produce assignment scores for each possible jet assignment to the parton-level decay products. Here too, a default of 3 layers is used for the DNNs, with exponentially decreasing dimensions.

- **Softmax Layer:** Finally, a softmax layer is applied to the assignment scores for each $b$-quark to obtain probabilities.

While the model has several hyperparameters, only the two most important (*number of transformer layers* and *embedding dimension*) are optimized. Due to the properties of the attention mechanism, the model is inherently permutation equivariant with respect to the jet ordering.
The structure of this architecture allows to easily concatenate additional features to each jet. The fixed structure of the inputs and the output, without any cross-attention mechanism, allows to easily add relational information between the jets and leptons through these additional features.

### 5.1.2. CrossAttentionTransformer

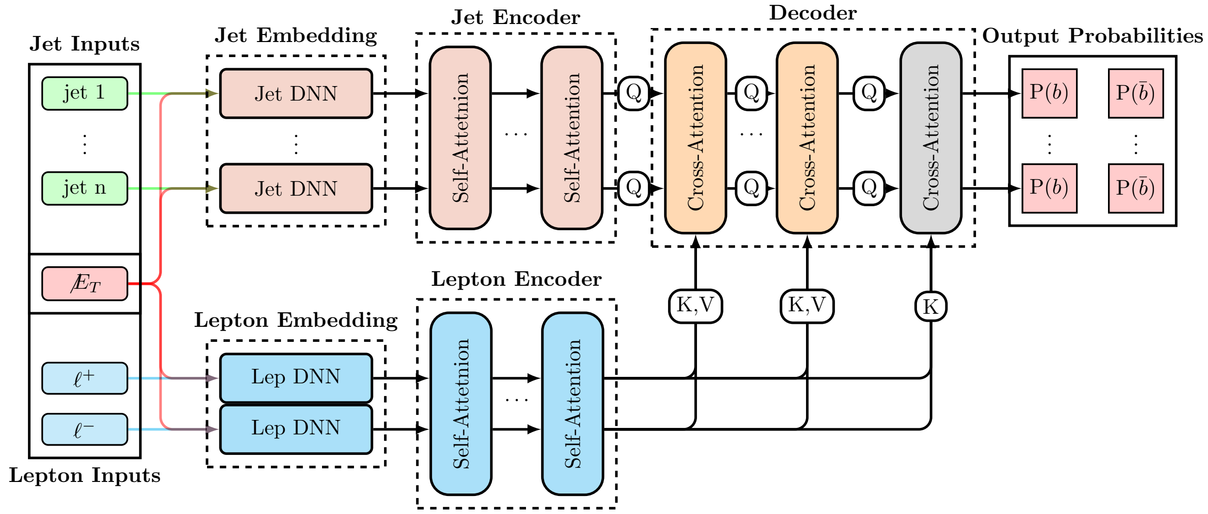


**Figure 5.2.:** Schematic of the **CrossAttentionTransformer** architecture. The annotations (Q:=Query, K:=Key, V:=Value) describe the inputs used for the in attention mechanism following the notation introduced in Section 3.2.3

The **CrossAttentionTransformer** architecture, illustrated in Figure 5.2, consists of the following components:

- **Input Features:**
  - **Jet Inputs:** The jet inputs are ($E$, $p_\mathrm{T}$, $\eta$, $\phi$), $b$-tag score working point, concatenated with event level features such as $\not{E}_\mathrm{T}$
  - **Lepton Inputs** The leptons inputs are their kinematic features ($E$, $p_\mathrm{T}$, $\eta$, $\phi$) and the $\not{E}_\mathrm{T}$ features.
- **Embedding DNNs:** The jet and lepton sequences are each embedded using their own shared DNN with 3 layers, that exponentially increase the dimension to match the *embedding dimension*.
- **Self-Attention Encoder:** Both the jet and lepton embeddings are processed through their own self-attention transformer encoders, each consisting of multiple layers. This allows the model to capture intra-object relationships within jets and leptons separately. The *number of encoder layers* is a hyperparameter of the model.
- **Cross-Attention Decoder:** The outputs from the jet and lepton encoders are then fed into a cross-attention decoder. Here, the jet embeddings serve as the Query (Q) inputs, while the lepton embeddings provide the Key (K) and Value (V) inputs. This mechanism enables the model to learn inter-object relationships between jets and leptons effectively. The *number of decoder layers* is a hyperparameter of the model. To simplify the architecture, the same number of layers is used for both the encoder and decoder.

The final layer of the cross-attention decoder outputs is used for the subsequent assignment.

This model has several hyperparameters, to keep the optimization feasible, only the two most important (*number of encoder/decoder layers* and *embedding dimension*) are optimized. Similar to the **FeatureConcatTransformer**, this model is also permutation equivariant with respect to the jet ordering, but due to the cross-attention mechanism it is additionally permutation equivariant with respect to the lepton ordering.
While, the two architectures share similarities, the **CrossAttentionTransformer** explicitly models the relationships between jets and leptons through the cross-attention mechanism, while the **FeatureConcatTransformer** relies on concatenated features to provide context. This structural difference may lead to varying performance depending on the complexity of the relationships in the data.

## 5.2. Training Procedure

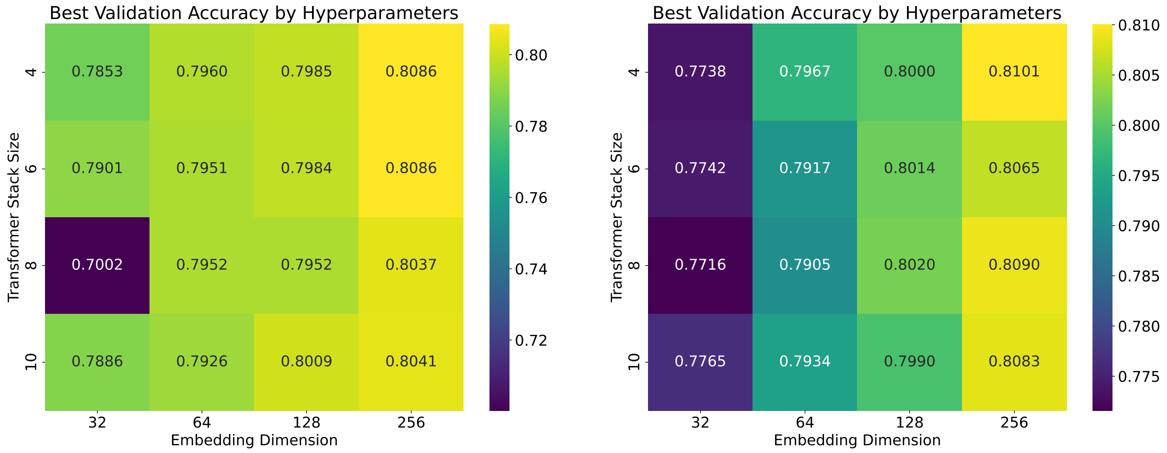
Both models are trained using supervised learning. The training dataset consists of simulated dileptonic $t\bar{t}$ events, where the true jet-to-parton assignments are known from the parton matching information. Both models are trained to minimize the categorical cross-entropy loss between the predicted assignment probabilities and the true assignments for each $b$-quark. The AdamW [63] optimizer is used with a learning rate of $10^{-4}$ and a weight decay of $10^{-4}$. Further, a dropout with a rate of $r = 0.1$ is applied to all internal forward passes during training. The models are trained for 50 epochs with a batch size of 1024 events, using early stopping based on the validation loss to prevent overfitting.

### 5.2.1. Metrics

The primary metric used to evaluate the performance of the jet assignment models is the **assignment accuracy**, defined as the fraction of events where both $b$-jets are correctly assigned to their respective parton-level $b$-quarks. Additionally, the **selection accuracy** is reported, which measures the fraction of events where the two selected jets correspond to the true $b$-quarks, regardless of the specific assignment.
Another useful metric, especially when comparing different methods across varying event conditions, is the **accuracy quotient**. This metric is the ratio of the assignment accuracy to the selection accuracy. It provides insight into how well a method performs the assignment task given that the correct jets have been selected, effectively normalizing out the jet selection performance.

## 5.3. Hyperparameter Optimization

Both architectures have multiple hyperparameters, that can be varied to tune the model performance. Increasing the model complexity generally improves the performance, but also increases the training time, memory requirements, and the risk of overfitting. While an extensive hyperparameter optimization is beyond the scope of this thesis. Given the



**(a) CrossAttentionTransformer** **(b) FeatureConcatTransformer**

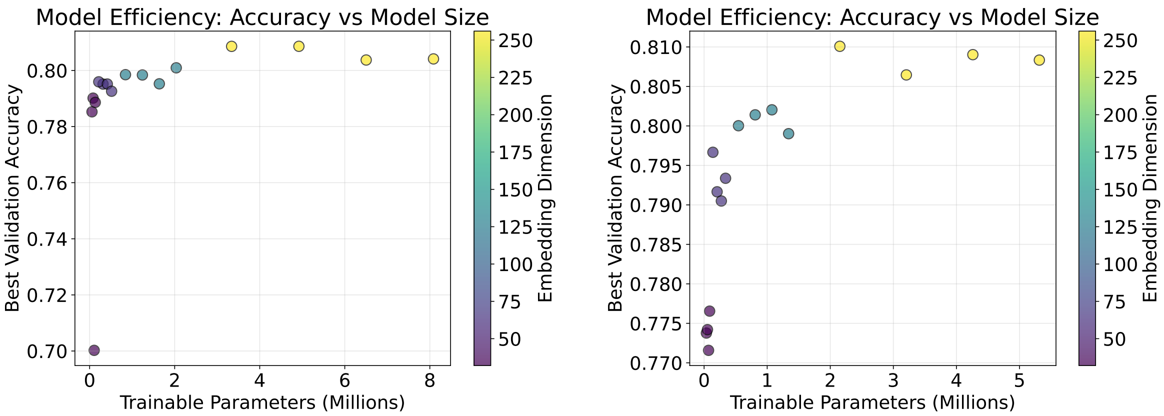
**Figure 5.3.:** Assignment accuracy on the validation dataset for different hyperparameter combinations for (a) the **CrossAttentionTransformer** and (b) the **FeatureConcatTransformer** architectures. Each square represents a model trained with the corresponding hyperparameter combination.

size of the dataset of about 20 million $t\bar{t}$ events in the nominal ATLAS sample, the computational limits could not be fully explored.

