Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hadronic Mono-Z Dark Matter Sensitivity with Flow Matching on CMS Open Data

Hitesh Rasineni

hitesh.23bce8825@vitapstudent.ac.in

VIT-AP University

Bhavishya Chebrolu

Mohan Babu University

Research Article

Keywords:

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10634384/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hadronic Mono-Z Dark Matter Sensitivity with Flow Matching on CMS Open Data

Hitesh Rasineni [*1,†] and Chebrolu Bhavishya [2,†]

[1]VIT-AP University, Amaravati, 522241, India ,
*hitesh.23bce8825@vitapstudent.ac.in
[2]Mohan Babu University, Tirupati, 517102, India ,
23102A041122@mbu.asia

Abstract

We present a projected sensitivity study for hadronic mono-Z dark-matter production using CMS Run 2015D HTMHT open data corresponding to 2.256382381 fb^{-1}, from which 1,439,523 events satisfy the hadronic mono-Z selection. Backgrounds are modelled with a conditional flow-matching continuous normalizing flow trained on the selected HTMHT events and evaluated on a held-out validation split reweighted to the full selected population. To mitigate artifacts from missing-object features and avoid in-sample scoring bias we apply sentinel imputation for undefined angular features, persist the train/validation split indices, and enforce a minimum reported background yield of 20 events when selecting the working point. A signal-side offline trigger proxy is applied to the simulated signal before scoring. Under this procedure the baseline analysis yields expected significances of 2.89σ, 7.62σ, and 7.41σ for three simplified-model benchmarks. An ablation study that removes the detailed extra-jet kinematics reduces the expected significance by 53–71%, indicating that extra-jet topology carries substantial discriminating power in the hadronic mono-Z channel. These results are projected sensitivities (no unblinding performed); the limitations and reproducibility of the study are discussed in Sections 8 and E.

*Corresponding author: hitesh.23bce8825@vitapstudent.ac.in
† These authors contributed equally and share first authorship.

1 Introduction

The existence of dark matter (DM) is inferred from a wide range of independent cosmological and astrophysical observations, yet its particle nature remains one of the most important open questions in fundamental physics. Collider experiments offer a complementary route to direct and indirect detection: DM may be pair-produced in high-energy proton–proton collisions, and the resulting momentum imbalance, quantified by the missing transverse momentum ($E_{\mathrm{T}}^{\mathrm{miss}}$), carries information about the DM couplings to Standard Model (SM) particles [1]. Among the so-called mono-X signatures, the mono-Z topology—a Z boson produced in association with a pair of stable DM particles—provides a distinctive handle on simplified models with s-channel mediators that radiate an on-shell Z boson [2, 3]. The Z boson can be reconstructed either through its leptonic decay $Z \to \ell^+\ell^-$ or its hadronic decay $Z \to q\bar{q}$. Although the leptonic channel offers a clean, low-background signature, the hadronic channel benefits from the substantially larger branching fraction of the Z to jets and is complementary to the mono-jet searches that dominate the early Run 2 DM programme.

In a preceding study [4], we demonstrated a density-based search for DM in the leptonic mono-Z final state using CMS Run 2015D open data. That analysis trained five Neural Spline Flows (NSFs) [5, 6]—two channel-specific SM background flows and three mediator-specific DM signal flows—and used the per-event log-likelihood ratio between the DM and SM den-

sities as the test statistic, followed by a simultaneous signal-region/validation-region profile-likelihood fit. The leptonic analysis was limited by a residual background-modelling discrepancy in the high-$E_{\mathrm{T}}^{\mathrm{miss}}$ tail, which inflated the observed-to-expected limit ratio by a factor of roughly 7–12 and prevented a clean interpretation of the fitted signal strength.

The present work extends this programme to the hadronic mono-Z channel and introduces a fundamentally different density-modelling strategy. Instead of closed-form normalizing flows with analytic densities, we model the background density with a conditional flow-matching continuous normalizing flow (CFM-CNF) [7, 8]. Whereas an NSF evaluates the density in closed form through a composition of invertible spline coupling layers, a CNF defines the density implicitly through a learned time-dependent vector field integrated along a probability-flow ODE, with the logarithmic density obtained from the divergence trace via Hutchinson's stochastic estimator [9]. This continuous-time formulation is trained with the simulation-free conditional flow-matching objective [7], which regresses the model onto the velocity of a straight-line probability path and avoids the architectural constraints of invertible coupling layers. The switch from a two-hypothesis likelihood-ratio score to a single background-density discriminant also changes the statistical treatment: rather than fitting signal-plus-background templates in a profile likelihood, we evaluate per-event background negative log-likelihoods on a held-out validation split and convert the resulting yields into an Asimov expected significance [10].

The hadronic channel introduces additional methodological challenges that motivate several of the technical choices in this paper. First, the reconstruction of a hadronic Z candidate from a dijet pair requires a dijet-mass window and dijet-angular selection, and the presence of additional jets (beyond the two Z-candidate jets) creates features that are physically undefined when no such jet exists. In the leptonic analysis these missing-object features were handled with binary jet-presence indicators and zero-imputation [4]; here we instead preserve the "does-not-exist" state as an explicit out-of-range sentinel value, avoiding the fabrication of a spurious point mass in the physical continuum. Second, because the background density is trained on real selected HTMHT events, we score only a persisted held-out validation split and reweight the scored events back to the full selected population, eliminating the in-sample scoring bias that would otherwise inflate the background yield. Third, since the DELPHES fast simulation does not emulate the CMS high-level trigger, we apply an offline trigger proxy to the simulated signal before scoring. Finally, to avoid the look-elsewhere bias of an unconstrained argmax over a noisy discriminant scan, we report the working point that maximises the expected significance subject to a minimum background yield of $B \geq 20$ events.

The analysis uses the CMS Run 2015D HTMHT MINIAOD open dataset [11], corresponding to a validated recorded integrated luminosity of 2.256382381 fb^{-1} [12, 13], and follows the open-data research practices advocated by the CMS open-data community [14, 15]. Dark-matter signal samples are generated with the `DMsimp_s_spin1` simplified-model implementation of an s-channel spin-1 mediator [1, 2] using a reproducible MADGRAPH5_AMC@NLO + PYTHIA8 + DELPHES toolchain [16, 17, 18]. Under the post-fix baseline procedure the analysis yields expected significances of 2.89σ, 7.62σ, and 7.41σ for three simplified-model benchmarks, and an ablation study that removes detailed extra-jet kinematics reduces the expected significance by 53–71%, indicating that extra-jet topology carries substantial discriminating power in the hadronic mono-Z channel.

The remainder of this paper is organised as follows. Section 2 describes the data and simulation, including the hadronic mono-Z selection and the signal-generation toolchain. Section 3 presents the conditional flow-matching background density model, the feature contract and preprocessing, and the training procedure. Section 4 summarises the validation studies, including closure checks and the before/after comparison of the sentinel-imputation fix. Section 5 documents the sensitivity methodology, and Section 6 reports the expected-sensitivity results. Section 7 presents the extra-jet feature ablation study, Section 8 discusses limitations, and Section 9 concludes.

2 Data and simulation

2.1 CMS Run 2015D HTMHT open data

The analysis uses the CMS Run 2015D HTMHT MINIAOD open dataset [11], corresponding to a validated recorded integrated luminosity of 2.256382381 fb^{-1} for runs 256630–260627 [12, 13]. Events are read in chunks via parallelised XRootD access with a custom feature-extraction pipeline built on `uproot` and `awkward`; per-file processing counts are recorded to verify completeness of the dataset traversal (Appendix C). In total 444 MINIAOD files containing 20,679,437 raw events were processed, of which 1,439,523 events satisfy the hadronic mono-Z selection described below, corresponding to a selection efficiency of 6.96%.

Events are accepted if they pass any of the five unprescaled HLT trigger paths listed in Appendix A, which comprise H_T-based, $E_{\text{T}}^{\text{miss}}$-based, and mono-jet trigger paths and accept 91.3% of the raw events. The offline hadronic mono-Z selection requires at least two jets with $p_T > 30$ GeV and $|\eta| < 2.4$. The two leading jets define the hadronic-Z candidate and must satisfy $70 < m_{jj} < 110$ GeV and $\Delta R_{jj} < 2.0$. A b-tag veto is applied to additional jets beyond the two Z-candidate jets. The selected events are stored in a single ROOT tree with a fixed 41-branch feature schema (kinematic, angular, and recoil observables; see Appendix B) and are used for the background-density training and the expected-sensitivity calculation described below.

The use of CMS open data in research brings specific challenges that have been documented by the CMS open-data team [14]. In particular, the authors note that the complexity of the CMS software (CMSSW) and the intrinsic complexity of the data format pose a significant hurdle for external users, and they emphasise the importance of providing full analysis benchmarks and simplified data formats such as MiniAOD and NanoAOD to improve usability [14]. The present analysis addresses these challenges by reading the HTMHT MiniAOD files directly with `uproot` and `awkward`, avoiding the need to run the full CMSSW analysis chain, and by recording per-file processing counts to verify the completeness of the dataset traversal. The trigger decision is reconstructed from the `filterLabels_` branch using a custom decoder, following the guidance of the open-data documentation for 2015 MiniAOD files whose `pathNames_` payload is empty [14]. The luminosity information is taken from the validated Run 2015D recorded value, following the open-data luminosity guidance [13].

2.2 Signal generation

Dark-matter signal samples are generated with MadGraph5_aMC@NLO, showered and hadronised with Pythia8, and passed through the CMS card of a parameterised Delphes detector simulation [16, 17, 18], with jets reconstructed using the anti-k_T algorithm [19]. The anti-k_T algorithm was chosen for its soft-resilient jet boundaries: as shown in the original paper [19], its jets are conical with radius R for isolated hard particles, and their boundaries are insensitive to soft radiation while remaining flexible to hard substructure. This property suppresses the back-reaction and area fluctuations that would otherwise smear the jet momenta entering the hadronic-Z reconstruction, and it is the algorithm adopted by the CMS and ATLAS collaborations for Run 2 analyses. Dark-matter production is modelled with the `DMsimp_s_spin1` simplified-model UFO implementation of an s-channel spin-1 mediator coupling the dark and visible sectors [1, 2], for the process $pp \to Z\bar{\chi}\chi$ with $Z \to jj$. Three benchmark points, listed in Table 1, are retained for the final analysis; each is generated with 5,000 events and a fixed MadGraph5 run-card seed (110020, 150200, and 101500, respectively) to ensure reproducibility, and the leading-order cross sections are taken from the generated LHE event samples. A fourth, light-scalar-mediator benchmark was generated but is excluded: the hadronic-Z reconstruction does not recover a usable dijet resonance for this point (the median closest-to-m_Z dijet mass remains at 145–178 GeV, far from $m_Z \simeq 91.2$ GeV).

