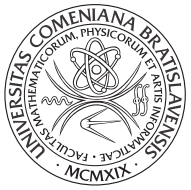
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



Oliver Majerský

Autoreferát dizertačnej práce

Charge asymmetry in boosted top quark pair production in *pp* collisions at the ATLAS experiment

na získanie akademického titulu philosophiae doctor

v odbore doktorandského štúdia:

Jadrová a subjadrová fyzika

Miesto a dátum:

Bratislava, 17.7. 2020

Dizertačná práca bola vypracovaná
v dennej forme doktorandského štúdia

na Katedre jadrovej fyziky a biofyziky

Predkladateľ: Mgr. Oliver Majerský
Katedra jadrovej fyziky a biofyziky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského
Mlynská dolina F1
842 48 Bratislava

Školiteľ: prof. RNDr. Jozef Masarik, DrSc.
Katedra jadrovej fyziky a biofyziky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského
Mlynská dolina F1
842 48 Bratislava

Študijný odbor: jadrová a subjadrová fyzika
Študijný program: jadrová a subjadrová fyzika (Jednoodborové štúdium, doktorandské III. st., denná forma)

Predseda odborovej komisie:

prof. RNDr. Jozef Masarik, DrSc.
Katedra jadrovej fyziky a biofyziky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského
Mlynská dolina F1
842 48 Bratislava

1 Introduction

The Standard Model (SM) of elementary particles encapsulates our current knowledge of the interactions of particles of the visible matter in the universe, successfully providing a huge amount of predictions which were experimentally confirmed over the course of its existence of roughly 50 years. Despite its success, the SM does not provide answers for several key questions on the nature of our universe, such as the explanation of the dark matter and dark energy, the baryon asymmetry in the universe, or the missing description of gravity at the quantum level. Therefore, many various experiments in the field of particle physics seek to find evidence for theories beyond SM (BSM), which could explain (some of) these phenomena.

The ATLAS experiment is one of the four experiments using proton-proton collisions of the Large Hadron Collider (LHC), the largest and most powerful proton-proton collider ever built. The ATLAS detector is a multi-purpose detector designed for studying various processes to both provide stringent tests of the SM predictions as well as to search for clues of BSM physics. From the year 2015 until 2018, the Large Hadron Collider operated at the unprecedented centre-of-mass energy $\sqrt{s} = 13\,\text{TeV}$. During this period, known as the Run-II, the ATLAS experiment has collected pp collision data of the approximate integrated luminosity of $140\,\text{fb}^{-1}$, surpassing the Run-I period of data taking in 2011-2012 by roughly a factor of seven. Combined with the increase of the collision energy from 8 TeV to 13 TeV, the LHC has allowed the ATLAS experiment to explore even rarer processes, and more extreme regions of phase space.

Many of the BSM theories predict the existence of massive particles decaying into top quarks or $W/Z/H$ bosons. The top quarks and the bosons are unstable particles further decaying into lighter particles. To study these decays, the particles must be reconstructed by the ATLAS detector to infer the properties of their decaying parent particles. In the momenta range of hundreds of GeV and more, the massive top quarks and the bosons become boosted, and thus their decay products highly collimated. Dedicated reconstruction techniques are necessary to identify these decays, referred to as *boosted* tagging. Part of this thesis is dedicated to investigation of the boosted tagging of top-quarks and W bosons, specifically the performance and accuracy of Monte Carlo simulations of various algorithms designed to identify these particles.

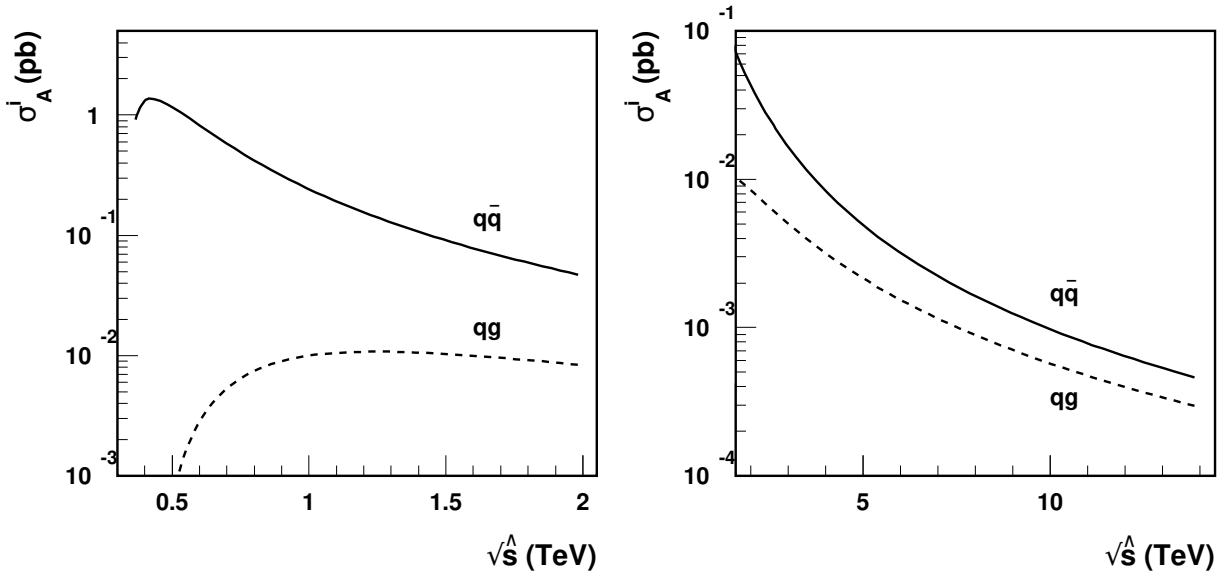
The large ATLAS Run-II dataset allows to measure very rare SM phenomena, one of which is the so-called charge asymmetry in the top-quark pair ($t\bar{t}$) production, describing a phenomenon in the production of the $t\bar{t}$ pairs, which leads to angular asymmetry of the distributions of top-quark and top anti-quark production with respect to the initial-state particles interacting in the pp collision. A measurement of the $t\bar{t}$ charge asymmetry at the $\sqrt{s} = 13\,\text{TeV}$ is performed in this thesis, with dedicated studies of highly-boosted top quarks, employing the boosted tagging techniques. Traditionally, this measurement is performed in the single-lepton channel, where one of the top-quark decays includes an electron or a muon. This is because the measurement of the charge asymmetry requires distinguishing top quark from top anti-quark, and the electric charge of the electron or muon can be used to infer the electric charge of the decaying top (anti-)quark with very high precision. A possibility of measuring $t\bar{t}$ charge asymmetry in the all-hadronic channel is investigated, which is the decay channel where both top-quarks decay hadronically, i.e. decay into quarks, thus having no prompt electrons or muons from the top-quark decay. In this study, alternative approaches to reconstructing the charge of the decaying top-quark are investigated, using the electric charge of the decay products of the hadronic top quark.

2 Top quark and the charge asymmetry

The top quark is the third-generation up-type quark. It was discovered in 1995 at the Tevatron accelerator by the CDF and D0 experiments, completing the three generations of quarks predicted by the SM. It is the heaviest known elementary particle with a mass of approximately 173 GeV [1]. Because of the large mass, the top quark has a very large predicted decay width $\Gamma = 1.32$ GeV [2], resulting in an extremely short mean life-time of $\approx 10^{-25}$ s. It is in fact order of magnitude less than the time it takes for the quark to hadronise ($\approx 10^{-23}$ s), i.e. produce shower of hadrons, bound states of quarks. This means that the properties of the top quark are passed directly onto its decay products, allowing for precise measurements of properties of a "pseudo-bare" quark.