But to get a first estimate of the behaviour of the models with respect to their hyperparameters, a grid search over two of the most important hyperparameters for each architecture was performed. For the **FeatureConcatTransformer**, the *number of transformer layers* and the *embedding dimension* were varied, while for the **CrossAttentionTransformer**, the *number of encoder/decoder layers* and the *embedding dimension* were varied. Again, due to computational limits, for this thesis, only one value for each hyperparameter could be tested. For the final thesis, a more extensive hyperparameter optimization is planned, using k-fold cross-validation to ensure robust performance estimates.

The results of the grid search for the **CrossAttentionTransformer** are depicted in Figure 5.3. Due to the single run per hyperparameter combination, the results are subject to statistical fluctuations. However, a clear trend of increasing accuracy with larger embedding dimensions is visible. The number of layers seems to have a smaller impact on the performance, with only a slight increase in accuracy for more layers and sometimes even a decrease for larger models. This can be explained, if ones considers, that the model performance is highly dependent on the number of parameters, which increases with the square of the embedding dimension but only linearly with the transformer stack size. A plot showing the number of trainable parameters for each hyperparameter combination is in the Appendix in Figure A.1.

In Figure 5.4 the accuracy for each hyperparameter combination is shown against the number of trainable parameters. Here, a clear trend of increasing accuracy with more parameters is visible for both architectures. For both models, it can be seen, that the performance gain starts to saturate for larger models, indicating that further increasing the model complexity might not yield significant improvements.



(a) **CrossAttentionTransformer**

(b) **FeatureConcatTransformer**

**Figure 5.4.:** Assignment accuracy on the validation dataset for different hyperparameter combinations for (a) the **CrossAttentionTransformer** and (b) the **FeatureConcatTransformer** architectures. Each dot represents a model trained with the corresponding hyperparameter combination. The dots are coloured according to the embedding dimension used in the model.

# 5.4. Performance Comparison

| Method | Assignment Accuracy | Selection Accuracy |
|---|---|---|
| CrossAttentionTransformer | $0.8041^{+0.0006}_{-0.0005}$ | $0.9563^{+0.0002}_{-0.0002}$ |
| FeatureConcatTransformer | $0.8083^{+0.0006}_{-0.0005}$ | $0.9561^{+0.0002}_{-0.0003}$ |
| FeatureConcatTransformer HLF | $0.8179^{+0.0004}_{-0.0005}$ | $0.9582^{+0.0002}_{-0.0003}$ |

**Table 5.1.:** Jet assignment accuracies and selection accuracies for each of the machine learning architectures. Each best performing hyperparameter combination from the grid search is used for the architectures. The **FeatureConcatTransformer + HLF** model includes additional high-level features as input. The uncertainties are statistical only and stem from bootstrap resampling. The validation sample size is about 2 million events, while the test sample size is about 4 million events.

For the overall performance comparison, the best performing hyperparameter combination from the grid search is selected for each architecture. The assignment accuracies and selection accuracies on the test dataset are summarized in Table 5.1[1]. It can be seen, that there is only a small difference in performance between the two architectures, with the **FeatureConcatTransformer** performing slightly better than the **CrossAttentionTransformer**. Note, that even though, these numbers come with errors, they

[1] Note, that the errors only account for the variance within the data samples itself. Meaning, they only serve to make statements on the performance of the isolated methods and especially their comparison. However, they do not allow to quantify the performance one would expect in an analysis pipline.

are statistical only and do not account for model variance. For the selection accuracies, both models achieve similar performance, indicating that they are equally effective at selecting the correct jets, even if the specific assignment differs. While both architectures provide promising performance, the **FeatureConcatTransformer** demonstrates a slight edge in assignment accuracy, making it the preferred choice for further studies in this thesis.

Additional plots evaluating the performance of both architectures are included in the Appendix in Figures A.3 to A.5.

### 5.4.1. High Level Input Features

As mentioned before, the variables used as input features for the models are primarily low-level kinematic variables of the reconstructed objects. However, high-level features, derived from physics insights, can provide additional context and improve model performance. Examples of such high-level features (HLF) include angular separations between jets and leptons $\Delta R(\ell, \text{jet})$, invariant masses of jet-lepton $m(\ell, \text{jet})$ pairs. These variables are known to be sensitive to the correct jet assignment in $t\bar{t}$ events. The corresponding distributions of the added high-level features for correct and incorrect assignments are shown in the Appendix in Figures A.6 and A.7.

The **FeatureConcatTransformer** architecture is particularly well-suited for incorporating these high-level features, as they can be easily concatenated to the jet feature vectors. The **CrossAttentionTransformer** cannot relate these features based on their position in the input features to the leptons, due to its symmetry in the lepton ordering. This limits the usability of these features in this architecture. To evaluate the impact of high-level features, the **FeatureConcatTransformer** model is retrained with $\Delta R$ and $m(\ell, \text{jet})$ added to each jet's feature vector. The model is trained and evaluated using the same procedure as before. The results, summarized in Table 5.1, indicate that the inclusion of high-level features leads to a noticeable improvement in assignment accuracy, while the selection accuracy remains relatively unchanged. This suggests that the high-level features provide valuable information that helps the model make more accurate assignments. Interestingly, the selection accuracy slightly decreases when high-level features are included, which could be due to the model focusing more on specific assignments rather than just selecting the correct jets. Overall, incorporating high-level features proves beneficial for enhancing the jet assignment performance of the **FeatureConcatTransformer** model. Plots displaying the hyperparameter optimization results with high-level features are included in the Appendix in Figure A.2. The training histories of both architectures, with and without high-level features, can be seen in Figure 5.5. It can be seen, that the **FeatureConcatTransformer** with high-level features converges faster and achieves a lower validation loss compared to the other models. This indicates that the high-level features help the model learn more effectively, leading to improved performance. Another intestering feature, can be seen in the training history of the **CrossAttentionTransformer**, where a significant gap between the training and validation accuracy is visible, indicating potential overfitting. In contrast, the **FeatureConcatTransformer** models deliver a better generalization to the validation data, suggesting that this architecture is more robust against overfitting in this context. For these models, the validation performs slightly better than

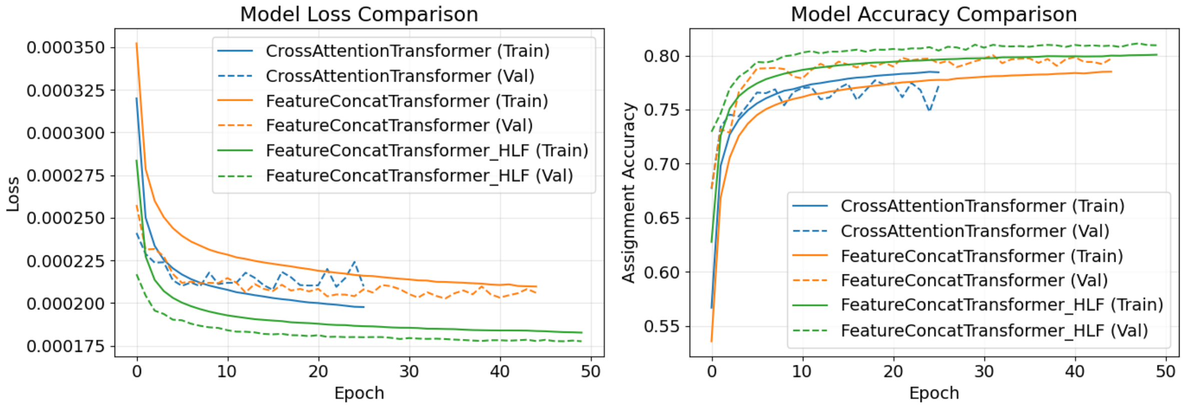
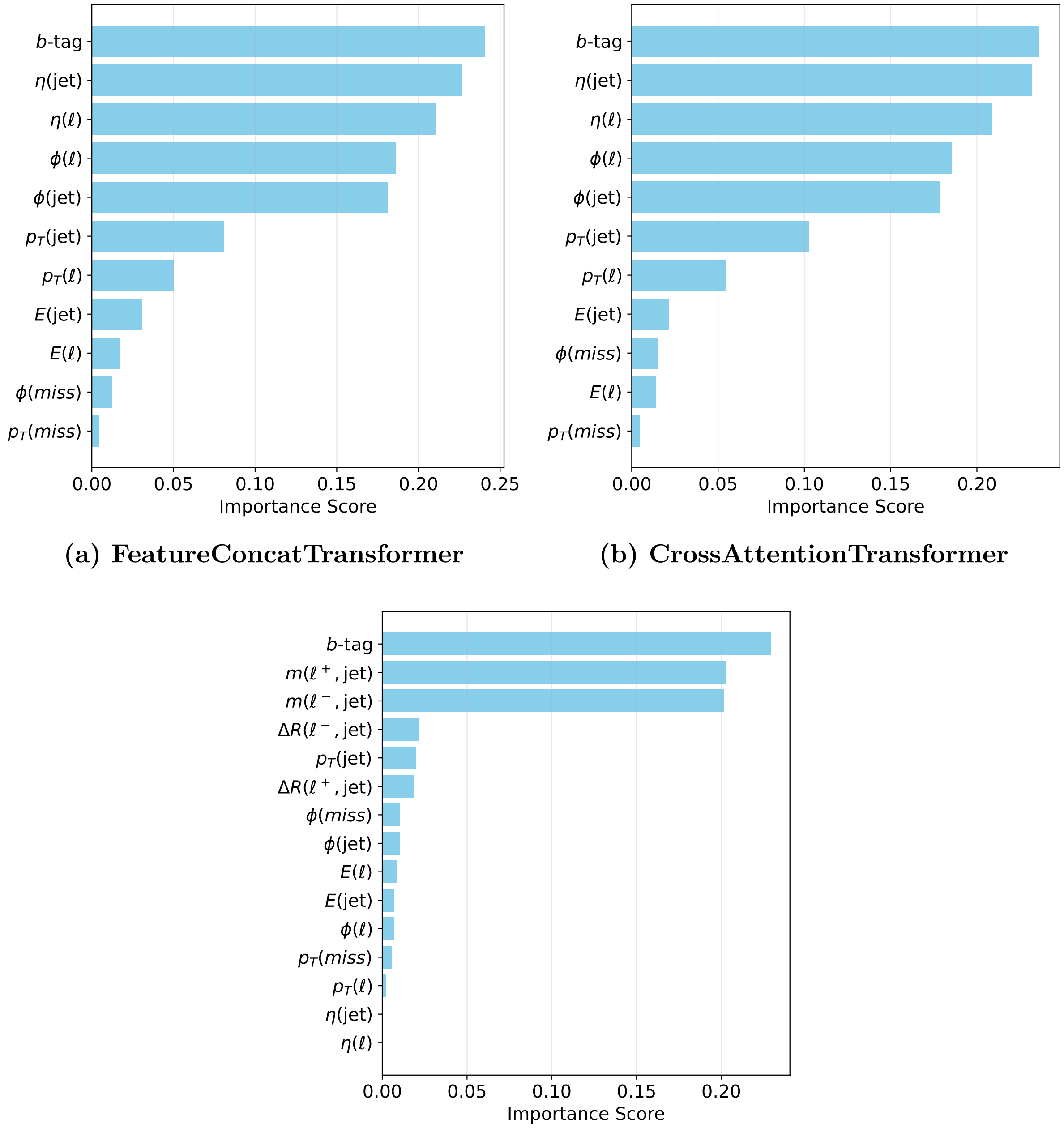


