Applying Machine Learning to the Matrix Element Method in Searches for Invisible Higgs Boson Decays

Author:

Luke

Johnson

Supervisors:

Dr. Sudan Paramesvaran,

Dr. Florian Bury

School of Physics

University of Bristol

A thesis submitted to the University of Bristol in accordance with the requirements of the Master of Science degree in Theoretical Physics

March 2025

Word count: 9189

Declaration

I, Luke Johnson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

This analysis is based on the 2018 Ultra Legacy dataset provided by the CMS Collaboration. All figures presented in this report were produced by myself, with the exception of the three neural framework diagrams (Figs. 7, 9, & 10), the Higgs potential (Fig. 6), and the Run 2 upper limits for the target process (Figs. 3–5) . The code for generating all the results plots was written by myself, except for the bias (Fig. 17) and classifier (Figs. 21 & 22) plots, which had source code written by Dr. Florian Bury.

My project partner, Tom Lenane, and I were jointly responsible for developing the preprocessing dataclasses used to rescale the dataset before feeding it into the generative models. The code for the classifiers and autoregressive Transfermer were provided to us by Dr. Florian Bury, with improvements made by Tom Lenane. However, training and inference were done by myself using the $t\bar{t}H$ dataset to create the corresponding results.

I was solely responsible for designing and writing all the code for the various Conditional Flow Matching (CFM) networks. This was made from scratch (using PyTorch) to construct an in-house framework encompassing a multitude of transport plans, bridging distributions, and optimal transport techniques. The corresponding notebooks used to train, run inference, and compare the CFM models were also made by myself.

Throughout the project, we benefited from invaluable guidance from Dr. Sudan Parameswaran and Dr. Florian Bury, who gave excellent ideas regarding the direction of the work. All the machine learning models interacted with the Mem-Flow library which was developed by Dr. Florian Bury and his colleagues at ETH Zurich.

Acknowledgements

I am deeply grateful to my supervisors, Dr. Sudan Paramesvaran and Dr. Florian Bury, for their unwavering support, expertise, and encouragement throughout this project. Your insightful guidance and enthusiasm have made the entire research process truly engaging and rewarding. My enjoyment of this project has fuelled my passion for particle physics research, which I am incredibly excited to pursue further through a PhD.

A special thank you to Dr. Florian Bury for his consistent and lightning-fast responses to my many queries – no matter the time or the day. This was not only impressive but also helped me learn so much while completing my research.

I am also greatly appreciative of my friends and peers at Bristol for many memorable moments over the years.

Finally, I apologise to anyone using the `gpu04` node on DICE while I was monopolising all six GPUs to train my generative models.

Abstract

This study explores using novel machine learning approaches to improve the Matrix Element Method for invisible Higgs boson decays in the $t\bar{t}H$ fully hadronic channel. The Transfer-function was encoded using both invertible (cINN) and flow-matching (CFM) conditional neural networks. It was found that CFM offers superior expressivity with its continuous transport map, whereas the cINN achieved 3.6x faster sampling speed. The inclusion of a Transformer provided much stronger estimates of reco-level topologies. Classifiers were successfully implemented to account for variable multiplicity and detector efficiency, with the latter achieving within 0.01% of the true acceptance rate.

Contents

1 Introduction

The Higgs boson (H) was first proposed in 1964 by Peter Higgs [1] and a group of theorists to explain the mechanism through which certain particles acquire mass. Its existence was confirmed in 2012 at CERN by the ATLAS [2] and CMS [3] Collaborations, with the most precise recording of its mass to date being 125.11 ± 0.11 GeV, measured by the ATLAS collaboration in 2023 [4]. This discovery has since led to the proliferation of research into the Higgs boson's properties with the aim of uncovering potential signs of new physics beyond the Standard Model (BSM). High-energy, proton-proton (pp) collisions at $\sqrt{s}$ = 7, 8, 13 TeV at the Large Hadron Collider (LHC) will enable researchers to probe the Higgs boson's self-coupling and its interactions with other standard model (SM) particles.

Higgs bosons are predominantly produced via the loop-induced gluon-gluon fusion (ggF) mechanism, which accounts for approximately 87% of its production [5]. Other production mechanisms include vector boson fusion (VBF), in which two quarks from the colliding protons exchange W or Z bosons that fuse to produce a Higgs boson, and VH production, where a Higgs boson is produced alongside a W or Z boson. This study focuses on the rare production of Higgs bosons in association with a top quark-antiquark pair ($t\bar{t}H$), a process that accounts for ∼1% of Higgs production [6].

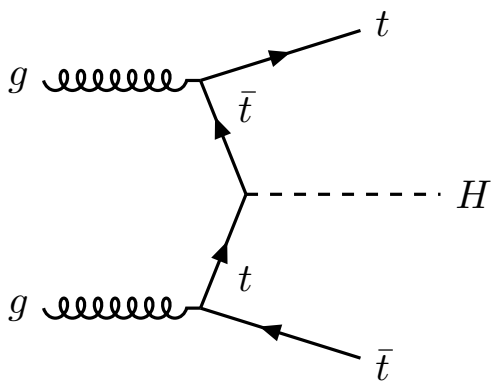

Figure 1: Leading-order Feynman diagram for the SM Higgs boson production mechanism $t\bar{t}H$.

The leading-order (LO) Feynman diagram for $t\bar{t}H$ production is shown in Fig. 1. According to the SM, the Higgs boson couples to fermions with a strength proportional to their mass [7]. Since the top quark is the heaviest known fermion, its Yukawa coupling to the Higgs boson is the strongest. This makes the $t\bar{t}H$ production mode particularly important for directly probing Higgs-fermion interactions. In 2018, the CMS Experiment observed these events with a significance of 5.2 standard deviations above the expectation from the background-only hypothesis [7]. This established the first direct evidence

of Higgs boson coupling to an up-type quark, marking a significant milestone in validating the SM description of Higgs-fermion interactions.

In the SM, the Higgs boson can only decay to an invisible final state via $H \to ZZ^* \to 4\nu$, with a predicted branching fraction of approximately 0.1% [8]. This low probability arises from the need for one Z^* boson to be off-shell, suppressing the available phase space for the decay. Moreover, since neutrinos in the SM are purely 'left-handed', they cannot participate in the Yukawa interaction with the Higgs, meaning they are massless and do not couple to the Higgs boson. Furthermore, each Z boson can decay into leptons, quark-antiquark pairs, or neutrinos, with each Z decaying to a pair of neutrinos only 20% of the time [9]. Hence, the probability of both Z bosons decaying exclusively into neutrinos is low, making the fully invisible final state incredibly rare.

Direct searches for $H \to \text{inv}$ have been conducted by general-purpose detectors (e.g. ATLAS and CMS) at the LHC [10–12] using data collected during Run 1 (2011-2012) and Run 2 (2015-2018). Run 3 commenced in 2022 with an improved beam energy of 6.8 TeV and integrated luminosity of 88.9 fb^{-1} for ATLAS and CMS as of September 2024 [13]. This run is expected to last until 2026 [14] and has scope for even more precise measurements.

Various BSM theories predict significant enhancements in the branching ratio $B_{H\to\text{inv}}$. One prominent example is the "Higgs-portal" dark matter model [15], where dark matter candidates interact with SM particles exclusively through their couplings with the Higgs sector. Provided these particles have a mass $m_{DM} < \frac{1}{2}m_H \approx 62.5$ GeV [10], the Higgs boson could acquire a new channel for invisible decays, as these dark matter particles would also be undetectable. This motivates further investigation into invisible Higgs decays, as any deviation from the SM's prediction for $B_{H\to\text{inv}}$ could reveal strong evidence for new physics.

The matrix element method (MEM) was initially developed for studies of W boson and top quark physics at the Tevatron collider [16]. The MEM is a powerful inference tool for computing the likelihoods of physical observables based on theoretical predictions. It has proven particularly effective for LHC analyses with limited event numbers. However, MEM calculations are computationally expensive, which precludes scalability for large datasets. Recent advancements in machine learning (ML) have enabled a more efficient means of approximating MEM calculations, offering the potential to significantly enhance the sensitivity of Higgs boson searches.

A key challenge of the MEM is the need to integrate over all possible parton-level configurations that could produce the observed detector-level event. This requires un-

folding detector measurements to infer parton-level quantities, as theoretical predictions are made at the hard-scattering level, whereas detectors record reconstructed objects such as jets and energy deposits. Traditional unfolding techniques such as matrix inversion [17], Bayesian unfolding [18], and Tikhonov regularisation [19] rely on response matrices derived from Monte Carlo (MC) simulations of the detector. These approaches are model-dependent, as they assume a specific theoretical model to build the response matrix, potentially introducing biases if the MC simulations are not physically accurate. In contrast, ML-based methods like conditional invertible neural networks (cINNs) [20] learn the unfolding process directly from data without relying on a predefined response matrix. These model-independent alternatives are less prone to bias, speed up numerical convergence, and improve the precision of likelihood calculations.

This paper presents the first-ever application of ML in the MEM for both $H \rightarrow$ inv decays and $t\bar{t}H$ production (with the fully hadronic final state). This pioneering research is conducted in collaboration between the University of Bristol, ETH Zurich, and the Machine Learning Group of CMS, and introduces a cutting-edge ML-framework which incorporates generative networks, acceptance networks [21], and a Transformer architecture [22]. The primary aim is to extract likelihood distributions for invisible Higgs boson decays produced via the ttH production mechanism.

2 The CMS Detector

The Compact Muon Solenoid (CMS) is a general-purpose detector at the Large Hadron Collider (LHC) located at CERN. It is capable of measuring photons, muons, electrons, and hadrons. The detector's central component is a superconducting solenoid magnet with a 6-meter internal diameter. Its interior is comprised of the strip tracker, electromagnetic calorimeter (ECAL), and hadron calorimeter (HCAL). Gas-ionisation chambers facilitate muon detection and measurements. The 'PARTICLE-FLOW' algorithm [23] is responsible for particle reconstruction and combines data from various detector subsystems. This builds detector-level observables such as missing transverse momentum (p_T^{miss}) and jets.

The Level-1 trigger system (L1T) leverages custom, low-latency hardware, such as field-programmable gate arrays (FPGAs), to perform online selection of the most interesting events. This is done at an impressive rate of around 100 kHz. The high-level trigger (HLT) uses more advanced event reconstruction to reduce this to a manageable rate of roughly 1 kHz [24] for further analysis. The High-Luminosity LHC (HL-LHC) [25, 26] era will require drastically increased computing power to handle the expected 200 pileup per event. This upgrade will come with new thresholds for the L1T acceptance rate (500

kHz) and HLT rate (5 kHz) [26]. For a more complete description of the CMS experiment see Ref. [27].

3 Decay Channels

The $H \to \text{inv}$ process can be accompanied by a variety of final states. Firstly, each top quark from Fig. 1 decays via

$$t \to W^+ b, \qquad \bar{t} \to W^- \bar{b}. \tag{1}$$

Hence, the final state depends on the decay modes of the W bosons. Fully-leptonic decays occur when both W bosons decay via

$$W^+ \to \ell^+ \nu, \qquad W^- \to \ell^- \bar{\nu}, \tag{2}$$

with a branching fraction $\mathcal{B}(W \to \ell\,\bar{\nu})$ of 10.83%, 10.94%, 10.77% for e, μ, τ respectively [28]. Alternatively, one W can decay leptonically while the other decays hadronically: $W \to q\bar{q}'$, where $\mathcal{B}(W \to q\bar{q}') = 67.46\%$ [28].

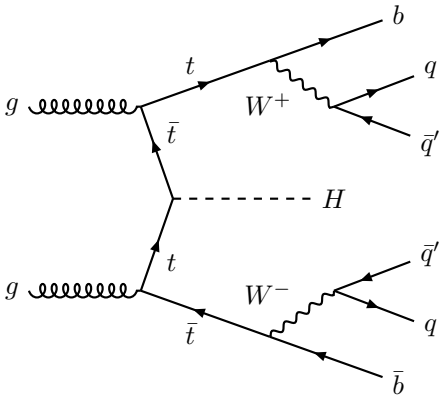

Figure 2: Feynman diagram showing the fully hadronic final state of $t\bar{t}H$ production.

This study targets the fully-hadronic decay channel of $H \to \text{inv}$, which is the most common and accounts for 45.5% of decays. Here both Ws decay to quarks which results in a total of 6 jets (2 b-jets from top decays and 4 light-flavour jets from W decays). The Feynman diagram for this process is shown in Fig. 2. The primary background sources include $t\bar{t}$ events, which yield the same final state but without a Higgs boson, and additional $Z \to \text{inv}$ decays, which also contribute missing energy.

The published estimates for $\mathcal{B}(H \to \text{inv})$ from the CMS Run 2 analysis are illustrated in Figs. 3 & 4. The former exclusively considers fully hadronic decays of the $t\bar{t}H$ production

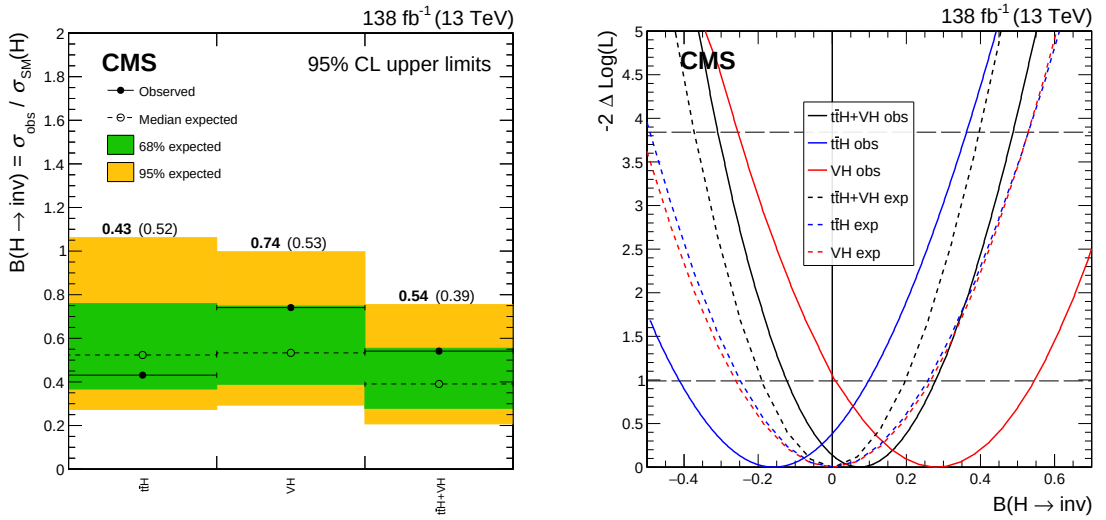

Figure 3: Observed and expected limits at 95% CL on $\mathcal{B}(H \to \text{inv})$ for the $t\bar{t}H$ and VH production mechanisms using Run 2 data (left). Profile likelihood scan showing the fit for various observed and expected ($\mathcal{B}(H \to \text{inv}) = 0$) limits (right) [11].

mode, which aligns closely with the focus of this study. The corresponding 95% confidence level (CL) upper limit (UL) for isolated hadronic $t\bar{t}H$ decays is $\mathcal{B}(H \to \text{inv}) = 0.43$ [11] . This is substantially higher than the SM target of 0.1%, although improvements are achieved when combining other production modes.

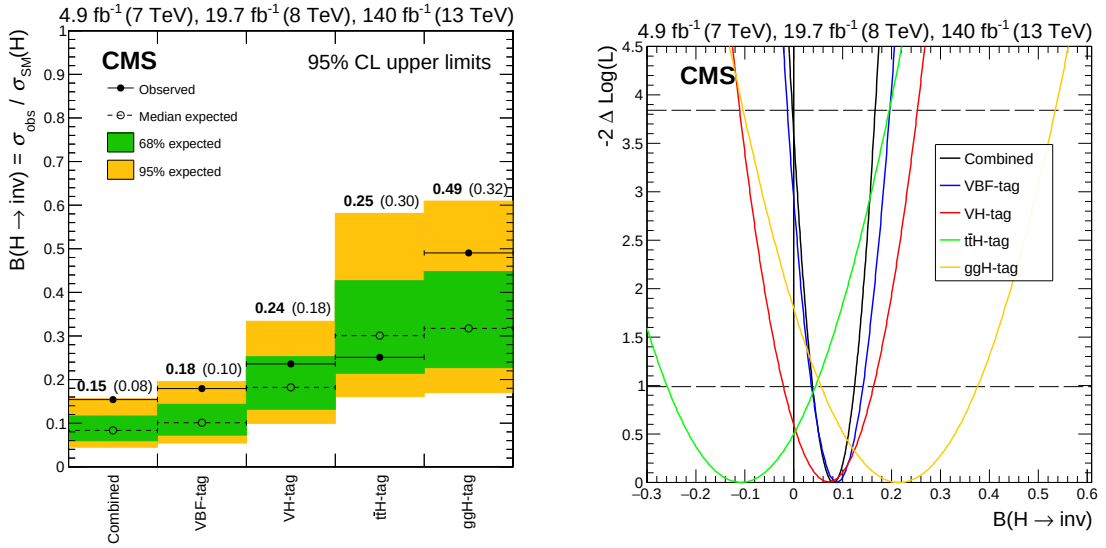

Figure 4: Upper limits at 95% CL on $\mathcal{B}(H \to \text{inv})$ for various Higgs boson production modes. Tagged by the CMS input analyses for Run 1 and Run 2 (left). Profile likelihood scan showing the fit for various observed and expected limits (right) [11].

