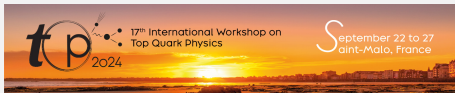
Machine learning in top quark physics at ATLAS and CMS

Matthias Komm, on behalf of the ATLAS and CMS collaborations

DESY, Hamburg, Germany

matthias.komm@cern.ch

Abstract

This note presents an overview of current and potential future applications of machine-learning-based techniques in the study of the top quark. The research community has developed a diverse set of ideas and tools, including algorithms for the efficient reconstruction of recorded collision events and innovative methods for statistical inference. Recent applications of some techniques by the ATLAS and CMS collaborations are also highlighted.

1 Introduction

In recent years, machine learning (ML) and artificial intelligence have revolutionized numerous fields. In top quark research, ML has already been a driving force for over a decade. Given the top quark's nearly exclusive decay to a b quark and a W boson, the continuous advancements in b jet identification through increasingly sophisticated ML algorithms is a prime example [1, 2]. Stretching from the discovery of single top quark production at the Tevatron [3, 4] to the recent observation of four-top-quark events by the ATLAS [5, 6] and CMS [7, 8] collaborations at the CERN LHC highlight the critical role of ML in achieving such milestones in the field.

This note reviews a collection of state-of-the-art ML algorithms addressing various aspects of top quark research, including top quark and event reconstruction, analysis strategies, and novel methods for statistical inference, which could offer valuable contributions to the field in the future. An outlook on the impact of ML on future research at the upcoming HL-LHC is also provided.

2 Top quark reconstruction

The reconstruction of top quark events typically involves two stages. First, the inference of the direction(s) of undetected neutrino(s) from leptonically decaying top quarks, $t \rightarrow b \ell \nu$, if

(a)   (b)

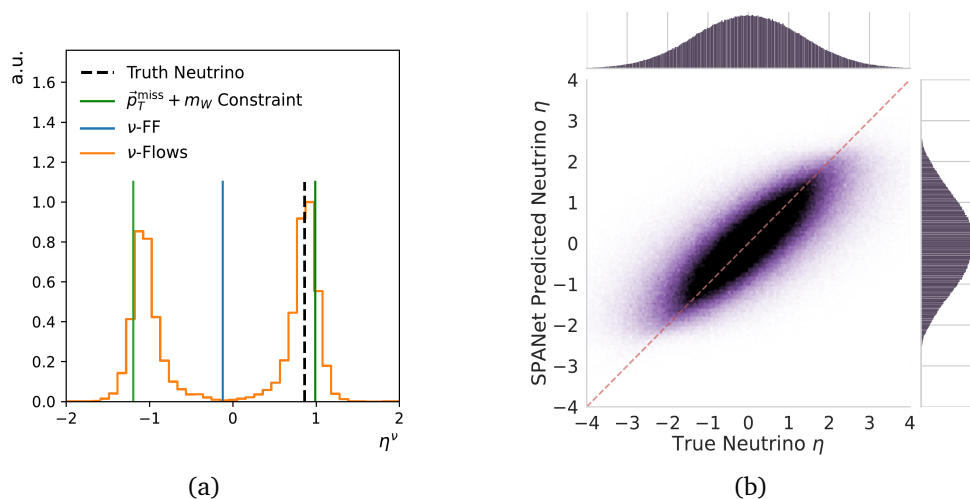



Figure 1: Inference of the pseudorapidity of the neutrino originating from the leptonically decaying top quark in semileptonic $t\bar{t}$ events: (a) ν-FLOW approach [9]; (b) SPANET approach [11].

present. Second, the association of the remaining decay products (leptons and jets) to each top quark in the event. Various ML algorithms have been developed to address both tasks.

The ν-FLOW approach [9,10] utilizes a normalizing flow neural network (NN) conditioned on reconstructed event observables. This network maps the true neutrino direction vector, derived from the event generator truth, to a three-dimensional normal distribution. By sampling from this distribution, the likelihood of possible neutrino directions can be inferred from data, as illustrated in Fig. 1a for a representative event, illustrating superior performance over regressing the values via a feed-forward NN or employing a W boson mass constraint.

An efficient solution for assigning all top decay products to reconstructed particles is provided by the SPANET approach [11]. This method employs an NN transformer architecture with over 10M parameters. The latest version of SPANET also incorporates auxiliary targets, such as the regression of the neutrino direction (as shown in Fig. 1b) and the discrimination of signal from background events.

A novel approach to associate the decay products to reconstructed particles is given by the HYPER method [12]. In this framework, the top quark decay products are represented as hypergraphs—a generalization of graph NNs, where each edge can connect more than two nodes. Despite using only 345k parameters, the NN achieves a performance comparable to SPANET.