**Figure 5.5.:** Training histories of both architectures and potentially high-level features. The left plot shows the value of the loss-function and the right plot shows the assignment accuracy. For each model, the dotted line corresponds to the validation data.

the training, which is explained by the dropout layers, which are only active during training. Lastly, the **FeatureConcatTransformer** without high-level features exhibits a more stable training process compared to the **CrossAttentionTransformer**, further highlighting its suitability for this task. But overall, the **FeatureConcatTransformer** with high-level features demonstrates the most stable and effective training behaviour, leading to superior performance in jet assignment tasks.

### 5.4.2. Feature Evaluation

To gain insights into the decision-making process of the models, permutation importance [64] is used to evaluate the importance of each input feature. This method involves randomly permuting each feature and measuring the decrease in assignment accuracy, providing a quantitative measure of how much each feature contributes to the model's performance. The feature importances for each of the three models are depicted in Figure 5.6. For all models, the $b$-tagging scores emerge as the most important features, underscoring their critical role in jet assignment tasks. Kinematic features such as $\phi$ and $\eta$ show significant importance, indicating that the model leverages these variables to distinguish between jets effectively. The missing transverse energy ($\not{E}_{\mathrm{T}}$) has a negligible importance, suggesting that its contribution to jet assignment is minimal in this context. When high-level features are included in the **FeatureConcatTransformer**, they exhibit substantial importance, particularly the angular separation $m(\ell, \mathrm{jet})$. This highlights that these derived features provide valuable context that enhances the model's ability to make accurate assignments. Additionally, the importance of the low-level features, especially the $\eta$ of both jets and leptons, is negligible. This suggests that the high-level features effectively capture the relevant information, reducing the reliance on certain low-level kinematic variables.



(a) **FeatureConcatTransformer**

(b) **CrossAttentionTransformer**

(c) **FeatureConcatTransformer** with HLF

**Figure 5.6.:** The feature importances for each of the three models evaluated using permutation importance [64]. The importance is calculated as the decrease in assignment accuracy when a specific feature is randomly permuted. Particle features are permutated over all particles. Higher values indicate more important features.

### 5.4.3. Summary

In conclusion, both the **FeatureConcatTransformer** and **CrossAttentionTransformer** architectures demonstrate promising performance in jet assignment tasks for dileptonic $t\bar{t}$ events. The **FeatureConcatTransformer** seems to have a slight edge in assignment accuracy, particularly when high-level features are incorporated. The training histories indicate that the **FeatureConcatTransformer** converges more effectively and generalizes better to validation data, suggesting its robustness against overfitting. Further, this model architecture allows to easily embed additional physics-inspired features, which can significantly enhance its performance. While, from a physics point of view, the **CrossAttentionTransformer** seems more appealing due to its symmetry properties, the **FeatureConcatTransformer** proves to be more effective in practice for this specific task. One reason for this could be, that the cross-attention mechanism is more limited in capturing broader event context, as it focuses on pairwise interactions between jets and leptons, potentially overlooking more complex relationships that the **FeatureConcatTransformer** can capture through its concatenated feature approach. Apart from that, even though from a physics perspective the process is symmetric with respect to the lepton charge, in practice, due to differences in how the detector can resolve leptons of different charges, this symmetry is not exact. Which is why many approaches involving leptons often include the charge as an additional feature. This would similarly break the symmetry in the lepton ordering, thus making the use of a model that does not have this symmetry less of a drawback.

# 6. Method Evaluation

This chapter presents the evaluation of the machine learning-based jet assignment methods described in Chapter 5 and compares their performance to the conventional baseline methods outlined in Section 4.2.2.

## 6.1. Accuracy Metrics

| Method | Assignment Accuracy | Selection Accuracy |
|---|---|---|
| $\Delta R(\ell, j)$-Method | $0.6344^{+0.0006}_{-0.0006}$ | $0.9232^{+0.0004}_{-0.0003}$ |
| $\chi^2$-Method($\nu^2$-Flows) | $0.7383^{+0.0006}_{-0.0005}$ | $0.9233^{+0.0004}_{-0.0004}$ |
| Transformer | $0.8179^{+0.0005}_{-0.0005}$ | $0.9582^{+0.0003}_{-0.0003}$ |

**Table 6.1.:** Reconstruction accuracies for the different jet assignment methods on the test dataset.

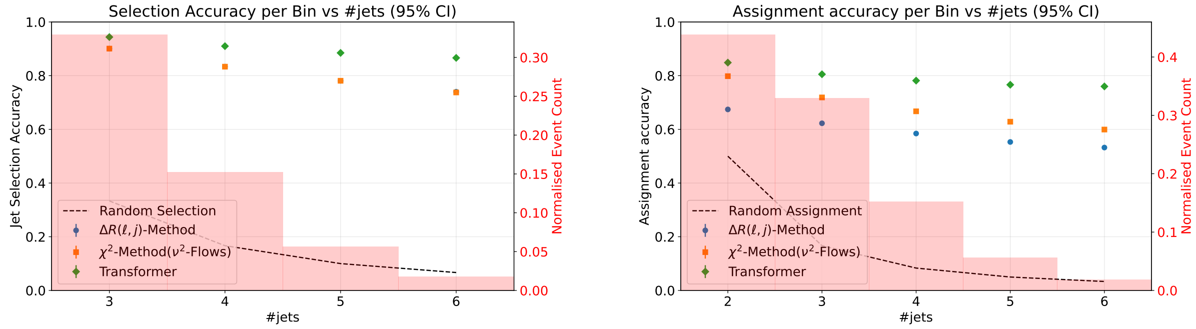
Since the jet assignment task is the core focus of this thesis, the primary evaluation metric is the assignment accuracy, as introduced in Section 5.2.1. Table 6.1 summarizes the assignment accuracies achieved by the different methods on the test dataset. The results show, that the transformer-based jet assignment model only slightly outperforms the conventional methods, in terms of selection accuracy. This is generally not surprising, all methods strongly rely on the $b$-tagging performance to select the two $b$-jet candidates. While it is possible for the transformer to learn patterns in the kinematic features of the jets to improve the selection, the $b$-tagging information is still the dominant factor. The transformer achieves a selection accuracy of 95.82%, compared to 92.33% for both conventional methods, resulting in an improvement of about 3.8%.

However, when looking at the assignment accuracy, which requires not only selecting the correct jets but also assigning them to the correct parton-level $b$-quarks, the transformer provides a significant improvement over the conventional methods. The transformer achieves an assignment accuracy of 81.79%, compared to 73.83% for the $\chi^2$-Method and 63.43% for the $\Delta R$-Method. The corresponds to improvements of about 10.7% and 28.9%, respectively.

Generally, this indicates, that while the conventional methods can effectively select the correct $b$-jets, they struggle with correctly assigning them to the parent partons. The transformer, on the other hand, is able to learn more complex patterns in the event kinematics, leading to a substantial improvement in assignment accuracy.

### 6.1.1. Variable Dependent Performance

To gain further insights into the performance of the jet assignment methods, the accuracy is evaluated as a function of key physics variables.



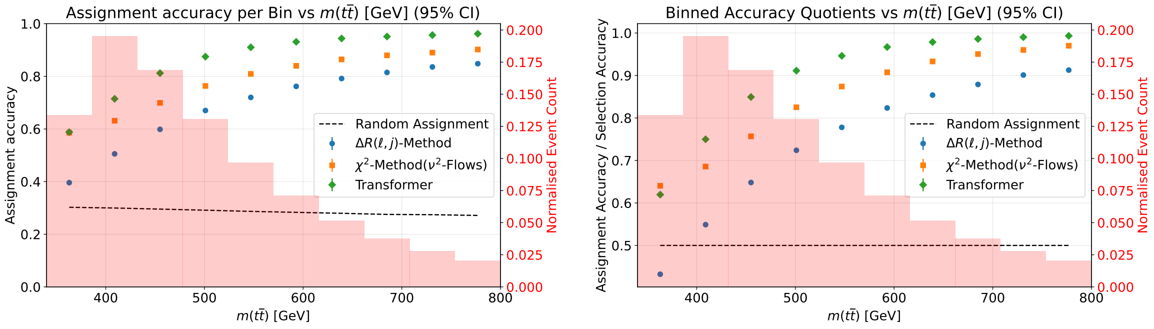
**(a)** Selection Accuracy

**(b)** Assignment Accuracy

**Figure 6.1.:** Accuracy metrics as a function of the number of jets in the event for the different jet assignment methods. The error bars represent the statistical uncertainty. The distribution of events across the bins is shown in the red distribution in the background. A dotted line, indicates the expected accuracy from random guessing based on the number of jet combinations in each bin.

As the combinatorial complexity of the jet assignment task increases with the number of jets in the event, it is expected that the accuracy decreases for events with more jets. Due to the nature of hadron colliders, however, background jets are abundant at the LHC. Figure 6.1 depicts the selection and assignment accuracies as a function of the number of jets in the event. Both accuracies decrease for events with a higher jet multiplicity using all examined methods. However, the transformer consistently outperforms the conventional methods across all jet multiplicities. Notably, the performance gap widens for events with more jets, highlighting the transformer's ability to handle increased combinatorial complexity more effectively.
While the overall selection accuracy showed no major improvement for the transformer over the conventional methods, it is evident from Figure 6.1a, that the transformer achieves a notably higher selection accuracy in events with a larger number of jets. This suggests that the transformer is better at discerning the correct $b$-jets in more complex event topologies, likely due to its ability to learn intricate patterns in the kinematic features of the jets. This allows analyses to utilise events with higher jet multiplicities more effectively, potentially increasing the statistical power of measurements and searches.