At hadron colliders, top quarks are most abundantly produced via $t\bar{t}$ pair production which is driven by strong interactions, though other production channels are also investigated, such as the single top quark production via weak interaction.

At the LHC, the $t\bar{t}$ is produced primarily via the gluon-gluon fusion, where each of the gluons originates from one of the colliding protons. The second most dominant contribution to the $t\bar{t}$ production is the $q\bar{q}$ annihilation ($q\bar{q} \to t\bar{t}$). The charge asymmetry in the context of pair production of quarks describes an anisotropy in the angular distributions of final-state quark and anti-quark production. The SM predicts that in the $q\bar{q} \to t\bar{t}$ production, top quarks are produced more frequently in the direction of the incoming quark q rather than $\bar{q}$. This effect is a result of higher-order perturbative corrections [3, 4] to the scattering amplitude of the $q\bar{q} \to t\bar{t}$ process, calculated in the framework of quantum field theory. An illustration of the dependence of the charge-asymmetric production cross-section vs the centre-of-mass energy of the colliding partons is shown in Fig. 1.

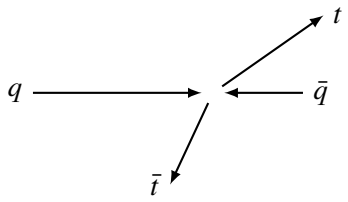


(a) (b)

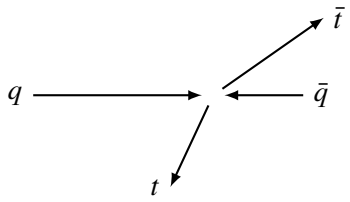
Fig. 1: Dependence of the asymmetric cross-section part on the partonic centre-of-mass energy [3], for $q\bar{q} \to t\bar{t}$ processes and for $qg \to t\bar{t}$.

At the LHC, measuring the charge asymmetry is challenging due to the symmetric initial state in the pp collision, thus there is no preferred direction in which a $q\bar{q} \to t\bar{t}$ process occurs. Nevertheless, the quark q is typically a valence quark from proton and the $\bar{q}$ is a sea quark from proton. Valence quarks on average carry more momentum than sea quarks, and hence the $t\bar{t}$ pair is produced boosted along the

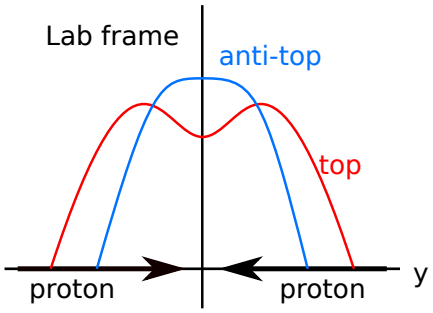
proton beam direction. This means that a positive charge asymmetry, where the top quark is produced more abundantly in the direction of the quark q, will result in top quarks being more often produced with larger rapidity[1] magnitude, i.e. in forward direction at the detector laboratory frame (Fig. 2a) and the $\bar{t}$ quarks will be produced more abundantly in the central region, as illustrated in Fig. 2c.



(a) t produced in direction of q:
$|y_t| > |y_{\bar{t}}| \Rightarrow \Delta|y| > 0$



(b) $\bar{t}$ produced in direction of q:
$|y_t| < |y_{\bar{t}}| \Rightarrow \Delta|y| < 0$



(c)

Fig. 2: Illustration of the effect of $q\bar{q}$ longitudinal boost in $q\bar{q} \to t\bar{t}$ on the rapidity magnitudes of t (a) and $\bar{t}$ (b) quarks and their rapidity distributions in the laboratory frame at the LHC under positive asymmetry assumption (c). If the asymmetry was zero, the rapidity distributions of t and $\bar{t}$ would coincide.

The difference between the rapidity magnitudes of top quark and top anti-quark can be used to quantify the charge asymmetry experimentally at the LHC:

$$A_C = \frac{N(\Delta|y| > 0) - N(\Delta|y| < 0)}{N(\Delta|y| > 0) + N(\Delta|y| < 0)}, \tag{2}$$

where $\Delta|y| = |y_t| - |y_{\bar{t}}|$ is the difference of the absolute rapidities of top quark and top anti-quark, respectively.

3 Reconstruction of $t\bar{t}$ events at ATLAS

Since the top quark is an unstable particle, the $t\bar{t}$ pairs decay into particles which are reconstructed and their properties measured, using the ATLAS detector. Depending on the final state, two decay channels are considered in this work. In the single-lepton channel one of the top quark decays semi-leptonically into a b-quark, electron or muon, and a corresponding neutrino, and the other top quark decays hadronically, into a pair of quarks and a b-quark. The neutrino passes through the ATLAS detector undetected, however, due

[1] Rapidity of a particle is defined using its energy E and the momentum projection along the proton beamline p_z:

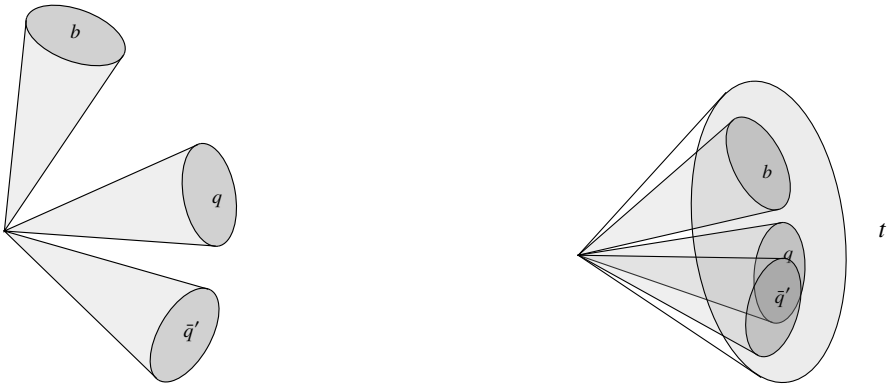
$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z}. \tag{1}$$

to the conservation of energy, the net transverse energy in the detector is expected to sum to zero[2]. The measurement of a non-zero negative vectorial sum of transverse energy is an indication of the presence of undetected particle, and is assumed to originate from the neutrino in the single-lepton channel. Due to the fact that longitudinal momentum of the initial state in the pp collision is not pre-determined, because the colliding quarks and gluons carry in general random fraction of momentum of the colliding protons, the longitudinal momentum component of the neutrino is not determined unambiguously and is estimated based on the constraints on the reconstructed mass of the semi-leptonically-decaying top-quark and the mass of the W boson from the top quark.

Aside from the single-lepton channel, the other channel considered is the all-hadronic channel, where both of the top quarks each decay into a pair of quarks and a b-quark. The obvious advantage of this final state is the absence of neutrino, which means the $t\bar{t}$ system is reconstructed completely, without the neutrino longitudinal momentum ambiguity.

In this work, two methods of $t\bar{t}$ reconstruction are distinguished. The first one is referred to as *resolved* topology, where the individual quarks from both top quark decays are reconstructed individually as showers of collimated particles within the ATLAS detector (referred to as *jets*), as shown in Fig. 3a. A multivariate technique using boosted decision trees (BDTs) is used to find the correct matching of jets to the quarks from the $t\bar{t}$ decay. The $t\bar{t}$ kinematics are inferred from the kinematics of the identified jets, the lepton kinematics, and from the missing transverse momentum, $E_{\mathrm{T}}^{\mathrm{miss}}$.