Fig. 4 depicts how the inclusion of leptonic decay channels within the $t\bar{t}H$ process improves the UL significantly, reducing it to $\mathcal{B}(H \to \text{inv}) = 0.25$. Combining all of the Higgs boson production modes and incorporating both Run 1 and Run 2 data substantially

enhances the statistical power, resulting in a final UL of $\mathcal{B}(H \to \text{inv}) = 0.15$ [11].

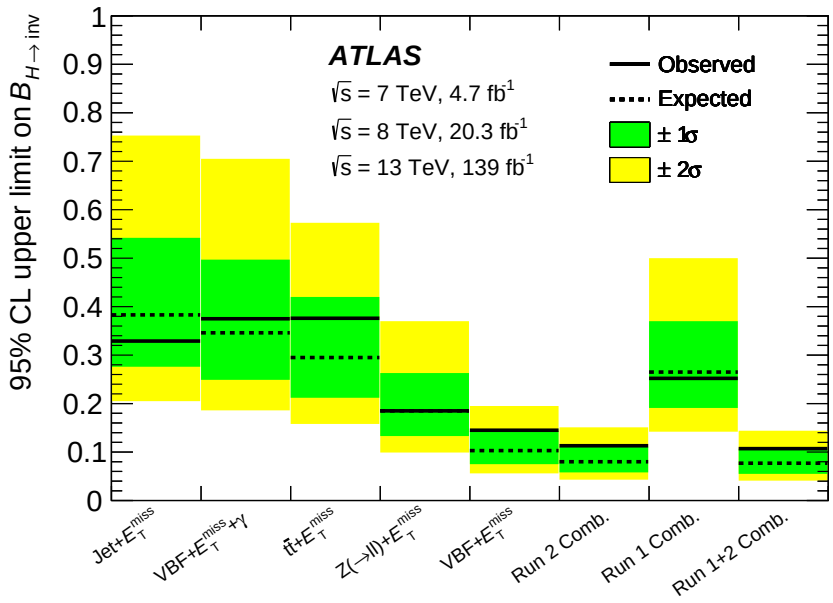

Figure 5: Upper limits at 95% CL on $\mathcal{B}(H \to \text{inv})$ for various Higgs boson production modes. Includes the combined ATLAS Run 1 + 2 result [12].

Invisible Higgs boson decays are also explored by the ATLAS experiment [10, 12]. Differences in detector signatures, background estimation, and selection criteria necessitate separate analyses. These consequently yield distinct upper limits for $\mathcal{B}(H \to \text{inv})$. The ATLAS Run 2 analysis observed a 95% CL upper limit of $\mathcal{B}(H \to \text{inv}) = 0.376$ when exclusively considering the $t\bar{t}H$ category. Merging the Run 1 and Run 2 data allowed ATLAS to surpass the CMS result with an observed UL of $\mathcal{B}(H \to \text{inv}) = 0.107$ [12].

The ML framework and methodology outlined in this work have been tested using simulated CMS samples from Run 2 (2018). Their proven efficacy within the MEM justifies their strong suitability for use in a real analysis based on Run 3 data collected at the LHC.

4 Higgs Mechanism

The Higgs mechanism is a fundamental concept in the SM that explains the mechanism through which gauge bosons acquire their mass. In 1964, François Englert, Robert Brout [29], and Peter Higgs [1] proposed that particles were massless during the early universe, gaining mass only through a process known as spontaneous symmetry breaking as the universe cooled after the Big Bang.

A theory is considered symmetric under a group G if transformations associated with

G leave the underlying laws unchanged and the Lagrangian, describing the dynamics of the system, invariant [30]. One important class of symmetries in quantum field theory is gauge symmetry, which involves local gauge transformations of the form

$$\psi'(\vec{r},t) = e^{-i\theta(\vec{r},t)}\psi(\vec{r},t), \tag{3}$$

where the phase θ of a wave function ψ depends on position $\vec{r}$ and time t. Gauge theories describe particle interactions via gauge fields, which mediate forces between particles. These gauge fields correspond to the force carriers in the SM.

The SM describes the electromagnetic and weak nuclear forces as different manifestations of a unified interaction via a Yang-Mills field. The underlying symmetry of this electroweak interaction is represented by the $SU(2)_L \times U(1)_Y$ gauge group. Here, $U(1)_Y$ represents the symmetry associated with weak hypercharge, a quantum number related to the electric charge. In contrast, $SU(2)_L$ governs the weak nuclear force and is related to weak isospin. The corresponding gauge bosons for these symmetries are the W^1, W^2, W^3 (for $SU(2)_L$), and B (for $U(1)_Y$) fields.

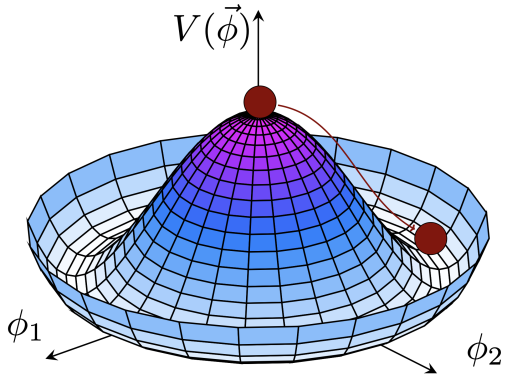

Figure 6: Higgs potential V as a function of the two-dimensional scalar field ϕ [31].

In 1967, Stephen Weinberg [32] and Abdus Salam [33] extended the unified electroweak theory developed by Sheldon Lee Glashow [34] by incorporating the Higgs mechanism. Their work led to the formulation of the Glashow-Weinberg-Salam (GWS) model, which now serves as a crucial part of the SM. Initially, the W^1, W^2, W^3, and B gauge bosons were massless. However, below a critical temperature, the Higgs field developed a non-zero vacuum expectation value (VEV), illustrated in Fig. 6. This spontaneously broke the $SU(2)_L \times U(1)_Y$ symmetry down to $U(1)_{EM}$, corresponding to the electromagnetic interaction.

This resulted in the emergence of three massive gauge bosons (W^+, W^-, Z^0) and one massless boson, the photon (γ), which mediates the electromagnetic force. Specifically, the charged gauge bosons W^1 and W^2 combined to form the massive $W^\pm$ bosons, while

the neutral bosons W^3 and B mixed to form the photon (γ) and Z^0 boson.

The GWS model also distinguishes fermions based on chirality. Left-handed fermions form weak isospin doublets under $SU(2)$ [35], while right-handed fermions are weak isospin singlets, reflecting why only left-handed particles and right-handed antiparticles experience the weak force.

The photon remains massless due to the unbroken $U(1)_{EM}$ symmetry. This ensures electromagnetic forces remain long-range. The masses of the W and Z bosons originate from the coupling constants of the gauge interactions and the VEV of the Higgs field. $W^{\pm}$ bosons acquire a mass of

$$M_W = \frac{1}{2} g v, \tag{4}$$

where g is the coupling constant for $SU(2)_L$ and v is the VEV of the Higgs field, approximately 246 GeV [36]. The Z boson mass is related to the W boson mass through

$$M_Z = \frac{M_W}{\cos\theta_w}, \tag{5}$$

where θ_w is the weak mixing angle, or Weinberg angle. This angle quantifies the mixing between the neutral components of the $SU(2)_L$ and $U(1)_Y$ gauge fields and is defined by

$$\tan\theta_W = \frac{g'}{g}, \tag{6}$$

where g and g' are the coupling constants for $SU(2)_L$ and $U(1)_Y$ respectively.

5 ML Matrix Element Method

In high-energy physics, the differential cross-section describes the hard-scattering process at the parton level. The MEM can be used to track the dependence of the hard-scattering cross-section on a given theory parameter α to extract likelihoods from collision data [21]. The differential cross-section can be factorised into

$$\frac{d\sigma(\alpha)}{dx_{\text{hard}}} = \sigma(\alpha)\, p(x_{\text{hard}} \mid \alpha), \tag{7}$$

where σ is the total cross section, $p(x_{\text{hard}} \mid \alpha)$ is a normalised probability density, and x_{hard} denotes the hard-scattering momenta. Eq.(7) can be rearranged for $p(x_{\text{hard}} \mid \alpha)$ to obtain the probability of a parton-level event corresponding to a given value of α.

Following the hard-scattering, additional processes such as parton showers, fragmen-

tation, detector effects, and event reconstruction can then be simulated to provide the reco-level cross section,

$$\frac{d\sigma_{\text{fid}}(\alpha)}{dx_{\text{reco}}} = \int dx_{\text{hard}}\, r(x_{\text{reco}} \mid x_{\text{hard}}) \frac{d\sigma(\alpha)}{dx_{\text{hard}}}. \tag{8}$$

Here, $r(x_{\text{reco}} \mid x_{\text{hard}})$ represents the transfer function that evolves hard-scattering processes to detector-level observations x_{reco}. It essentially represents the probability density of reconstructing a detector-level event x_{reco} given a hard-level event x_{hard}. The fiducial cross section $\sigma_{\text{fid}}(\alpha)$ accounts for detector acceptance and selection criteria that restrict the observed phase space compared to the parton-level phase space.

The transfer function $r(x_{\text{reco}} \mid x_{\text{hard}})$ must encapsulate the possibility that some parton-level events may not be successfully reconstructed in the detector (e.g. due to trigger requirements or acceptance cuts). Therefore, the response function r seen in Eq.(8) can be replaced with

$$r(x_{\text{reco}} \mid x_{\text{hard}}) = \epsilon(x_{\text{hard}}) p(x_{\text{reco}} \mid x_{\text{hard}}), \tag{9}$$

where $\epsilon(x_{\text{hard}})$ is efficiency of the detector, quantifying the probability of an event being accepted and successfully reconstructed. $p(x_{\text{reco}} \mid x_{\text{hard}})$ is the normalised transfer probability of observing a reconstructed event given a specific hard-scattering process. Eq.(9) can now be substituted into Eq.(8) to dispense an expression for the differential cross section,

$$\frac{d\sigma_{\text{fid}}(\alpha)}{dx_{\text{reco}}} = \int dx_{\text{hard}}\, \epsilon(x_{\text{hard}}) p(x_{\text{reco}} \mid x_{\text{hard}}) \frac{d\sigma(\alpha)}{dx_{\text{hard}}}. \tag{10}$$

By integrating Eq.(10) over the reco-level phase space, the fiducial cross section $\sigma_{\text{fid}}(\alpha)$ can be obtained and used to represent the final expression for the reco-level likelihood

$$p(x_{\text{reco}} \mid \alpha) = \frac{1}{\sigma_{\text{fid}}(\alpha)} \int dx_{\text{hard}}\, \frac{d\sigma(\alpha)}{dx_{\text{hard}}}\, \epsilon(x_{\text{hard}})\, p(x_{\text{reco}} \mid x_{\text{hard}}). \tag{11}$$

In this study, the theory parameter α parameterises the presence of the $t\bar{t}H$ hypothesis from Fig. 1. The complete architecture of the MEM integrator can be seen in Fig. 7, with the theory and framework largely following Ref. [22].

5.1 Sampling-cINN

The MEM integral requires considering all possible hard-scattering phase space configurations that could produce the observed detector-level state. Since the differential cross section spans several orders of magnitude, the integration in Eq.(11) becomes extremely

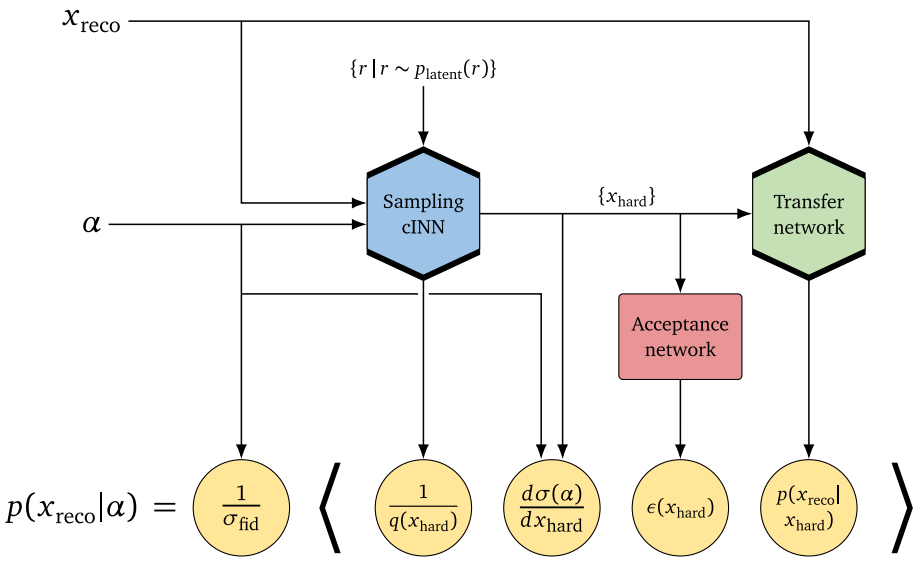

Figure 7: The neural framework used to evaluate the MEM integral through sampling the response function r. The Sampling-cINN is conditioned on the $t\bar{t}H$ process (α) and detector-level event x_{reco}. The Transfer network is conditioned on the parton-level configuration x_{hard}. The acceptance network computes $\epsilon(x_{\text{hard}})$ to account for detector efficiency [22].

demanding. Furthermore, the transfer probability $p(x_{\text{reco}} \mid x_{\text{hard}})$ forms narrow peaks in phase space due to propagator enhancements in the matrix element, such as resonances and collinear singularities. In contrast, detector effects and jet smearing reduce the precision of the reconstruction process, with the jet energy resolution actually introducing significant broadening, particularly at lower transverse momenta. These effects mean that only specific x_{hard} configurations contribute meaningfully to x_{reco}.

Traditional MC integration relies on randomly sampling points in phase space to approximate the integral. Unfortunately, due to the narrow phase space peaks, most configurations of x_{hard} are superfluous and contribute negligibly. Consequently, this naive approach converges very slowly, wasting lots of effort sampling from irrelevant regions. The Sampling-cINN seen in Fig. 7 circumvents this challenge by employing a heuristic sampling strategy that prioritises the most relevant regions of phase space.

If Eq.(11) is rewritten with

$$p_{\text{fid}}(x_{\text{hard}} \mid \alpha) = \frac{1}{\sigma_{\text{fid}}(\alpha)} \frac{d\sigma(\alpha)}{dx_{\text{hard}}} \epsilon(x_{\text{hard}}), \tag{12}$$

the reco-level likelihood $p(x_{\text{reco}} \mid \alpha)$ can be expressed as

$$p(x_{\text{reco}}|\alpha) = \int dx_{\text{hard}}\, p_{\text{fid}}(x_{\text{hard}}|\alpha)\, p_\theta(x_{\text{reco}}|x_{\text{hard}}), \tag{13}$$

where $p_{\text{fid}}(x_{\text{hard}} \mid \alpha)$ is the modified parton-level likelihood, accounting for acceptance cuts. Essentially, it re-weights the previous probability distribution $p(x_{\text{hard}} \mid \alpha)$, which contains all theoretically possible events, to reflect those that can actually be observed in the detector. $p_\theta(x_{\text{reco}}|x_{\text{hard}})$ is the transfer probability learned by the neural network.

The Sampling-cINN learns an importance distribution $q(x_{\text{hard}} \mid x_{\text{reco}}, \alpha)$, which approximates the optimal distribution for minimising variance in the MEM integral. This distribution is proportional to the conditional parton-level probability and causes the integral's variance to vanish when

$$q(x_{\text{hard}} \mid x_{\text{reco}}, \alpha) \propto p(x_{\text{hard}} \mid x_{\text{reco}}, \alpha). \tag{14}$$

However, in practice the distribution is approximated by the product of the learned transfer probability p_θ and the fiducial cross-section density p_{fid},

$$q(x_{\text{hard}} \mid x_{\text{reco}}, \alpha) \propto p_\theta(x_{\text{reco}} \mid x_{\text{hard}}) p_{\text{fid}}(x_{\text{hard}} \mid \alpha). \tag{15}$$

The Sampling-cINN trains a normalizing flow to map random latent variables r from a simple Gaussian distribution to parton-level events x_{hard}, conditioned on α and x_{reco},

$$r \sim p_{\text{latent}}(r), \xleftrightarrow{\text{Sampling-cINN}} x_{\text{hard}}(r) \sim q_\phi(x_{\text{hard}} \mid x_{\text{reco}}, \alpha), \tag{16}$$

where ϕ represents the trainable parameters of the network.

The network aims to minimise the variance of the MC estimator by iteratively refining the importance distribution to emulate the true distribution. This negates sampling configurations with trivial $p(x_{\text{reco}} \mid x_{\text{hard}})$, thereby expediting numerical efficiency and convergence.

5.2 Transfer Network

The primary objective of the Transfer network in Fig. 7 is to establish a bijective mapping between a latent Gaussian distribution and the detector-level phase space, conditioned on x_{hard}. It is trained on pairs of parton-level and reco-level events (x_{hard}, x_{reco}) that pass detector selection criteria. This captures the transfer probability from Eq.(9),

$$x_{\text{reco}} \sim p_\theta(x_{\text{reco}} \mid x_{\text{hard}}) \xleftrightarrow{\text{Transfer network}} r \sim p_{\text{latent}}(r), \tag{17}$$

where r is a latent vector drawn from a simple Gaussian distribution, and θ represents the trainable parameters.

The Transfer network serves as a generative model; it learns the unknown distribution of reco-level events $p(x_{\text{reco}} \mid x_{\text{hard}})$. If implemented as a cINN using a normalizing flow architecture, it would also be invertible, facilitating the generation of synthetic detector-level events based on conditional parton-level inputs. A key feature of the network is its use of a Transformer architecture with self-attention mechanisms at its core. This improves the model's ability to encode more complex particle interactions without relying as heavily on localised features.