As mentioned before, a lot of the interesting physics in dileptonic $t\bar{t}$ events happens close to the production threshold of the top-quark pair. Therefore, it is crucial for jet assignment methods to perform well in this region. Figure 6.2 shows the selection and assignment accuracies as a function of the true $t\bar{t}$ invariant mass (after final state radiation (FSR)). Both metrics and all methods yield a decreasing accuracy for lower $m(t\bar{t})$ values. This behaviour can be understood by considering the event kinematics at different invariant



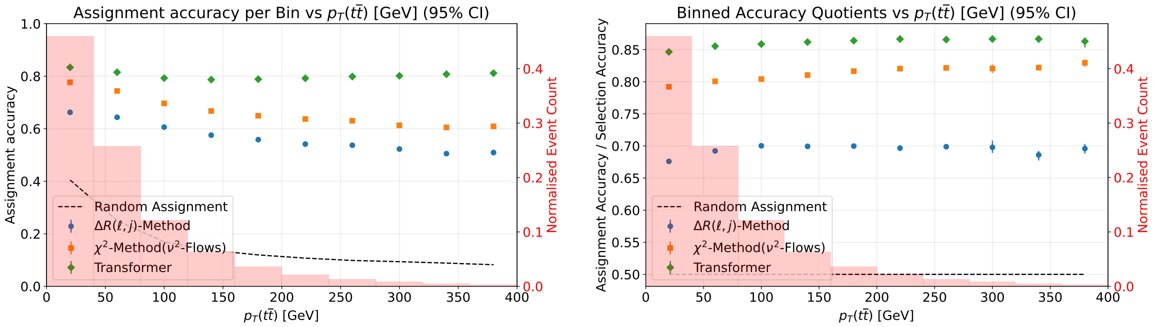
**(a)** Assignment Accuracy

**(b)** Accuracy Quotient

**Figure 6.2.:** Accuracy metrics as a function of the reconstructed $t\bar{t}$ invariant mass for the different jet assignment methods. The error bars represent the statistical uncertainty. The distribution of events across the bins is shown in the red distribution in the background. A dotted line, indicates the expected accuracy from random guessing based on the number of jet combinations in each bin.

mass scales. Close to the production threshold, the top quarks are produced nearly at rest, leading to more isotropic decay products. This results in a higher likelihood of overlapping jets and leptons, complicating the jet assignment task. In contrast, at higher invariant masses, the top quarks are more boosted, causing their decay products to be more collimated and easier to distinguish. This can be seen in a lower $\eta$-$\phi$-separation of the leptons or smaller difference between correct and incorrect lepton jet pairing. The plots, outlining this trend, can be found in the Appendix in Figures A.6 to A.9. This affects the $\Delta R$-method the most, as it only relies on spatial separation for the assignment. While the drop of the transformer is less pronounced, it still suffers from a significant decrease to about 60% compared to way above 90% at higher invariant masses. Given the investigating in Section 5.4.2, the transformer is also highly reliant on these variables. Explaining, why it exhibits a similar drop in performance. In the lowest bin, the transformer and $\chi^2$-Method seem to be on par. When looking at the accuracy quotient in Figure 6.2b, however, it becomes evident, that the transformer solely outperforms the $\chi^2$-Method due to is increased jet-selection. When factoring out the selection performance, the $\chi^2$-Method can actually outperform the transformer in this regime.

This highlights, that especially close to the production threshold, the neutrino kinematic reconstruction used in the $\chi^2$-Method provides valuable information for the jet assignment task, which the transformer is currently not able to fully exploit. This points towards a promising avenue for future improvements, by combining the transformer with neutrino reconstruction methods to provide a more holistic event interpretation. Interestingly, when looking at the accuracy quotient, the $\Delta R$-Method performs worse than random guessing in the lowest bin. This indicates, that at very low $m(t\bar{t})$ values, the spatial separation between jets and leptons becomes so ambiguous, that the $\Delta R$-Method is systematically misassigning jets. While the other methods still perform above random guessing, this highlights the challenges posed by events near the production threshold.



**(a)** Assignment Accuracy

**(b)** Accuracy Quotient

**Figure 6.3.:** Accuracy metrics as a function of the reconstructed $t\bar{t}$ invariant mass for the different jet assignment methods. The error bars represent the statistical uncertainty. The distribution of events across the bins is shown in the red distribution in the background. A dotted line, indicates the expected accuracy from random guessing based on the number of jet combinations in each bin.

Another important variable to consider is the transverse momentum of the $t\bar{t}$ system, $p_T(t\bar{t})$. This variable is sensitive to additional QCD radiation and can impact the event topology. It was also used in previous ATLAS analyses [36] to categorise events. Figure 6.3 shows the assignment accuracy and accuracy quotient as a function of the true $p_T(t\bar{t})$ (after FSR). The baseline methods exhibit a decreasing accuracy for higher $p_T(t\bar{t})$ values, which can be attributed to the increased jet activity and more complex event topologies associated with higher transverse momenta. This directly relates to the aforementioned challenges with higher jet multiplicities, as additional jets from initial and final state radiation reduces the selection accuracy, which in turn impacts the assignment accuracy. The transformer, however, maintains a relatively stable assignment accuracy across the entire $p_T(t\bar{t})$ spectrum.

When looking at the accuracy quotient in Figure 6.3b, all methods show an increase for higher $p_T(t\bar{t})$ values. This indicates, that while the overall assignment accuracy decreases for the baseline methods, they are still able to effectively assign the jets once the correct ones have been selected. The transformer is able to compensate for the drop in selection accuracy at higher $p_T(t\bar{t})$ values, maintaining a consistent assignment performance.

## 6.2. Impact on Physics Observables

To investigate the possible impact of the improved jet assignment on physics analyses, the reconstruction of key observables is studied. For this purpose, the resolution of several reconstructed variables is compared between the different jet assignment methods. The (relative) resolution is defined as the standard (relative) deviation of the reconstructed and true value of the observable, normalized to the true value. Where the true value is taken from the parton-level information after final state radiation (FSR). Since the main focus of this thesis is on the jet assignment task, the neutrino momenta are reconstructed using

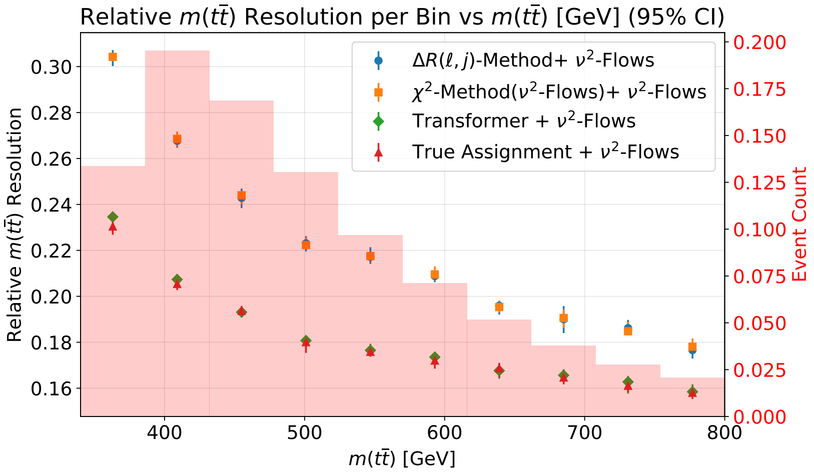


**Figure 6.4.:** Relative resolution of the reconstructed $t\bar{t}$ invariant mass for the different jet assignment methods. The error bars represent the statistical uncertainty.

the same $\nu^2$-Flows method for all jet assignment methods to ensure a fair comparison.

Figure 6.4 shows the relative resolution of the reconstructed $t\bar{t}$ invariant mass for the different jet assignment methods. Notably, this variable is only sensitive to the jet selection, as the invariant mass calculation is symmetric with respect to the two $b$-jets. Albeit the transformer only gains a slight improvement in selection accuracy over the conventional methods, this translates to a significant enhancement in the mass resolution (The binned plot of the selection accuracy can be found in the Appendix in Figure A.10). This difference is most strongly pronounced for events with a low $t\bar{t}$ invariant mass. Also, it can be seen, that the resolution achieved by the transformer almost matches the ideal case of perfect jet selection, which is interesting given that the transformer still makes incorrect selection in about 4% of the events. But, this is very close difference in selection accuracy between the transformer and the conventional methods, which results in a large difference in mass resolution. This can be, due to the way the incorrect selections are distributed. If the conventional methods tend to select jets that are kinematically very different from the true $b$-jets when they make a mistake, this would lead to a larger degradation in mass resolution compared to the transformer. This highlights the importance of not only the selection accuracy but also the nature of the mistakes made by the jet assignment methods.