The second reconstruction approach, referred to as *boosted* topology, is used for hadronically-decaying top quarks with very high transverse momentum $p_{\mathrm{T}} > 350\,\mathrm{GeV}$, where the decay products become highly collimated, resulting in difficulty in resolving individual jets from the three quarks from the decay. Instead, the decay products are reconstructed as a single jet with large radius (large-R jet), as shown in Fig. 3b, and the internal structure of the jet is quantified using jet substructure observables. Various algorithms for the tagging of large-R jets originating from top quarks are studied in the thesis, where most of them apply some selection on substructure observables. These observables discriminate between top-quark jets and the jets originating from multijet background, the dominant background process.



(a) Resolved topology (b) Boosted topology

Fig. 3: Illustration of resolved and boosted topologies of reconstructed top quarks. In the resolved topology (a), the top-quark decay products are reconstructible as individual jets. At sufficiently high p_{T}, all of the top quark decay products can be contained within a single large-R jet as shown in (b).

[2]The transverse momenta of all the reconstructed particles in the ATLAS detector are summed as vectors in the plane perpendicular to the LHC beamline. This vectorial sum is zero, if no undetected particles are present in the final state. In practice, this quantity can fluctuate around zero due to experimental resolution of the various ATLAS detector components.

In Run-II, novel top-quark tagging and W-boson tagging algorithms were developed, which employ machine learning techniques, such as deep neural networks (DNNs) or boosted decision trees (BDTs). These techniques allow to combine many observables sensitive to the substructure of the large-R jets, in order to discriminate signal jets from multijet background. To compare the performance of various tagging algorithms, the signal efficiency ϵ_{sig} and the background efficiency ϵ_{bckg} are used, defined as follows:

$$\epsilon_{\text{sig}} = \frac{N_{\text{signal}}^{\text{tagged}}}{N_{\text{signal}}^{\text{tagged \& untagged}}}, \qquad \epsilon_{\text{bckg}} = \frac{N_{\text{background}}^{\text{tagged}}}{N_{\text{background}}^{\text{tagged \& untagged}}}, \tag{3}$$

with $N_{\text{signal}}^{\text{tagged}}$ $\left(N_{\text{background}}^{\text{tagged}}\right)$ denoting the number of tagged jets which are labelled as signal (background) and $N_{\text{signal}}^{\text{tagged \& untagged}}$ $\left(N_{\text{background}}^{\text{tagged \& untagged}}\right)$ denoting the total number of jets labelled as signal (background), both tagged and untagged. In addition to ϵ_{bckg}, it is common to use the reciprocal quantity $1/\epsilon_{\text{bckg}}$, referred to as background rejection.

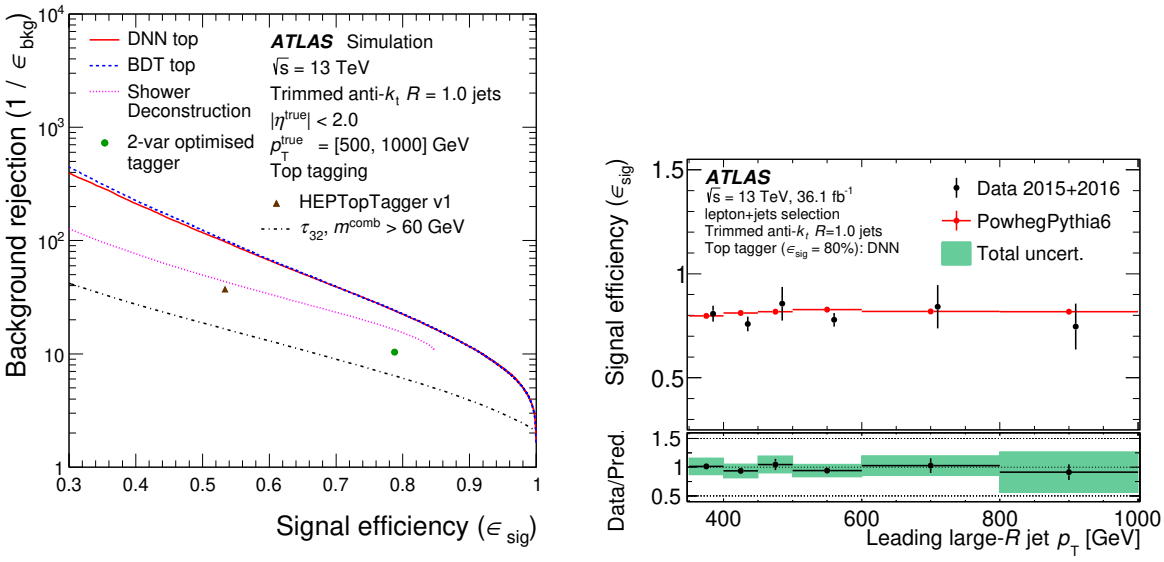
Fig. 4a shows the comparison of dependence of background rejection on signal efficiency based on MC simulations for various top-quark tagging algorithms. These performance comparisons demonstrate that the DNN and BDT outperform alternative approaches in terms of background rejection. In the thesis, these tagging algorithms are also studied using data collected by the ATLAS experiment. Their signal efficiency is measured in $t\bar{t}$ events in the single-lepton channel, and compared with the MC-simulated predictions. In Fig. 4b the measurement of the signal efficiency of the neural-network top-quark tagger is shown. Good agreement between the measurement and MC prediction within the uncertainties is observed.

4 Charge asymmetry measurement in the single-lepton channel

One of the main results presented in the thesis is the measurement of the charge asymmetry in the $t\bar{t}$ production in the single-lepton channel, combining both resolved and boosted topologies as separate selection regions. The measurement is performed using the full Run-II dataset of LHC pp collisions, collected by the ATLAS detector in years 2015–2018, with an integrated luminosity of 139 fb^{-1}.

The selected collision events are required to fire a single-electron or single-muon trigger of the ATLAS trigger and data acquisition system. The events must contain at least one isolated electron or muon with $p_{\text{T}} > 28\,\text{GeV}$, and $E_{\text{T}}^{\text{miss}} > 30\,\text{GeV}$ and the transverse mass of the W boson $M_{\text{T}}^{W} > 30\,\text{GeV}$ for events with an electron, whereas for events containing a muon, $E_{\text{T}}^{\text{miss}} + M_{\text{T}}^{W} > 60\,\text{GeV}$ is required. In addition, events are required to contain at least one anti-k_t $R = 0.4$ jet which is identified as originating from a b-quark. After this pre-selection, the resolved and boosted regions are defined.