5.3 Acceptance & Multiplicity Networks

The acceptance $\epsilon(x_{\text{hard}})$ from Fig. 7 can be encoded using a simple binary classifier. The network estimates the probability of an event being accepted and successfully reconstructed in the detector. It learns to distinguish between accepted (labelled 1) and rejected (labelled 0) events by employing a binary cross-entropy loss function to enforce well-calibrated output probabilities. Once trained, this model provides an efficient means of event re-weighting within the MEM calculation.

The multiplicity network serves as a multiclass classification model and is responsible for predicting the number of reconstructed jets in a given event. Events may contain additional jets from QCD radiation, while others may be lost due to acceptance and trigger cuts. Incorporating this into the Transfer network is crucial for accurate event generation. Categorical cross-entropy loss is used to correctly encode the number of jets,

$$p(x_{\text{reco}}, n \mid x_{\text{hard}}) = p(n \mid x_{\text{hard}}) p(x_{\text{reco}} \mid x_{\text{hard}}, n). \tag{18}$$

Here, $p(n \mid x_{\text{hard}})$ is the probability of an event containing n final state particles. Once trained, it is possible to initially sample the number of jets from the classifier, and then dynamically generate the reco-level particles using the generative models.

6 Normalizing Flows

Normalizing flows (NFs) are a class of generative models that learn a complex target distribution by transforming a simple base distribution with a sequence of invertible transformations. Let the invertible map $T : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable and

the sequence of samples $z^{(1)}, \ldots, z^{(n)} \in \mathbb{R}^d$ follow $z \sim p_z$ [37] which is chosen to be a simple Gaussian $\mathcal{N}(0, I_d)$. Now, the conditional target distribution $p(x_{\text{reco}} \mid x_{\text{hard}})$ can be induced by sampling

$$x_{\text{reco}} = T(z; x_{\text{hard}}), \qquad z = T^{-1}(x_{\text{reco}}; x_{\text{hard}}). \tag{19}$$

The probability density of x_{reco} can be obtained with the change-of-variables formula

$$\begin{aligned} p(x_{\text{reco}} \mid x_{\text{hard}}) &= p_z(T^{-1}(x_{\text{reco}}; x_{\text{hard}})) \left|\det J_{T^{-1}}(x_{\text{reco}}; x_{\text{hard}})\right| \\ &= \frac{p_z(z)}{\left|\det J_T(z; x_{\text{hard}})\right|}, \end{aligned} \tag{20}$$

where $J_{T^{-1}}(x_{\text{reco}}; x_{\text{hard}})$ is the Jacobian of the inverse map and $J_{T^{-1}} \in \mathbb{R}^{d \times d}$.

A simple linear transformation would be inadequate for the generative model to approximate the complex reco-level distribution. A neural network can be used to learn a flexible transformation T_θ with learnable parameters θ [38–40]. This can be broken down into a series of simpler maps $T_\theta = \phi_K \circ \ldots \circ \phi_1$. Each transformation ϕ_K is designed to be invertible and have a tractable Jacobian determinant in order to satisfy Eq.(20). Using the chain rule yields the following result for the log-likelihood

$$\begin{aligned} \log p_\theta(x_{\text{reco}} \mid x_{\text{hard}}) &= \log p_z(\phi_1^{-1} \circ \ldots \circ \phi_K^{-1}(x_{\text{reco}}; x_{\text{hard}})) + \sum_{k=1}^{K} \log\left|\det J_{\phi_k^{-1}}(z_k)\right| \\ &= \log p_z(z_K) + \sum_{k=1}^{K} \log\left|\det J_{\phi_k^{-1}}(z_k)\right|, \end{aligned} \tag{21}$$

where $z_k = \phi_k(z_{k-1})$ for $k \in \{1, \ldots, K\}$ and $z_K = T^{-1}(x_{\text{reco}}; x_{\text{hard}})$. Evaluating the probability density at any point in the target distribution requires efficient computation of the Jacobian determinant in each transformation step.

The cINN Transfer network is then trained on pairs of events (x_{reco}, x_{hard}) using Maximum Likelihood Estimation (MLE),

$$\begin{aligned} \mathcal{L}_{\text{NF}}(\theta) &= -\prod_i^N p_\theta(x_{\text{reco}}^{(i)} \mid x_{\text{hard}}^{(i)}) \\ \log \mathcal{L}_{\text{NF}}(\theta) &= -\sum_i^N \log p_\theta(x_{\text{reco}}^{(i)} \mid x_{\text{hard}}^{(i)}). \end{aligned} \tag{22}$$

The training objective is therefore to minimise the negative log-likelihood $\mathcal{L}_{\text{NF}}(\theta)$ of the

data under the model. This is achieved by optimising the model parameters θ

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}_{\text{NF}}(\theta). \tag{23}$$

7 Conditional Flow Matching

The Transfer-cINN [21] discussed in Sec. 6 can be replaced by a different type of neural network to encode the conditional transfer probability $p(x_{\text{reco}} \mid x_{\text{hard}})$ from Fig. 7. While cINNs excel in their stability and speed, they require strict bijectivity and tractable Jacobian determinants, which can hinder their flexibility.

Conditional Flow Matching (CFM) models are a new and exciting phenomenon, first appearing in publications in 2023 [41–43]. Their superior expressivity and more powerful transport map have since led to their proliferation in particle physics studies at the LHC [44]. In some ways, CFM networks are similar to Denoising Diffusion Probabilistic Models (DDPMs) in the way that they transform between data and noise with a time-evolving scheduler [45]. However, while DDPMs rely on a stochastic diffusion process that must be reversed via score-matching, CFM models follow a continuous ordinary differential equation (ODE) to parametrise sample evolution

$$\frac{dx(t)}{dt} = v(x(t), t) \quad \text{with} \quad x(t=0) = x_0, \quad x(t=1) = x_1. \tag{24}$$

Here, $v(x(t), t)$ defines a time-dependent velocity field, $x_0 \sim \mathcal{N}(0, I)$ is a sample drawn from a Gaussian prior, and x_1 represents a sample from the target distribution we are modelling. Consequently, $v(x(t), t)$ can be thought of as a vector field that transports prior samples x_0 into their targets x_1.

Fig. 8 visualises the CFM generative process for a single b-jet. Initially, the samples follow a Gaussian prior, but as time progresses, the velocity field from Eq.(24) continuously warps their distribution towards the target. These trajectories are graphically plotted in the bottom row.

This transport process governs a corresponding probability density path $p(x, t)$ which obeys the continuity equation

$$\frac{\partial p(x,t)}{\partial t} + \nabla_x \left[p(x,t) v(x,t) \right] = 0. \tag{25}$$

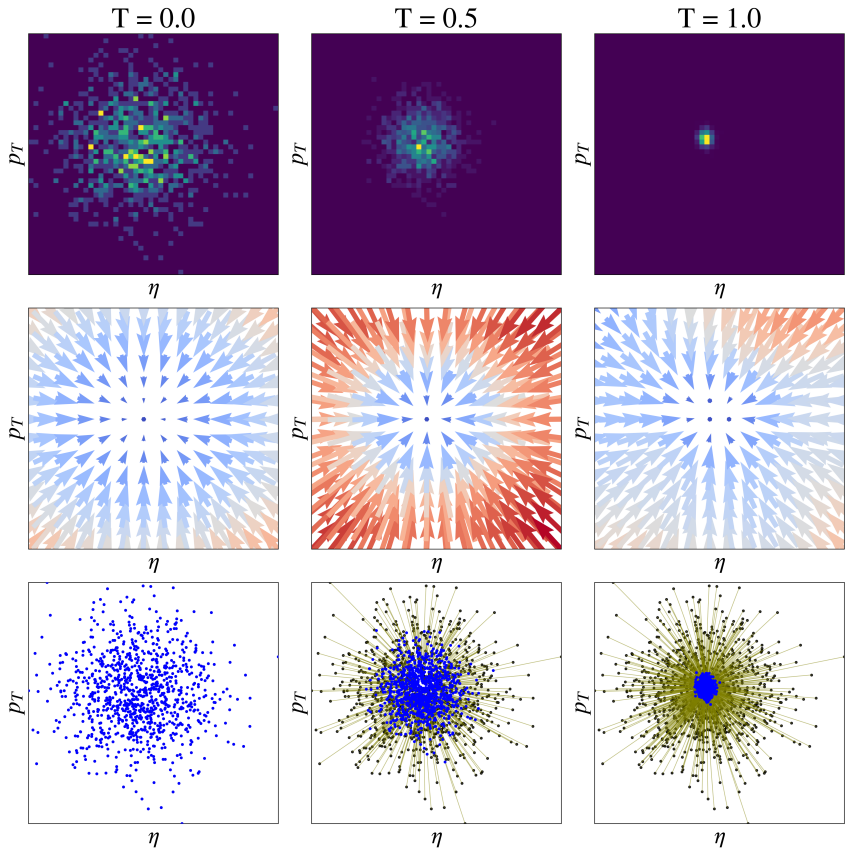

Figure 8: Time evolution of 500 samples from an initial Gaussian prior (T = 0) to the true target distribution (T = 1) using conditional flow matching. Shows the evolving sample density (top), learned velocity field (middle), and the individual trajectories (bottom). Data corresponds to a single b-jet from a $t\bar{t}H$ decay, with feature values (omitted) represented in the model's preprocessed space.

The velocity field from Eq.(24) evolves $p(x,t)$ and so it satisfies the boundary conditions

$$p(x,t) = \begin{cases} p_{\text{prior}}(x) = \mathcal{N}(x;0,I) & t = 0 \\ p_\theta(x) \approx p_{\text{data}}(x) & t = 1, \end{cases} \tag{26}$$

where the model is tasked with learning p_θ, an approximation of the unknown target distribution p_{data}.

A wide range of interpolation strategies can be employed to define the path between x_0 and x_1. A linear function is adequate in describing an effective deterministic transport equation [46, 47]

$$x(t \mid x_0, x_1) = (1-t)x_0 + t x_1 \rightarrow \begin{cases} x_0 & t \rightarrow 0 \\ x_1 & t \rightarrow 1. \end{cases} \tag{27}$$

This guarantees that samples are initialised from the Gaussian prior $x_0 \sim \mathcal{N}(0, I)$ at $t = 0$, and smoothly approach the target $x_1 \approx p_{\text{data}}(x)$ as $t \to 1$.

Currently, Eq.(27) is fully deterministic. While this provides a clear transport path, it can limit the model's ability to generalise effectively. Introducing a level of stochasticity can make the model more robust to variations in the data and reduce overfitting. Additionally, if the model even slightly deviates from the deterministic path, it will likely get stuck due to a lack of information on how to recover. By perturbing the trajectory with noise, the model will encounter a more diverse range of possible paths, improving stability and letting it learn how to recover better during training. This can be achieved by modifying Eq.(27) to incorporate a degree of randomness

$$x(t \mid x_0, x_1) = (1-t)x_0 + tx_1 + \sigma\epsilon, \tag{28}$$

where $\epsilon \sim \mathcal{N}(0, I)$ injects some random noise and σ is a scalar that controls the amount [48]. To obtain the associated velocity field that follows the ODE from Eq.(24), we can differentiate Eq.(28)

$$\begin{aligned} v(x(t \mid x_0, x_1), t \mid x_0, x_1) &= \frac{d}{dt}[(1-t)x_0 + tx_1 + \sigma\epsilon] \\ &= x_1 - x_0. \end{aligned} \tag{29}$$

This result indicates that for the linear trajectory defined in Eq.(28), the velocity field is simply the difference between the target and prior samples. In contrast, the generative model's velocity field is given by [41]

$$\begin{aligned} v(x,t) &= \int dx_0 dx_1 v(x, t \mid x_0, x_1) p(x, t \mid x_0, x_1) p_{\text{data}}(x_1) p_{\text{prior}}(x_0) \\ &= \int dx_0 dx_1 (x_1 - x_0) p(x, t \mid x_0, x_1) p_{\text{data}}(x_1) p_{\text{prior}}(x_0). \end{aligned} \tag{30}$$

This represents the expected velocity predicted by the model at any point (x, t), marginalising over all possible pairs of (x_0, x_1).

The network is trained by minimising the discrepancy between the predicted velocity $v_\theta(x, t)$ and the true velocity $v(x, t \mid x_0, x_1)$, leading to the loss function

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, x_0, x_1} \left\| v_\theta(x, t) - v(x, t \mid x_0, x_1) \right\|^2. \tag{31}$$

These expectations are taken over $t \sim \mathcal{U}[0, 1]$, ensuring the field is well defined across all time values, as well as over the initial sample $x_0 \sim p_{\text{prior}}$, and target sample $x_1 \sim p_{\text{data}}$. Once the model is trained, new samples can be generated by integrating the ODE in

Eq.(24) between $t = 0 \to 1$ using the learned velocity field.

In practice, the Transfer-CFM is trying to learn the distribution of $p(x_{\text{reco}} \mid x_{\text{hard}})$. Using the fact that x_1 represents the reco-level momenta to be generated and x_{hard} is just some extra information given to the model to help it evolve x_0, the time evolution of the log-likelihood can be computed using the change of variables formula [49]

$$\frac{d \log p(x, t \mid x_{\text{hard}})}{dt} = -\nabla_x \cdot v_\theta(x, t \mid x_{\text{hard}}). \tag{32}$$

Integrating this expression from $t = 0 \to t = 1$ gives the final conditional likelihood of the generated reco-level sample

$$p(x_{\text{reco}} \mid x_{\text{hard}}) = p_{\text{prior}}(x_0) \exp\left(- \int_0^1 \nabla_x \cdot v_\theta(x, t \mid x_{\text{hard}}) dt \right). \tag{33}$$

Unfortunately, to gain the final likelihood result, Eq.(33) must be evaluated $\mathcal{O}(100)$ times for all components of the velocity with respect to the inputs. This repeated evaluation of both the divergence and the ODE makes it extremely computationally expensive compared to Eq.(21) for the cINN. The hyperparameters of the Transfer-CFM network are highlighted in Tab. 2.

8 Jet Combinatorics

In order for the model architectures in Secs. 6 & 7 to accurately model the target distribution $p(x_{\text{reco}} \mid x_{\text{hard}})$, they must assimilate contextual information from the underlying hard-scattering event. This is imperative for guiding the generative process. Jet combinatorics is a formidable challenge intrinsic to the Transfer network [50], wherein reconstructed jets must be correctly assigned to their corresponding partonic progenitors.

Transformers provide a powerful means to capture complex interdependencies among elements in a set. Popularised by their aptitude in natural language processing (NLP) [51], they have demonstrated remarkable promise in high-energy physics applications [44, 52, 53] such as jet-flavour tagging and tracking [54, 55].

Unlike conventional neural networks which impose a fixed input order, Transformers natively accommodate permutation-invariance. This correctly treats jets as an unordered set rather than a sequence. Self-attention mechanisms ensure that the model can attend between all particles simultaneously, removing any impact of particle ordering during the encoder stage. These learned embeddings serve as contextual inputs in the cINN or

CFM models when transforming samples into reco-level momenta.

8.1 Variable Multiplicity Transfermer

A Transformer cannot directly supplant the cINN architecture introduced in Sec. 6 due to its inherent lack of invertibility and inability to obtain a tractable Jacobian determinant [56–58]. These limitations arise from how self-attention mechanisms induce highly non-linear dependencies across the input sequence. Moreover, the Transformer's many-to-one mapping breaks invertibility as identical inputs may propagate into disparate outputs.

By making the reco-level sampling autoregressive [22] and appending a lightweight cINN to the Transformer output, invertibility can be restored. The cINN is conditioned on the Transformer output and maps these embeddings to physically valid, continuous reco-level momenta values. The use of Transformer in the Transfer network will be referred to as the Transfermer [22], and represents the Transformer + cINN framework in Fig. 9.

To make the sampling autoregressive, the transfer probability can be factored into

$$p(x_{\text{reco}} \mid x_{\text{hard}}) = \prod_{i=1}^{n} p(x_{\text{reco}}^{(i)} \mid c(e_{\text{reco}}^{(0)}, \ldots, e_{\text{reco}}^{(i-1)}, e_{\text{hard}})), \tag{34}$$

where c denotes the Transformer conditioning and e_{hard} is the global-embedding for the whole hard-scattering event, obtained from the Transformer encoder. $e_{\text{reco}}^{(i)}$ corresponds to the particle-wise embedding for the reconstructed momenta, iteratively refined in the Transformer decoder. $e_{\text{reco}}^{(0)}$ is a special starting token needed to initialise the autoregressive sequence. As each output is generated, the decoder input sequence is shifted along by one position, and a triangular mask is applied to the self-attention matrix to enforce strict causality. This ensures that each reco-level particle is conditioned exclusively on previously generated ones. Once trained, samples are drawn in a sequential fashion via

$$p(x_{\text{reco}}^{(i)} \mid x_{\text{hard}}) = p(x_{\text{reco}}^{(i)} \mid c(e_{\text{reco}}^{(0)}, \ldots, e_{\text{reco}}^{(i-1)}, e_{\text{hard}})). \tag{35}$$

This autoregressive decomposition allows the conditioning vector $c^{(i)}$ to be computed only once per particle, reducing training and likelihood evaluation overhead.

After obtaining the Transformer conditioning, x_{reco} must be parametrised into a valid 4-momenta, accommodating both massive and massless particles. This can be procured by decomposing it into a product of distributions

$$p(x_{\text{reco}}^{(i)} \mid c^{(i)}) = p(p_T^{(i)}, \eta^{(i)}, \phi^{(i)} \mid c^{(i)}) \times p(m^{(i)} \mid p_T^{(i)}, \eta^{(i)}, \phi^{(i)}, c^{(i)}). \tag{36}$$

This formulation ensures the mass term remains disentangled from the generation of the other fields. The corresponding cINN architecture is illustrated in the right-panel of Fig. 9.