The three retained benchmarks follow the benchmark framework recommended by the ATLAS/CMS Dark Matter Forum [1], which codified the s-channel spin-1 mediator models with vector and axial-vector couplings together

with a recommended grid in the (m_χ, M_{med}) plane for early Run 2 searches. In particular, the axial points $(m_\chi, M_{\text{med}}) = (10, 20)$ GeV and $(50, 200)$ GeV lie in the threshold and on-shell regions of the Forum's parameter grid, respectively, while the vector point $(1, 500)$ GeV is an on-shell benchmark with a heavy mediator. The Forum's parameter scans further showed that the shapes of the $E_{\text{T}}^{\text{miss}}$ and boson p_T distributions are insensitive to the coupling choices g_q and g_χ within the scanned range, with only the production cross section changing; the kinematic emphasis of the present study on the hadronic-Z reconstruction and extra-jet topology is therefore expected to hold across the recommended coupling choices. Our use of the `DMsimp` implementation follows the Forum's recommendation of a standardised generator implementation for these models, at leading-order accuracy for the $V + E_{\text{T}}^{\text{miss}}$ final states [2, 1], which is appropriate for a projected-sensitivity study of this type.

The leading-order cross sections used for signal normalisation should be interpreted in light of the higher-order QCD predictions for these s-channel mediator models [2]. The DMsimp implementation was validated and made available at next-to-leading order (NLO) in QCD within the FeynRules/MadGraph5_aMC@NLO framework [2], and the NLO corrections to DM pair production in association with jets were found to be significant for light mediators and light DM, with K factors up to ∼1.8 for $(m_Y, m_X) = (10, 1)$ GeV, decreasing to $K \sim 1.1$–1.2 for heavy mediators and heavy DM [2]. For the benchmark points retained here, the axial $(10, 20)$ GeV and $(50, 200)$ GeV points and the vector $(1, 500)$ GeV point, the NLO corrections are expected to be moderate ($K \sim 1.1$–1.4), and the leading-order cross sections quoted in Table 1 should be regarded as lower estimates of the true production rates. The NLO study also showed that the shapes of the $E_{\text{T}}^{\text{miss}}$ and jet p_T distributions are largely preserved at NLO, with the largest corrections in the low-$E_{\text{T}}^{\text{miss}}$ region [2], which is consistent with the kinematic emphasis of the present analysis on the hadronic-Z reconstruction and extra-jet topology rather than on the absolute signal normalisation.

After the offline hadronic mono-Z selection, 116, 487, and 848 events remain for the three benchmarks, respectively. Because DELPHES does not emulate the CMS high-level trigger, an offline trigger proxy reconstructing the logic of the five HLT paths used on data (Appendix A; §5.5) is applied to the simulated samples before scoring; all events surviving the offline selection also satisfy the proxy, i.e. the trigger-proxy efficiency is 100% for each point.

2.3 Z-candidate reconstruction and feature contract

For the data sample the hadronic-Z candidate is formed from the two leading jets, while for simulated signal the candidate is chosen as the dijet pair closest to m_Z among the leading eight jets, consistent with the signal-generation validation. Both choices use the same jet p_T and η thresholds and produce a common 41-branch feature schema, locked via the extractor schema contract so that data and signal share identical feature definitions. Undefined angular observables that arise when no additional jet exists (for example the $\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{jet}_{1,2})$ and extra-jet η features) are preserved as explicit out-of-range sentinel values rather than imputed into the physical continuum; this handling and the final training feature set are described in the methods section.

3 Methods

This section describes the background-density model, the preprocessing of the extracted features, the training procedure and hyperparameters, and the evaluation of per-event densities used in the sensitivity calculation.

3.1 Background density model

The background density is modelled with a conditional flow-matching continuous normalizing flow (CFM-CNF) [7]. The "conditional" refers to the standard conditional flow-matching construction, in which the flow is conditioned on its target data point in the training objective; the model is not conditioned on auxiliary covariates. A continuous normalizing flow [8, 6] maps a base Gaussian $\mathbf{z}_0 \sim \mathcal{N}(0, I_D)$ onto standardised feature vectors $\mathbf{x} \in \mathbb{R}^D$ by integrating a learned time-conditioned vector field

Table 1: Retained dark-matter signal benchmarks. `DMsimp_s_spin1` is an s-channel spin-1 mediator; m_χ and m_{med} are the dark-matter and mediator masses. Cross sections are leading-order MadGraph5 values.

Signal point	Mediator	$(m_\chi,\, m_{\text{med}})$ [GeV]	σ [pb]	N_{sel}
`axial_mx10_mv20`	axial	$(10, 20)$	1.204	116
`axial_mx50_mv200`	axial	$(50, 200)$	1.050	487
`vector_mx1_mv500`	vector	$(1, 500)$	0.544	848

$v_\theta : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$,

$$\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t} = v_\theta(\mathbf{z}(t), t)\,, \qquad \mathbf{z}(1) = \mathbf{x}, \quad \mathbf{z}(0) = \mathbf{z}_0, \tag{1}$$

and the per-event density is obtained from the change of variables along this ODE. In contrast to closed-form normalizing-flow architectures such as the neural spline flows used in previous work on this channel [5, 4], the CNF density is evaluated by numerically integrating Eq. (1), which is computationally more expensive but retains a fully continuous-time density model.

The choice of a CNF over the NSF architecture used in the leptonic predecessor [4] reflects several methodological differences. An NSF evaluates the density in closed form through a composition of invertible rational-quadratic spline coupling layers, which is fast and exact but constrains the model to a fixed, layer-wise transformation. A CNF instead defines the density implicitly through a learned time-dependent vector field, which is more flexible but requires numerical ODE integration and a stochastic divergence-trace estimate for each density evaluation. In the leptonic analysis the NSF score was constructed as a two-hypothesis log-likelihood ratio between a DM flow and a channel-specific SM flow, and the final interpretation used a simultaneous signal-region/validation-region profile-likelihood fit with per-channel normalisation nuisances. Here we instead train a single background-density model and use the per-event negative log-likelihood as a one-sided discriminant, converting the yields above a threshold into an Asimov expected significance [10]. This simplification is appropriate for a projected-sensitivity study in which no unblinding is performed, and it avoids the background-modelling residual that dominated the leptonic analysis's high-$E_{\text{T}}^{\text{miss}}$ tail [4].

The continuous-time formulation adopted here follows the continuous normalizing flow construction introduced in the neural-ODE framework [8], which showed that the change in log density under a continuous transformation is governed by the trace of the vector-field Jacobian rather than the log determinant of a finite transformation. This replaces the cubic-cost determinant computation of discrete flows with a trace that can be estimated in linear time, at the price of numerical ODE integration for both the state and the accumulated log density [8]. The normalizing-flow review of Papamakarios et al. [6] classifies flow architectures into finite compositions of invertible transformations and continuous-time (infinitesimal) flows, and notes that the two families differ fundamentally in how they trade off expressivity, invertibility, and computational cost. The present work sits in the continuous-time family: unlike the finite-composition NSF used in the leptonic predecessor [5, 4], the CNF does not require a partitioning or ordering of the feature dimensions and can be trained by maximum likelihood without the architectural constraints of coupling or autoregressive layers [8, 6]. The review also emphasises that the choice between evaluating the forward or inverse transformation is not a crucial implementation decision for continuous-time flows, since both directions have the same computational complexity [6]; this symmetry is exploited here by evaluating the background density directly through the probability-flow ODE.

The two building blocks of the leptonic predecessor are also worth contrasting explicitly. The NSF architecture of Durkan et al. [5] parameterises the elementwise transformations in coupling or autoregressive layers as monotonic rational-quadratic splines, which are analytically invertible and have a closed-form Jacobian determinant. This provides a powerful but finite-composition flow whose flexibility is

bounded by the number and width of the coupling layers. The present work instead follows the FFJORD construction of Grathwohl et al. [9], which showed that the continuous-time instantaneous change of variables of Chen et al. [8] can be scaled to high-dimensional problems by estimating the Jacobian trace with Hutchinson's stochastic estimator [9]. Because the trace estimator has a cost of $\mathcal{O}(D)$ per evaluation rather than the $\mathcal{O}(D^2)$ cost of an exact trace or the $\mathcal{O}(D^3)$ cost of a Jacobian determinant, the CFM-CNF is free to use an unrestricted network architecture for the vector field. This is the same design choice adopted here: the residual-MLP vector field described below is not constrained by coupling-layer or autoregressive structure, and the density is evaluated with a single Rademacher projection per ODE evaluation, as in FFJORD [9]. The learned residual-MLP vector field can therefore capture correlations between all feature dimensions without the partitioning or ordering requirements of the NSF coupling layers used in the leptonic analysis [5, 4].

The vector field is trained with the simulation-free conditional flow-matching objective [7], which regresses the model onto the velocity of the straight-line probability path between a Gaussian sample $\mathbf{x}_0$ and a data point $\mathbf{x}_1$,

$$\begin{aligned}\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\Big[\|v_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\Big], \\ t \sim U[0,1],\ \mathbf{x}_0 \sim \mathcal{N}(0, I),\ \mathbf{x}_1 \sim p, \\ \mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1.\end{aligned} \tag{2}$$

so the training loss is a mean-squared-error regression on target velocities, evaluated per mini-batch. At inference, the logarithmic density is

$$\log p(\mathbf{x}) = \log p_0(\mathbf{z}_0) - \int_0^1 \mathrm{Tr}\left(\frac{\partial v_\theta}{\partial \mathbf{z}}\right) \mathrm{d}t, \tag{3}$$

where p_0 is the unit Gaussian density and the divergence trace is estimated with Hutchinson's stochastic estimator using a single Rademacher projection per ODE evaluation [9]. The probability-flow ODE is solved with a fixed-step RK4 solver (step size 0.05, absolute and relative tolerances 10^{-3}); density values are accumulated in batches of 16 events.

3.2 Feature contract and preprocessing

From the shared 41-branch data/signal schema, 32 features are used for training. The remaining nine columns are excluded because they are either audit bookkeeping columns (`source_file_idx`, `source_entry`), trigger or veto flags (`trigger_pass`, `btag_pass`, `n_bjets`), or columns whose data/MC semantics are known to differ (`max_btag_csvv2` and `btag_wp_medium`, because Delphes b-tagging is a binary flag rather than a continuous CSVv2 discriminator; and `n_vertices` and `rho`, because the signal simulation does not include a pileup overlay while the data carries real pileup conditions).

Undefined extra-jet angular observables—`jet1_eta`, `jet2_eta`, `dphi_met_jet1`, and `dphi_met_jet2`, which are NaN when the corresponding extra jet does not exist—are mapped to an explicit out-of-range sentinel value of -999 before standardisation, rather than median-imputed into the physical continuum, which would fabricate a spurious point mass. This sentinel handling is applied identically to background, validation, and signal samples. Any remaining non-finite values are filled with training-set medians, and all features are standardised using the training-set mean and standard deviation; the saved preprocessor is applied unchanged to all downstream samples. Events are split 80/20 into training and validation subsets using a fixed random seed (20260730), and the split indices are persisted so that all density and sensitivity evaluations use the exact same held-out validation split.