In the resolved region, at least four anti-k_t [6] jets with radius parameter $R = 0.4$ with $p_{\text{T}} > 25\,\text{GeV}$ and $|\eta| < 2.5$ are required, and the identification of the correct jet-to-quark assignment is attempted using the BDT multivariate technique. In the boosted region, a single anti-k_t $R = 0.4$ jet close to the lepton must be present ($\Delta R(\text{jet}, \ell) < 1.5$) and a single anti-$k_t$ $R = 1.0$ jet with $p_{\text{T}} > 350\,\text{GeV}$ and $|\eta| < 2.0$, isolated from the leptonically-decaying top quark by additional angular cuts is required. Namely, ΔR(large-R jet, small-R jet) > 1.5 and $\Delta\phi$(large-R jet, ℓ) > 2.3. To identify the large-R jet as originating from hadronically-decaying top quark, a tagger applying selection on the mass [7] of the jet and the N-subjettiness [8] ratio τ_{32} is used. The mass of the jet is expected to peak around the value of top-quark



(a) (b)

Fig. 4: Simulation performance of selected boosted top-quark taggers as a function of the large-R jet transverse momentum ($p_{\mathrm{T}}^{\mathrm{true}}$), for $p_{\mathrm{T}}^{\mathrm{true}} \in [500, 1000]$ GeV (a) [5]. The lines indicate the dependence of background rejection for a given signal efficiency for various taggers. The two top most lines are the machine-learning based neural network (DNN) and boosted decision tree (BDT) taggers. Additionally, the green dot in the plot shows the 2-variable optimised $m_{\mathrm{comb}} + \tau_{32}$ tagger designed to give 80 % efficiency to tag top quarks. In (b) the measured signal efficiency in single-lepton $t\bar{t}$ events is shown (black points), compared with the predicted signal efficiency (red points), as a function of the large-R jet p_{T}. The green band in the bottom plot shows the MC statistical and systematic uncertainties summed in quadrature. The error bars on black points show the statistical uncertainty of the data.

mass of approximately 173 GeV, while the τ_{32} quantifies the compatibility of the jet of containing radiation from three quarks. Both variables discriminate against jets from multijet background, where such jets have a mass much smaller than the mass of the top quark and appear to have radiation pattern from a single quark with soft wide-angle emissions inside the large-R jet.

4.1 Analysis strategy

After the previously-described selection, the events are divided into four regions, resolved 1b-inclusive, resolved 2b-exclusive, boosted 1b-inclusive and boosted 2b-exclusive, where 1b and 2b denote the multiplicity of jets identified as originating from a b-quark. A distribution of $\Delta|y|$ (shown in Fig. 5), $\Delta|y|$ vs $t\bar{t}$ invariant mass $m_{t\bar{t}}$ and $\Delta|y|$ vs longitudinal Lorentz boost $\beta_{z,t\bar{t}}$ is constructed for each of the four regions. From the $\Delta|y|$ distribution, it is possible to measure the value of A_{C}, and also differential measurement of the A_{C} vs $m_{t\bar{t}}$ and $\beta_{z,t\bar{t}}$ for the $\Delta|y|$ vs $m_{t\bar{t}}$ and $\Delta|y|$ vs $\beta_{z,t\bar{t}}$ double-differential distributions. The actual approach is slightly more complicated due to the limited acceptance and response of the detector and reconstruction, which cause dilution of the reconstructed A_{C}. In order to compare the A_{C} measurement with theoretical predictions, an unfolding procedure is used.

The unfolding attempts to determine the binned true distribution $\mathbf{T}$ of the observable of interest from an observed binned reconstructed distribution $\mathbf{D}$, assuming the knowledge of detector-level distribution of

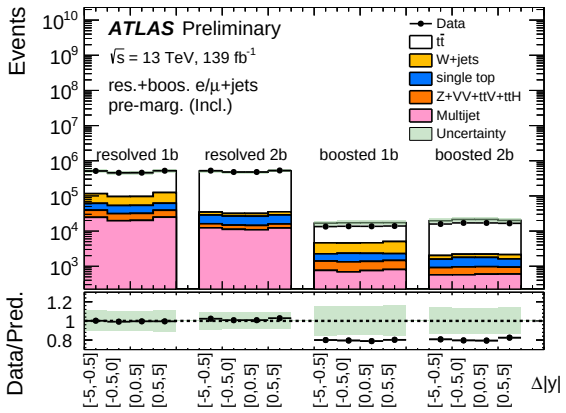


Fig. 5: The comparison of data to prediction for the $\Delta|y|$ distribution [9]. The x-axis labels show the bin edges of the $\Delta|y|$ bins as optimised for the unfolding. All four regions are shown in the plot, denoted by the labels above the distributions. The black points in the bottom plot show the ratio of data to total prediction, with the green band showing the total uncertainty, including both statistical and systematic uncertainties on the prediction in quadrature. In the boosted regions, the predicted signal normalisation is overestimated by approximately 20 % due to a known mis-modelling [10], which is corrected for in this analysis using data-driven normalisation factor.

backgrounds processes $\mathbf{B}$ and the response matrix $\mathcal{M}$, encoding the detector response and acceptance:

$$\mathbf{D} = \mathbf{S} + \mathbf{B} \equiv \mathcal{M} \cdot \mathbf{T} + \mathbf{B}, \tag{4}$$

In this analysis, the backgrounds are determined via MC simulations and data-driven methods and the response matrix is determined from the MC simulations. The bins of the matrix include the information about the probability that an event from true bin t_j is reconstructed and migrates into reconstructed distribution bin r_i. A naive solution by using matrix inversion in Eq. 4 can rarely be used, as the original problem is ill-defined, resulting in amplifications of the statistical fluctuations in data, yielding wildly oscillating solution. In this thesis, the Fully-Bayesian unfolding (FBU) [11] method is used, which applies Bayes' theorem, yielding a full posterior probability distribution of the unfolded spectra. It employs likelihood formalism, allowing for a natural inclusion of systematic uncertainties as nuisance parameters. Via Bayesian marginalisation, this approach allows to reduce total uncertainty by accounting for correlations between the nuisance parameters and by constraining the magnitude of the systematic variations.

In the analysis, the reconstructed data and background distributions are the $\Delta|y|$ distributions for the four regions, stacked side-by-side as a single distribution, that is unfolded to a single unfolded $\Delta|y|$ distribution. The inclusive $\Delta|y|$ distribution as well as for each bin of the differential $\Delta|y|$ vs $m_{t\bar{t}}$ and $\beta_{z,t\bar{t}}$, respectively, four bins in $\Delta|y|$ are used for each region. This minimises the statistical uncertainty of individual bins, and at the same time allows for a more accurate mapping of migrations, rather than having a single $\Delta|y| > 0$ and $\Delta|y| < 0$ bin. From the unfolded $\Delta|y|$ distribution the A_C is computed according to Eq. 2.

4.2 Results

The measured A_C in data is presented in Table 1 and Fig. 6 and 7, compared to the SM theory prediction calculated up to NNLO QCD + NLO EWK accuracy [12, 13]. The results are found to agree with the SM prediction, showing no significant deviations. Additionally, for the first time at a hadron collider, the evidence of a non-zero A_C with the significance of $4\,\sigma$ is found in the inclusive measurement. The results in Table 1 show that the measurement precision could still be further improved by increasing the amount of statistics, as in particular the kinematic region of high $m_{t\bar{t}}$ is statistically dominated. This is also the region of phase space, where the contribution of charge-asymmetric $q\bar{q} \to t\bar{t}$ is the largest, and hence has highest potential sensitivity to BSM physics contributions.

Table 1: Estimate of the systematic component in the FBU marginalisation. The unfolding is performed without nuisance parameters (Stat. column) and with full nuisance parameter marginalisation (Stat.+Syst. column). The Syst. column is calculated as $(\sigma^2_{\text{tot.}} - \sigma^2_{\text{stat.}})^{1/2}$ and gives an approximate estimate of the relative impact of statistical and systematic uncertainties.