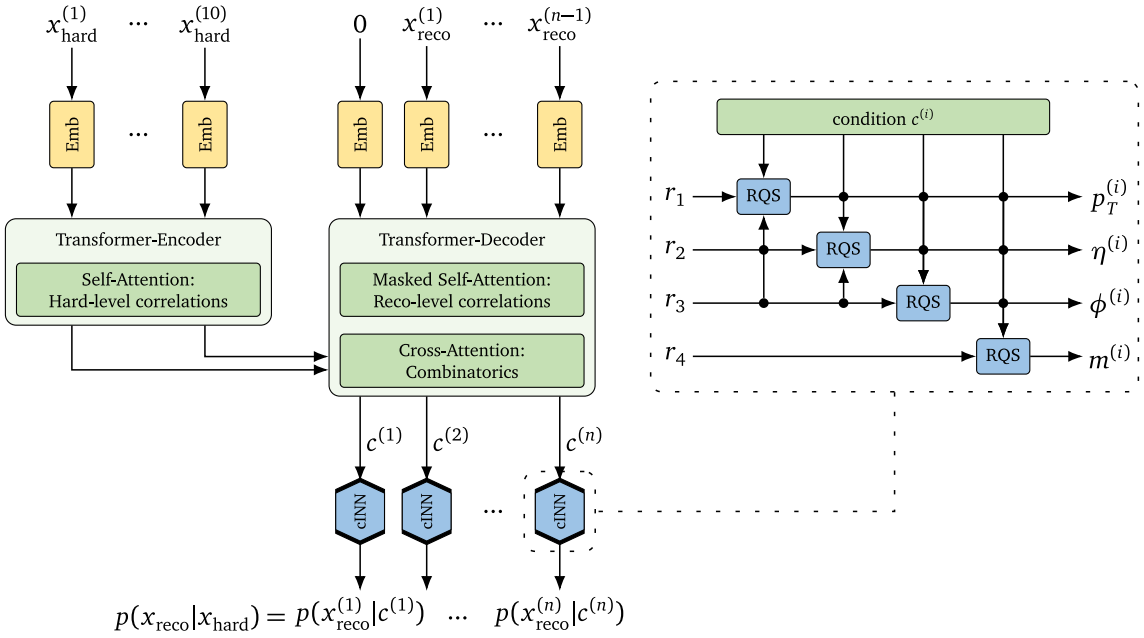

Figure 9: Transformer with cINN architecture used for density estimation (left). cINN setup with autoregressive feature sampling (right) [22]. r denotes the latent variables transformed by the normalizing flow. $x_{\text{hard}}^{(1:10)}$ represents the 10 final state particles produced in Fig. 2 where H decays invisibly.

The Transfermer can be extended to handle variable-jet final states using the multiplicity network described in Sec. 5.3. To achieve this, Eq.(18) can be modified to evaluate the transfer probability autoregressively in the same way as Eq.(34),

$$p(x_{\text{reco}}, n \mid x_{\text{hard}}) = p(n \mid x_{\text{hard}}) p(x_{\text{reco}}^{(1:n_{\text{min}})} \mid x_{\text{hard}}, n) \prod_{i=n_{\text{min}}+1}^{n} p(x_{\text{reco}}^{(i)} \mid x_{\text{reco}}^{(1:i-1)}, x_{\text{hard}}, n). \quad (37)$$

Here, n is the number of final state particles, $x_{\text{reco}}^{(1:k)}$ denotes the first k reco-level momenta and n_{min} is the minimum number of reconstructed particles in an accepted event. The term $p(n \mid x_{\text{hard}})$, which encodes the probability of an event containing n final state particles, is explicitly learned by the multiplicity classifier in Sec. 5.3.

To ensure consistency between the vastly different data ranges of kinematic observables, the input features underwent specialised transformations. Transverse momentum and mass were log-transformed ($\log p_T$, $\log m$) and standardised to zero mean and unit variance. Similarly, pseudorapidity (η) and azimuthal angle (ϕ) were rescaled into uniform latent spaces, respecting signal-region (SR) cuts. ϕ required special handling for its periodicity and therefore leveraged periodic RQS splines [59].

The Transfermer was trained by minimising the negative log-likelihood from Eq.(22) and used the standard Transformer module in PyTorch [60]. The hyperparameters for the Transfermer setup are detailed in Tab. 1.

8.2 Permutation-Invariant Transfusion

The Transfer-CFM discussed in Sec. 7 currently lacks the ability to handle events with varying multiplicity. It harnesses a Transformer encoder to generate context from the hard-level processes, but is devoid of the decoder setup used by the Transfermer in Sec. 8.1. The Transfermer combatted this issue by autoregressively generating its output sequence, allowing it to dynamically determine the cardinality.

The cINN in the Transfermer setup from Sec. 8.1 could be supplanted with a lightweight CFM completely analogously. However, unlike cINNs, CFM models do not require invertibility, which engenders an alternate method of generating reco-level momenta. The autoregressive methodology can be dispensed for parallel generation of all x_{reco}. This is feasible because the continuous velocity field in Eq.(30) is inherently multi-dimensional and capable of evolving all particle 4-momenta concurrently.

At every step t, the Transformer decoder now ingests all reco-level states t $x_{\text{reco}}^{(1,\dots,n)}(t \mid x_0)$. These are each defined by Eq.(28) across all n particles. The encoder still generates the hard-level correlations as before. Consequently, the Transformer now outputs one temporally-conditioned embedding per particle, all of which are passed into the CFM network. The particle-wise (i^{th}) component of the velocity field is then

$$v^{(i)}(c(e_{\text{reco}}(t), e_{\text{hard}}, t), t), \tag{38}$$

where the contextual embedding c is derived from the current state representations $e_{\text{reco}}(t)$ and the hard-level process e_{hard}. The time t is also explicitly fed into the decoder as an additional input. Note, because the samples are evolved in parallel without autoregression, this model is permutation-invariant. Thus, the ordering of particles will not affect the 4-momenta generation, and the causal (triangular) mask in the decoder is no longer needed. The combination of this permutation-invariant Transformer and CFM network will be referred to as Parallel Transfusion [22], whose architecture is shown in Fig. 10.

The primary advantage of CFM networks over their cINN counterparts lies in their superior expressivity from their more powerful continuous transport map. This motivates the Parallel Transfusion paradigm over a simple substitution of the Transfermer's

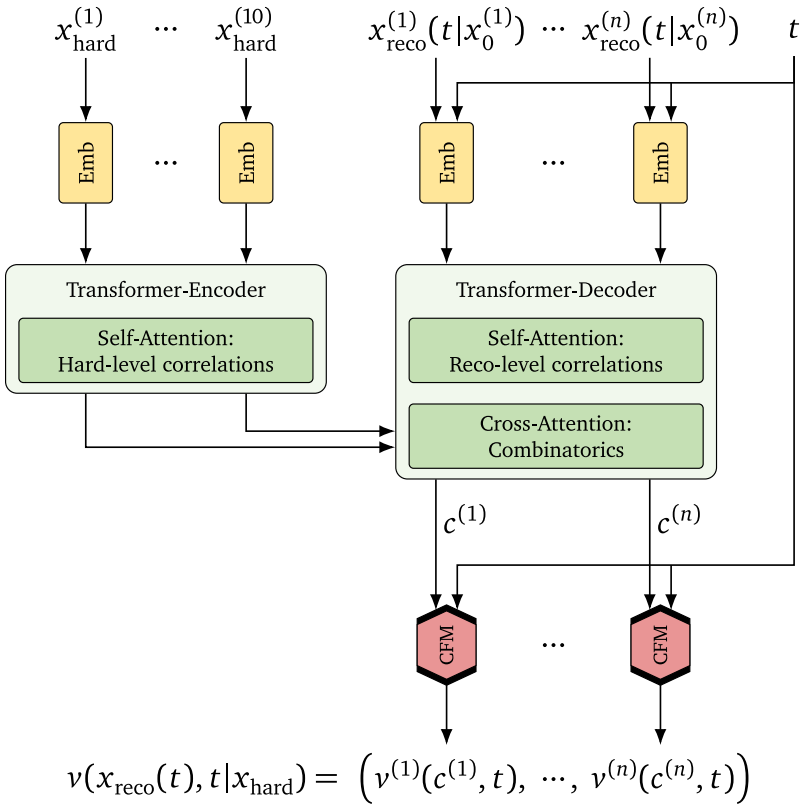

Figure 10: Architecture of the permutation-invariant Parallel Transfusion model used to evolve the latent space vector r into x_{reco} [22]. The Transformer is conditioned on time dependent states instead being autoregressive like in the Transfermer.

cINN with a lightweight CFM. By harnessing a high-dimensional vector field to evolve all features jointly, interdependencies between particles and kinematic variables are better preserved. In contrast, an autoregressive CFM network would necessitate a strictly one-dimensional velocity field, significantly limiting its capacity to capture correlations between features during the evolution process. Allowing a single velocity field with increased dimensions, encourages a more holistic representation of particle interactions and properties in line with Fig. 10. This ultimately increases the efficacy of the Parallel Transfusion architecture in modelling more complex densities.

The hyperparameters associated with the Parallel Transfusion architecture are presented in Tab. 1.

9 Generative Network Results

This analysis uses the CMS 2018 Ultra Legacy dataset based on the GEANT4 toolkit [61–63] with events generated at next-to-leading-order (NLO) precision. A cut was applied to

only consider events in the signal region (SR) [64]. This imposed a lower p_T threshold of 200 GeV for missing transverse energy (MET) and 30 GeV for jets to suppress most of the background. The forward hadronic calorimeter (HF) extended the total pseudorapidity coverage to $|\eta| < 5.0$ [65]. Only events with at least five total jets and one b-tagged jet were considered. Super-rare events where a top quark decayed to anything other than a W boson and a b quark were also ignored. The first two jets were ordered by b-tag score, leaving the remainder to be sorted by p_T, after which the first six were selected. This meant each event contained either five or six jets.

Three main generative models are compared in this section. To be clear, Transfermer refers to the autoregressive Transformer and cINN architecture in Fig. 9, Parallel Transfusion is the permutation-invariant Transformer and CFM network in Fig. 10, and Transfer-CFM denotes the encoder only CFM setup in Sec. 7. The model hyperparameters used to obtain all results are outlined in Appendix A.

While this study builds upon the proof-of-concept architecture presented in Ref. [22], a direct comparison of results may initially suggest a degradation in performance. However, this can be attributed to several key factors. Firstly, $H \to \gamma\gamma$ (the target process in Ref. [22]) is much easier to reconstruct than $H \to$ inv, as photons are detectable in the ECAL, unlike neutrinos which have no direct signature. Secondly, the reference paper employed DELPHES [66, 67] for a much simpler detector simulation compared to the full CMS simulation used in this analysis. Although full simulation leads to much more realistic detector inefficiencies and responses, it complicates the jet and MET reconstruction and precludes the Transfer network's task. Lastly, the reference study relies on LO precision for MEM calculations, whereas this study uses NLO corrections. These higher-order contributions introduce additional jets and partonic activity that further complicate event reconstruction and selection. It is worth noting that the CFM theory and equations are also more recent and align with the latest literature [46].

9.1 Kinematic Observables

The easiest way to compare the fidelity of the generative models is to visualise the distributions of individual components of the 4-momenta. Fig. 11 depicts the reconstructed distributions for all jets across all events. Immediately, it is clear that the Parallel Transfusion network exhibits superior expressivity in capturing the intricate densities of the reconstructed jet kinematics.

All three networks display accurate estimations of the jet p_T distribution. The errors increase as expected at higher transverse momentum values, due to low statistics. As

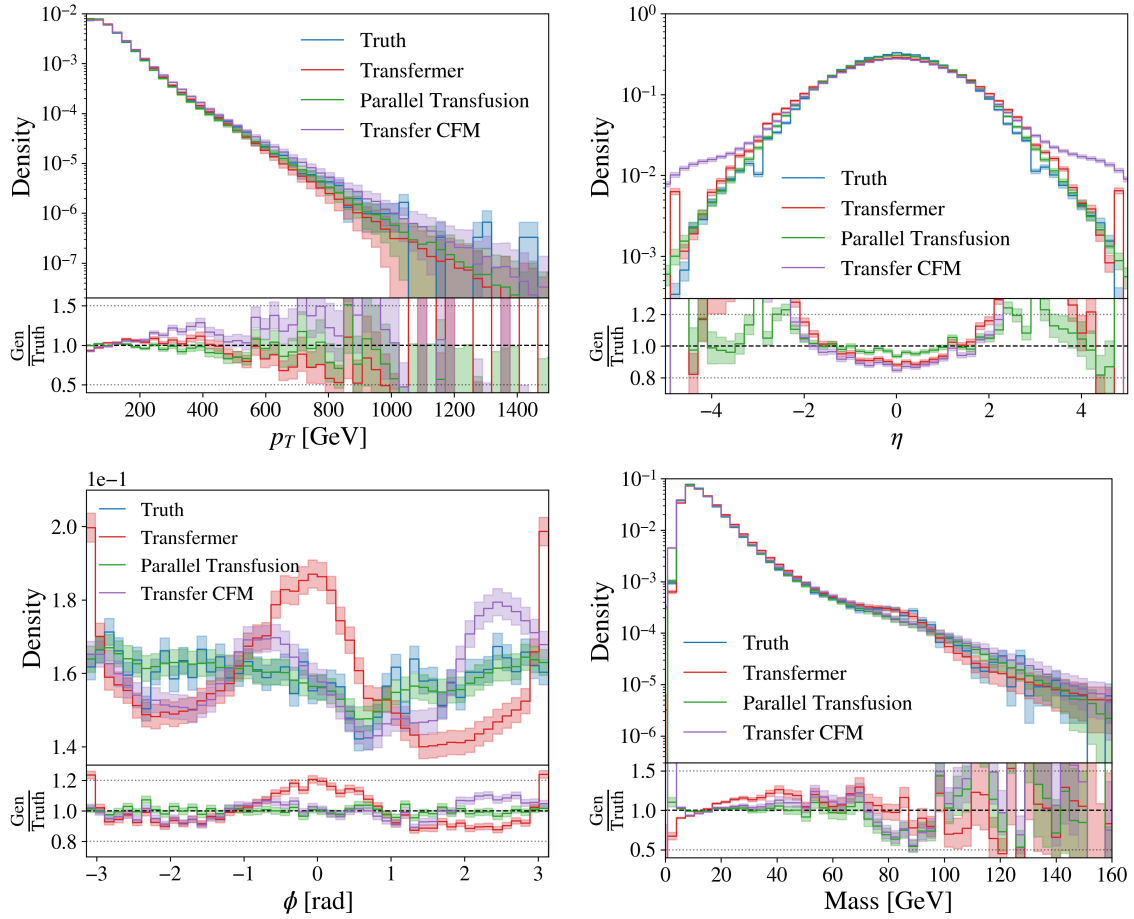

Figure 11: Reco-level distributions of 4-momentum components for all jets. Compares Transfer-CFM conditioned on hard-level momenta, autoregressive Transfermer and permutation-invariant Parallel Transfusion models. Truth (blue) corresponds to 18,354 events from the validation dataset. Generative models sampled 100 times per event and are normalised. Includes Poisson errors (shaded).

predicted, the Transfer-CFM network underperforms relative to the two Transformer-based models due to the absence of a Transformer decoder. This leaves it unable to capture reco-level correlations, unlike the other two models which can. Conversely, the Transfermer and Parallel Transfusion models exhibit comparable performance, with the latter benefiting from the extra power of its CFM transport map, engendering marginally better results.

The Transfer-CFM network struggles at extreme $|\eta|$ values in the forward regions of the detector [68]. The selection bias of requiring a b-tagged jet favours the $t\bar{t}H$ signal and enhances jet reconstruction in the barrel region. This is because b-tagging requires the CMS silicon tracker, which only extends to $|\eta| \leq 2.5$ [69]. On the other hand, processes such as QCD radiation and proton splitting do not discriminate between η values. This increases the proportion of background jets in the far-forward regions [70–72], thereby

making it harder to disentangle the primary vertex (PV) from pileup.

More importantly, the ECAL and HCAL calorimeters suffer from limited granularity in the end-cap regions, compounding reconstruction difficulties. At large $|\eta|$, the Tracker contains fewer layers [73] and track multiplicity per unit area increases, degrading jet-parton assignment. Overall, this combination of lower jet resolution at high $|\eta|$ values and the lack of cross-attention functionality explains the shortcomings of the Transfer-CFM network in estimating the η density in Fig. 11. The Parallel Transfusion and Transfermer models showcase more robustness to these detector effects.

The conspicuous artefact in the Transfermer's ϕ predictions is not entirely understood and remains an active area of research for ETH Zurich and the University of Bristol. Particle ordering matters for autoregressive models, which likely explains why the Transfermer is particularly impacted. One hypothesis is that the Transfermer becomes confused by the cylindrical symmetry of the detector (properties should not change under a rotation about the beam-axis). Again, the Parallel Transfusion model upholds a high level of accuracy across all ϕ values, consistently staying within the Poisson uncertainty of the truth data. The largest observed deviation relative to the truth was merely 7.5%.