3.3 Architecture and training

The vector field is a residual MLP: a sinusoidal time embedding of dimension 32 is concatenated to the input and mapped by an input layer to a hidden width of 192; the result passes through four residual blocks, each consisting of `LayerNorm`, `SiLU`, a 192-wide linear layer, `SiLU`, `Dropout(0.05)`, and a second 192-wide linear layer, followed by a residual connection; the output head is `LayerNorm`, `SiLU`, and a final linear layer back to the feature dimension D.

Training minimises Eq. (2) with `AdamW` (`lr` =

2×10^{-4}, weight decay 10^{-5}), a batch size of 2048, and gradient clipping at norm 5.0, for at most 500 epochs. Because the flow-matching loss is not a direct density objective, convergence is monitored every five epochs on the held-out validation negative log-likelihood from Eq. (3), with early stopping after six consecutive checks without improvement. The baseline fit used $D = 32$ features on 1,439,523 selected background events (1,151,618 training, 287,905 validation) and its best checkpoint is at epoch 245 (validation NLL ≈ -64.97). The ablation fit is identical except that the six extra-jet features (`jet1_pt`, `jet1_eta`, `jet2_pt`, `jet2_eta`, `dphi_met_jet1`, `dphi_met_jet2`) are removed, leaving $D = 26$ features; its best checkpoint is at epoch 165 (validation NLL ≈ -43.63). The smaller validation-NLL magnitude of the ablation fit is expected from the lower-dimensional density and does not indicate worse training.

3.4 Inference and sensitivity evaluation

Per-event negative log-likelihoods (NLL) from Eq. (3) serve as the discriminant between background and signal. Sensible comparison requires the background NLL distribution to be evaluated on events not used for training; we therefore score only the persisted held-out validation split and reweight the scored events back to the full selected background population, as detailed in Section 5. Signal samples are scored only after the offline trigger proxy described in Section 5.5. The expected-significance calculation, the minimum-background-yield working point, and the diagnostic closure studies are described in Sections 5 and 4.

4 Validation

This section summarises the validation studies used to verify the background density model and to demonstrate that the sentinel-imputation fix removed the dominant imputation artifact without introducing regressions in other features.

4.1 Closure plots (representative set)

Closure assesses how faithfully the trained flow reproduces the held-out validation data distribution, which is never used for training. We compare histograms of the real validation events with samples drawn from the fitted flow after inverse standardisation, using the same binning in both cases. Representative panels are shown in Figure 1, and the full per-feature closure set is provided in Appendix D.

The hadronic-Z mass panel (top left of Figure 1) shows that the flow reproduces the Z-peak structure inside the 70–110 GeV window and tracks the validation data across the full displayed range. A localised excess is visible in the flow sample near $m_{jj} \approx 91.6$ GeV (density ~ 0.038 versus ~ 0.024 in the held-out data, a $\sim$50% discrepancy in that bin), consistent with the continuous-density model slightly over-concentrating probability at the sharp Z pole; this is a sub-bin-shape effect rather than an order-of-magnitude mismatch. The $E_{\mathrm{T}}^{\mathrm{miss}}$ panel (top right) and the ΔR_{Zjj} panel (bottom right) confirm good agreement for the missing-transverse-momentum and dijet-angular observables that drive the mono-Z discrimination, while the H_T panel (bottom left) verifies closure for the overall event scale. In all panels the model histogram follows the data within statistical fluctuations of the validation sample, with no order-of-magnitude mismatches.

Figure 2 shows additional closure panels for the reconstructed Z transverse momentum, the hadronic recoil p_T, the $E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{H_T}$ ratio, and $|m_{jj} - m_Z|$. Good closure for these derived observables indicates that the flow captures not only the individual jet distributions but also the correlations encoded in the Z kinematics and the recoil system, which are the primary inputs to the per-event density score used in Section 5.

4.2 Before/after imputation-fix comparison

Auditing the original preprocessing revealed that four undefined extra-jet angular features—`jet1_eta`, `jet2_eta`, `dphi_met_jet1`, `dphi_met_jet2`—were median-imputed when no extra jet exists. Because

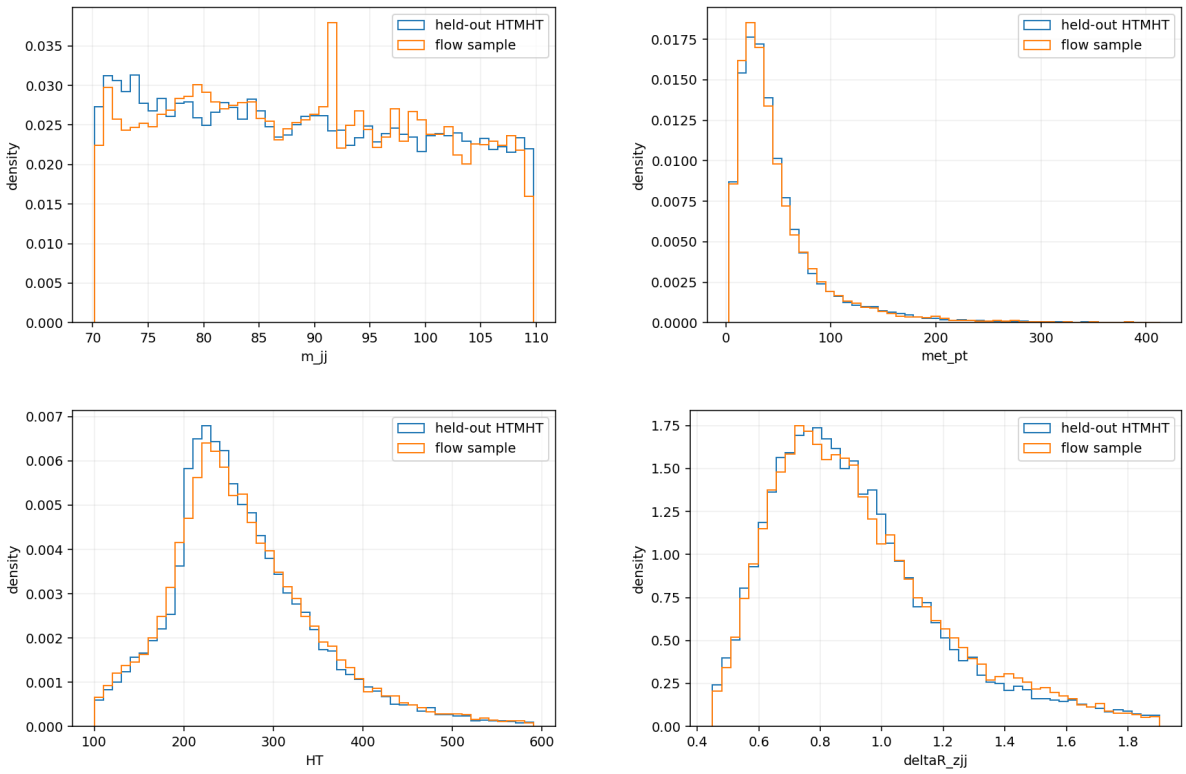


Figure 1: Representative closure plots comparing held-out validation data (filled histogram) to samples drawn from the fitted flow (step histogram). Top left: hadronic-Z mass m_{jj}. Top right: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution. Bottom left: event scalar H_T. Bottom right: ΔR_{Zjj}. The full per-feature closure set is provided in Appendix D.

these observables are physically undefined for events without additional jets, the median fill collapsed the "no-extra-jet" state into a spurious point mass inside the physical continuum, producing artificial spike-blending artifacts in the closure plots. The post-fix pipeline maps these undefined values to an explicit out-of-range sentinel value of -999 (`nan_sentinel_value` $= -999.0$), applied identically to background, validation, and signal samples before standardisation.

Figure 3 shows the post-fix closure panels for all four fixed features. Each panel now exhibits two cleanly separated populations: a sharp isolated bar at the sentinel value (the "does-not-exist" state) and a smooth histogram over physical values, both reproduced by the fitted flow with reasonable closure. In particular, the `dphi_met_jet2` panel previously showed a flow spike at density $\sim$10.9 against a data value of $\sim$5.45—an order-of-magnitude mismatch from the median fill—and now shows a substantially reduced but still visible sentinel-bar mismatch (flow density ~ 0.023 versus data ~ 0.033, a $\sim$30% residual discrepancy) together with reasonable agreement over the physical $\Delta\phi$ range. A comparable residual sentinel-bar mismatch is visible in the `jet2_eta` panel, while `jet1_eta` and `dphi_met_jet1` show closer agreement.

4.3 Training convergence and validation NLL

Because the flow-matching loss is a regression objective rather than a direct density loss, convergence is monitored on the held-out validation negative log-likelihood obtained from the ODE density evaluation (Eq. (3)). Figure 4 shows the validation-NLL curves for the baseline ($D = 32$) and ablation ($D = 26$) fits. Both descend smoothly and plateau cleanly over the final $\sim$30–50 epochs, with no evidence that training stopped while the model was still improving. The baseline selects its best checkpoint at epoch 245 (validation NLL ≈ -64.97) and the ablation at epoch 165 (validation NLL

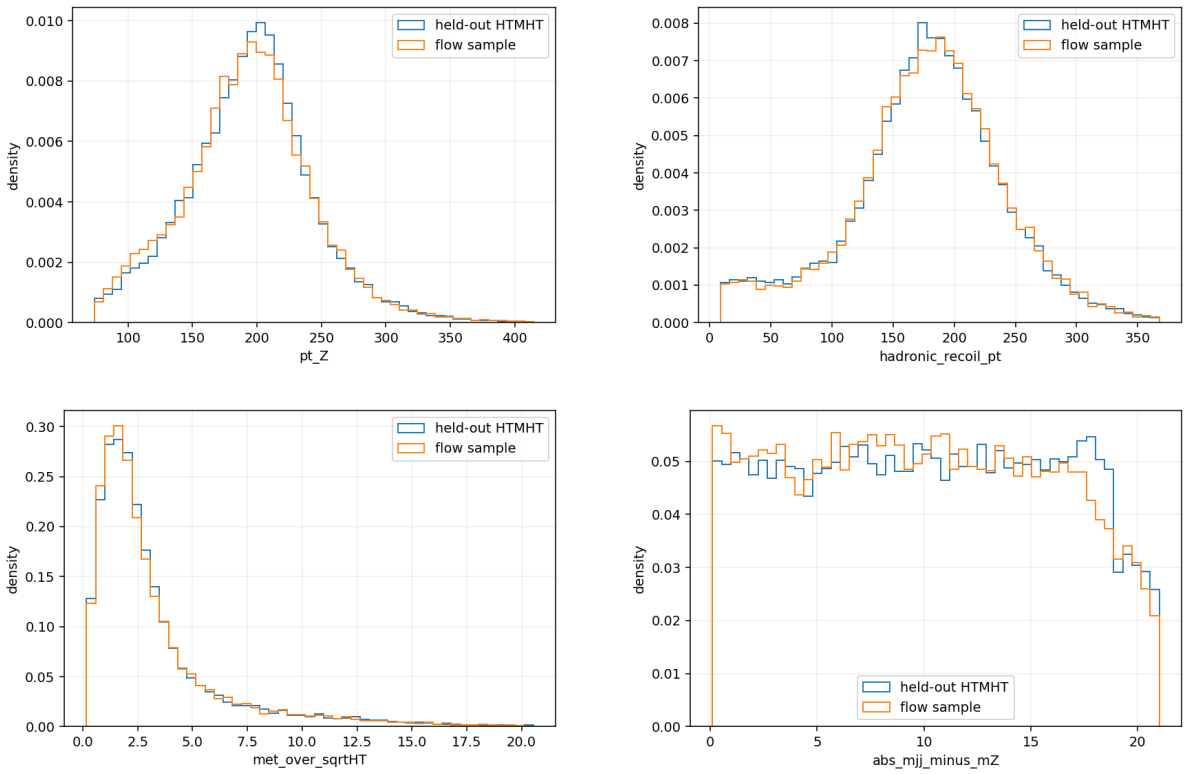


Figure 2: Additional closure panels for kinematic observables entering the density score. Top left: reconstructed Z transverse momentum $p_{T,Z}$. Top right: hadronic recoil p_T. Bottom left: $E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{H_T}$. Bottom right: $|m_{jj} - m_Z|$. The full per-feature closure set is provided in Appendix D.