	A_C	Stat.	Syst.	Stat.+Syst.
Inclusive	0.0060	0.0011	0.0009	0.0014
$m_{t\bar{t}}$				
< 500 GeV	0.0045	0.0028	0.0034	0.0044
[500, 750] GeV	0.0051	0.0020	0.0021	0.0029
[750, 1000] GeV	0.0100	0.0049	0.0046	0.0067
[1000, 1500] GeV	0.0169	0.0072	0.0027	0.0077
> 1500 GeV	0.0121	0.0277	0.0150	0.0315
$\beta_{z,t\bar{t}}$				
[0, 0.3]	0.0007	0.0040	0.0032	0.0051
[0.3, 0.6]	0.0085	0.0031	0.0025	0.0040
[0.6, 0.8]	0.0014	0.0029	0.0033	0.0044
[0.8, 1.0]	0.0100	0.0026	0.0042	0.0049

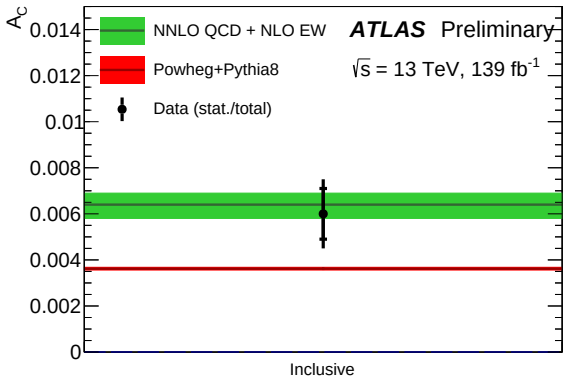
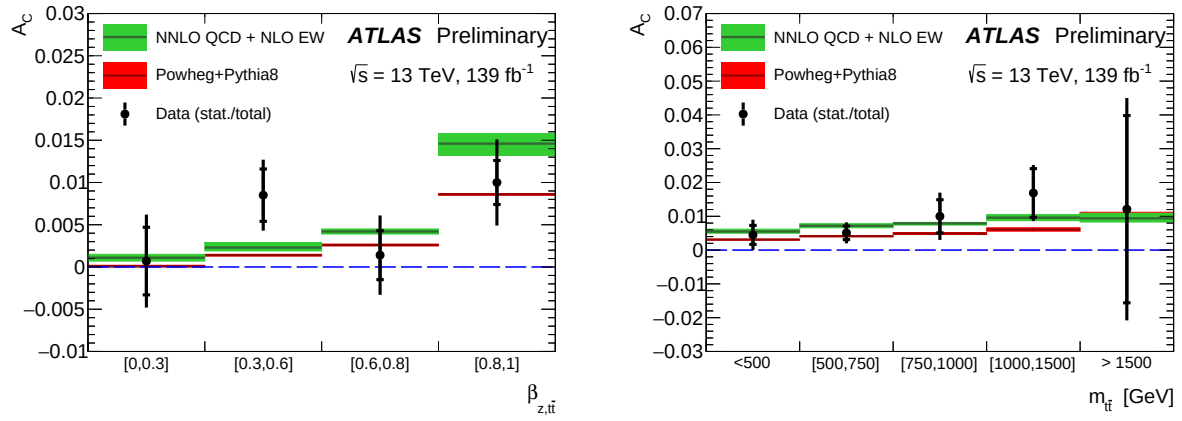


Fig. 6: Comparison of the unfolded A_C in data with the Asimov prediction obtained from POWHEG+PYTHIA8 simulation (dark red line) and with the NNLO QCD + NLO EWK prediction [12, 13] (green band), for the inclusive A_C measurement. The black error bars show the total uncertainty, while the smaller black error bars denoted by the vertical ticks show the statistical uncertainty only.



(a) (b)

Fig. 7: Comparison of the unfolded A_C in data with the Asimov prediction obtained from POWHEG+PYTHIA8 simulation (dark red line) and with the NNLO QCD + NLO EWK prediction [12, 13] (green band), for the differential A_C measurement as a function of $\beta_{z,t\bar{t}}$ (a) and $m_{t\bar{t}}$ (b). The error bars follow the same convention as in Fig. 6.

5 Charge asymmetry in the boosted all-hadronic channel

The limited statistical precision of the high-$m_{t\bar{t}}$ differential A_C measurement in the single-lepton channel motivates the investigation of other $t\bar{t}$ decay channels. The all-hadronic channel has the largest branching ratio of approximately 46 % [1] and therefore is the most suitable candidate. In addition, in the kinematic regime of interest, using boosted top quark reconstruction techniques, the reconstruction of the all-hadronic $t\bar{t}$ production is simpler than in the single-lepton channel, requiring the reconstruction and identification of two large-R jets originating from the $t\bar{t}$ pair production. The absence of a neutrino in the final state means that it is possible to more precisely reconstruct the final state, reducing the migrations due to poor measurement of the $m_{t\bar{t}}$ or the top quark and top anti-quark rapidities. The major challenge in this channel is the absence of the isolated prompt lepton from top-quark decay. It is therefore necessary to identify the electric charge of the top quark and top anti-quark large-R jets using different approach. In this thesis, we investigate the discrimination of the top-quark electric charge using the electric charges of the reconstructed charged particle tracks in the ATLAS inner detector, originating from the hadronisation showers of the top-quark decay products. For this task, neural networks are used to combine several charge-sensitive observables into a single discriminant. This discriminant is applied in a subsequent analysis to determine approximate estimate of the statistical and systematic uncertainties for a potential measurement of the A_C in the all-hadronic channel.

The feasibility of the A_C measurement is studied using MC simulations and a data-driven background estimate of the multijet background. The events of interest are required to fire a single-jet trigger and contain at least one anti-k_t $R = 1.0$ jet with $p_T > 500\,\text{GeV}$ and another jet with $p_T > 350\,\text{GeV}$. Both jets must pass the criteria of a novel neural-network-based boosted tagging algorithm [5], and both jets must be identified as containing b-quarks, to be assumed as top-quark candidate jets. The two jets are subsequently used for the calculation of the event $\Delta|y|$ observable. A double-differential $\Delta|y|$ vs $m_{t\bar{t}}$ distribution is obtained for the total expectation, which is the sum of the predicted signal and background contributions, obtained from the MC simulations and the data-driven multijet background estimate. This total expectation is referred to as the Asimov dataset, which is subsequently unfolded using the same

Bayesian technique as in the single-lepton channel, to obtain the Asimov estimate of the A_C as a function of $m_{t\bar{t}}$. The uncertainties on the unfolded Asimov A_C values are compared with the single-lepton channel.

5.1 Top quark charge discrimination using neural networks

Several measurements in the past have used observables based on weighted average of charge of tracks matched to a jet, to infer the charge of the parton initiating the jet. The measurement of top-quark charge at the CDF experiment [14] used the following observable to calculate the charge of a b-tagged jet:

$$Q_{\text{jet}} = \frac{\sum_i q_i \left(\vec{p}_i \cdot \vec{p}_{\text{jet}}\right)^\kappa}{\sum_i \left(\vec{p}_i \cdot \vec{p}_{\text{jet}}\right)^\kappa}, \tag{5}$$

where q_i is the charge of the individual tracks, $\vec{p}_i$ their momentum vector, and $\vec{p}_{\text{jet}}$ is the momentum vector of the jet. The κ exponent is a free parameter. The sums in numerator and denominator run over all tracks matched to the jet. The observable in Eq. 5 was also used in the measurement of the $b\bar{b}$ forward-backward asymmetry at the CDF experiment [15], as well as the measurement of the top-quark charge at the ATLAS experiment [16]. The observable was used to distinguish the b-jet from $\bar{b}$-jet, where a jet with negative Q_{jet} was assigned to originate from b quark, and a jet with positive Q_{jet} was assigned to originate from $\bar{b}$ quark. The κ parameter was chosen to be 0.5, optimised to maximise the separation power of Q_{jet}. In general $\kappa > 0$ values are used, because such values guarantee the infra-red safety of the definition [17].