3   Analysis strategy

A crucial component of many physics analyses is determining the rate of background events. This is particularly challenging for analyses focusing on multijet events, a common feature of top quark analyses, which have to deal with a non-negligible fraction of events arising from pure QCD interactions. The precise yield and distribution of these interactions are difficult to predict accurately through simulation.

One solution is the so-called ABCD matrix method, where the background fraction is extrapolated using data from orthogonal regions defined by two independent observables. The DISCO method [13] introduces an NN classifier to construct suitable observables that are uncorrelated and simultaneously separate the signal from background processes. This is achieved

by introducing an additional penalty term during the training to suppress distance correlations either between the scores of two NN classifiers or between one score and an auxiliary observable.

Another approach was showcased in a recent analysis by the CMS collaboration searching for all-hadronic four-top-quark events [14]. An auto-regressive normalizing flow was employed to transform data events from a background-enriched region into the signal region.

4 Statistical inference

In addition to improving event reconstruction and analysis strategies, new ML-based tools have been studied to enhance the statistical inference, going beyond the traditional (binned)-likelihood approach.

Likelihood-free inference, also known as simulation-based inference, focuses on using the output score s of a classifier as a test statistic directly. This approach exploits the fact that s can provide a likelihood ratio $H_1/H_0 = s/(1-s)$, which is essential for hypothesis testing, ie. rejecting the null hypothesis H_0 in favour of an alternative hypothesis H_1. Examples of tools employing this method include INFERNO [15] and SALLY [16], both of which can account for sources of systematic uncertainties as well.

Another domain where ML proves valuable is unfolding of differential cross sections, enabling direct comparisons with theoretical predictions or results from other experiments. However, it is crucial to note that no ML algorithm can overcome the inherent ill-posed nature of the unfolding problem (i.e., inverting a Fredholm integral equation), necessitating some form of regularization. The OMNIFOLD approach [17] facilitates unbinned and multidimensional unfolding through an iterative procedure in which differences between simulation and data are learned via a classifier and subsequently reweighted to match the distributions in the simulated signal sample to data. The number of iterations controls the degree of regularization.

This method has been successfully demonstrated by the ATLAS and CMS collaborations for unfolding Drell-Yan [18] and minimum bias events [19], respectively. Selected results are shown in Fig. 2. Notably, the unbinned nature of OMNIFOLD enables the unfolding of novel quantities, such as taking the average of the jet mass as a function of another observable.

(a)   (b)

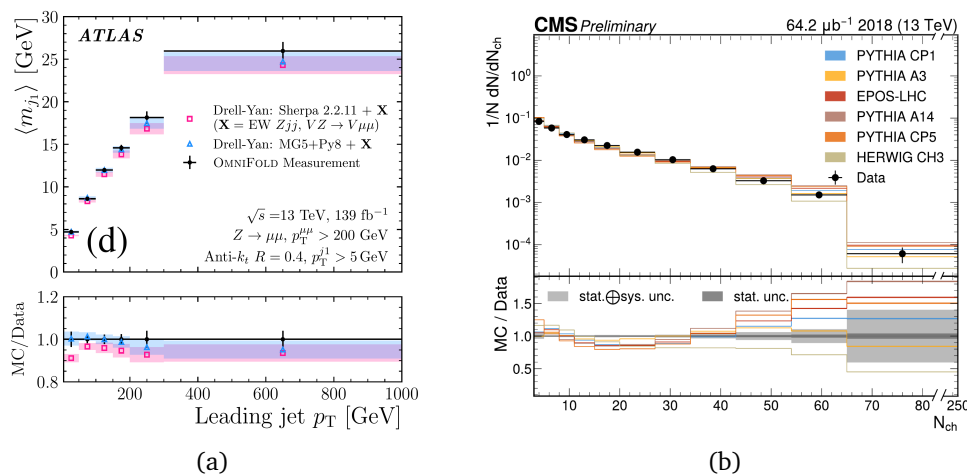



Figure 2: Examples of unbinned, multidimensional unfolding with OMNIFOLD [17]: (a) average jet mass as a function of the jet p_T in Drell-Yan events [18] by ATLAS; (b) number of charged jet constituents in minimum bias events [19] by CMS.

5 HL LHC outlook

The upcoming high-luminosity (HL) phase of the LHC will significantly increase computing demands for analyzing the much larger datasets. An actively researched area is the use of ML to reduce the reliance on simulated samples or enhance the speed of detector simulations.