≈ -43.63). The smaller magnitude of the ablation value is expected for a lower-dimensional density and does not indicate worse training; the closure checks above confirm that the ablation model remains of comparable quality on its retained features.

4.4 Known residual limitation: jet p_T delta-plus-continuum

A remaining methodological caveat is the inability of a continuous normalizing flow to exactly represent a discrete point mass (the $p_T = 0$ spike for absent extra jets) combined with a continuous distribution for present jets within the same feature. The sentinel mapping resolves the undefined-angular features, but the extra-jet p_T features carry a legitimate physical $p_T = 0$ sentinel (zero-fill) alongside a continuum, and the CNF necessarily smooths the point mass. As shown in Figure 5, the closure for `jet1_pt` and `jet2_pt` retains a mild undershoot of the zero-spike height (for example, flow density ~ 0.071 versus data ~ 0.142 for `jet1_pt`). This is a structural limitation of continuous-density models for this feature type, not a training defect; it is stated explicitly as a caveat in Section 8.

5 Sensitivity methodology

This section documents the statistical and procedural choices used to convert the per-event background-density scores into the expected significances reported in Section 6. The per-event negative log-likelihood (NLL) from Eq. (3) serves as the discriminant: signal-like events are expected to populate the high-NLL tail of the background distribution, and a threshold on this score defines a counting region in which the expected signal and background yields S and B are evaluated.

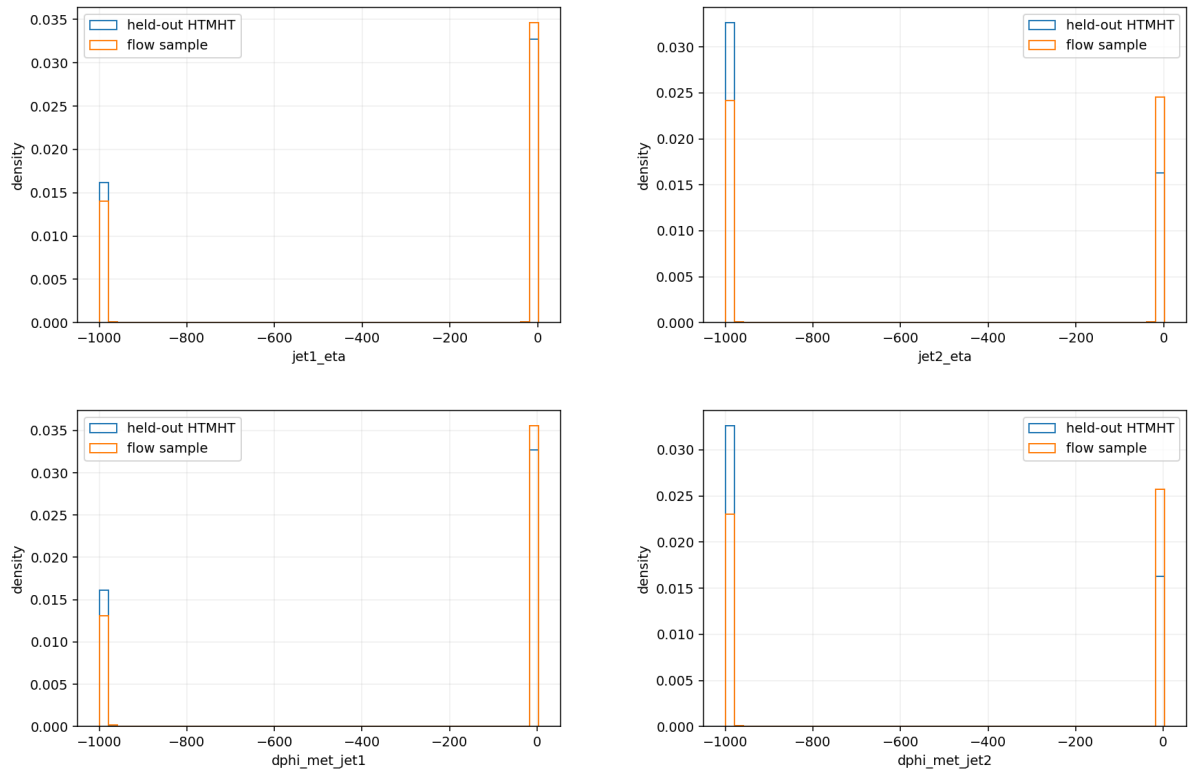


Figure 3: Post-fix closure panels for the four undefined extra-jet angular features. Top left: `jet1_eta`. Top right: `jet2_eta`. Bottom left: `dphi_met_jet1`. Bottom right: `dphi_met_jet2`. Each panel shows a distinct sentinel bar at -999 (the "no extra jet" state) plus the physical continuum, both reproduced by the flow.

5.1 Asimov significance

Expected significances are computed with the Asimov counting formula [10],

$$Z_A = \sqrt{2\left[(S+B)\ln\left(1+\frac{S}{B}\right) - S\right]}, \quad (4)$$

where S and B are the weighted signal and background yields above the chosen NLL threshold. This is a projected, expected significance: it is evaluated from the simulated signal and the background-density model alone, without comparing real data to a background prediction, and therefore corresponds to the median significance that would be obtained in an ensemble of background-only pseudo-experiments.

The use of the Asimov counting formula follows the asymptotic framework of Cowan et al. [10], which derived closed-form sampling distributions for profile-likelihood-ratio test statistics and introduced the Asimov data set as a tool for estimating the median expected significance without Monte Carlo simulation. In that framework, the Asimov data set is constructed with all parameters equal to their expectation values, and the median discovery significance for a signal hypothesis is approximated by $\mathrm{med}[Z_0|\mu'] = \sqrt{q_{0,A}}$ [10]. The counting formula of Eq. (4) is the binomial-expansion limit of this result for a single-bin counting experiment with negligible background uncertainty; it corresponds to the median significance one would obtain from the profile-likelihood-ratio statistic q_0 in the large-sample limit [10]. The full formal treatment with nuisance parameters would generalise Eq. (4) to a profile likelihood over the background-density model and the signal-yield normalisation, at the cost of substantially greater computational complexity. For a projected-sensitivity study the counting approximation is a well-motivated simplification, and the $B \geq 20$ working-point floor ensures that the Gaussian approximations underlying the asymptotic formula remain reasonable, as recommended in [10].

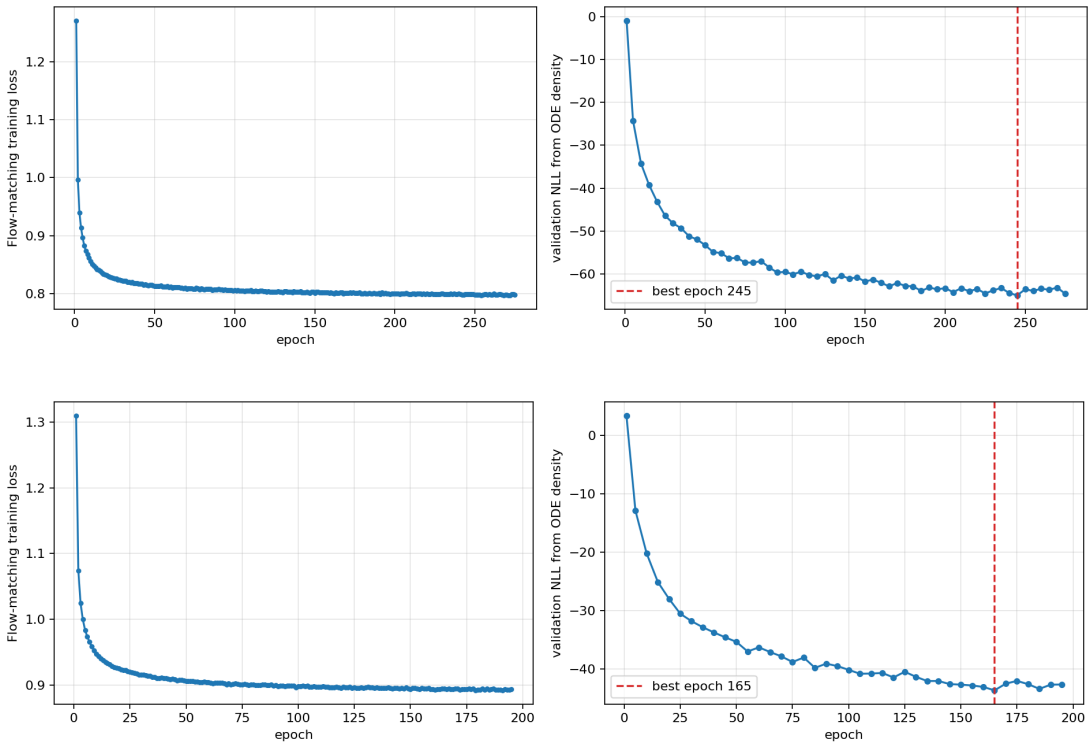


Figure 4: Validation negative-log-likelihood curves for the baseline (top) and extra-jet-drop ablation (bottom) fits. The red dashed line marks the best epoch selected by early stopping.

5.2 Signal-yield normalization

The expected signal yield above a threshold is obtained by normalising the number of simulated signal events that pass the threshold to the integrated luminosity. For each benchmark, the per-event weight is

$$w_{\mathrm{sig}} = \frac{\sigma\,\mathcal{L}\,\epsilon_{\mathrm{trig}}}{N_{\mathrm{gen}}}, \tag{5}$$

where σ is the leading-order cross section, $\mathcal{L}$ is the integrated luminosity, ϵ_{trig} is the signal-side trigger-proxy efficiency, and N_{gen} is the number of generated signal events used for normalisation. The signal yield above a threshold is then $S = w_{\mathrm{sig}}\, N_{\mathrm{sig}}^{>\mathrm{cut}}$, where $N_{\mathrm{sig}}^{>\mathrm{cut}}$ is the number of simulated signal events whose NLL exceeds the threshold. The cross sections and generated-event counts are taken from the signal-generation manifest (Table 1), and the luminosity is the validated Run 2015D recorded value of 2.256382381 fb^{-1} [12, 13].