An extension of the jet charge approach has been investigated within a study in Ref. [18] based on ATLAS MC simulations, by exploiting additional information contained within the B-hadron decays in the b-jet combined into a neural network (NN) classifier, referred to as *jet vertex charge* (JVC). This algorithm uses tracks matched to a small-R calorimeter jet to define a set of high-level observables used in the NN. In addition to the charge of all tracks matched to a b-tagged jet, a charge of tracks originating from displaced secondary and tertiary vertices reconstructed by *JetFitter* algorithm [19] is calculated.

In this thesis, the identification of the charge of top quark is taken one step further, by including the charged tracks originating from the quarks from hadronic decay of the W boson from top quark. In the boosted topology of a decay, it is expected that all three quarks from top-quark decay are ideally contained within the single large-R jet. To distinguish the individual quarks, tracks matched to the large-R jets are clustered into small track-jets with $R = 0.2$ radius parameter. The charge calculated similarly as in Eq. 5 of the three highest-p_{T} track jets is considered, out of which one must be identified as originating from b-quark. In addition, the charge of vertices from b-quark decay in the b-tagged track jets are considered, inspired by the approach taken in the JVC. All of the observables are combined into a neural network.

The NN is trained using SM $gg \to t\bar{t}$ MC-simulated events. The discriminant is shown in Fig. 8, compared between the set used to train the NN, and an independent set of simulated events to cross-check against potential over-training.

The performance of the NN is evaluated in the MC-simulated all-hadronic $t\bar{t}$ events, by using the NN to determine which of the two large-R jet candidates is originating from top quark and from top anti-quark, respectively. The two jets are denoted as J_1 and J_2, with their values of input variables for the NN $\mathbf{x}_{J_1}$ and $\mathbf{x}_{J_2}$, respectively. They are identified as the top-quark jet, J_t, and top anti-quark jet, $J_{\bar{t}}$, by comparing the

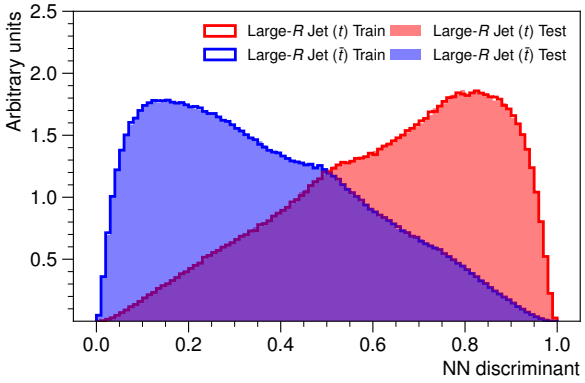


Fig. 8: The comparison of the NN discriminant for the training and testing set, for the large-R jets originating from top quarks (red) and from top anti-quarks (blue), respectively.

NN discriminant values $f(\mathbf{x}_{J_1})$ and $f(\mathbf{x}_{J_2})$, where $f(\mathbf{x})$ is the NN discriminant function:

$$\begin{aligned} f(\mathbf{x}_{J_1}) > f(\mathbf{x}_{J_2}) \Rightarrow J_t \equiv J_1, \quad J_{\bar{t}} \equiv J_2, \\ f(\mathbf{x}_{J_1}) < f(\mathbf{x}_{J_2}) \Rightarrow J_t \equiv J_2, \quad J_{\bar{t}} \equiv J_1. \end{aligned} \tag{6}$$

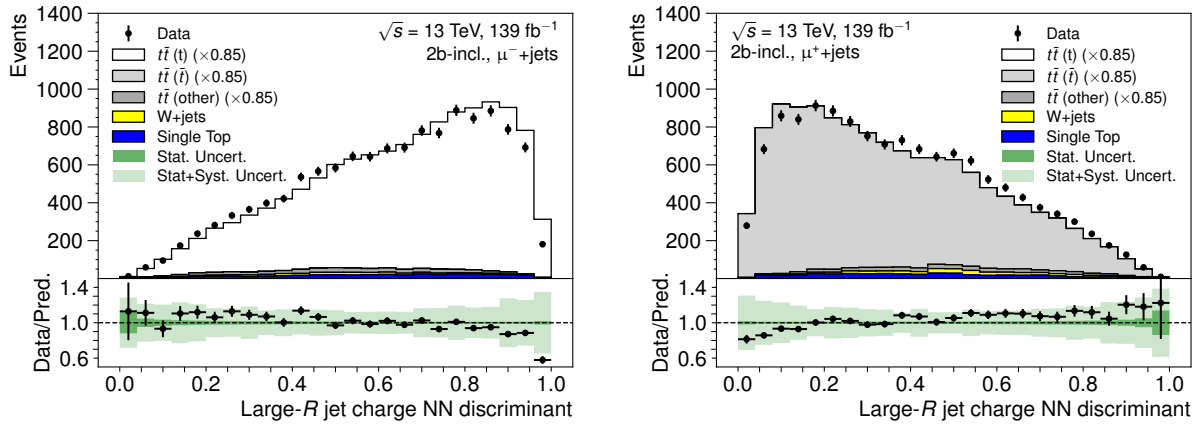
The main source of dilution in the all-hadronic channel A_C measurement is expected to be the mis-identification of the jets in Eq. 6, resulting in wrong determination of the sign of $\Delta|y| = |y_t| - |y_{\bar{t}}|$, where $|y_t|$ is the absolute rapidity of J_t and $|y_{\bar{t}}|$ is the absolute rapidity of $J_{\bar{t}}$. Therefore, the figure of merit when comparing the NN with alternative approaches, is the $\Delta|y|$ sign assignment purity, labelled as $\mathcal{P}_{\Delta|y|}$. It is defined as the fraction of events where the sign of reconstructed $\Delta|y|$ matches the sign of $\Delta|y|^{\text{parton}} = |y_t^{\text{parton}}| - |y_{\bar{t}}^{\text{parton}}|$, calculated from the true absolute rapidities of the true top-quark and top anti-quark based on information from the MC simulation (referred to as *parton* level). The results of the comparison are shown in Table 2, as a function of the invariant mass of the parton-level $t\bar{t}$ pair, $m_{t\bar{t}}^{\text{parton}}$. Compared to other previously-examined approaches relying on the charge of the b-tagged jet, a significant boost in performance is observed, mostly attributed to the inclusion of the jet charges from quarks from the W boson decay. The single-lepton channel achieves higher $\mathcal{P}_{\Delta|y|}$ compared to all-hadronic channel, primarily due to the low probability of incorrectly measuring the electric charge of the isolated prompt lepton. Thus, the dilution of $\Delta|y|$ sign is driven primarily by mis-reconstruction of object kinematics, such as incorrect neutrino longitudinal component determination, or incorrect jet-to-parton assignment.

Table 2: Comparison of the $\Delta|y|$ charge assignment purity $\mathcal{P}_{\Delta|y|}$ in the all-hadronic channel obtained using the NN and by using b-tagged track-jet charge only, and in the single-lepton channel, in bins of parton-level $m_{t\bar{t}}$.