The CMS collaboration has recently addressed the former challenge by introducing the DCTR method [20, 21] for reweighting simulated samples. The study comprised the reweighting of samples to emulate parameter shifts used for estimating the impact by systematic uncertainties. An example variation is shown in Fig. 3a involving the h_{damp} parameter of the POWHEG event generator [22], which regulates the energy of additional radiations. Its effect on the sample cannot be described analytically but is approximated by the NN. The good agreement observed after reweighting demonstrates that this technique could replace the need for generating dedicated samples in the future, resulting in improved sustainability by skipping the computational needs of classical detector simulations.

Another application of the DCTR method is reweighting samples to achieve higher-order accuracy. In Fig. 3b, an NLO sample has been successfully reweighted to match an NNLO prediction.

6 Conclusion

For over a decade, ML has been a driving force in top quark research. Novel ML-based approaches in top quark reconstruction, background estimation techniques, and innovative tools for statistical inference are also setting the stage for the upcoming high-precision era of the HL LHC.

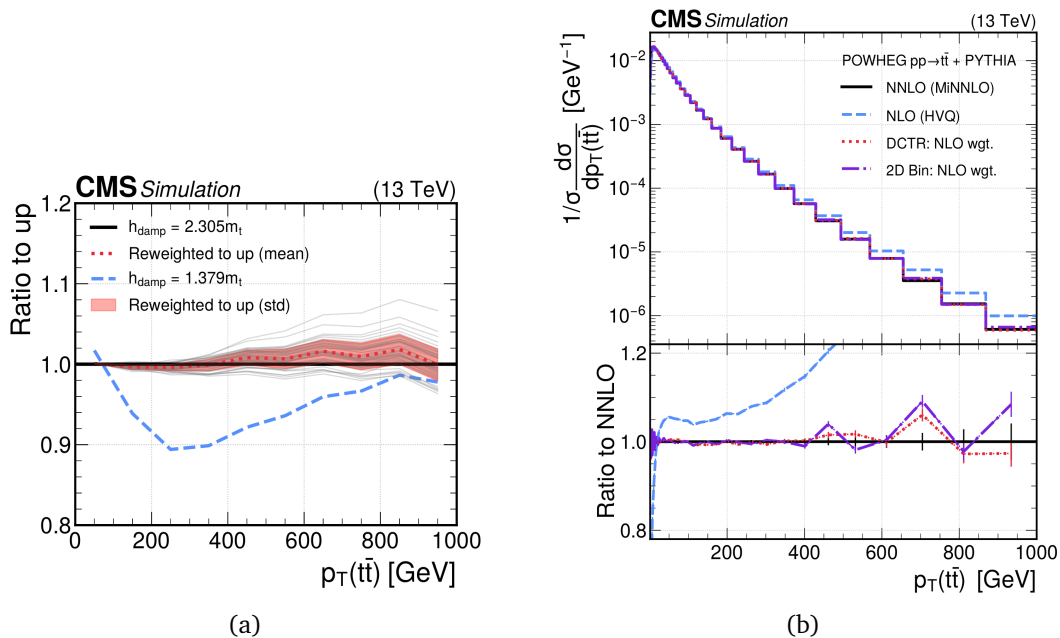



Figure 3: Reweighting of simulated events using the DCTR method [20] by CMS: (a) reweight to emulate a variation of the h_{damp} parameter of the POWHEG event generator; (b) reweight to emulate NNLO accuracy from an NLO sample. Figures are taken from Ref. [21].

References

[1] ATLAS collaboration: G. Aad et al., *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79**, 970 (2019), doi:10.1140/epjc/s10052-019-7450-8 [preprint doi:10.48550/arXiv.1907.05120].

[2] CMS collaboration: A. M. Sirunyan et al., *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, J. Instrum. **13**, P05011 (2018), doi:10.1088/1748-0221/13/05/P05011 [preprint doi:10.48550/arXiv.1712.07158].

[3] CDF collaboration: T. Aaltonen et al., *Observation of single top quark production and measurement of $|V_{tb}|$ with CDF*, Phys. Rev. D **82**, 112005 (2010), doi:10.1103/PhysRevD.82.112005 [preprint doi:10.48550/arXiv.1004.1181].

[4] D0 collaboration: V. M. Abazov et al., *Observation of single top-quark production*, Phys. Rev. Lett. **103**, 092001 (2009), doi:10.1103/PhysRevLett.103.092001 [preprint doi:10.48550/arXiv.0903.0850].

[5] ATLAS collaboration: G. Aad et al., *The ATLAS experiment at the CERN Large Hadron Collider*, J. Instrum. **3**, S08003 (2008), doi:10.1088/1748-0221/3/08/S08003.

[6] ATLAS collaboration: G. Aad et al., *Observation of four-top-quark production in the multilepton final state with the ATLAS detector*, Eur. Phys. J. C **83**, 496 (2023), doi:10.1140/epjc/s10052-023-11573-0 [preprint doi:10.48550/arXiv.2303.15061].