5.3 Discriminant scan and working-point selection

The NLL threshold is scanned over the combined set of background and signal scores, and at each threshold the yields S and B are evaluated and Z_A is computed from Eq. (4). During early audits we found that reporting the unconstrained argmax of this scan can select thresholds with vanishing background yields (raw counts as small as 3–12 events in the full selected population), producing unstable and biased headline numbers because the optimisation effectively performs a look-elsewhere search on stochastic background fluctuations in the tail. To avoid this pitfall, the primary reported working point is the best threshold satisfying a minimum weighted background yield of $B \geq 20$; the unconstrained optimum is still recorded for diagnostic comparison but is not used as the headline result. The $B \geq 20$ floor ensures the reported Z_A is evaluated at a working point with sufficient background statistics that the Gaussian approximations underlying the asymptotic formula are reasonable and that the optimisation is not dominated by rare-tail

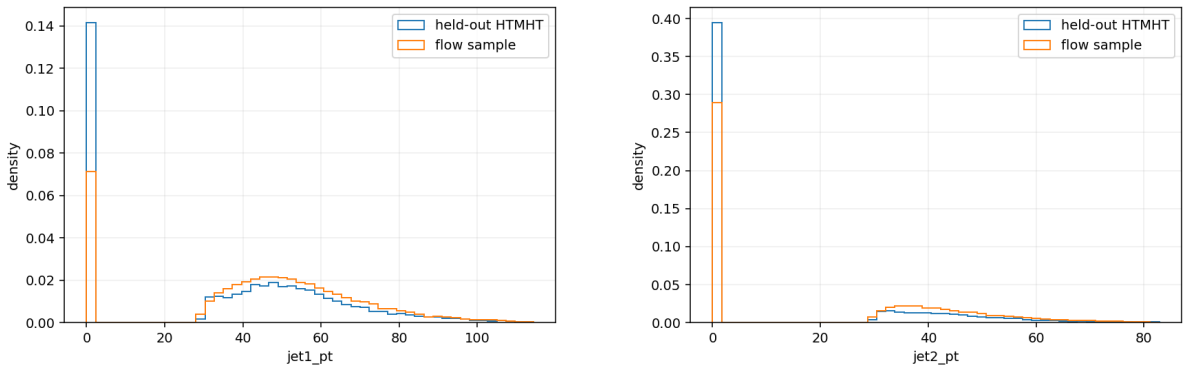


Figure 5: Closure panels for the extra-jet p_T features `jet1_pt` (left) and `jet2_pt` (right), illustrating the residual delta-plus-continuum limitation: the flow smooths the physical $p_T = 0$ sentinel spike, undershooting its height while preserving the continuum.

sampling noise. As a diagnostic, the relative Poisson uncertainty on the background yield, $1/\sqrt{B}$, is also recorded at the selected working point.

5.4 Held-out-only background scoring with reweighting

To avoid in-sample scoring bias, the background NLL distribution is evaluated only on the held-out validation split, which is never used for training. The scored events are then reweighted back to the full selected background population via

$$w_{\text{pop}} = \frac{N_{\text{selected}}}{N_{\text{scored}}}, \tag{6}$$

where $N_{\text{selected}} = 1{,}439{,}523$ is the post-selection population and $N_{\text{scored}} = 287{,}905$ is the held-out validation count in the baseline run. The background yield above a threshold is $B = w_{\text{pop}}\, N_{\text{bg}}^{>\text{cut}}$, where $N_{\text{bg}}^{>\text{cut}}$ is the number of held-out background events whose NLL exceeds the threshold. This provides a held-out estimate of the full-population background yield while avoiding the positive bias introduced by scoring training-set rows.

5.5 Signal-side trigger proxy

Because Delphes does not emulate the CMS HLT turn-on, the signal samples are processed with an offline trigger proxy approximating the five HLT paths used on the recorded background. The adopted proxy applied to signal is $H_T > 350$ GeV OR $E_{\text{T}}^{\text{miss}} > 120$ GeV OR ($E_{\text{T}}^{\text{miss}} > 90$ GeV and leading jet $p_T > 80$ GeV). After the hadronic mono-Z offline selection, 116, 487, and 848 events remain for the axial 10/20, axial 50/200, and vector 1/500 benchmarks, respectively; all satisfy the proxy, so the trigger-proxy efficiency is 100% for each point and $\epsilon_{\text{trig}} = 1$ in Eq. (5). This proxy is conservative but does not replace a full HLT efficiency measurement, which would require a data-driven turn-on parameterisation.

6 Results

The following presents the post-fix expected-sensitivity results and supporting robustness checks. The study is an expected-sensitivity calculation: no unblinding of a signal region against a background prediction was performed, and all significances reported below are projections computed using simulated signal together with a background density model trained on held-out HTMHT data.

The background model is a conditional flow-matching continuous normalizing flow trained on the CMS Run 2015D HTMHT MINIAOD open-data sample [11]. After the hadronic mono-Z preselection the local extraction selects 1,439,523 events from 20,679,437 raw events in 444 MINIAOD files (selection efficiency 6.96%). The integrated luminosity used for normalization is 2.256382381 fb^{-1}, taken from the validated Run2015D recorded luminosity provenance [12, 13].

Table 2: Summary of the processed MINIAOD dataset used for extraction.

Item	Value
Processed MINIAOD files	444
Total raw events	20,679,437
Selected events (post-selection)	1,439,523
Selection efficiency	6.96%

Full per-file counts (entries and selected events) are provided in Appendix C and summarized above.

Significances are computed with the Asimov counting formula (Eq. (4)) described in Section 5, where S and B denote the weighted signal and background yields above a threshold on the negative-log-likelihood score. To avoid reporting optimised cut points driven by very small background populations, the reported working point is chosen as the best cut satisfying $B \geq 20$; the unconstrained optimum (with $B \ll 20$) is retained for diagnostic comparison but is not used as the headline result. Background yields are estimated from the held-out 20% validation split and reweighted to the full selected HTMHT population to avoid in-sample scoring bias, and the offline HT/MET trigger proxy described in §5.5 is applied to simulated signal prior to scoring.

Table 3 reports the post-fix baseline numbers. The light axial benchmark ($m_\chi = 10$ GeV, $m_{\rm med} = 20$ GeV) yields an expected significance of 2.89σ, while the heavier axial ($m_\chi = 50$ GeV, $m_{\rm med} = 200$ GeV) and the vector ($m_\chi = 1$ GeV, $m_{\rm med} = 500$ GeV) benchmarks reach 7.62σ and 7.41σ, respectively. The two heavier benchmarks therefore provide a clear projected discovery-level sensitivity in this channel, whereas the light axial point is only moderately sensitive. In all three cases the reported working point sits at $B \simeq 25$ weighted background events, corresponding to a relative Poisson background uncertainty of $1/\sqrt{B} \simeq 0.20$; the signal efficiency at the working point is 0.20–0.25, indicating that the selected region retains a substantial fraction of the signal while keeping the background yield at the $B \geq 20$ floor.

The same analysis also records the unconstrained optima (which occur at very small background yields) for comparison; these diagnostic values are significantly larger (e.g. $Z_A \approx 4.96, 10.81, 10.70$ for the three benchmarks) but are not used as the principal results due to their instability under validation-sample fluctuations.

6.1 Comparison with the leptonic NSF predecessor

It is instructive to compare the present hadronic flow-matching analysis with the leptonic NSF study on the same Run 2015D open-data era [4]. The two analyses differ in three principal respects: the Z decay channel, the density-modelling architecture, and the statistical interpretation.

First, the leptonic analysis selected $Z \to \ell^+\ell^-$ events with a dilepton-mass window $60 < m_{\ell\ell} < 120$ GeV, a signal region defined by $E_{\rm T}^{\rm miss} \geq 50$ GeV, $|\Delta\phi(E_{\rm T}^{\rm miss}, Z)| > 2.5$, and $n_{\rm jets} \leq 1$, and trained the SM background flows on a control region with $E_{\rm T}^{\rm miss} < 50$ GeV. The present hadronic analysis instead reconstructs the Z from a dijet pair with $70 < m_{jj} < 110$ GeV and $\Delta R_{jj} < 2.0$, and trains the background density directly on the full selected HTMHT population. The hadronic channel therefore exploits the larger $Z \to q\bar{q}$ branching fraction at the cost of a substantially more complex background, which is precisely the regime in which a flexible learned density model is most valuable.

Second, the density models differ fundamentally. The leptonic analysis used closed-form NSFs with analytic densities and a two-hypothesis log-likelihood-ratio score, whereas the present work uses a continuous flow-matching CNF with an ODE-integrated density and a one-sided background-NLL discriminant. The CNF's continuous-time formulation avoids the architectural constraints of invertible coupling layers and is trained with a simulation-free regression objective, but requires numerical ODE integration and a stochastic divergence-trace estimate at inference [7, 8, 9].

Third, the statistical treatment differs. The leptonic analysis performed a simultaneous signal-region/validation-region profile-likelihood fit with per-channel normalisation nuisances and reported observed and ex-

Table 3: Primary post-fix expected sensitivity for the 32-feature flow-matching model. The signal trigger proxy is applied before scoring; the reported working point requires $B \geq 20$. The final column is the relative Poisson uncertainty on the background yield, $1/\sqrt{B}$.

Signal point	Mediator benchmark	S	B	Best cut	Signal eff.	Z_A	$1/\sqrt{B}$
axial_mx10_mv20	axial, $m_\chi = 10$ GeV, $m_{\text{med}} = 20$ GeV	15.76	25.00	77.05	0.250	2.89	0.20
axial_mx50_mv200	axial, $m_\chi = 50$ GeV, $m_{\text{med}} = 200$ GeV	46.90	25.00	67.43	0.203	7.62	0.20
vector_mx1_mv500	vector, $m_\chi = 1$ GeV, $m_{\text{med}} = 500$ GeV	45.38	25.00	64.97	0.218	7.41	0.20

pected 95% CL upper limits on the signal strength. That fit was dominated by a residual background-modelling discrepancy in the high-$E_{\text{T}}^{\text{miss}}$ tail ($E_{\text{T}}^{\text{miss}} \geq 100$ GeV), which inflated the observed-to-expected limit ratio by a factor of roughly 7–12 and prevented a clean interpretation of the fitted signal strength [4]. The present analysis instead reports projected Asimov expected significances [10] computed from a held-out background score distribution, with a $B \geq 20$ working-point floor to avoid the look-elsewhere bias of an unconstrained discriminant scan. This choice reflects the projected-sensitivity nature of the present study and avoids the tail-modelling residual that limited the leptonic interpretation.