Channel, method	$\mathcal{P}_{\Delta\|y\|}$ [%] vs $m_{t\bar{t}}^{\text{parton}}$ [GeV]	
	[1000, 1500]	> 1500
All-hadronic, NN discriminant	**82.3**	**82.3**
All-hadronic, b-jet charge	65.9	64.8
Single-lepton	91.4	90.2

The accuracy of the MC simulation of the NN classifier distribution as well as the input observables

used for the NN is examined in data, using a sample enriched in $t\bar{t}$ events with a single isolated prompt lepton, using selection criteria similar to those for the boosted region in the single-lepton charge asymmetry measurement. Only events with at least two b-tagged jets and isolated prompt muon are considered, because such selection yields very high purity of the sample, with almost negligible background contributions. The comparison of data to prediction for the NN discriminant distribution is shown in Fig. 9. The comparisons are performed separately in two regions defined by the sign of the electric charge of the prompt muon. The charge of the muon is a proxy to the charge of the hadronically-decaying top quark reconstructed as the large-R jet, on which the NN is tested. In general, some tension in the modelling and the data is observed in the distributions, and hence a dedicated systematic uncertainty is derived in the single-lepton channel to account for the discrepancies, which is subsequently applied in the all-hadronic channel.



(a) (b)

Fig. 9: Comparison of data and prediction for the NN discriminant of the hadronic top-quark large-R jet candidate, in the μ^- region (a) and μ^+ region (b). The $t\bar{t}$ contribution is split according to the true charge of the top quark matched to the large-R jet, into t, $\bar{t}$ and other (failed matching) category. The dark green band shows the statistical uncertainty of the total prediction, and the light green band shows the total prediction statistical and systematic uncertainty summed in quadrature. The error bars on black points show the statistical uncertainty of data. The $t\bar{t}$ prediction is scaled to normalise the total prediction yield to data.

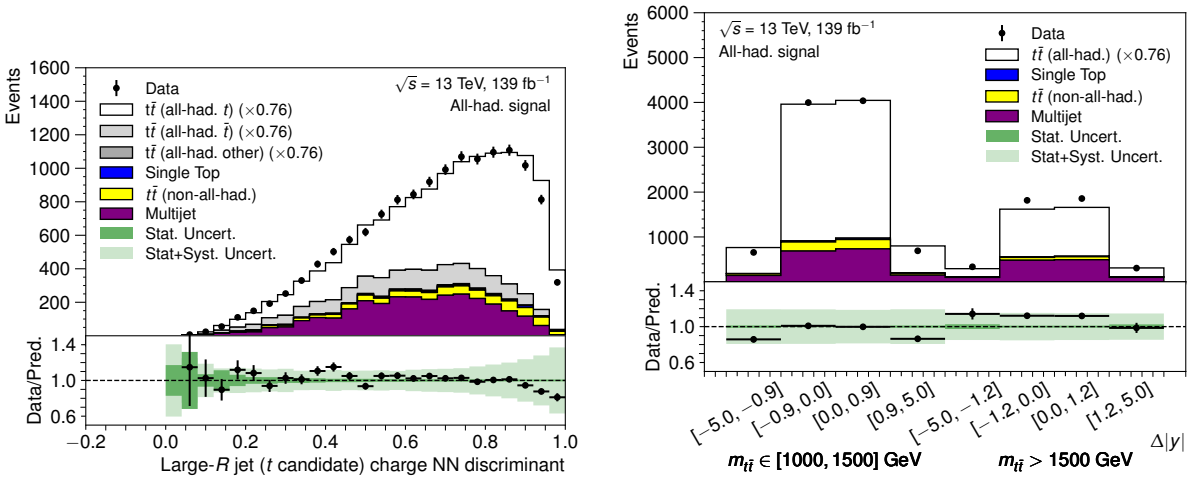
5.2 Results of the A_C sensitivity estimate

The modelling of the MC simulations and the data-driven estimates is checked by performing comparisons of the Run-II data with the Asimov prediction. The data-to-prediction comparison of the top-quark candidate large-R jet NN charge tagger discriminant is shown in Fig. 10a and the constructed double-differential $\Delta|y|$ vs $m_{t\bar{t}}$ observable is shown in Fig. 10b. Both of the observables are found to agree with data within the statistical and systematic uncertainties.

The results of the Asimov unfolding for the all-hadronic channel and the single-lepton channel are shown in Table 3, showing the obtained A_C for the two highest-$m_{t\bar{t}}$ bins corresponding to the single-lepton channel. The unfolded A_C values in the all-hadronic channel agree well with the true MC-predicted value. The true values between the all-hadronic and single-lepton channel MC predictions differ, showing some tension, but the differences can be explained by statistical nature. Physics-related differences are not expected, as the A_C does not depend on the decay channel. The precision of the all-hadronic channel in the $m_{t\bar{t}} \in [1000, 1500]$ GeV is limited by the available statistics in contrast to the lepton channel. In this

kinematic region, the all-hadronic channel is limited by the leading jet p_{T} cut, dictated by the limited throughput of the ATLAS jet trigger used in this analysis. For the $m_{t\bar{t}} > 1500\,\mathrm{GeV}$, the precision of the all-hadronic channel is closer to the single-lepton channel, suggesting that a sizeable reduction the A_{C} uncertainty could be achieved by combining the single-lepton and the all-hadronic channel.

The total A_{C} uncertainty in the all-hadronic channel is dominated by the statistical uncertainties, thus motivating further studies to enhance the amount of statistics by loosening the selection criteria. Multiple approaches are discussed in the thesis, focused on employing improved b-tagging algorithms, that can achieve similar background rejection at a higher signal efficiency, or by employing di-jet triggers and triggers with additional criteria on jet mass, which have a lower jet p_{T} cut, thus selecting more events.



(a) (b)

Fig. 10: The comparison of data to prediction of the distribution of top-quark candidate large-R jet NN discriminant, in the validation regions K (a) and N (b). In the figures, the all-hadronic $t\bar{t}$ contribution is scaled to match the prediction yield to data. The dark-green band in the ratio plot shows the statistical uncertainty on the prediction, while the light-green band shows the statistical and systematic uncertainty on the prediction, summed in quadrature. The error bars on the black points show the statistical uncertainty of data.

Table 3: The comparison of the unfolded full-phase-space and true (parton-level) A_{C} and its uncertainty for the Asimov dataset, between the all-hadronic and the single-lepton channel. The parton-level true A_{C} uncertainty is statistical-only. The unfolded A_{C} uncertainty is the total statistical and systematic uncertainty. Additionally, the NNLO QCD + NLO EWK SM theory prediction [12, 13] is shown for reference of a higher-order prediction.

	A_{C} vs $m_{t\bar{t}}$[GeV]	
	[1000, 1500]	> 1500 GeV
All-hadronic unfolded	0.0047 ± 0.0302	0.0045 ± 0.0393
All-hadronic true	0.0047 ± 0.0008	0.0047 ± 0.0014
Single-lepton unfolded	0.0063 ± 0.0084	0.0114 ± 0.0297
Single-lepton true	0.0061 ± 0.0006	0.0100 ± 0.0013
NNLO QCD + NLO EWK SM	$0.0096^{+0.0009}_{-0.0009}$	$0.0094^{+0.0015}_{-0.0011}$

6 Conclusions

In the ATLAS Run-II period, the machine learning techniques were applied in the field of identification of boosted hadronically-decaying top quarks and W bosons, showing substantial improvements in the multijet background suppression, especially for the top-quark tagging, where the rich decay structure provides many features exploitable in a multivariate technique. It is demonstrated in this thesis, that these novel techniques are well simulated and any discrepancies with respect to data can be corrected using a data-driven calibration. The results were published in Ref. [5].

The benefits of the large Run-II dataset have greatly helped push the precision top-quark physics frontier, particularly for rare phenomena such as the charge asymmetry. For the first time at a hadron collider, the evidence of non-zero A_C in $t\bar{t}$ production was observed in the inclusive measurement with a significance of $4\,\sigma$, as published in a ATLAS conference note in Ref. [9], where previous measurements were both compatible with zero asymmetry and the SM prediction. Both the inclusive and differential A_C measurements show good agreement with the SM prediction.