The jet mass distribution is also visible in Fig. 11. This analysis used $\Delta R < 0.4$ for jet-clustering which suppresses boosted W-jet signatures. Nonetheless the 'bump' around 80 GeV in the mass distribution infers that the constituents from W decays are still occasionally clustered into a single jet [74, 75]. However, the majority of jets have a much lower mass, with the broad peak around 10-20 GeV conforming to the mass distributions of light-flavoured jets from unboosted W decays. The wide tail between these two ranges reflects the expected mass distribution of b-jets arising from top decays [76]. While highly boosted top decays comprised of all three jets are even rarer with the current choice of ΔR, they are still visible in the dataset through a notable peak at ~170 GeV (although not plotted due to axis cut and low statistics in validation dataset).

Interestingly, the Transfermer outperforms Parallel Transfusion in predicting higher jet masses. This is a function of the autoregressive sampling paradigm in Fig. 9, wherein mass is computed last. Mass is often highly correlated with p_T, η, and ϕ, hence the Transfermer can condition on all these features to perform a well-informed estimation of each jet mass. In contrast, the Parallel Transfusion model evolves all features simultaneously. While this can help preserve correlations during interpolation, the mass ends up being transformed without the full final-state information available to the Transfermer.

Similar density estimation plots for the MET are displayed in Fig. 12. The Parallel Transfusion model yields near-perfect results across low and high p_T domains. As antici-

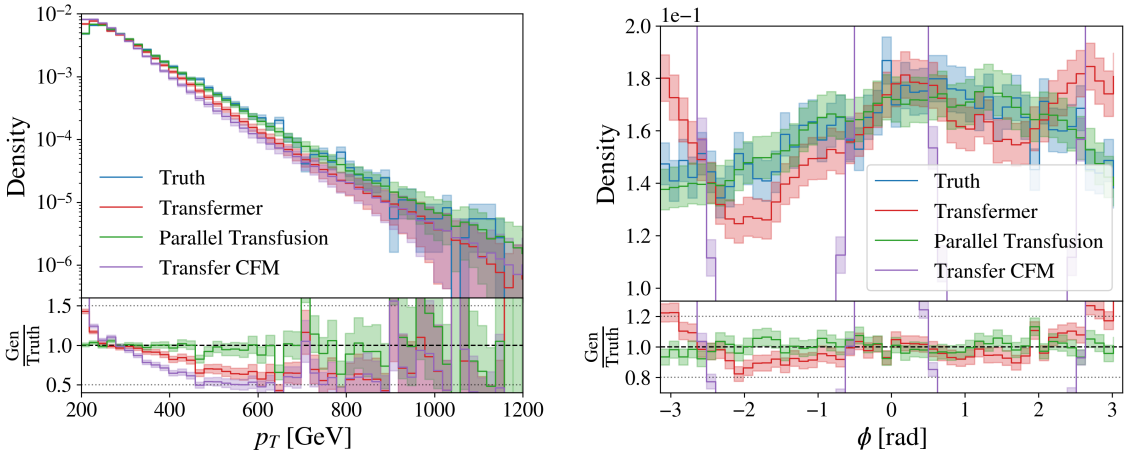

Figure 12: Reco-level distributions of MET transverse momentum (left) and azimuthal angle (right). Compares Transfer-CFM conditioned on hard-level momenta, autoregressive Transformer and permutation-invariant Parallel Transfusion. Truth (blue) corresponds to 18,354 events from the validation dataset. Generative models sampled 100 times per event and are normalised. Includes Poisson errors (shaded).

pated, this is followed by the Transformer and Transfer-CFM models for the same reasons as before. At $p_T > 600$ GeV, the Transfer-CFM seems to perform almost on par with the Transformer.

A lower threshold of 200 GeV was imposed on MET p_T during preprocessing. Any underestimations below this cut are re-projected into the 200 GeV bin, explaining the artificial surplus of predictions here for the less powerful Transformer and Transfer-CFM models.

For the MET ϕ density, the Parallel Transfusion model once again achieves remarkable agreement with the truth, consistently remaining within the Poisson uncertainty bands. While the Transformer yields improved predictions for the MET ϕ compared to jets, the anomalous artefact observed in Fig. 11 persists, albeit less profound.

The energy of the jet with the highest b-tag score (j_1), is calculated using both its mass and momentum. As previously seen in Fig. 11, it was clear that at high masses, the Transformer model performed the best because of how it autoregressively generated this feature last. This complies with the E_{j_1} distribution in Fig. 13 where above 500 GeV, the best performing model is the Transformer. Referring back to Fig. 11, the simpler Transfer-CFM network tended to overestimate p_T and mass at higher values, explaining why it deviates from the true E_{j_1} at around 400 GeV.

A similar trend is perceived in the H_T distribution at the bottom of Fig. 13, from the same systematic p_T overestimation in the Transfer-CFM network. Deviations from the

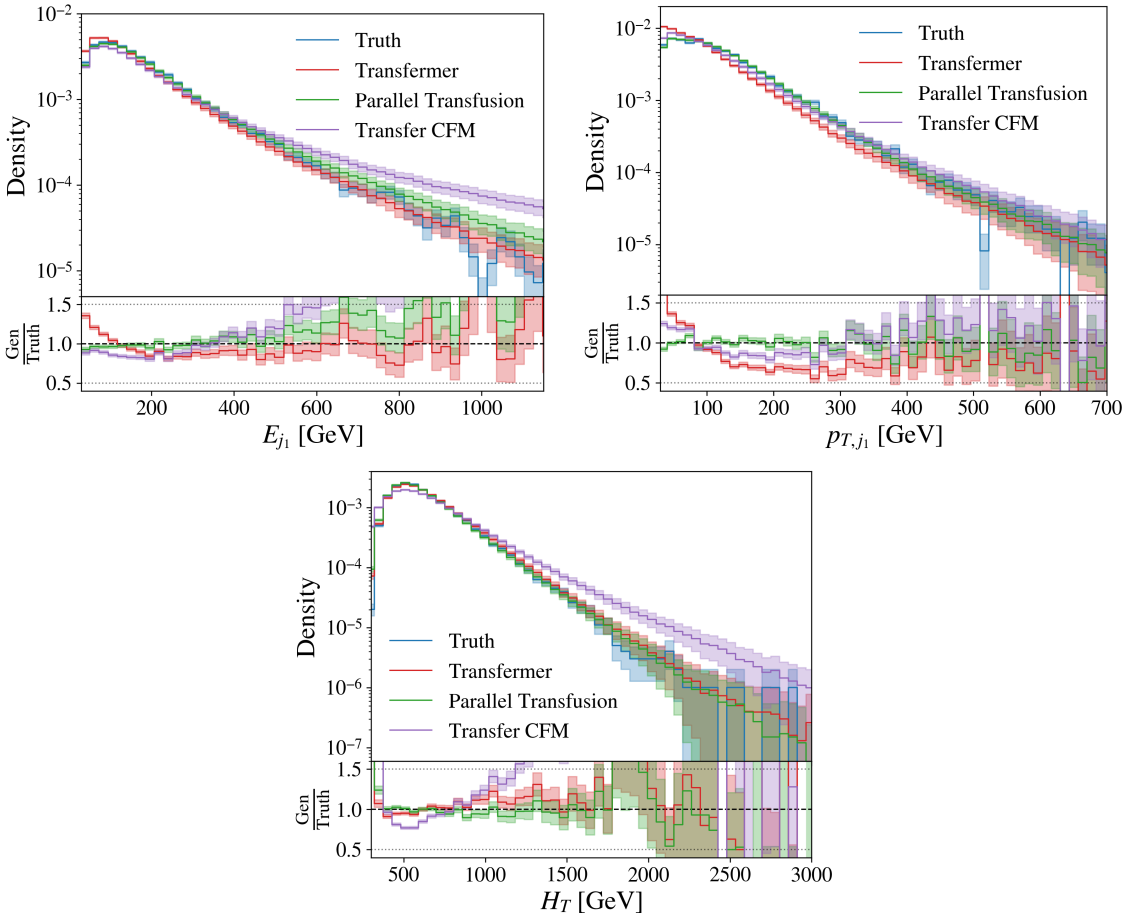

Figure 13: Reco-level distributions of jet energy (top-left), transverse momenta (top-right), and the total scalar sum of jet-p_T in each event (bottom). j_1 denotes the jet with the leading b-tag score. Compares Transfer-CFM, Transfermer and Parallel Transfusion generative models. Truth (blue) corresponds to 18,354 events from the validation dataset. Each model was sampled 100 times per event and normalised accordingly. Includes Poisson errors (shaded).

truth at $H_T > 1500$ GeV can be attributed to a lack of breadth in the truth data.

Intriguingly, while the Parallel Transfusion model delivers a near-optimal estimation of p_{T,j_1} density as expected, the ostensibly less sophisticated Transfer-CFM network actually outperforms the Transfermer on average. The conjecture for this counter-intuitive result is that while the Transfermer can predict the complete event topology (as demonstrated in Fig. 11), it struggles to discern the two leading b-jets. Consequently, it appears inferior when restricting the distribution to only j_1. Event-level sampling plots visualise this phenomenon in more detail in Sec. 9.2.

To rigorously vindicate the fidelity of the generative models beyond individual jet and MET features, high-level kinematic observables derived from the quantities in Figs. 11 & 12 are examined in Fig. 14. These observables encapsulate more intricate correlations

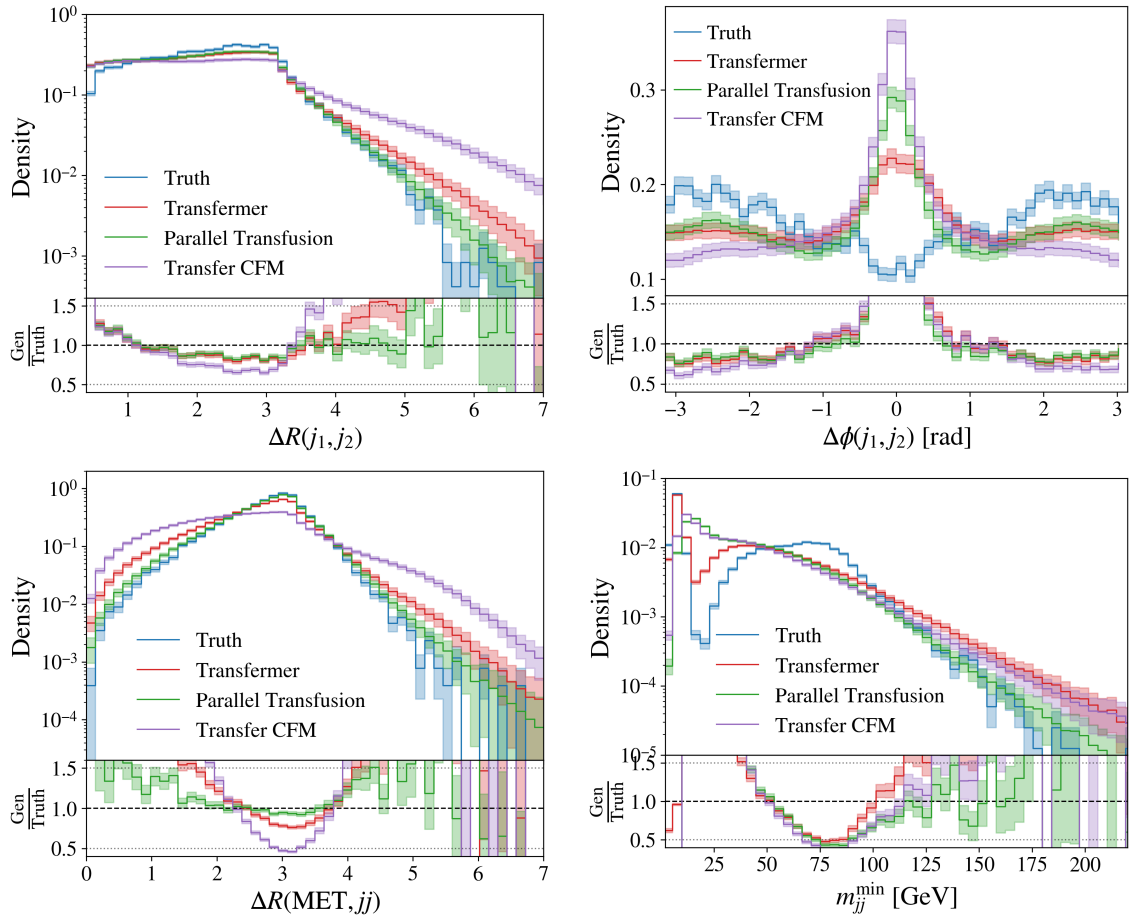

Figure 14: High-level variables derived from the reco-level jet and MET 4-momenta. j_1 and j_2 denote the two b-jets (ordered by b-tag score). Compares Transfer-CFM, Transfermer and Parallel Transfusion generative networks. Truth (blue) corresponds to 18,354 events from the validation dataset. Each model was sampled 100 times per event and normalised accordingly. Includes Poisson uncertainty bands (shaded).

and serve as a stringent test of each model's ability to replicate complex event topologies.

j_1 and j_2 denote the jets with the highest b-tag score in each event. In the overwhelming majority of cases, these will be the b-jets directly produced by the top quarks in Fig. 2. However, sometimes true b-jets might be missed by the tagging algorithm. Likewise, some light-jets from the W decays might be erroneously classified as false b-jets. This analysis employs the DeepJet algorithm for jet-flavour tagging with a 'medium' working point (WP) corresponding to a threshold of 1% in the discriminator distribution [77, 78]. This means that 1% of light jets are misidentified as a b-jet and corresponds to a b-jet efficiency of ∼80%. The full distribution of b-jet efficiencies using DeepJet can be seen in Ref. [77].

Linking this to the results, the identity of j_1 and j_2 may not always be correct. Although this realism is desirable for a more representative dataset, it introduces additional noise

from tagging inefficiencies and ultimately makes the model harder to train. This could be an underlying cause for the Transfer-CFM network's inability to accurately match the $\Delta R(j_1, j_2)$ distribution. To mitigate this issue, there are several improvements that could be made. Firstly, angular constraints could be introduced to improve the jet-parton matching by identifying the two jets from each W as the pair closest in ΔR. This would require careful implementation to handle cases where the b-quark from the top is highly collinear with a quark from the W boson decay. Secondly, some more sophisticated analyses use a Transformer architecture known as SPA-NET to perform the jet-parton assignment [79, 80]. Adopting these improvements could enhance the quality of the truth dataset and subsequently the performance of the generative models.

The azimuthal angle difference, $\Delta\phi(j_1, j_2)$, relates the angular separation of the two leading b-jets. Upon inspecting this observable's distribution in Fig. 14, it is obvious that all three models exhibit a spurious enhancement at $\Delta\phi(j_1, j_2) = 0$, absent in the truth data. What seems to happen is that the models unintentionally learn a joint distribution between the two b-jets. When repeatedly sampling j_1, this causes two solutions to emerge: one correctly centred on the true ϕ_{j_1} value, with the other centred on the ϕ corresponding to j_2 (and vice versa when sampling ϕ_{j_2}). As a result, one of three things transpires for any given event: (i) ϕ_{j_1} and ϕ_{j_2} could both be correctly sampled. This would lead to an accurate estimate of $\Delta\phi(j_1, j_2)$. (ii) ϕ_{j_1} is incorrectly sampled and ϕ_{j_2} is correctly sampled. This is effectively the same as sampling ϕ_{j_2} twice. Therefore, calculating their $\Delta\phi$, gives ~0 (as this is just ϕ_{j_2} - ϕ_{j_2}). The inverse case of correctly sampling ϕ_{j_1} and incorrectly sampling ϕ_{j_2} leads to the same result. (iii) ϕ_{j_1} and ϕ_{j_2} are both incorrectly sampled. This would lead to the correct $\Delta\phi(j_1, j_2)$ measurement but with a sign error.

In essence, the models struggle to disentangle the ϕ distributions of the jets for the reasons mentioned in case (ii), causing the central spike in Fig. 14. It also explains why this error remained obscured during earlier plots of the combined distributions in Fig. 11. This degeneracy can be seen on a per-event basis in more detail in Sec. 9.2.

The $\Delta R(\text{MET}, jj)$ distribution, which measures the spatial separation between the missing transverse energy (MET) and the dijet system (jj), is well matched by the Parallel Transfusion model across the entire range. This is followed by the Transfermer and then the Transfer-CFM. The latter exhibits disparities consistent with its suboptimal ϕ and η predictions seen earlier in Figs. 11 & 12.

The variable m_{jj}^{min} defines the minimum invariant mass among all possible jet pairs within an event. This metric is particularly relevant in hadronic final states, where specific mass peaks can correspond to certain resonances, such as the W boson. The main peak drop-off at ~ 80 GeV in the truth data coincides with the mass of the W boson [75],

indicating that jet pairs originating from W decays, often reconstruct close to the W mass. The secondary peak, around 10 GeV is more complex to interpret. It is possible that low-energy jets originating from final-state radiation (FSR), or pileup are the likely culprits. FSR occurs when a high-energy parton radiates a softer quark or gluon in the final state. Pileup jets, on the other hand, originate from additional proton-proton collisions in the same event.

The invariant mass is contingent on both the spatial proximities and transverse momenta of the constituent particles. This lets even high-energy particles have a low invariant mass if they are sufficiently close. FSR jets tend to be highly collinear with the original jet, which naturally leads to small m_{jj} values, consistent with Fig. 14. These effects are challenging for the models to learn, making m_{jj}^{min} an exigent test of model fidelity. ML models often perform well in the most common scenarios, so shifting focus to these difficult parameters can aid with assessing their veracity.

Clearly, none of the models can coherently predict this variable. FSR jets are not included in the hard-scattering matrix element as they occur at the parton shower stage. Thus, they cannot be included in the x_{hard} conditioning, which is a vital asset in the generative process for all three models. For this reason, the models struggle to capture the m_{jj}^{min} distribution.