7 Robustness: extra-jet feature ablation

To assess the contribution of extra-jet kinematics to discrimination, an ablation study was performed in which six padded extra-jet columns were removed from the training feature set: the leading and subleading extra-jet p_T and η (`jet1_pt`, `jet1_eta`, `jet2_pt`, `jet2_eta`) and the two corresponding $\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{jet})$ variables (`dphi_met_jet1`, `dphi_met_jet2`). This reduces the feature dimension from $D = 32$ to $D = 26$. The ablation model was trained from scratch with the same architecture, hyperparameters, and held-out scoring procedure as the baseline; its best checkpoint is at epoch 165 (validation NLL ≈ -43.63), and the closure validation (Section 4) confirms that the ablation model is not undertrained—every retained feature closes as well as in the baseline.

Table 4 shows that removing detailed extra-jet information reduces the expected significance at all three benchmark points, with relative drops of 53–71%. The largest impact is on the light axial benchmark ($m_\chi = 10$ GeV, $m_{\text{med}} = 20$ GeV), whose significance collapses from 2.89σ to 0.83σ; in the ablation model the $B \geq 20$ floor forces the working point to a higher background yield ($B \simeq 50$) at a lower signal efficiency, indicating that the signal NLL distribution has lost separation from the background. This is visible in the signal-versus-background NLL overlap plot (Figure 7): in the ablation model the `axial_mx10_mv20` signal NLL distribution sits almost entirely on top of the background peak, barely distinguishable, whereas the other two benchmarks retain a visible (if reduced) tail past the background peak.

This behavior supports the interpretation that the extra-jet kinematics carry genuine discriminating information for the hadronic mono-Z topology, plausibly arising from initial-state-radiation jet topology differences between mono-Z dark-matter production and generic multijet HTMHT background, rather than being an artifact of the (now-fixed) sentinel imputation. The count feature `n_extra_jets` alone is insufficient to recover this discriminating power; the detailed per-jet kinematics matter.

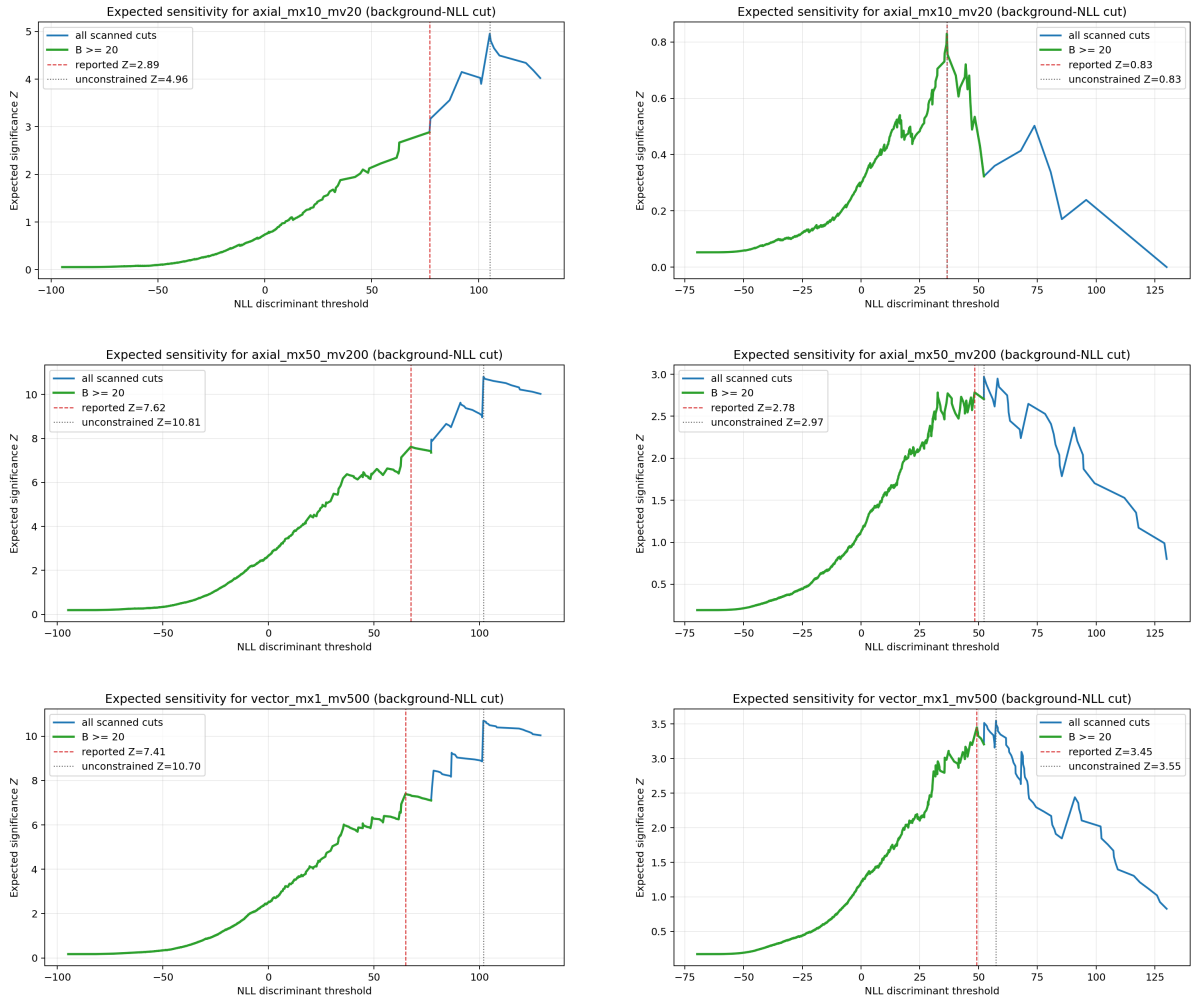


Figure 6: Expected-significance scans versus the NLL discriminant threshold for the three benchmark points. Each row pairs the baseline (32-feature) model (left) with the extra-jet-drop ablation (26-feature) model (right) for one signal point: axial $m_\chi = 10/m_{\text{med}} = 20$ GeV (top row), axial $m_\chi = 50/m_{\text{med}} = 200$ GeV (middle row), and vector $m_\chi = 1/m_{\text{med}} = 500$ GeV (bottom row). The green curve marks the $B \geq 20$ constrained scan used for the reported working point, the red dashed line the reported optimum, and the grey dotted line the unconstrained optimum.

8 Limitations

Several caveats should be emphasized when interpreting these results.

Projected sensitivity, not observation. The study reports projected, expected significances rather than observed excesses. The background density is trained on real selected HTMHT events, but S, B, and Z_A are obtained by combining the held-out background score distribution with simulated signal. No unblinding of a signal region against a background prediction was performed at any point.

Trigger proxy. The signal trigger treatment is an offline proxy implemented to approximate the dominant HT/MET/monojet HLT logic in the recorded dataset. Delphes does not emulate the full HLT turn-on, and a complete experimental reinterpretation would require trigger-efficiency parameterization and detector systematic uncertainties [18]. The proxy efficiency is 100% for all three benchmarks after offline selection, which is likely an overestimate of the true trigger acceptance for the lighter benchmarks.

No background systematic uncertainties. Only the statistical (Poisson) background uncertainty, $1/\sqrt{B}$, is propagated. No systematic

Table 4: Effect of dropping the extra-jet kinematic features. The ablation working point is selected with the same $B \geq 20$ floor as the baseline.

	Baseline ($D = 32$)			Ablation ($D = 26$)			
Signal point	S	B	Z_A	S	B	Z_A	Rel. drop
axial 10/20	15.76	25.00	2.89	5.98	50.00	0.83	71.3%
axial 50/200	46.90	25.00	7.62	15.16	25.00	2.78	63.5%
vector 1/500	45.38	25.00	7.41	19.13	25.00	3.45	53.5%

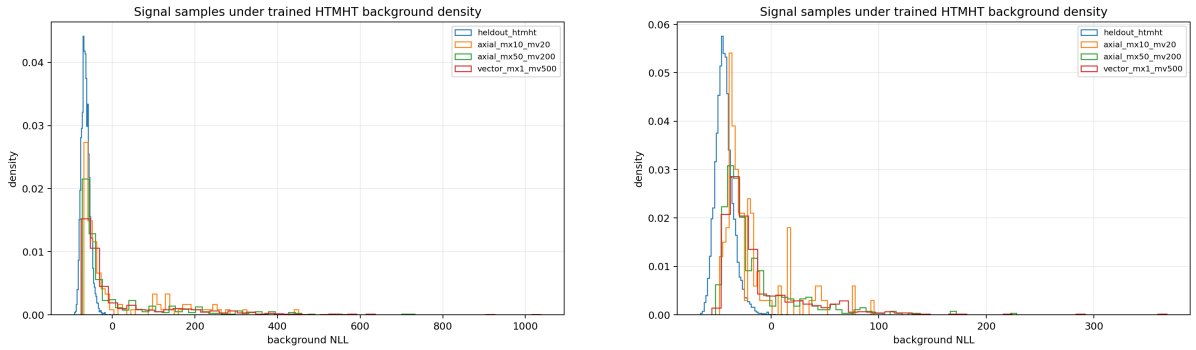


Figure 7: Signal-versus-background NLL distributions for the baseline (left, $D = 32$) and extra-jet-drop ablation (right, $D = 26$) models. In the ablation model the `axial_mx10_mv20` signal distribution (orange) overlaps almost entirely with the background peak (blue), explaining its collapse to sub-1σ significance.

uncertainty on the background density model—for example from the choice of training features, the ODE solver tolerance, or the finite validation sample—is included. This is an expected-sensitivity study, not a full profile-likelihood limit-setting analysis, and the reported Z_A values should be interpreted accordingly.

Signal simulation gaps. The Delphes signal simulation does not include a pileup overlay, so the `n_vertices` and `rho` features are zero placeholders for signal and are excluded from training. The b-tag discriminator semantics also differ between Delphes (binary flag) and data (continuous CSVv2 score), so the b-tag-related columns are excluded from the feature set. A full treatment would require pileup overlay and a Delphes card with a compatible b-tag discriminator output.

Delphes fast-simulation scope. The signal samples rely on the Delphes 3 fast simulation [18], which is designed for phenomenological studies rather than for reproducing the full detector response of a specific experiment. As documented in the Delphes 3 paper, the framework parameterises the detector response with smearing functions and a simplified particle-flow reconstruction, and it does not simulate fake rates for leptons and photons or a full trigger emulation [18]. The Delphes validation against CMS and ATLAS shows good agreement for jet and $E_{\mathrm{T}}^{\mathrm{miss}}$ resolutions at $p_T > 30$ GeV [18], which is the regime relevant to the hadronic mono-Z selection used here. Nevertheless, the absence of a pileup overlay in the CMS card and the parametric b-tagging are inherent limitations of the fast-simulation approach, and the corresponding features are therefore excluded from the density model rather than treated as mismodelled inputs. This is consistent with the Delphes authors' own statement that the framework is not meant for advanced detector studies, for which more accurate tools are required [18].

Incomplete signal cutflow. Per-stage signal-side cutflow counters (trigger proxy, jet kinematics, m_{jj} window, ΔR, b-veto) were added to the signal-generation notebook but the resulting per-stage CSVs have not been reviewed to diagnose why the `axial_mx10_mv20` benchmark has a lower raw selection efficiency than

the other two points. This diagnosis remains open.