In Figure 6.5 the resolution of the reconstructed top-quark spin correlation observable $c_{\mathrm{hel}}$ for the different jet assignment methods is plotted. This variable is sensitive to both the jet selection and assignment, because the correct pairing of $b$-jets to leptons is crucial for accurately reconstructing the top-quark decay kinematics. The latter are used to boost the leptons into the top-quark restframe, which is required to compute $c_{\mathrm{hel}}$ (See Section 2.2.4). This makes $c_{\mathrm{hel}}$ an interesting observable to study the impact of jet assignment performance. In contrast to $m(t\bar{t})$, the transformer provides a more moderate improvement in resolution over the conventional methods for $c_{\mathrm{hel}}$. Further, its resolution

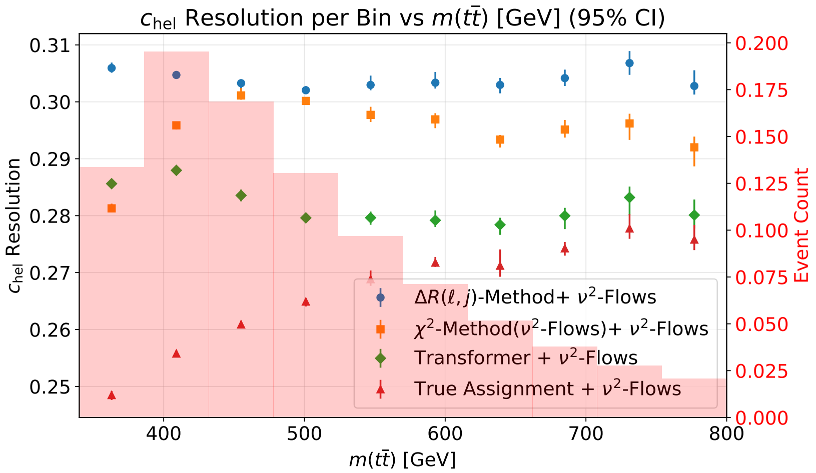


**Figure 6.5.:** Resolution of the reconstructed top-quark spin correlation observable $c_{\rm hel}$ for the different jet assignment methods. The error bars represent the statistical uncertainty. Additional to the different methods, the resolution for the case of perfect jet assignment is shown for reference.

is noticeably worse than the ideal case of perfect jet assignment. This is expected, given that $c_{\rm hel}$ is sensitive to both jet selection and assignment, and the transformer, while significantly better at assignment, still makes mistakes in both tasks. Also, in the lowest bin, the $\chi^2$-Method seems to outperform the transformer. Since both methods achieved a similar assignment accuracy in this regime, this indicates that the $\chi^2$-Method makes its mistakes in a way that has a less detrimental impact on the $C_{\rm han}$ resolution. This might be since the mistakes made by the $\chi^2$-Method more often involve selection a wrong jet, while the transformer might misassign the selected jets more frequently. Since $c_{\rm hel}$ is sensitive to the correct pairing of jets and leptons, misassignments seem to have a larger negative impact on its resolution compared to incorrect selections. The gap between the different assignment methods and the perfect assignment case widens significantly towards the lower $m(t\bar{t})$ values. This is consistent with the findings in Figure 6.2, where all methods suffered a significant drop in assignment accuracy near the production threshold. This highlights the challenges posed by events close to the production threshold, where the kinematics are less distinct, making especially the jet assignment more difficult. It also underscores room for further improvements in jet assignment methods to improve the reconstruction of spin-sensitive observables in this challenging regime. A similar trend can be observed for the resolution of the observable $C_{\rm han}$, which is included in the Appendix in Figure A.13. Since $C_{\rm han}$ also relies on the correct pairing of jets and leptons for its calculation, it exhibits similar sensitivities to jet assignment performance as $c_{\rm hel}$. The transformer again provides a moderate improvement over the conventional methods, but still falls short of the ideal case of perfect jet assignment, especially at low $m(t\bar{t})$ values.

## 6.3. Near-Threshold Optimized Model

Given the challenges associated with jet assignment in events near the $t\bar{t}$ production threshold, a specialized transformer model is trained focusing exclusively on this regime. For this purpose, a dedicated dataset with events from the toponium-resonance ($m(t\bar{t}) < 390$ GeV) is used[1]. One of the mayor advantages of machine learning-based methods is their flexibility to be adapted to specific event categories or kinematic regimes. To test this, the transformer architecture is retrained on specialized datasets, while keeping the hyperparameters fixed to those found during the optimization in Section 5.3. This Section explores the performance of three models:

- **Nominal Trained Transformer**: The transformer model trained on the full dataset, as described in Chapter 5 and evaluated in the previous sections.
- **Toponium Trained Transformer**: A transformer model trained exclusively on events with $m(t\bar{t}) < 390$ GeV.
- **Mixed Trained Transformer**: A transformer model trained on a combination of both the full dataset and the near-threshold dataset, which contains equal parts of events from both regimes.

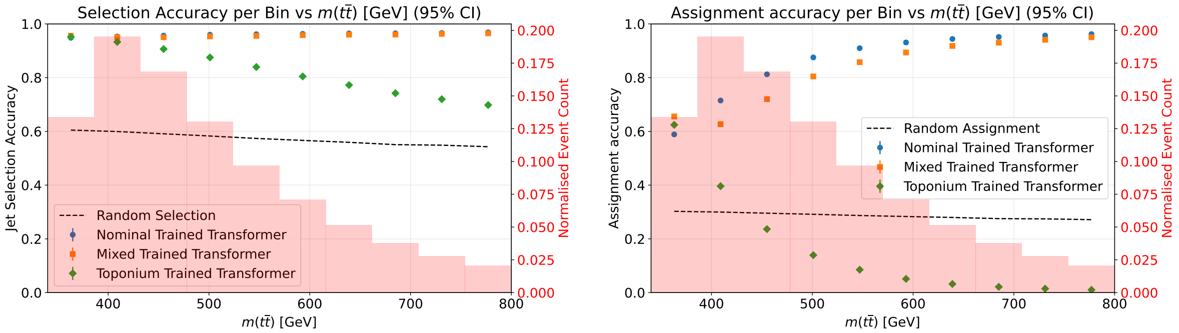
| Method | Assignment Accuracy | Selection Accuracy |
|---|---|---|
| Nominal Trained Transformer | $0.8179^{+0.0006}_{-0.0006}$ | $0.9582^{+0.0003}_{-0.0003}$ |
| Mixed Trained Transformer | $0.7745^{+0.0007}_{-0.0006}$ | $0.9552^{+0.0002}_{-0.0003}$ |
| Toponium Trained Transformer | $0.2348^{+0.0006}_{-0.0006}$ | $0.8588^{+0.0004}_{-0.0004}$ |

**Table 6.2.:** Reconstruction accuracies for the different jet assignment methods on the nominal test dataset.

The performance of the three models is evaluated on the nominal test dataset, which contains events across the entire $m(t\bar{t})$ spectrum. Table 6.2 summarizes the selection and assignment accuracies achieved by the different models. The results show, that the **Toponium Trained Transformer** significantly underperforms compared to the other two models when evaluated on the full dataset. Given its still high selection accuracy of 85.9%, this indicates, the assignment accuracy of only 23.48% is much lower, than the accuracy randomly assigning the correctly selected jets would yield. This suggests, that the model has picked up a special set of features or patterns specific to the near-threshold events, which do not generalize well to the broader event population. This highlights, the kinematic differences between near-threshold events and those at higher invariant masses. The **Mixed Trained Transformer** performs much better, achieving

[1] Note, that the use of a sample containing events from the toponium-resonance, might introduce additional biases from the diverging angular correlation in these events compared to the nominal sample. It was used, because it was the only Monte Carlo sample containing only low $m(t\bar{t})$ events available at the moment.

an assignment accuracy of 77.5%, which is closer to the nominal model's performance of 81.79%. Its selection accuracy is almost equal to that of the nominal model.



**(a)** Selection Accuracy

**(b)** Assignment Accuracy

**Figure 6.6.:** Accuracy metrics as a function of the reconstructed $t\bar{t}$ invariant mass for the differently trained models. The error bars represent the statistical uncertainty. The distribution of events across the bins is shown in the red distribution in the background. A dotted line, indicates the expected accuracy from random guessing based on the number of jet combinations in each bin.

To investigate how the models behave across the $m(t\bar{t})$ spectrum, Figure 6.6 shows the selection and assignment accuracies as a function of the true $t\bar{t}$ invariant mass for the different models. While the selection accuracy of the **Toponium Trained Transformer** slowly decreases for higher invariant masses, the two other models maintain a very stable selection performance across the entire spectrum. This indicates, that the features learned by the **Toponium Trained Transformer** for jet selection do not generalize well to higher mass events, while the other two models have learned more robust selection criteria.
For the assignment accuracy, the **Toponium Trained Transformer** declines steeply in performance as the invariant mass increases, further highlighting its lack of generalization. Its performance drops below that of random guessing for a large part of the spectrum. In contrast, the **Mixed Trained Transformer** closely follows the performance of the nominal model across the entire mass range, albeit with a slightly lower accuracy in the region above 400 GeV. In the lowest $m(t\bar{t})$, however, it performs best. This suggests, that while the model is able to pick up the specific features of both ends of the spectrum, it can only do so at the cost of a small decrease in performance in the intermediate mass region. This can be understood, as the model has difficulty reconciling the differing kinematic patterns present in the two regimes, leading to a compromise in performance. Especially in the intermediate region, where neither set of features is dominant, and the model struggles to tell from which regime the event originates.

# 7. Conclusion

## 7.1. Summary

In this work, a novel machine learning-based approach for the assignment of $b$-jets to their parent top quarks in dileptonic $t\bar{t}$ events has been presented. By leveraging transformer architectures, significant improvements over traditional methods have been achieved, enhancing the accuracy of event reconstruction in this challenging decay channel. Two primary model architectures were explored: one exploiting maximum symmetry between the input particles, and another incorporating physics-inspired features. Both models demonstrated superior performance compared to existing techniques, with the physics-inspired model showing a slight edge in assignment accuracy. The models were trained and evaluated on simulated datasets, with careful consideration of hyperparameter tuning and training strategies. Extensive performance evaluation and comparison with baseline methods confirmed the effectiveness of the proposed approach. The transformer models could improve the selection efficiency of correct $b$-jets and significantly enhance the assignment accuracy. It was shown, that the improved jet assignment leads to better reconstruction of key physics observables, such as the $t\bar{t}$ invariant mass and angular distributions, which are crucial for precision measurements and searches for phenomena beyond the SM.

The dependence of the model performance on the physics regions was also investigated, revealing that the models' performance significantly drops in phase space regions with low $m(t\bar{t})$. While the models still outperform traditional methods in these regions, this observation highlights the need for further research to enhance model robustness across all kinematic regimes. Especially, since these regions are of particular interest for current measurements.

Based on these findings, the influence of the underlying physics on the model's performance was studied. For this, events were categorized based on various kinematic variables, and the accuracy of the methods was analysed within these categories. This analysis provided insights into the strengths and limitations of the model, guiding future improvements. The importance of the input features was also assessed using permutation importance techniques, revealing why the model struggles in certain kinematic regions.

To address the performance loss, it was investigated how emphasizing these regions during training can improve the performance of the model. By adjusting the loss function to give more weight to events in these challenging regions, a noticeable improvement in assignment accuracy was achieved, demonstrating the potential of targeted training strategies.