Finally, a study of the possibility of A_C measurement in the boosted all-hadronic channel has been performed. This study was motivated by the fact that the boosted all-hadronic channel would potentially contribute in the region of $m_{t\bar{t}} > 1\,\text{TeV}$, where the A_C sensitivity to BSM effects is higher due to larger $q\bar{q} \to t\bar{t}$ contribution. The signal-to-noise ratio in this channel is improved compared to previous measurements [20] in the boosted all-hadronic $t\bar{t}$ topology by using the aforementioned novel neural network boosted top-quark tagging techniques [5]. The main challenge of this measurement is the necessity to distinguish reconstructed top quark from top anti-quark. It is demonstrated, that using neural networks and tracking information from the inner detector, it is possible to perform this charge identification with a precision much higher than previously applied approaches for top-quark charge measurement not relying on semi-leptonic quark decays. The first estimate of A_C in this channel is presented, and is found to be statistically limited, however, with a room for further improvement in the selection efficiency at a small to no reduction in the background rejection. Given the presented results, a combination with the single-lepton channel would most benefit the $m_{t\bar{t}} > 1.5\,\text{TeV}$ region, with further improvements possible at lower $m_{t\bar{t}}$ if event selection re-optimisation is investigated.

References

[1] P. D. G. Collaboration, *Review of Particle Physics*, Phys. Rev. D **98** (2018) 030001.

[2] J. Gao, C. S. Li, and H. X. Zhu, *Top-Quark Decay at Next-to-Next-to-Leading Order in QCD*, Phys. Rev. Lett. **110** (2013) 042001.

[3] J. H. Kühn and G. Rodrigo, *Charge asymmetry of heavy quarks at hadron colliders*, Physical Review D **59** (1999) 054017.

[4] W. Bernreuther and Z.-G. Si, *Top quark and leptonic charge asymmetries for the Tevatron and LHC*, Phys. Rev. D **86** (2012) 034026.

[5] ATLAS Collaboration, *Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC*, Eur. Phys. J. C **79** (2019) 375.

[6] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k(t) jet clustering algorithm*, JHEP **04** (2008) 063.

[7] ATLAS Collaboration, *Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV*, Eur. Phys. J. C **76** (2016) 154.

[8] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, JHEP **03** (2011) 015.

[9] ATLAS Collaboration, *Inclusive and differential measurement of the charge asymmetry in $t\bar{t}$ events at 13 TeV with the ATLAS detector*, ATLAS-CONF-2019-026, 2019, https://cds.cern.ch/record/2682109.

[10] ATLAS Collaboration, G. Aad et al., *Measurements of top-quark pair differential and double-differential cross-sections in the ℓ+jets channel with pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector*, Eur. Phys. J. C **79** (2019) 1028.

[11] G. Choudalakis, *Fully Bayesian Unfolding*, arXiv:1201.4612 [hep-ex].

[12] M. Czakon, et al., *Top-quark charge asymmetry at the LHC and Tevatron through NNLO QCD and NLO EW*, Phys. Rev. D **98** (2018) 014003.

[13] M. Czakon, et al., *Differential distributions for top-pair production in NNLO QCD + NLO EW*, http://www.precision.hep.phy.cam.ac.uk/results/ttbar-nnloqcd-nloew/. Accessed: 2019-06-03.

[14] CDF Collaboration Collaboration, CDF Collaboration, *Exclusion of exotic top-like quarks with $-4/3$ electric charge using jet-charge tagging in single-lepton $t\bar{t}$ events at CDF*, Phys. Rev. D **88** (2013) 032003.

[15] CDF Collaboration, *First measurement of the forward-backward asymmetry in bottom-quark pair production at high mass*, Phys. Rev. D **92** (2015) 032006.

[16] ATLAS Collaboration, *Measurement of the top quark charge in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, JHEP **11** (2013) 031.

[17] ATLAS Collaboration, *Measurement of jet charge in dijet events from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector*, Phys. Rev. D **93** (2016) 052003.

[18] ATLAS Collaboration, *A new tagger for the charge identification of b-jets*, Tech. Rep. ATL-PHYS-PUB-2015-040, CERN, Geneva, Sep, 2015. https://cds.cern.ch/record/2048132.

[19] ATLAS Collaboration Collaboration,, *Topological b-hadron decay reconstruction and identification of b-jets with the JetFitter package in the ATLAS experiment at the LHC*, Tech. Rep. ATL-PHYS-PUB-2018-025, CERN, Geneva, Oct, 2018. https://cds.cern.ch/record/2645405.

[20] ATLAS Collaboration, *Measurements of $t\bar{t}$ differential cross-sections of highly boosted top quarks decaying to all-hadronic final states in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector*, Phys. Rev. D **98** (2018) 012003.

Selection of publications of the thesis author

[1] O. Majerský on behalf of the ATLAS Collaboration, *Identifying hadronically-decaying vector bosons and top quarks in ATLAS*, Journal of Physics: Conference Series **1525** (2020) 012118.

[2] O. Majerský on behalf of the ATLAS Collaboration, *Highlights of top-quark properties measurements at the ATLAS experiment*, Nuclear and Particle Physics Proceedings **309-311** (2020) 22–29.

[3] ATLAS Collaboration, *Inclusive and differential measurement of the charge asymmetry in $t\bar{t}$ events at 13 TeV with the ATLAS detector*, ATLAS-CONF-2019-026, 2019, `https://cds.cern.ch/record/2682109`.

[4] ATLAS Collaboration, *Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC*, Eur. Phys. J. C **79** (2019) 375.

[5] ATLAS Collaboration, *Measurements of top-quark pair spin correlations in the $e\mu$ channel at $\sqrt{s} = 13$ TeV using pp collisions in the ATLAS detector*, `arXiv:1903.07570 [hep-ex]`.

[6] ATLAS Collaboration, *Measurement of the top quark mass in the $t\bar{t} \to$ lepton+jets channel from $\sqrt{s} = 8$ TeV ATLAS data and combination with previous results*, Eur. Phys. J. C **79** (2019) 290.

[7] ATLAS Collaboration, *Measurements of top-quark pair differential and double-differential cross-sections in the ℓ+jets channel with pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector*, Eur. Phys. J. C **79** (2019) 1028.

[8] ATLAS Collaboration, *Measurement of the top-quark mass in $t\bar{t}$ + 1-jet events collected with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV*, JHEP **11** (2019) 150.

[9] ATLAS Collaboration, *Search for top-quark decays $t \to Hq$ with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **05** (2019) 123.

[10] ATLAS Collaboration, *Measurements of $t\bar{t}$ differential cross-sections of highly boosted top quarks decaying to all-hadronic final states in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector*, Phys. Rev. D **98** (2018) 012003.

[11] ATLAS Collaboration, *Search for flavor-changing neutral currents in top quark decays $t \to Hc$ and $t \to Hu$ in multilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **98** (2018) 032002.

[12] ATLAS Collaboration, *Search for $W' \to tb$ decays in the hadronic final state using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Lett. B **781** (2018) 327.

[13] ATLAS Collaboration, *Measurement of colour flow using jet-pull observables in $t\bar{t}$ events with the ATLAS experiment at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **78** (2018) 847.

[14] CDF Collaboration, *Measurement of the forward-backward asymmetry in low-mass bottom-quark pairs produced in proton-antiproton collisions*, Phys. Rev. D **93** (2016) 112003.