Continuous-flow structural limitation. Continuous normalizing flows cannot represent exact discrete-plus-continuous mixtures (for example, the point mass at zero for padded extra-jet p_T combined with a continuous distribution for present jets). The sentinel-imputation fix applied to undefined angular features removes the largest imputation artifact, but the p_T point-mass limitation remains a structural caveat of the continuous-density approach and would require a discrete-plus-continuous hybrid architecture to resolve fully.

9 Conclusion

We have presented a projected sensitivity study for hadronic mono-Z dark matter production using CMS Run 2015D HTMHT open data corresponding to 2.256382381 fb^{-1}. A conditional flow-matching continuous normalizing flow is trained on 1,439,523 selected background events to model the hadronic mono-Z background density, and per-event negative-log-likelihood scores serve as the discriminant between background and simulated dark-matter signal. The methodology includes sentinel imputation for undefined extra-jet angular features, held-out-only background scoring with population reweighting, a signal-side offline trigger proxy, and a minimum-background-yield ($B \geq 20$) working-point floor to avoid look-elsewhere bias in the NLL-cut scan.

Under this procedure the post-fix baseline analysis yields expected significances of 2.89σ, 7.62σ, and 7.41σ for the three retained simplified-model benchmarks (axial $m_\chi = 10/m_{\text{med}} = 20$, axial $m_\chi = 50/m_{\text{med}} = 200$, and vector $m_\chi = 1/m_{\text{med}} = 500$ GeV). An ablation study that removes the six detailed extra-jet kinematic features reduces the expected significance by 53–71%, demonstrating that extra-jet topology—plausibly from initial-state radiation—carries substantial genuine discriminating power in the hadronic mono-Z channel beyond what the jet multiplicity alone provides.

This work also extends our earlier leptonic mono-Z study [4] in three ways. Methodologically, it replaces the closed-form neural spline flow with a continuous flow-matching normalizing flow, trading exact analytic densities for a more flexible continuous-time model trained with a simulation-free objective. Statistically, it replaces the two-hypothesis likelihood-ratio score and profile-likelihood fit with a single background-density discriminant evaluated on a held-out validation split and converted into an Asimov expected significance, avoiding the high-$E_{\text{T}}^{\text{miss}}$ tail-modelling residual that dominated the leptonic interpretation. Physically, it moves from the clean leptonic Z decay to the higher-rate hadronic channel, where the substantially more complex multijet background makes a flexible learned density model particularly valuable and where the extra-jet topology provides a genuine discriminating handle.

Future work should extend this study to a full CLs/profile-likelihood treatment with background systematic uncertainties, replace the offline trigger proxy with a data-driven HLT efficiency measurement, add pileup overlay to the signal simulation, diagnose the per-stage signal cutflow for the light axial benchmark, and perform a direct head-to-head comparison of the flow-matching and neural-spline-flow density models on the same hadronic channel and dataset [4].

Declarations

Funding: The authors declare that no funding was received for this work.

Ethics, Consent to Participate, and Consent to Publish declarations: Not applicable.

Clinical trial registration: Not applicable. This study does not involve clinical trials.

Conflict of interest: The authors declare no competing interests.

Data availability: The datasets analysed are publicly available CMS Open Data.

Appendix

A HLT paths and trigger details

The primary HLT paths used to select the background sample (and used to construct the signal-side offline trigger proxy) are listed here for completeness:

- `HLT_PFHT350_v`
- `HLT_PFHT475_v`
- `HLT_PFMETNoMu90_PFMHTNoMu90_IDTight_v`
- `HLT_PFMETNoMu120_PFMHTNoMu120_IDTight_v`
- `HLT_MonoCentralPFJet80_PFMETNoMu90_PFMHTNoMu90_IDTight_v`

In the 2015 MiniAOD files the `selectedPatTrigger.pathNames_` branch is often empty, so the trigger decision is reconstructed from the `filterLabels_` branch using the following HLT filter tokens: `hltPFHT`, `hltPFMET`, `hltPFMHT`, and `hltMonoCentralPFJet`. The trigger audit confirms that 18,875,596 of 20,679,437 raw events (91.3%) pass at least one of the five HLT paths, all identified via the filter-label fallback.

The signal offline trigger proxy applied to Delphes signal samples is $H_T > 350$ GeVOR $E_{\mathrm{T}}^{\mathrm{miss}} > 120$ GeVOR ($E_{\mathrm{T}}^{\mathrm{miss}} > 90$ GeVand leading jet $p_T > 80$ GeV), as defined in Section 5.5.

The b-tag veto uses the CSVv2 medium working point at a discriminator value of 0.800, corresponding to the 2015/76X working point for the `pfCombinedInclusiveSecondaryVertexV2BJetTags` discriminator. The veto is applied only to extra jets beyond the two Z-candidate jets; the Z-daughter jets are not subject to the b-tag requirement.

B Feature definitions and extraction formulas

This appendix tabulates the 41-branch data/signal feature schema and the formulas used to compute each branch. Notation: $\vec{p} = (p_x, p_y, p_z)$, $p_T = \sqrt{p_x^2 + p_y^2}$, and $\Delta\phi(a, b)$ denotes the azimuthal separation wrapped to $(-\pi, \pi]$.

The offline selection parameters are: leading and subleading Z-candidate jet $p_T > 30$ GeV, $|\eta| < 2.4$, $70 < m_{jj} < 110$ GeV, $\Delta R_{jj} < 2.0$, and a b-tag veto on extra jets at CSVv2 medium (0.800). The 41-branch data/signal schema is reduced to 32 training features by excluding nine columns: `source_file_idx`, `source_entry` (audit bookkeeping), `trigger_pass`, `btag_pass`, `n_bjets` (trigger/veto flags), `max_btag_csvv2`, `btag_wp_medium` (b-tag schema mismatch), and `n_vertices`, `rho` (pileup mismatch).

Table 5: Feature schema (41 branches) and extraction formulas. Nine columns excluded from training are annotated "(not trained)".

Feature	Formula
`met_pt`	$E_{\mathrm{T}}^{\mathrm{miss}} = \sqrt{p_{x,\mathrm{miss}}^2 + p_{y,\mathrm{miss}}^2}$
`met_phi`	$\mathrm{atan2}(p_{y,\mathrm{miss}}, p_{x,\mathrm{miss}})$

Continued on next page

Feature	Formula
min_dphi_met_jets	$\min_i \lvert\Delta\phi(\mathrm{MET}, \mathrm{jet}_i)\rvert$
dphi_met_jet1	$\Delta\phi(\mathrm{MET}, \mathrm{jet}_1)$
dphi_met_jet2	$\Delta\phi(\mathrm{MET}, \mathrm{jet}_2)$
met_over_sqrtHT	$E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{HT}$
hadronic_recoil_pt	$\sqrt{\mathrm{recoil}_x^2 + \mathrm{recoil}_y^2}$
m_jj	$\sqrt{(E_1+E_2)^2 - \lvert\vec{p}_1+\vec{p}_2\rvert^2}$
pt_Z	$p_{T,Z} = \sqrt{p_{x,Z}^2 + p_{y,Z}^2}$ with $p_Z = p_1 + p_2$
eta_Z_pseudo	$\eta_Z = -\ln\tan(\theta_Z/2)$, $\theta_Z = \arctan 2(p_{T,Z}, p_{z,Z})$
phi_Z	$\mathrm{atan2}(p_{y,Z}, p_{x,Z})$
dphi_met_Z	$\Delta\phi(\mathrm{MET}, Z)$
met_over_ptZ	$E_{\mathrm{T}}^{\mathrm{miss}}/p_{T,Z}$
abs_met_minus_ptZ	$\lvert E_{\mathrm{T}}^{\mathrm{miss}} - p_{T,Z}\rvert$
abs_mjj_minus_mZ	$\lvert m_{jj} - m_Z\rvert$, with $m_Z = 91.1876$ GeV
u_parallel	$\mathrm{recoil}_x z_{\mathrm{unit},x} + \mathrm{recoil}_y z_{\mathrm{unit},y}$
u_perp	$-\mathrm{recoil}_x z_{\mathrm{unit},y} + \mathrm{recoil}_y z_{\mathrm{unit},x}$
u_parallel_over_ptZ	$u_{\parallel}/p_{T,Z}$
u_perp_over_ptZ	$u_{\perp}/p_{T,Z}$
zjet1_pt	$p_{T,\text{leading Z-daughter jet}}$
zjet1_eta	$\eta_{\text{leading Z-daughter jet}}$
zjet2_pt	$p_{T,\text{subleading Z-daughter jet}}$
zjet2_eta	$\eta_{\text{subleading Z-daughter jet}}$
deltaR_zjj	$\sqrt{(\eta_{j1}-\eta_{j2})^2 + \Delta\phi(\phi_{j1},\phi_{j2})^2}$
deltaphi_zjj	$\Delta\phi(\phi_{j1}, \phi_{j2})$
n_jets	N_{jets}
n_extra_jets	$N_{\mathrm{jets}} - 2$
jet1_pt	$p_{T,\text{leading extra jet}}$
jet1_eta	$\eta_{\text{leading extra jet}}$ (sentinel when absent)
jet2_pt	$p_{T,\text{subleading extra jet}}$
jet2_eta	$\eta_{\text{subleading extra jet}}$ (sentinel when absent)
HT	$HT = \sum_{\text{jets}} p_{T,\text{jet}}$
n_bjets	$N_{\text{b-jets}}$, count of extra jets with CSVv2 ≥ 0.800; not trained (veto flag)
n_vertices	N_{vtx}, number of reconstructed primary vertices from `recoVertexs_offlineSlimmedPrimaryVertices`; not trained (pileup mismatch)
rho	ρ_{FastJet} = `double_fixedGridRhoFastjetAll` (pileup energy density); not trained (pileup mismatch)
source_file_idx	Integer index of source MINIAOD file in traversal order; not trained (audit)
source_entry	Entry index within source MINIAOD file; not trained (audit)
trigger_pass	Binary HLT decision: $\mathbb{1}$[event fires ≥ 1 path in App. A] (via `pathNames_` or `filterLabels_` fallback); not trained (trigger flag)
btag_pass	Binary selection flag: $\mathbb{1}$[no extra jet has CSVv2 ≥ 0.800]; not trained (veto flag)
max_btag_csvv2	$\max_{i\in\text{extra jets}} \mathrm{CSVv2}_i$; -1.0 if no extra jets exist; not trained (b-tag schema mismatch)
btag_wp_medium	Constant 0.800, the CSVv2 medium working-point threshold (`BTAG_MEDIUM_WP`); not trained (b-tag schema mismatch)

C Per-file event counts and audit

During dataset traversal, per-MINIAOD-file raw and selected event counts were tracked. For the dataset used in this study 444 MINIAOD files were processed, yielding 20,679,437 raw events and 1,439,523 selected events after the analysis selection (overall selection efficiency 6.96%). No chunks were skipped and no events were lost to processing errors. The trigger audit records 18,875,596 events passing at least one HLT path (91.3% of raw events), all identified via the filter-label fallback. The b-tag decoder audit records 204,250,006 jets decoded across all events, with 14 discriminator labels observed in the `pairDiscriVector_` payload; the primary discriminator `pfCombinedInclusiveSecondaryVertexV2BJetTags` has a score range of $[-10, 1]$. Per-file raw entries vary from a few hundred to a few tens of thousands; typical per-file counts are of order 10^3–10^4. For brevity the manuscript omits the full per-file listing; complete per-file counts are available with the analysis code (Section E).