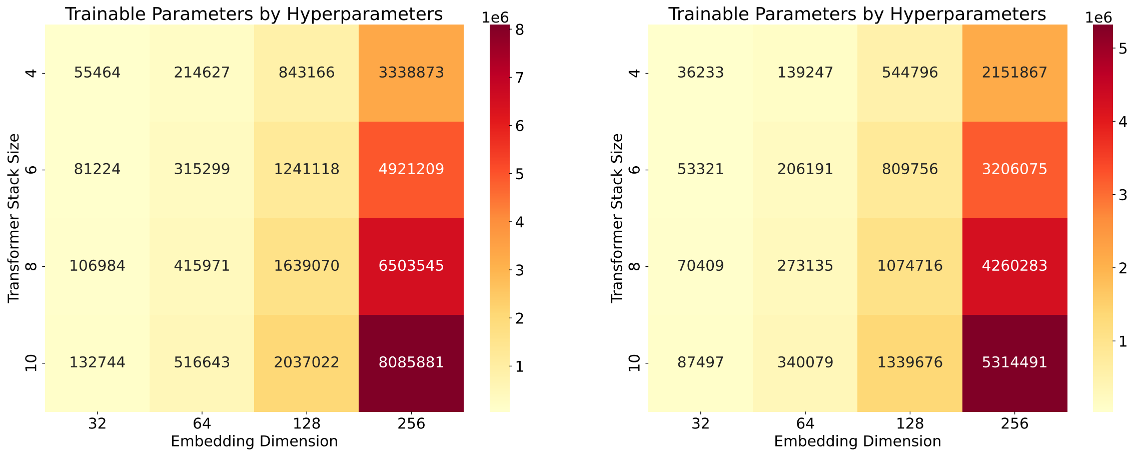
Overall, this work establishes a strong foundation for the application of advanced machine learning techniques in the reconstruction of dileptonic $t\bar{t}$ events, paving the way for future developments in this area.

## 7.2. Outlook

Building on the promising results of this study, several avenues for future research and development can be pursued to further enhance the performance and applicability of machine learning models in dileptonic $t\bar{t}$ event reconstruction.
For once, the integration of the jet assignment model with neutrino momentum regression techniques will be explored to create a comprehensive event reconstruction pipeline. This promises to enable a more holistic approach to reconstructing dileptonic $t\bar{t}$ events, potentially leading to even greater improvements in accuracy and resolution of physics observables. Apart from that, the shared information processing for both processes can lead to an increased efficiency and reduced computational requirements. While the increase in performance was one of the major benefits of using machine learning in the all-hadronic channel, the machine learning based approach for the dilepton channel is computationally much more expansive than traditional methods. A combined model might help to mitigate this issue.
Additionally, further refinement of the model architectures and training strategies will be investigated. This includes exploring more sophisticated transformer variants, experimenting with different loss functions, and incorporating additional physics-inspired features to enhance model performance.
Lastly, the application of the developed models to real experimental data from the LHC could be pursued. This would involve validating the models' performance in a real-world setting and assessing their impact on precision measurements and new physics searches in dileptonic $t\bar{t}$ events.

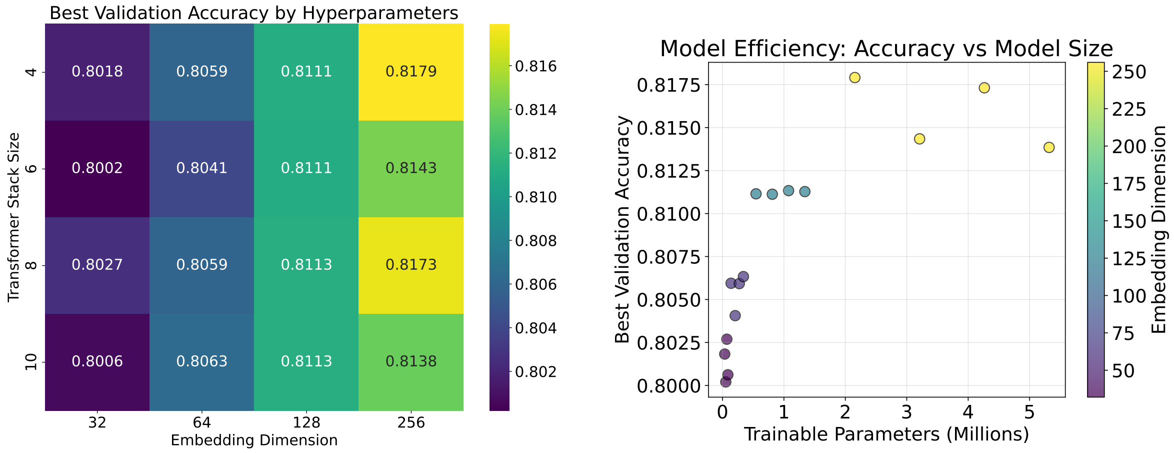
# A. Additional Plots

## A.1. Model and Hyperparameter Studies



**(a) CrossAttentionTransformer** **(b) FeatureConcatTransformer**

**Figure A.1.:** Number of trainable parameters for different hyperparameter combinations for (a) the **CrossAttentionTransformer** and (b) the **FeatureConcatTransformer** architectures. Each square represents a model trained with the corresponding hyperparameter combination.



**(a)** Assignment Accuracy

**(b)** Performance Efficiency

**Figure A.2.:** **FeatureConcatTransformer** with additional high-level features (HLF): (a) Assignment accuracy and (b) performance efficiency on the validation dataset for different hyperparameter combinations. Each dot represents a model trained with the corresponding hyperparameter combination. The dots are coloured according to the embedding dimension used in the model.

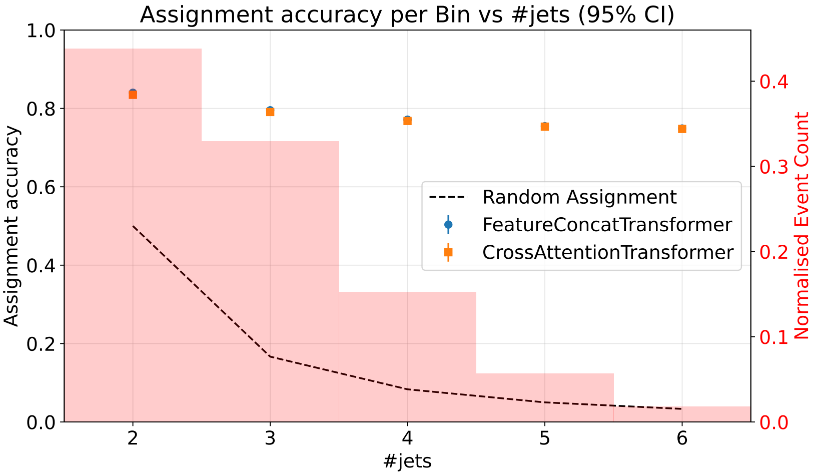


**Figure A.3.:** Assignment Accuracay for the two model architectures binned for different numbers of jets in the event.

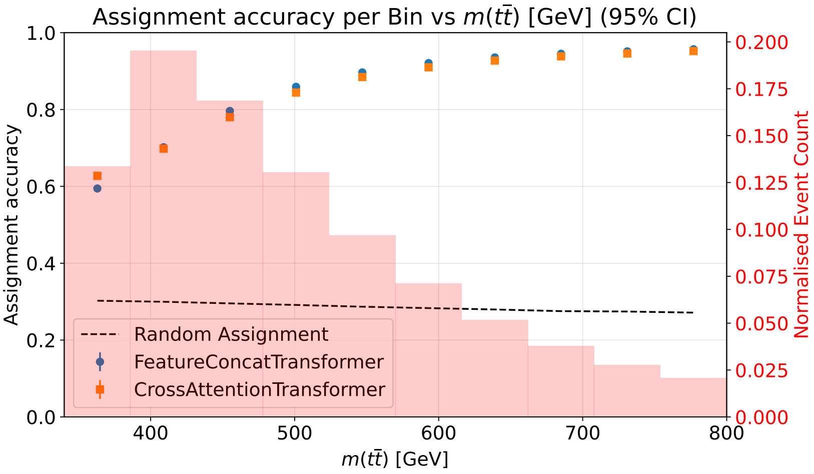


**Figure A.4.:** Assignment Accuracay for the two model architectures binned in true $m(t\bar{t})$.

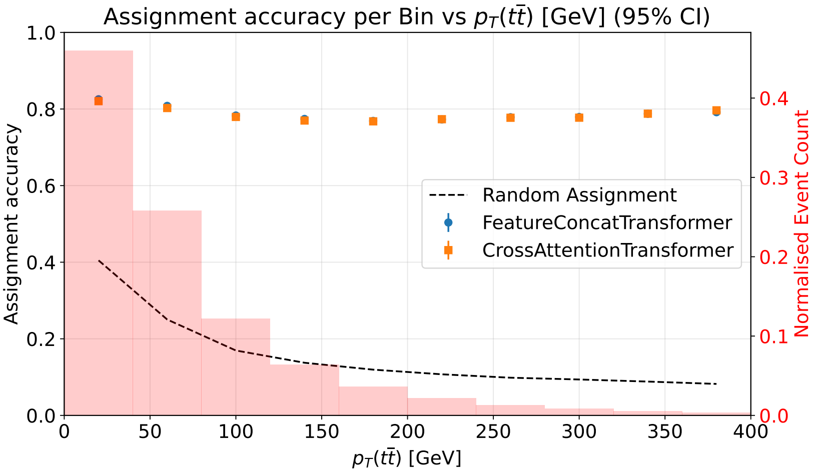
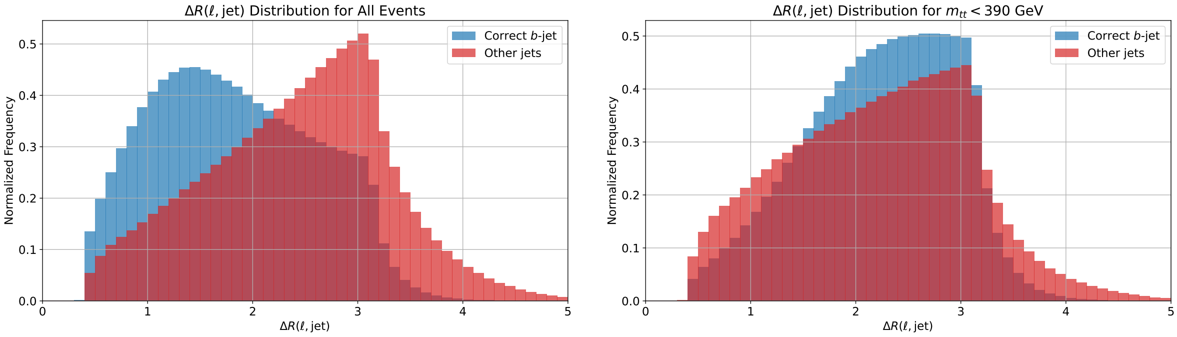