[7] CMS collaboration: S. Chatrchyan et al., *The CMS experiment at the CERN LHC*, J. Instrum. **3**, S08004 (2008), doi:10.1088/1748-0221/3/08/S08004.

[8] CMS collaboration: A. Hayrapetyan et al., *Observation of four top quark production in proton-proton collisions at 13 TeV*, Phys. Lett. B **847**, 138290 (2023), doi:10.1016/j.physletb.2023.138290 [preprint doi:10.48550/arXiv.2305.13439].

[9] M. Leigh, J. A. Raine, K. Zoch and T. Golling, *ν-flows: Conditional neutrino regression*, SciPost Phys. **14**, 159 (2023), doi:10.21468/SciPostPhys.14.6.159 [preprint doi:10.48550/arXiv.2207.00664].

[10] J. A. Raine, M. Leigh, K. Zoch and T. Golling, *Fast and improved neutrino reconstruction in multineutrino final states with conditional normalizing flows*, Phys. Rev. D **109**, 012005 (2024), doi:10.1103/PhysRevD.109.012005 [preprint doi:10.48550/arXiv.2307.02405].

[11] M. J. Fenton, A. Shmakov, H. Okawa, Y. Li, K.-Y. Hsiao, S.-C. Hsu, D. Whiteson and P. Baldi, *Reconstruction of unstable heavy particles using deep symmetry-preserving attention networks*, Commun. Phys. **7**, 139 (2024), doi:10.1038/s42005-024-01627-4 [preprint doi:10.48550/arXiv.2309.01886].

[12] C. Birch-Sykes, B. Le, Y. Peters, E. Simpson and Z. Zhang, *Reconstructing short-lived particles using hypergraph representation learning*, Phys. Rev. D **111**, 032004 (2025), doi:10.1103/PhysRevD.111.032004 [preprint doi:10.48550/arXiv.2402.10149].

[13] G. Kasieczka, B. Nachman, M. D. Schwartz and D. Shih, *Automating the ABCD method with machine learning*, Phys. Rev. D **103**, 035021 (2021), doi:10.1103/PhysRevD.103.035021 [preprint doi:10.48550/arXiv.2007.14400].

[14] CMS collaboration: A. Tumasyan et al., *Evidence for four-top quark production in proton-proton collisions at 13 TeV*, Phys. Lett. B **844**, 138076 (2023), doi:10.1016/j.physletb.2023.138076 [preprint doi:10.48550/arXiv.2303.03864].

[15] P. de Castro and T. Dorigo, *INFERNO: Inference-aware neural optimisation*, Comput. Phys. Commun. **244**, 170 (2019), doi:10.1016/j.cpc.2019.06.007 [preprint doi:10.48550/arXiv.1806.04743].

[16] R. Barrué, P. Conde Muíño, V. Dao and R. Santos, *Simulation-based inference in the search for CP violation in leptonic WH production*, J. High Energy Phys. **04**, 014 (2024), doi:10.1007/JHEP04(2024)014 [preprint doi:10.48550/arXiv.2308.02882].

[17] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman and J. Thaler, *OmniFold: A method to simultaneously unfold all observables*, Phys. Rev. Lett. **124**, 182001 (2020), doi:10.1103/PhysRevLett.124.182001 [preprint doi:10.48550/arXiv.1911.09107].

[18] ATLAS collaboration: G. Aad et al., *Simultaneous unbinned differential cross-section measurement of twenty-four Z + jets kinematic observables with the ATLAS detector*, Phys. Rev. Lett. **133**, 261803 (2024), doi:10.1103/PhysRevLett.133.261803 [preprint doi:10.48550/arXiv.2405.20041].

[19] CMS collaboration, *Measurement of event shapes in minimum bias events from pp collisions at 13 TeV*, Tech. Rep. CMS-PAS-SMP-23-008, CERN, Geneva, Switzerland (2024), https://cds.cern.ch/record/2899591.

[20] A. Andreassen and B. Nachman, *Neural networks for full phase-space reweighting and parameter tuning*, Phys. Rev. D **101**, 091901 (2020), doi:10.1103/PhysRevD.101.091901 [preprint doi:10.48550/arXiv.1907.08209].

[21] CMS collaboration: A. Hayrapetyan et al., *Reweighting simulated events using machine-learning techniques in the CMS experiment*, Eur. Phys. J. C **85**, 495 (2025), doi:10.1140/epjc/s10052-025-14097-x [preprint doi:10.48550/arXiv.2411.03023].

[22] S. Alioli, P. Nason, C. Oleari and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: The POWHEG BOX*, J. High Energy Phys. **06**, 043 (2010), doi:10.1007/JHEP06(2010)043 [preprint doi:10.48550/arXiv.1002.2581].