It is worth noting that initial state radiation (ISR) was omitted from the hard-scattering event data and therefore does not contribute to the results. Similarly, the overwhelming majority of soft QCD jets are likely filtered out by the $p_T > 20$ GeV threshold cut, also negating their impact.

9.2 Event-Level Sampling

To scrutinize the generative models in more detail, it is beneficial to look at the sampling performance on a per-event basis. Fig. 15 visualises the feature-space sampling performance for a single b-jet across all three architectures, offering a unique perspective on their consistency.

The order of performance concurs with the conclusions drawn from Sec. 9.1, with the Parallel Transfusion model providing the greatest accuracy. The ϕ sampling artefact identified in Fig. 14 can now be further dissected at the event level. All three models struggle to resolve the individual ϕ components of the jets. The Parallel Transfusion model appears more resilient, with the issue manifesting less frequently across events.

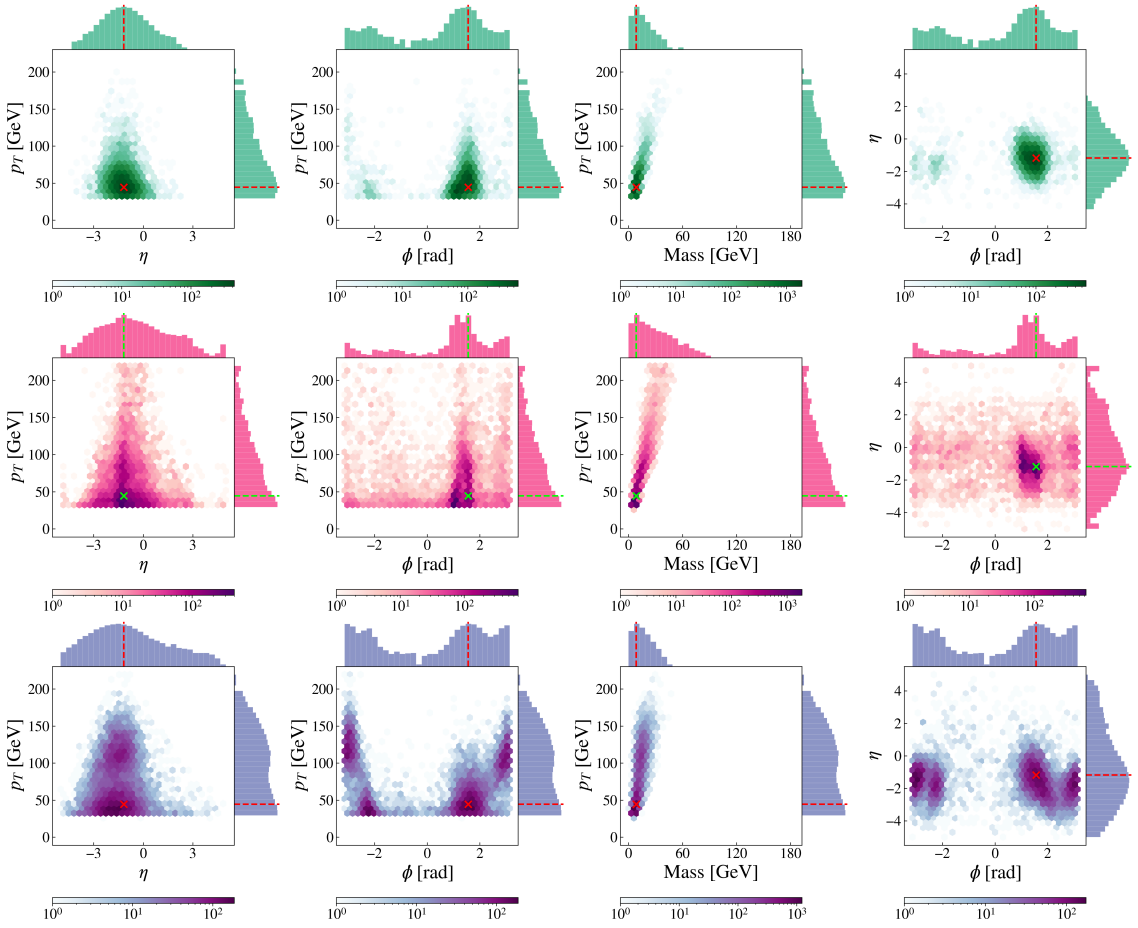

Figure 15: Event-level sampling for a single b-jet in the validation dataset. Compares Parallel Transfusion (top), Transfermer (middle) and Transfer-CFM (bottom) networks. Dashed line (1D histograms) and 'x' symbol (2D histograms) correspond to the truth. Generative models sampled 10,000 times per object.

Consider the Transfermer's result in η, ϕ space (middle-row, right histogram). Degenerate peaks in ϕ are found, one correctly aligned with the true value, and the other offset. Interestingly, an identical bimodal structure is observed in a similar plot for the other b-jet within the same event. However, for that jet, the truth is centred over the opposite peak. This suggests that the Transfermer struggles to disentangle the two leading b-jets, the reason for which is not fully understood. The current conjecture is that detector symmetries in ϕ can cause degenerate solutions. This ambiguity is not an issue when considering the joint distributions between all the jets.

Similar plots for the missing transverse energy are illustrated in Fig. 16. The Parallel Transfusion and Transfermer models both showcase impressive performance and it is likely that these results are very close to the true $p(x_{\text{reco}} \mid x_{\text{hard}})$. It is imperative to acknowledge that the same hard-level event can lead to different reconstructed outcomes due to detector smearing, hadronization, and jet clustering uncertainties. This inherent

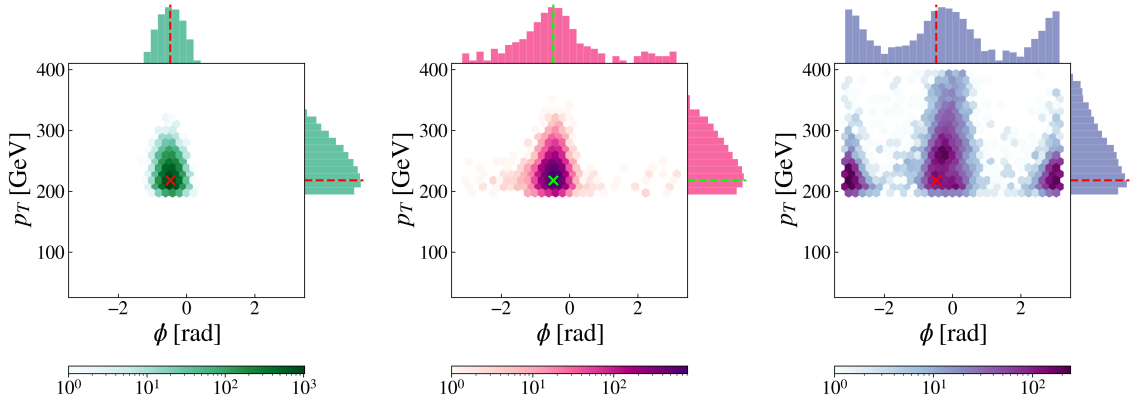

Figure 16: Event-level sampling for missing transverse energy from the validation dataset. Compares Parallel Transfusion (left), Transfermer (middle) and Transfer-CFM (right) networks. Dashed line (1D histograms) and 'x' symbol (2D histograms) corresponds to the true value. Generative models sampled 10,000 times per object.

stochasticity is what the models are attempting to capture, instead of just learning a deterministic solution. Consequently, a Gaussian-like spread around the true value for each event is desired.

Unfortunately, the Transfer-CFM failed to learn the MET ϕ distribution, corroborating the findings from Fig. 12. It seems to converge towards an average of the two most common values, a phenomenon which usually occurs when a model lacks the complexity to model higher-level distributions. This is unsurprising given that it lacks a Transformer decoder to provide vital reco-level context when generating samples.

9.3 Model Bias

Model bias refers to systematic deviations between a model's predicted values and the truth. Understanding these biases is imperative to prevent erroneous interpretations of underlying physical processes. Variance is defined by the spread of the models predictions, offering crucial insight into the reliability and stability of the generative models. An ideal model would have both minimal bias and variance.

Fig. 17 provides a quick analysis of the Parallel Transfusion model's efficacy. Previous sections have established that this model provides the most capable estimations of $p(x_{\text{reco}} \mid x_{\text{hard}})$. The 2D histograms show a strong linear correlation along the diagonal, suggesting very strong conformity with the truth. However, closer inspection of the bottom plots reveals that the bias increases at low p_T values. This can be partially attributed to the proportional nature of the y-axis, meaning that small absolute discrepancies in the low-energy regime manifest as significant relative errors. At $p_T > 200$ GeV, the bias re-

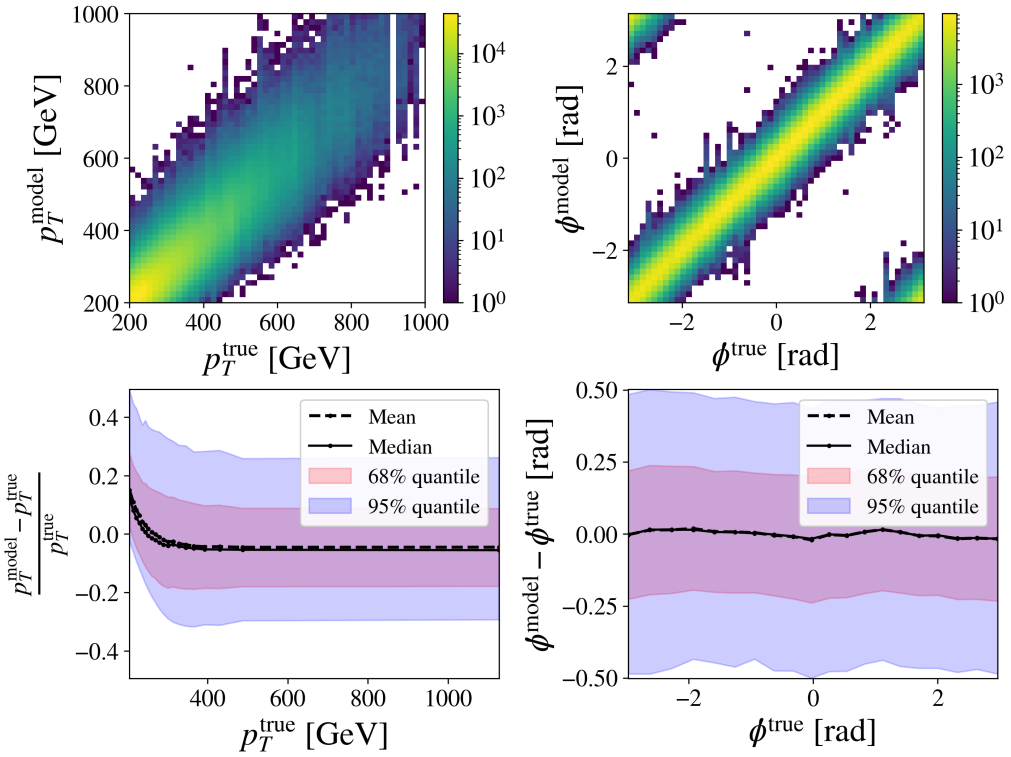

Figure 17: Direct comparison between the true feature values and the Parallel Transfusion model's predictions of missing transverse energy (top). Includes bias and variance plots (bottom). Mean (dashed) and median (solid) residuals are plotted in black, with the shaded regions indicating the spread. Model sampled 100 times for each of the 18,354 events in the validation dataset.

mains centred just below zero. This suggests that, on average, the model tends to slightly underestimate MET p_T.

The quantile bands give a compelling visualisation of the model's variance, showing that the Parallel Transfusion network consistently achieves 20% accuracy in around 60% of cases. Moreover, the model's ϕ_{MET} predictions are commendable, boasting residuals below 0.5 radians in 95% of instances. This low variance is further endorsed by a negligible bias across the entire ϕ spectrum.

Fig. 18 extends this analysis to include all three Transfer networks. Normalising the residuals by the standard deviation of the best model (Parallel Transfusion) allows for intuitive comparison of the bias and variance across architectures. The inclusion of a standard normal distribution provides a reference metric for assessing the uncertainty calibration of each model.

For the leading b-jet, the Transfer-CFM sees the broadest variance across all 4-momentum components, evidenced by its wider distribution and diminished peak height. Again, a

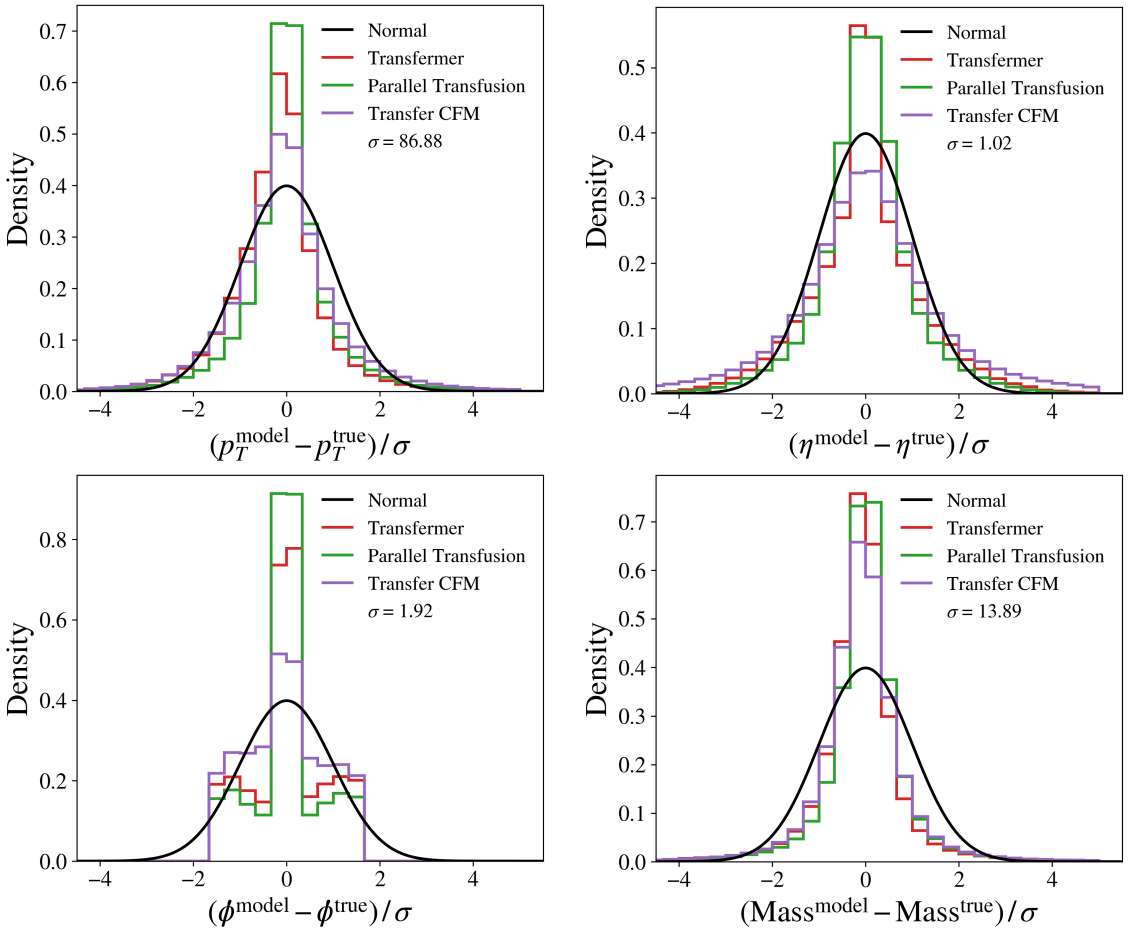

Figure 18: Standardised residual distributions for the 4-momentum components of the leading b-jet. Represents the difference between the model prediction and the truth, normalised by the standard deviation of the Parallel Transfusion network σ. A standard normal distribution is overlaid (black) for reference. Compares the different Transfer network architectures. Generative models sampled 100 times for each of the 18,354 events in the validation dataset.

lack of a Transformer decoder prohibits the procurement of critical reco-level context. All models exhibit marginal bias and are well centred around the origin. The mass residuals demonstrate particularly narrow spreads across all three models, considerably outperforming the reference normal distribution.

Fig. 19 displays the model performance in reconstructing the missing transverse energy across a multitude of events. These plots are also standardised by the standard deviation of the Parallel Transfusion model. The lower central peak in the p_T^{MET} residuals suggests that the MET is harder to generate than the b-jet, possibly due to heightened susceptibility to detector effects. Unlike individual jet measurements, MET is a global event-level observable, computed from the vector sum of all reconstructed particles. This makes it highly sensitive to small fluctuations in any of the jet energies. In contrast, single jets

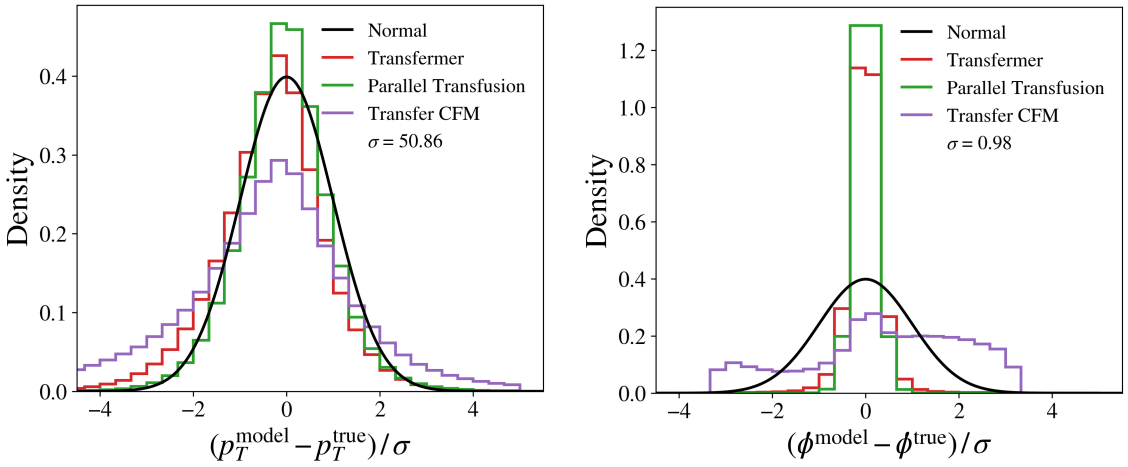

Figure 19: Standardised residual distributions for components of the missing transverse energy. Represents the difference between the model prediction and the truth, normalised by the standard deviation of the Parallel Transfusion network σ. A standard normal distribution is overlaid (black) for reference. Compares the different Transfer network architectures. Generative models sampled 100 times for each of the 18,354 events in the validation dataset.

are less affected as an energy mismeasurement is isolated to that specific object. The Transfer-CFM exhibits the largest σ. The other two models show a narrower spread, but with the Transfermer having a subtle left-shifted bias.