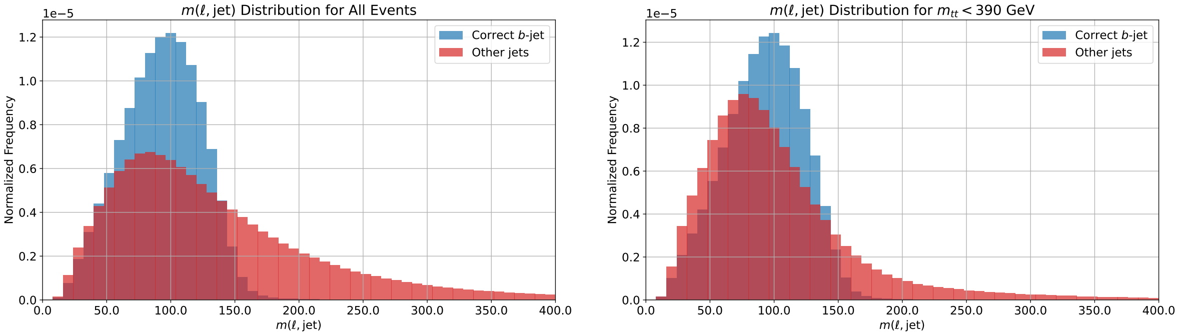
**Figure A.5.:** Assignment Accuracay for the two model architectures binned in true $p_T(t\bar{t})$.

## A.2. Feature Distributions



**(a)** Inclusive **(b)** Near Threshold ($m_{t\bar{t}} < 390$ GeV)

**Figure A.6.:** Distribution of the angular distance $\Delta R$ between the $b$-jets and leptons for the correct and incorrect jet-lepton assignments. (a) shows the inclusive distribution, while (b) focuses on events near the $t\bar{t}$ production threshold.



**(a)** Inclusive **(b)** Near Threshold ($m_{t\bar{t}} < 390$ GeV)

**Figure A.7.:** Distribution of the invariant mass $m(\ell, \text{jet})$ of the jet-lepton pairs for the correct and incorrect jet-lepton assignments. (a) shows the inclusive distribution, while (b) focuses on events near the $t\bar{t}$ production threshold.

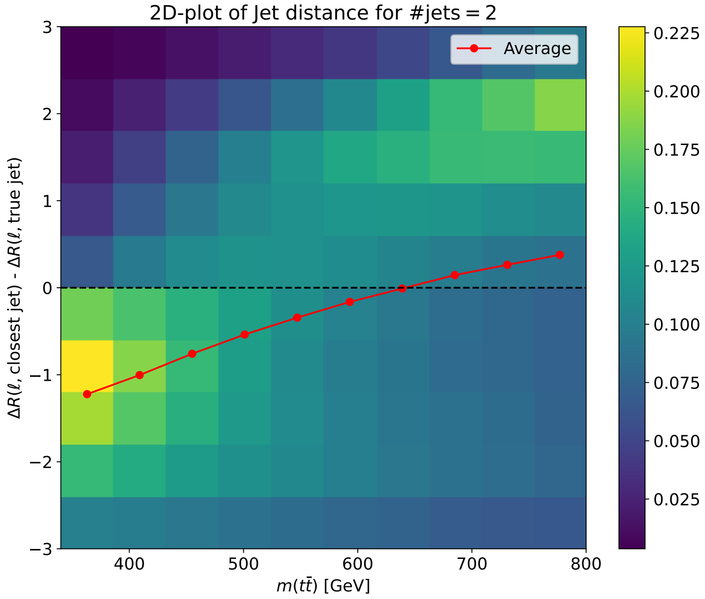


**Figure A.8.:** Correlation between the $t\bar{t}$ invariant mass and difference in angular between the correct and closest incorrect lepton-jet pairing $\Delta R_{\text{lep, non-match}}$. Events where this difference is negative indicate that the incorrect pairing is closer in $\Delta R$ than the correct one. This occurs more frequently at low $t\bar{t}$ invariant masses, highlighting the challenge of jet assignment in this regime. Here, only events with exactly two jets are considered to avoid confounding effects from additional jets.

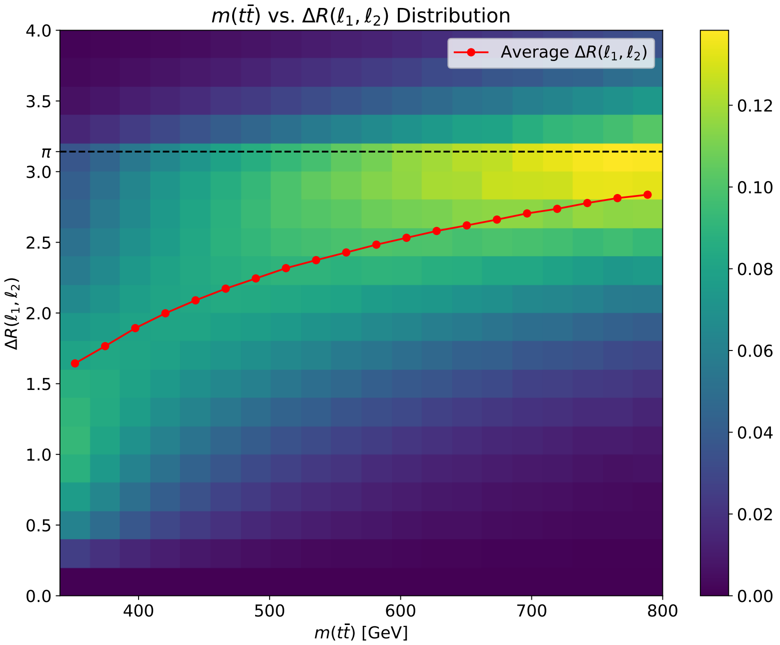


**Figure A.9.:** Correlation between the $t\bar{t}$ invariant mass and the angular distance between the two leptons $\Delta R(\ell^+, \ell^-)$. At low $t\bar{t}$ invariant masses, the leptons tend to be closer together.

## A.3. Performance Evaluation Plots

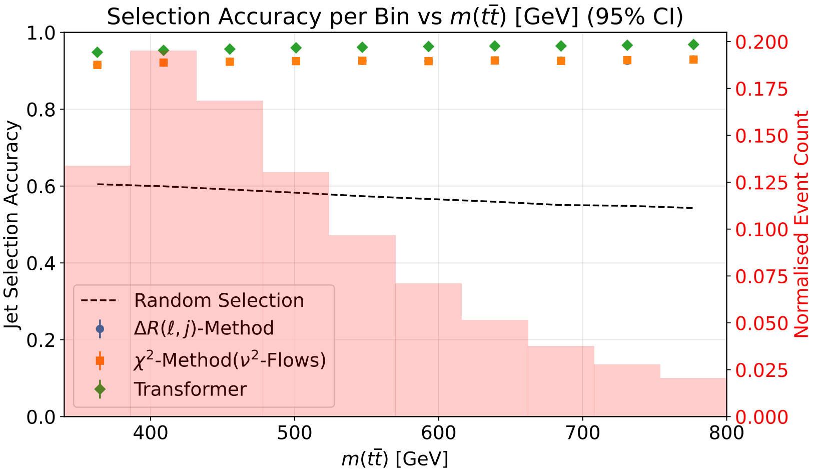


**Figure A.10.:** Selection Accuracy for the different methods binned in true $m(t\bar{t})$.

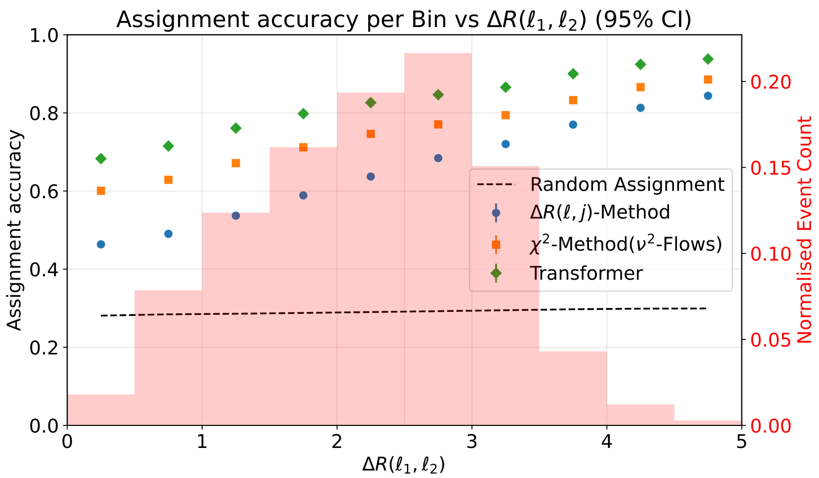
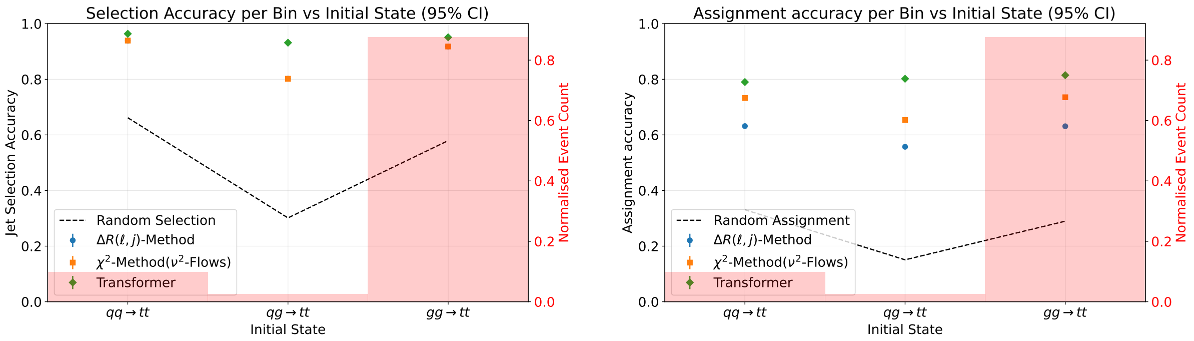


**Figure A.11.:** Assignment Accuracy for the different methods binned in $\Delta R(\ell^+, \ell^-)$.



**(a)** Selection Accuracy

**(b)** Assignment Accuracy

**Figure A.12.:** Assignment and Selection Accuracy for the different methods binned in the initial state of the event. The initial state is categorized based on the flavours of the incoming partons: gluon-gluon (gg), quark-antiquark (qqbar), and quark-gluon (qg).

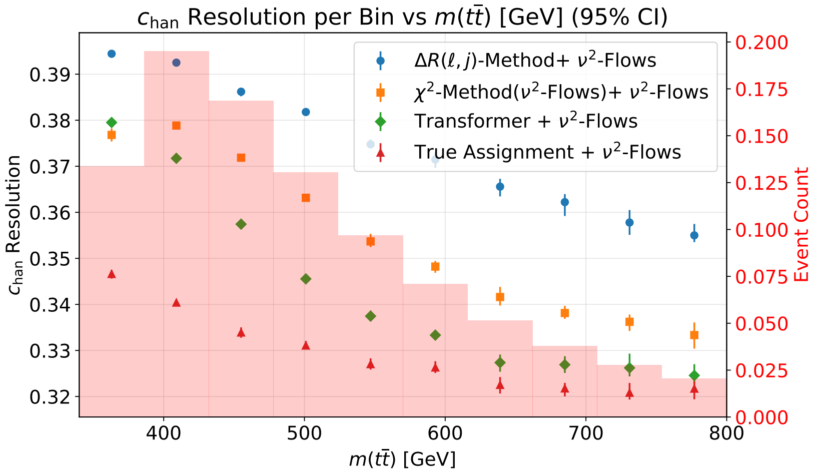


**Figure A.13.:** Resolution of the reconstructed top quark spin correlation observable $C_{\mathrm{han}}$ for the different jet assignment methods. The error bars represent the statistical uncertainty. Additionally to the different methods, the resolution for the case of perfect jet assignment is shown for reference.

# Bibliography

[1] ATLAS Collaboration, *Observation of quantum entanglement with top quarks at the ATLAS detector*, Nature **633**, 542 (2024)

[2] ATLAS Collaboration, *Observation of a cross-section enhancement near the t bart production threshold in sqrts* = 13 *TeV pp collisions with the ATLAS detector*, Technical Report ATLAS-CONF-2025-008, ATLAS (2025)

[3] CMS Collaboration, *Observation of a pseudoscalar excess at the top quark pair production threshold*, Rep. Prog. Phys. **88**, 087801 (2025)

[4] M. Jeżabek et al., *V - A tests through leptons from polarised top quarks*, Physics Letters B **329(2)**, 317 (1994)

[5] A. Brandenburg et al., *QCD-corrected spin analysing power of jets in decays of polarized top quarks*, Physics Letters B **539(3)**, 235 (2002)

[6] A. Shmakov et al., *SPANet: Generalized permutationless set assignment for particle physics using symmetry preserving attention*, SciPost Phys. **12**, 178 (2022)

[7] S. Navas et al. (Particle Data Group Collaboration), *Review of Particle Physics*, Phys. Rev. D **110**, 030001 (2024)

[8] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22**, 579 (1961)

[9] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19**, 1264 (1967)

[10] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. C **680519**, 367 (1968)

[11] D. J. Gross et al., *Asymptotically Free Gauge Theories*, Phys. Rev. D **8**, 3633 (1973)

[12] H. D. Politzer, *Reliable Perturbative Results for Strong Interactions*, Phys. Rev. Lett. **30**, 1346 (1973)

[13] G. 't Hooft, *Renormalizable Lagrangians For Massive Yang-Mills Fields*, Nucl. Phys. B **35**, 167 (1971)

[14] P. W. Higgs, *Broken Symmetries, Massless Particles and Gauge Fields*, Phys. Lett. **12**, 132 (1964)

[15] F. Englert et al., *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13**, 321 (1964)

[16] G. S. Guralnik et al., *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13**, 585 (1964)

[17] T. D. Lee et al., *Question of Parity Conservation in Weak Interactions*, Phys. Rev. **104**, 254 (1956)

[18] C. S. Wu et al., *Experimental Test of Parity Conservation in Beta Decay*, Phys. Rev. **105**, 1413 (1957)

[19] N. Cabbibo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10**, 531 (1963)

[20] M. Kobayashi et al., *CP Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49**, 652 (1973)

[21] H. D. Politzer, *Asymptotic Freedom: An Approach to Strong Interactions*, Phys. Rept. **14**, 129 (1974)

[22] J. C. Collins et al., *Heavy Particle Production in High-Energy Hadron Collisions*, Nucl. Phys. B **263**, 37 (1986)

[23] T. Sjöstrand, *A model for initial state parton showers*, Physics Letters B **157(4)**, 321 (1985)

[24] G. Marchesini et al., *Simulation of QCD jets including soft gluon interference*, Nuclear Physics B **238(1)**, 1 (1984)

[25] T. Sjöstrand et al., *PYTHIA*, Comput. Phys. Commun. **135**, 238 (2001)

[26] B. Andersson et al., *Parton fragmentation and string dynamics*, Physics Reports **97(2)**, 31 (1983)

[27] T. Sjöstrand, *Jet fragmentation of multiparton configurations in a string framework*, Nuclear Physics B **248(2)**, 469 (1984)

[28] J. J. Aubert et al., *Experimental Observation of a Heavy Particle J*, Phys. Rev. Lett. **33**, 1404 (1974)

[29] F. Abe et al. (CDF), *Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab*, Phys. Rev. Lett. **74**, 2626 (1995)

[30] S. Abachi et al. (DØ), *Observation of the Top Quark*, Phys. Rev. Lett. **74**, 2632 (1995)

[31] I. Bigi et al., *Production and decay properties of ultra-heavy quarks*, Physics Letters B **181(1)**, 157 (1986)

[32] G. Mahlon et al., *Spin correlation effects in top quark pair production at the LHC*, Phys. Rev. D **81**, 074024 (2010)

[33] E. Schrödinger, *Discussion of Probability Relations between Separated Systems*, Mathematical Proceedings of the Cambridge Philosophical Society **31(4)**, 555 (1935)

[34] A. Peres, *Separability Criterion for Density Matrices*, Phys. Rev. Lett. **77**, 1413 (1996)

[35] W. K. Wootters, *Entanglement of Formation of an Arbitrary State of Two Qubits*, Phys. Rev. Lett. **80**, 2245 (1998)

[36] Y. Afik et al., *Entanglement and quantum tomography with top quarks at the LHC*, The European Physical Journal Plus **136(9)**, 907 (2021)

[37] Y. Kiyo et al., *Top-quark pair production near threshold at LHC*, Eur. Phys. J. C **60**, 375 (2009)

[38] Y. Sumino et al., *Bound-state effects on kinematical distributions of top quarks at hadron colliders*, Journal of High Energy Physics **2010(9)**, 34 (2010)

[39] K. Hagiwara et al., *Bound-state effects on top quark production at hadron colliders*, Physics Letters B **666(1)**, 71 (2008)

[40] B. Fuks et al., *Signatures of toponium formation in LHC run 2 data*, Phys. Rev. D **104**, 034023 (2021)

[41] B. Fuks et al., *Simulating toponium formation signals at the LHC*, The European Physical Journal C **85(2)**, 157 (2025)

[42] L. Sonnenschein, *Analytical solution of $t\bar{t}$ dilepton equations*, Phys. Rev. D **73**, 054015 (2006)

[43] V. Abazov et al., *Measurement of the top quark mass in the dilepton channel*, Physics Letters B **655(1)**, 7 (2007)

[44] V. M. Abazov et al. (The D0 Collaboration), *Measurement of the top quark mass in final states with two leptons*, Phys. Rev. D **80**, 092006 (2009)

[45] J. A. Raine et al., *Fast and improved neutrino reconstruction in multineutrino final states with conditional normalizing flows*, Phys. Rev. D **109**, 012005 (2024)

[46] W. Bernreuther et al., *Top quark pair production and decay at hadron colliders*, Nucl. Phys. B **690**, 81 (2004)

[47] W. Bernreuther et al., *A set of top quark spin correlation and polarization observables for the LHC: Standard Model predictions and new physics contributions*, J. High Energy Phys. **2015(12)**, 1 (2015)

[48] C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal **27(3)**, 379 (1948)

[49] F. Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain.*, Psychological Review **65(6)**, 386 (1958)

[50] V. Nair et al., *Rectified linear units improve restricted boltzmann machines*, in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, Omnipress, Madison, WI, USA (2010)

[51] D. E. Rumelhart et al., *Learning representations by back-propagating errors*, Nature **323(6088)**, 533 (1986)

[52] J. L. Elman, *Finding structure in time*, Cognitive Science **14(2)**, 179 (1990)

[53] S. Hochreiter et al., *Long Short-Term Memory*, Neural Computation **9(8)**, 1735 (1997)

[54] A. Vaswani et al., *Attention is All you Need*, in I. Guyon et al., editors, *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc. (2017)

[55] J. Kiefer et al., *Stochastic Estimation of the Maximum of a Regression Function*, The Annals of Mathematical Statistics **23(3)**, 462 (1952)

[56] D. P. Kingma et al., *Adam: A Method for Stochastic Optimization*, International Conference on Learning Representations (ICLR) (2015)

[57] N. Srivastava et al., *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, Journal of Machine Learning Research **15(56)**, 1929 (2014)

[58] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, J. High Energy Phys. **11**, 040 (2004)

[59] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, J. Instrum. **3**, S08003 (2008)

[60] M. Cacciari et al., *The anti-$k_t$ jet clustering algorithm*, J. High Energy Phys. **04**, 063 (2008)

[61] ATLAS Collaboration, *Transforming jet flavour tagging at ATLAS*, Technical report, CERN, Geneva (2025), submitted to: Nature Communications, 2505.19689

[62] S. Mondal et al., *Machine learning in high energy physics: a review of heavy-flavor jet tagging at the LHC*, The European Physical Journal Special Topics **233(15)**, 2657 (2024)

[63] I. Loshchilov et al., *Decoupled Weight Decay Regularization*, in *International Conference on Learning Representations* (2019)

[64] L. Breiman, *Random Forests*, Machine Learning **45(1)**, 5 (2001)