The Transfer-CFM's inability to model the MET ϕ distribution is reaffirmed in Fig. 19. This failure stems from its encoder-only setup, which prevents it from leveraging reco-level dependencies during sampling.

9.4 Sampling Speed

Numerically solving the CFM ODEs in Eq.(30) takes a large number of evaluations to achieve good precision. Generating samples can quickly become impractical when considering a large number of events.

Fig. 20 elucidates the computational cost of sampling from the various generative architectures. Smaller batch sizes fail to saturate the 60 streaming multiprocessors on the Nvidia RTX 4070 Ti [81], causing the GPU to be underutilised. This explains why linear scaling with time only emerges at batch sizes exceeding $\mathcal{O}(10^3)$ events. Below this threshold, smaller systems see no performance benefit.

The Transfer-CFM is the most computationally efficient model at lower batch sizes due to lack of a complex transformer architecture. The cINN's transport map is much more

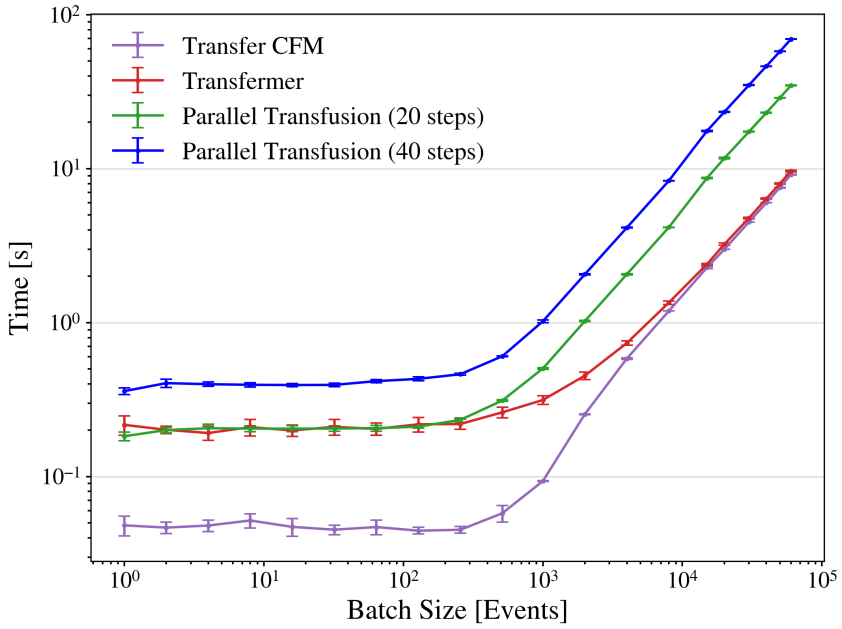

Figure 20: Sampling time as a function of batch size for different Transfer network architectures. Evaluated on an Nvidia RTX 4070 Ti GPU using the hyperparameters from Appendix A. Values are averages from 10 repeats with error bars representing one standard deviation.

efficient to evaluate than the CFM models. This is evident by the 3.6x faster inference speed over the Parallel Transfusion setup at batch size $= 10^5$ (20 integration steps).

The Transfermer requires autoregressive sampling, precluding the simultaneous evolution of all particles. This inadvertently results in similar sampling times to the Transfer-CFM for larger batches. This proves that the Transformer decoder requires a similar computational effort to the velocity field ODE by chance.

The Parallel Transfusion framework, in principle, exploits the powerful advantage of evolving all samples concurrently, theoretically offering superior scalability. However, this comes with an unfortunate trade-off: the expensive Transformer conditioning must be re-evaluated at every integration step. This makes inference heavily dependent on the number of velocity field evaluations. While more steps is desirable, inference time scales linearly with this parameter, confirmed by a 2.0x slower execution time when doubling the number of integration steps from 20 to 40.

To make things worse, this overhead is amplified when computing the likelihood ODE in Eq.(33). This function involves calculating the gradients of all components of the velocity with respect to the inputs. As a result the CFM likelihood calculation is significantly slower than sampling and can become difficult to scale at large event numbers. In contrast, cINNs are a score-based model, requiring only a single forward pass through

the network to evaluate the likelihood. This makes them $\mathcal{O}(30)$ times faster than CFM-based methods for computing the transfer probability [22]. Prospective techniques such as diffusion distillation [82–84] might eventually make CFM models more practical in this regard.

9.5 Future Work

While CFM has demonstrated its superior capacity for handling complex transport maps compared to cINNs, the current implementation still has scope for improvement.

Optimal Transport (OT) is a mathematical framework for finding the most efficient means of transforming one probability distribution into another [85]. Fundamentally, OT seeks to minimise the cost (or 'effort') of redistributing probability mass [86]. This can facilitate more precise and stable sampling compared to simpler generative approaches.

Vanilla OT assumes strict mass conservation between source and target distributions. This means the total probability in both distributions is exactly equal (e.g. normalised). This stringent one-to-one mapping can be inappropriate for the jet combinatorics issue in Sec. 8. Recent advancements in ML have proposed novel 'unbalanced' OT plans that are capable of relaxing the strict mass preservation constraint [43, 86–89]. This facilitates mass creation and destruction during the bridging process, which can reduce the impact of outliers [89] caused by variable multiplicity. This conforms with how jets are lost due to detector inefficiencies, or enhanced from soft QCD, FSR, and ISR.

This study has developed a suite of complex generative networks, demonstrating exciting promise in sampling reco-level events. An immediate and simple next step would be to generate the likelihood $p(x_{\text{reco}} \mid x_{\text{hard}})$ for the CFM models, through evaluating Eq.(33). This will provide the transfer probability required in Fig. 7.

Optionally, it could be beneficial to compare the efficacy of an autoregressive CFM model. Such an network can be implemented by simply replacing the lightweight cINN in the Transfermer with a small CFM network. This would allow for sequential generation of reco-level momenta instead of the simultaneous evolution found in the Parallel Transfusion model.

For a full MEM likelihood evaluation, a Sampling-cINN must be developed for more efficient sampling of hard-level events. The remaining networks (multiplicity and acceptance classifiers) are discussed in Sec. 10. With these additions, the full ML-framework in Fig. 7 will be ready for a final likelihood estimate in a real analysis.

10 Classifier Results

The acceptance classifier discussed in Sec. 5.3 is used to compute the detector efficiency $\epsilon(x_{\text{hard}})$ term in the overall MEM integral from Eq.(11). The hyperparameters governing the classifier's training and results are listed in Tab. 3. Overall, the true global efficiency for this process is 12.08%. The network was able to achieve a learned acceptance of 12.09%, accurately matching the true distribution across the dataset.

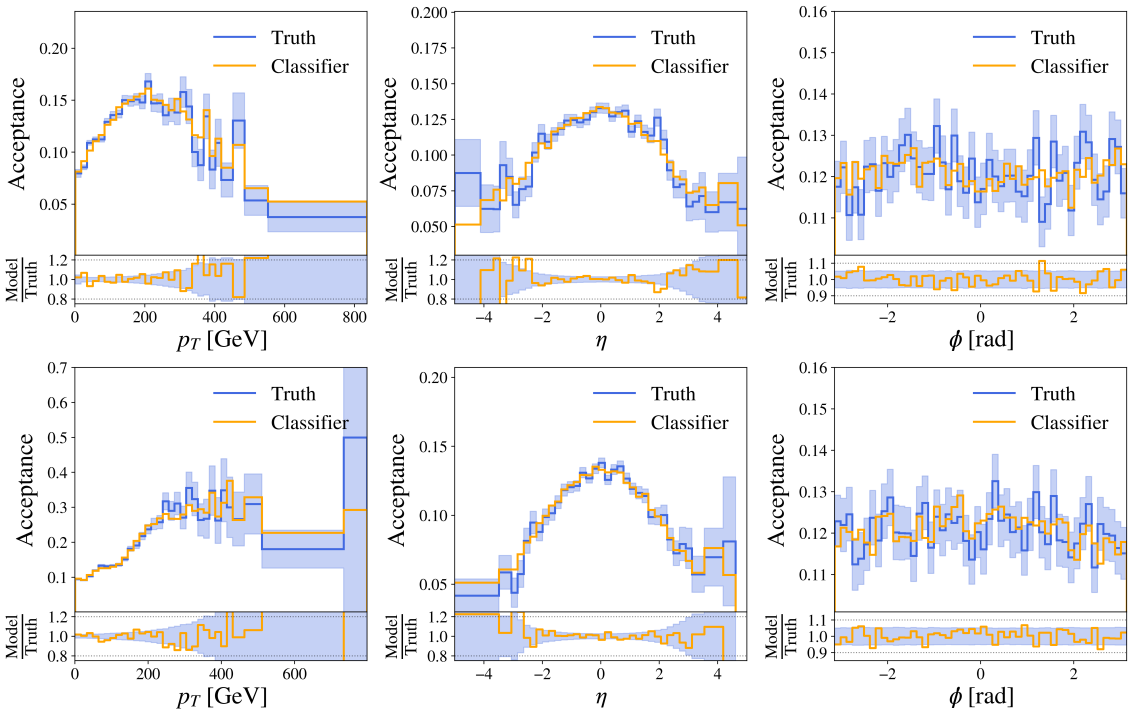

Figure 21: Performance of the acceptance classifier for a b-quark (top) and a neutrino (bottom) from the $t\bar{t}H$ process as a function of different kinematic observables. Truth (blue) corresponds to acceptance across 151,329 events from the validation dataset and includes Poisson uncertainty bands (shaded).

The results in Fig. 21 demonstrate the performance of the acceptance classifier in reproducing the true distributions for b-quarks and neutrinos. Overall, the network successfully captures detector acceptance across a variety of kinematic observables, firmly staying within the Poison errors in the truth. Similar results were observed for all final-state particles. Future work will involve incorporating this acceptance network alongside the Transfer network in the final MEM likelihood calculation.

The purpose of the multiplicity classifier is to generalise the Transfer network to handle a variable number of jets. In order to pass this information to the generative models, the classifier needs to be capable of predicting how many jets are reconstructed in the detector. Fig. 22 demonstrates great performance from the classifier in the most populous

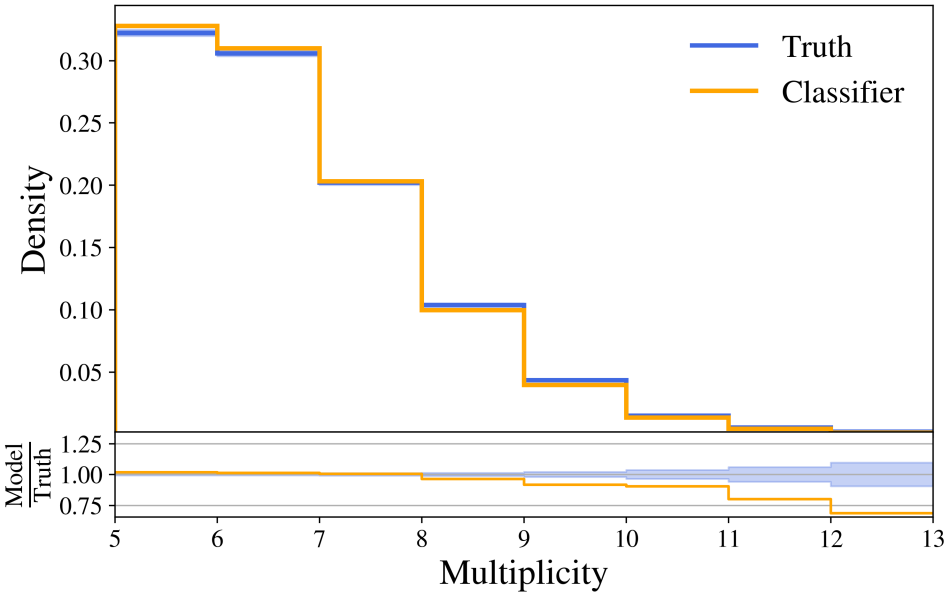

Figure 22: Multiplicity classifier performance showing the distribution of jet counts across events. Truth (blue) corresponds to 50,561 events and includes Poisson uncertainty bands (shaded). Results obtained using the hyperparameters listed in Tab. 3.

bins with lower multiplicities. However, in the lower statistics regions, where it becomes rarer to see events with high particle counts, the performance degrades.

Initial-state radiation (ISR) was excluded from the training data for this model. Since ISR can contribute additional jets to the final state, its omission might lead to an underestimation of jet multiplicity in real collision events. A logical improvement would be to include ISR in the training data, allowing for more realistic predictions.

A natural next step would be to extend the Transfer networks to incorporate the multiplicity predictions from this classifier. This can be achieved by simply appending the multiplicity network output to the x_{hard} embeddings in one-hot encoded form. The final transfer probability $p(x_{\text{reco}} \mid x_{\text{hard}}, n)$ can then be modelled in line with Eq.(18). It is important to note that while the Transfermer and Parallel Transfusion models both feature a Transformer decoder, enabling them to dynamically adjust the number of reco-level particles to generate, the Transfer-CFM lacks this mechanism. Therefore, the Transfer-CFM inherently operates with a fixed multiplicity assumption, limiting its flexibility in modelling final-state topologies.

11 Conclusion

Overall, this study successfully applied novel ML methods to key components of the MEM integral. Ultimately, this aims to improve searches for invisible Higgs boson decays produced via the $t\bar{t}H$ mechanism, targeting the fully hadronic final state.

Three distinct generative architectures were deployed to evaluate the transfer probability $p(x_{\text{reco}} \mid x_{\text{hard}})$. The findings demonstrated significant improvements in replicating reco-level event topologies when using Transformer-based models. The Parallel Transfusion network provided superior expressivity, primarily due to the enhanced flexibility of its continuous transport map. However, the cINN displayed practical advantages through its 3.6x faster sampling speed and much simpler likelihood formula. The autoregressive Transfermer excelled in modelling complex jet mass distributions but faced limitations related to particle ordering dependencies. Additionally, all three models struggled to disentangle individual ϕ_{jet} values due to detector symmetries.

The acceptance network boasted a learned efficiency of 12.09%, closely matching the true 12.08% value for $\epsilon(x_{\text{hard}})$. While the multiplicity network made accurate predictions at lower jet counts, the inclusion of ISR data for this process would lead to a more representative result. The Transfer network also awaits implementation of the multiplicity classifier to handle variable numbers of jets.

The methods outlined in this paper used simulated CMS samples from Run 2. This lays the groundwork for an extension to LHC Run 3 data. The enhanced luminosity and energy of this period could present an exciting opportunity for real-world validation of the proposed methods.

A Model Hyperparameters

Parameter	Value
Optimiser	Adam
Learning rate	0.0001
LR schedule	Reduce-on-plateau
Minimum LR	1e-7
Batch size	1024
Epochs	500
Number of heads	8
Number of encoder layers	6
Number of decoder layers	8
Embedding dimension	64
Transf. feed-forward dim	256
Flow transformations	5
Number of subnet layers	3
Subnet hidden nodes	256
Subnet activation function	ReLU
RQS spline bins	16
Training samples	73,455
Validation samples	18,364
Trainable parameters	4.1 M

Parameter	Value
Optimiser	Adam
Learning rate	0.0001
LR schedule	Cosine-annealing
Batch size	1024
Epochs	500
Number of heads	8
Number of encoder layers	6
Number of decoder layers	8
Embedding dimension	64
Transf. feed-forward dim	256
Velocity net layers	8
Velocity net hidden nodes	512
Velocity net activation	SiLU
Bridging noise strength, σ	0.1
ODE solver	Runge-Kutta 4
Solver step-size	0.05
Training samples	73,455
Validation samples	18,364
Trainable parameters	2.7 M

Table 1: Hyperparameters of the autoregressive Transfermer (left) and Parallel Transfusion (right) models.

Parameter	Value
Optimiser	Adam
Learning rate	0.0001
LR schedule	Cosine-annealing
Batch size	1024
Epochs	500
Number of heads	8
Number of encoder layers	6
Embedding dimension	64
Transf. feed-forward dim	256
Velocity net layers	8
Velocity net hidden nodes	512
Velocity net activation	SiLU
Bridging noise strength, σ	0.1
ODE solver	Runge-Kutta 4
Solver step-size	0.05
Training samples	73,455
Validation samples	18,364
Trainable parameters	2.2 M

Table 2: Hyperparameters of the Transfer-CFM model.

Parameter	Value
Optimiser	RAdam
Learning rate	0.0001
LR schedule	Reduce-on-plateau
Minimum LR	1e-7
Batch size	1024
Epochs	50
Transf. encoder layers	3
Transf. heads	8
Transf. hidden nodes	256
Activation function	GeLU
MLP layers	3
Hidden nodes	128, 64, 32
Loss	Categorical cross-entropy
Training samples	117,975
Validation samples	50,561
Trainable parameters	2.0 M

Parameter	Value
Optimiser	RAdam
Learning rate	0.00001
LR schedule	Reduce-on-plateau
Minimum LR	1e-7
Batch size	1024
Epochs	180
Transf. encoder layers	6
Transf. heads	8
Transf. hidden nodes	512
Activation function	GeLU
MLP layers	4
Hidden nodes	128, 64, 64, 32
Loss	Binary cross-entropy
Training samples	605,313
Validation samples	151,329
Trainable parameters	3.2 M

Table 3: Hyperparameters of the classifiers used to learn the jet multiplicity (left) and detector efficiency $\epsilon(x_{\text{hard}})$ (right).

References

[1] P. W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (16 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[2] G. Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.020.

[3] S. Chatrchyan et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.021.

[4] G. Aad et al. "Combined Measurement of the Higgs Boson Mass from the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ Decay Channels with the ATLAS Detector Using $\sqrt{s} = 7$, 8, and 13 TeV pp Collision Data". In: *Phys. Rev. Lett.* 131 (25 2023), p. 251802. DOI: 10.1103/PhysRevLett.131.251802.

[5] A. Collaboration. *Exploring the Higgs boson 'discovery channels', ATLAS releases full Run 2 measurements of Higgs boson properties*. EPS-HEP. 2019.

[6] J. Mellenthin and on behalf of theATLAS Collaboration. "Observation of the $t\bar{t}H$ production". In: *Journal of Physics: Conference Series* 1390.1 (2019), p. 012042. DOI: 10.1088/1742-6596/1390/1/012042.

[7] A. M. Sirunyan et al. "Observation of $t\bar{t}H$ Production". In: *Phys. Rev. Lett.* 120 (23 2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801.

[8] P. D. Group et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2022.8 (2022), p. 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptac097.

[9] A. Blondel, A. de Gouvêa, and B. Kayser. "Z-boson decays into Majorana or Dirac heavy neutrinos". In: *Phys. Rev. D* 104 (5 2021), p. 055027. DOI: 10.1103/PhysRevD.104.055027.

[10] M. Aaboud et al. "Combination of Searches for Invisible Higgs Boson Decays with the ATLAS Experiment". In: *Phys. Rev. Lett.* 122 (23 2019), p. 231801. DOI: 10.1103/PhysRevLett.122.231801.

[11] A. Tumasyan et al. "A search for decays of the Higgs boson to invisible particles in events with a top-antitop quark pair or a vector boson in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *The European Physical Journal C* 83.10 (2023), p. 933. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-023-11952-7.

[12] G. Aad et al. "Combination of searches for invisible decays of the Higgs boson using 139 fb^1 of proton-proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment". In: *Physics Letters B* 842 (2023), p. 137963. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2023.137963.

[13] R. Steerenberg. *Accelerator Report: LHC Run 3 achieves record-breaking integrated luminosity*. CERN. 2024.

[14] *New schedule for CERN's accelerators*. CERN. 2024.

[15] G. Arcadi, A. Djouadi, and M. Raidal. "Dark Matter through the Higgs portal". In: *Physics Reports* 842 (2020). Dark Matter through the Higgs portal, pp. 1–180. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2019.11.003.

[16] K. Kondo. "Dynamical Likelihood Method for Reconstruction of Events with Missing Momentum. I. Method and Toy Models". In: *Journal of the Physical Society of Japan* 57.12 (1988), pp. 4126–4140. DOI: 10.1143/JPSJ.57.4126.

[17] Schmitt, Stefan. "Data Unfolding Methods in High Energy Physics". In: *EPJ Web Conf.* 137 (2017), p. 11008. DOI: 10.1051/epjconf/201713711008.

[18] G. D'Agostini. *Improved iterative Bayesian unfolding*. 2010. arXiv: 1010.0632 [physics.data-an].

[19] R. A. Willoughby. "Solutions of Ill-Posed Problems (A. N. Tikhonov and V. Y. Arsenin)". In: *SIAM Review* 21.2 (1979), pp. 266–267. DOI: 10.1137/1021044.

[20] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe. *Guided Image Generation with Conditional Invertible Neural Networks*. 2019. arXiv: 1907.02392 [cs.CV].

[21] A. Butter, T. Heimel, T. Martini, S. Peitzsch, and T. Plehn. "Two invertible networks for the matrix element method". In: *SciPost Phys.* 15 (2023), p. 094. DOI: 10.21468/SciPostPhys.15.3.094.

[22] T. Heimel, N. Huetsch, R. Winterhalder, T. Plehn, and A. Butter. "Precision-machine learning for the matrix element method". In: *SciPost Phys.* 17 (2024), p. 129. DOI: 10.21468/SciPostPhys.17.5.129.

[23] A. Sirunyan et al. "Particle-flow reconstruction and global event description with the CMS detector". In: *Journal of Instrumentation* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003.

[24] V. Khachatryan et al. "The CMS trigger system". In: *Journal of Instrumentation* 12.01 (2017), P01020. DOI: 10.1088/1748-0221/12/01/P01020.

[25] G. Arduini et al. "High Luminosity LHC: challenges and plans". In: *Journal of Instrumentation* 11.12 (2016), p. C12081. DOI: 10.1088/1748-0221/11/12/C12081.

[26] T. R. F. P. Tomei. *The High-Level Trigger for the CMS Phase-2 Upgrade*. 2022. arXiv: 2211.03684 [hep-ex].

[27] T. C. Collaboration et al. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[28] A. Tumasyan et al. "Precision measurement of the W boson decay branching fractions in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Physical Review D* 105.7 (2022). ISSN: 2470-0029. DOI: 10.1103/physrevd.105.072008.

[29] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (9 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[30] A. Zee. *Group Theory in a Nutshell for Physicists*. In a Nutshell. Princeton University Press, 2016. ISBN: 9780691162690.

[31] M. Turtola. *Electroweak Symmetry Breaking and Higgs Pair Production in Effective Field Theories*. Tech. rep. Course Thesis, 1FA599. Uppsala University, 2024.

[32] S. Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (21 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[33] A. Salam and J. C. Ward. "Weak and Electromagnetic Interactions". In: *Il Nuovo Cimento (1955-1965)* 11 (4 1959), pp. 568–577. ISSN: 1827-6121. DOI: 10.1007/BF02726525.

[34] S. L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2.

[35] S. Raychaudhuri. "The Higgs Boson and its physics: an overview". In: *Indian Journal of Physics* 97.11 (2023), pp. 3189–3224. ISSN: 0974-9845. DOI: 10.1007/s12648-023-02718-8.

[36] A. Abokhalil. *The Higgs Mechanism and Higgs Boson: Unveiling the Symmetry of the Universe*. 2023. arXiv: 2306.01019 [hep-ph].

[37] I. Kobyzev, S. J. Prince, and M. A. Brubaker. "Normalizing Flows: An Introduction and Review of Current Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2021), pp. 3964–3979. DOI: 10.1109/TPAMI.2020.2992934.

[38] E. Tabak and T. Cristina. "A Family of Nonparametric Density Estimation Algorithms". In: *Communications on Pure and Applied Mathematics* 66 (2013). DOI: 10.1002/cpa.21423.

[39] D. J. Rezende and S. Mohamed. *Variational Inference with Normalizing Flows*. 2016. arXiv: 1505.05770 [stat.ML].

[40] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. *Normalizing Flows for Probabilistic Modeling and Inference*. 2021. arXiv: 1912.02762 [stat.ML].

[41] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. *Flow Matching for Generative Modeling*. 2023. arXiv: 2210.02747 [cs.LG].

[42] M. S. Albergo and E. Vanden-Eijnden. *Building Normalizing Flows with Stochastic Interpolants*. 2023. arXiv: 2209.15571 [cs.LG].

[43] X. Liu, C. Gong, and Q. Liu. *Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow*. 2022. arXiv: 2209.03003 [cs.LG].

[44] A. Butter et al. *Jet Diffusion versus JetGPT – Modern Networks for the LHC*. 2024. DOI: https://doi.org/10.21468/SciPostPhysCore.8.1.026. arXiv: 2305.10475 [hep-ph].

[45] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[46] A. Tong et al. "Improving and generalizing flow-based generative models with minibatch optimal transport". In: *Trans. Mach. Learn. Res.* 2024 (2023).

[47] A. Tong et al. *Simulation-free Schrödinger bridges via score and flow matching*. 2024. arXiv: 2307.03672 [cs.LG].

[48] M. Skreta, L. Atanackovic, A. J. Bose, A. Tong, and K. Neklyudov. *The Superposition of Diffusion Models Using the Itô Density Estimator*. 2025. arXiv: 2412.17762 [cs.LG].

[49] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. *Neural Ordinary Differential Equations*. 2019. arXiv: 1806.07366 [cs.LG].

[50] B. M. Dillon et al. "Symmetries, safety, and self-supervision". In: *SciPost Phys.* 12 (2022), p. 188. DOI: 10.21468/SciPostPhys.12.6.188.

[51] A. Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[52] V. Mikuni and F. Canelli. "ABCNet: an attention-based method for particle tagging". In: *The European Physical Journal Plus* 135.6 (2020), p. 463. DOI: 10.1140/epjp/s13360-020-00497-3.

[53] T. Finke, M. Krämer, A. Mück, and J. Tönshoff. "Learning the language of QCD jets with transformers". In: *Journal of High Energy Physics* 2023.6 (2023), p. 184. DOI: 10.1007/JHEP06(2023)184.

[54] S. V. Stroud et al. "Secondary vertex reconstruction with MaskFormers". In: *The European Physical Journal. C, Particles and Fields* 84 (2023).

[55] S. V. Stroud et al. *Transformers for Charged Particle Track Reconstruction in High Energy Physics*. 2024. arXiv: 2411.07149 [hep-ex].

[56] D. P. Kingma et al. *Improving Variational Inference with Inverse Autoregressive Flow*. 2017. arXiv: 1606.04934 [cs.LG].

[57] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. *Neural Autoregressive Flows*. 2018. arXiv: 1804.00779 [cs.LG].

[58] M. Patacchiola, A. Shysheya, K. Hofmann, and R. E. Turner. *Transformer Neural Autoregressive Flows*. 2024. arXiv: 2401.01855 [cs.LG].

[59] J. Ackerschott, R. K. Barman, D. Gonçalves, T. Heimel, and T. Plehn. "Returning CP-observables to the frames they belong". In: *SciPost Phys.* 17 (2024), p. 001. DOI: 10.21468/SciPostPhys.17.1.001.

[60] A. Paszke et al. "PyTorch: an imperative style, high-performance deep learning library". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[61] S. Agostinelli et al. "Geant4—a simulation toolkit". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8.

[62] J. Allison et al. "Geant4 developments and applications". In: *IEEE Transactions on Nuclear Science* 53.1 (2006), pp. 270–278. DOI: 10.1109/TNS.2006.869826.

[63] J. Allison et al. "Recent developments in Geant4". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 835 (2016), pp. 186–225. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2016.06.125.

[64] A. Hayrapetyan et al. "Measurement of the $t\bar{t}H$ and tH production rates in the $H \to b\bar{b}$ decay channel using proton-proton collision data at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2025.2 (2025). ISSN: 1029-8479. DOI: 10.1007/jhep02(2025)097.

[65] Y. Guler. *Radiation Damage Monitoring System of The CMS HF Detector*. 2020. arXiv: 2009.10172 [physics.ins-det].

[66] J. de Favereau et al. "DELPHES 3: a modular framework for fast simulation of a generic collider experiment". In: *Journal of High Energy Physics* 2014.2 (2014), p. 57. ISSN: 1029-8479. DOI: 10.1007/JHEP02(2014)057.

[67] A. Mertens. "New features in Delphes 3". In: *Journal of Physics: Conference Series* 608.1 (2015), p. 012045. DOI: 10.1088/1742-6596/608/1/012045.

[68] V. Khachatryan et al. “Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV”. In: *Journal of Instrumentation* 12.02 (2017), P02014–P02014. ISSN: 1748-0221. DOI: 10.1088/1748-0221/12/02/p02014.

[69] E. Focardi et al. “The CMS silicon tracker”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 453.1 (2000). Proc. 7th Int. Conf on Instrumentation for colliding Beam Physics, pp. 121–125. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(00)00616-1.

[70] The ATLAS Collaboration. “Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1”. English. In: *European Physical Journal C: Particles and Fields* 77 (2017), pp. 1–73. ISSN: 1434-6044. DOI: 10.1140/epjc/s10052-017-5004-5.

[71] M. Aaboud et al. “Identification and rejection of pile-up jets at high pseudorapidity with the ATLAS detector”. In: *The European Physical Journal C* 77.9 (2017), p. 580. DOI: 10.1140/epjc/s10052-017-5081-5.

[72] F. Abe et al. “Jet pseudorapidity distribution in direct photon events in $p\overline{p}$ collisions at $\sqrt{s} = 1.8$TeV”. In: *Phys. Rev. D* 57 (3 1998), pp. 1359–1365. DOI: 10.1103/PhysRevD.57.1359.

[73] E. Butz. “Operation and performance of the CMS silicon detectors”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1064 (2024), p. 169377. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2024.169377.

[74] A. Tumasyan et al. “Measurement of the differential $\mathrm{t\bar{t}}$ production cross section as a function of the jet mass and extraction of the top quark mass in hadronic decays of boosted top quarks”. In: *The European Physical Journal C* 83.7 (2023), p. 560. DOI: 10.1140/epjc/s10052-023-11587-8.

[75] G. Aad et al. “Measurement of the W-boson mass and width with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 7$ TeV”. In: *The European Physical Journal C* 84.12 (2024), p. 1309. DOI: 10.1140/epjc/s10052-024-13190-x.

[76] A. M. Sirunyan et al. “Measurement of the Jet Mass Distribution and Top Quark Mass in Hadronic Decays of Boosted Top Quarks in pp Collisions at $\sqrt{s} = 13$ TeV”. In: *Phys. Rev. Lett.* 124 (20 2020), p. 202001. DOI: 10.1103/PhysRevLett.124.202001.

[77] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia. “Jet flavour classification using DeepJet”. In: *Journal of Instrumentation* 15.12 (2020), P12012. DOI: 10.1088/1748-0221/15/12/P12012.

[78] T. C. collaboration et al. “A new calibration method for charm jet identification validated with proton-proton collision events at $\sqrt{s} = 13$ TeV”. In: *Journal of Instrumentation* 17.03 (2022), P03014. DOI: 10.1088/1748-0221/17/03/P03014.

[79] M. J. Fenton et al. “Permutationless many-jet event reconstruction with symmetry preserving attention networks”. In: *Phys. Rev. D* 105 (11 2022), p. 112008. DOI: 10.1103/PhysRevD.105.112008.

[80] A. Shmakov et al. “SPANet: Generalized permutationless set assignment for particle physics using symmetry preserving attention”. In: *SciPost Phys.* 12 (2022), p. 178. DOI: 10.21468/SciPostPhys.12.5.178.

[81] *NVIDIA GeForce RTX 4070 Ti Specs*. TechPowerUp. 2025.

[82] T. Salimans and J. Ho. *Progressive Distillation for Fast Sampling of Diffusion Models*. 2022. arXiv: 2202.00512 [cs.LG].

[83] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. *Consistency Models*. 2023. arXiv: 2303.01469 [cs.LG].

[84] E. Buhmann et al. *CaloClouds II: Ultra-Fast Geometry-Independent Highly-Granular Calorimeter Simulation*. 2024. arXiv: 2309.05704 [physics.ins-det].

[85] S. T. Rachev and L. Rüschendorf. *Mass transportation problems*. 1998.

[86] C. Bunne, A. Krause, and M. Cuturi. *Supervised Training of Conditional Monge Maps*. 2023. arXiv: 2206.14262 [cs.LG].

[87] G. Kerrigan, G. Migliorini, and P. Smyth. *Dynamic Conditional Optimal Transport through Simulation-Free Flows*. 2024. arXiv: 2404.04240 [cs.LG].

[88] M. S. Albergo, M. Goldstein, N. M. Boffi, R. Ranganath, and E. Vanden-Eijnden. *Stochastic interpolants with data-dependent couplings*. 2024. arXiv: 2310.03725 [cs.LG].

[89] M. Gazdieva, A. Asadulaev, A. Korotin, and E. Burnaev. *Light Unbalanced Optimal Transport*. 2025. arXiv: 2303.07988 [cs.LG].

Certificate of Copyright Ownership

This Project Report is presented as part of, and in accordance with, the requirements for the degree of MSci at the University of Bristol, Faculty of Science.

I hereby assert that I own exclusive copyright in the item named below. I give permission to the University of Bristol Library to add this item to its stock and to make it available for consultation in the library, and for inter-library lending for use in another library. I also give consent for this report to made available electronically to staff and students within the University of Bristol. It may be copied in full or in part for any bona fide library or research work. No quotation and no information derived from it may be published without the author's prior consent.

Author	Luke Johnson
Title	Applying Machine Learning to the Matrix Element Method in Searches for Invisible Higgs Boson Decays
Date of submission	20/03/2025

I agree that submission of this report constitutes signing of this declaration.

This project/dissertation is the property of the University of Bristol and may only be used with due regard to the rights of the author. Bibliographical references may be noted, but no part may be copied for use or quotation in any published work without the prior permission of the author. In addition, due acknowledgement for any use must be made.