C.1 Data-side cutflow

Table 6 summarises the data-side event selection stages. The trigger stage applies the five unprescaled HLT paths of Appendix A; the offline stage applies the hadronic mono-Z selection described in Section 2 (jet p_T and η requirements, the m_{jj} window, ΔR_{jj}, and the b-tag veto on extra jets). The overall selection efficiency is the ratio of the final selected population to the raw event count.

Table 6: Data-side cutflow for the CMS Run 2015D HTMHT open dataset.

Stage	Events	Efficiency
Raw events processed	20,679,437	100%
Pass $\geq$1 HLT path	18,875,596	91.3%
Pass offline hadronic mono-Z selection	1,439,523	6.96%

The signal-side cutflow is not tabulated here because the per-stage signal-generation counters have not yet been reviewed (see Section 8); the final offline post-selection yields for the three benchmarks are 116, 487, and 848 events, respectively.

D Full closure figures

The full set of per-feature closure plots not shown in the main text is collected below. Representative closure plots for m_{jj}, $E_{\mathrm{T}}^{\mathrm{miss}}$, H_T, and ΔR_{Zjj} appear in Figure 1; additional kinematic closure panels appear in Figure 2; the sentinel-imputation panels appear in Figure 3; the validation-NLL curves appear in Figure 4; and the sensitivity and signal-versus-background NLL figures appear in Figures 6 and 7. The remaining per-feature closure panels are collected below.

Remaining baseline closure figures

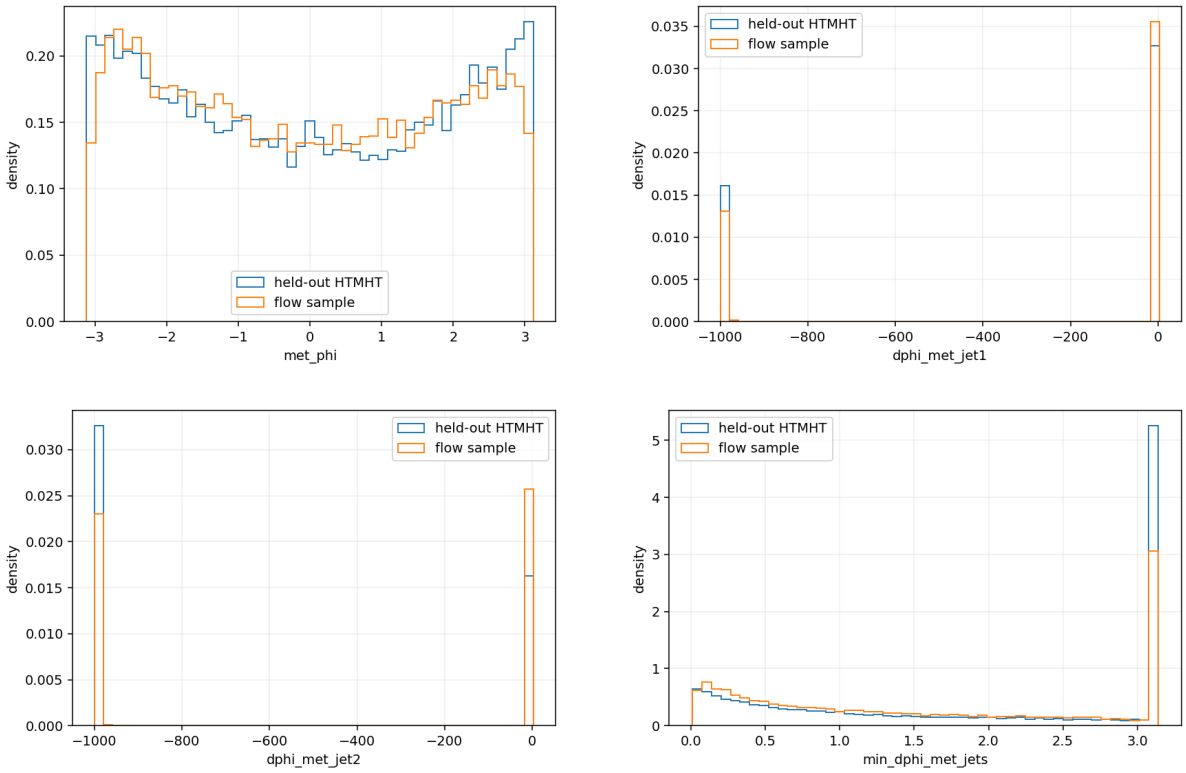


Figure 8: Baseline closure plots for the missing-momentum observables and the undefined extra-jet angular features.

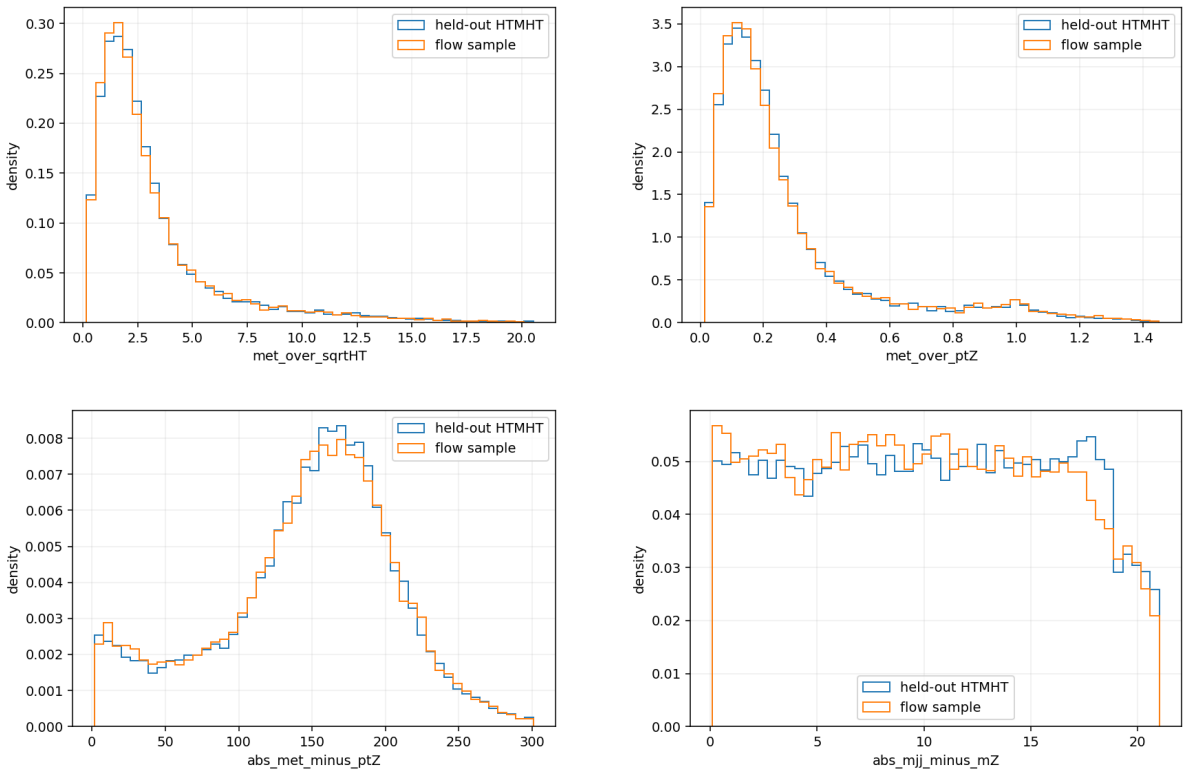


Figure 9: Baseline closure plots for the MET-to-scale ratios and angular proximity summary.

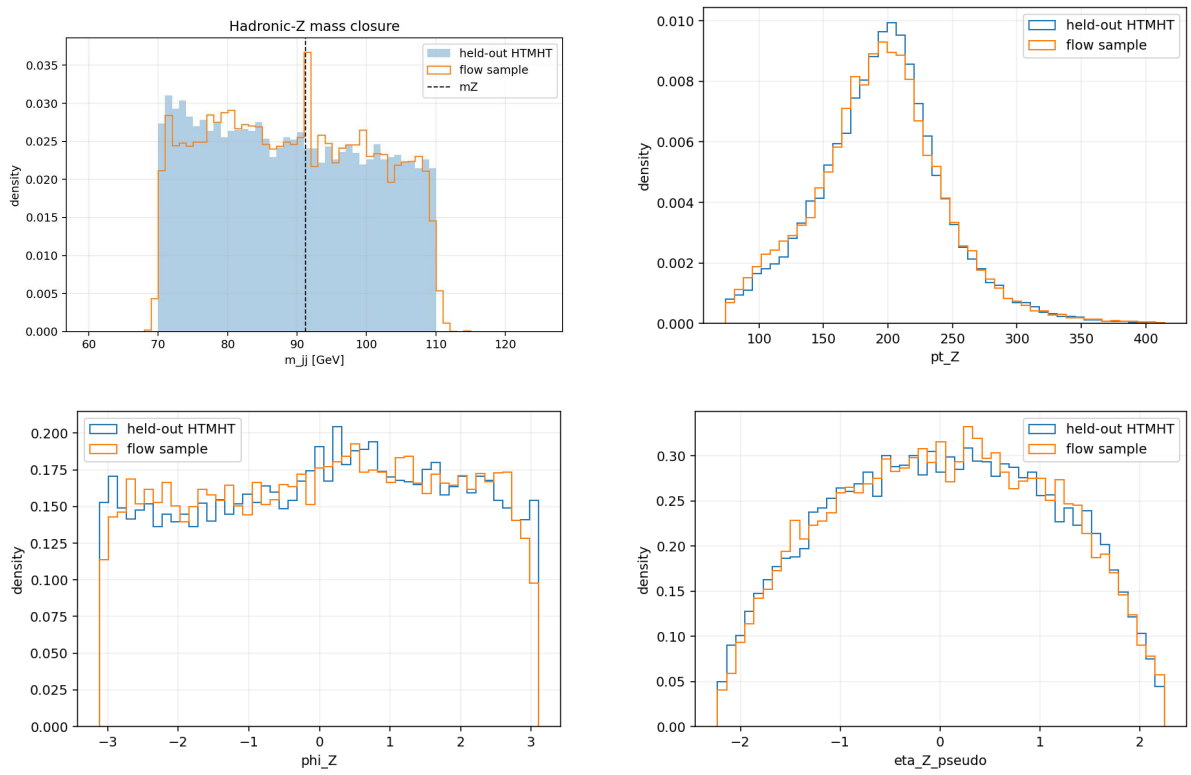


Figure 10: Baseline closure plots for the hadronic-Z mass window observables and reconstructed Z kinematics.

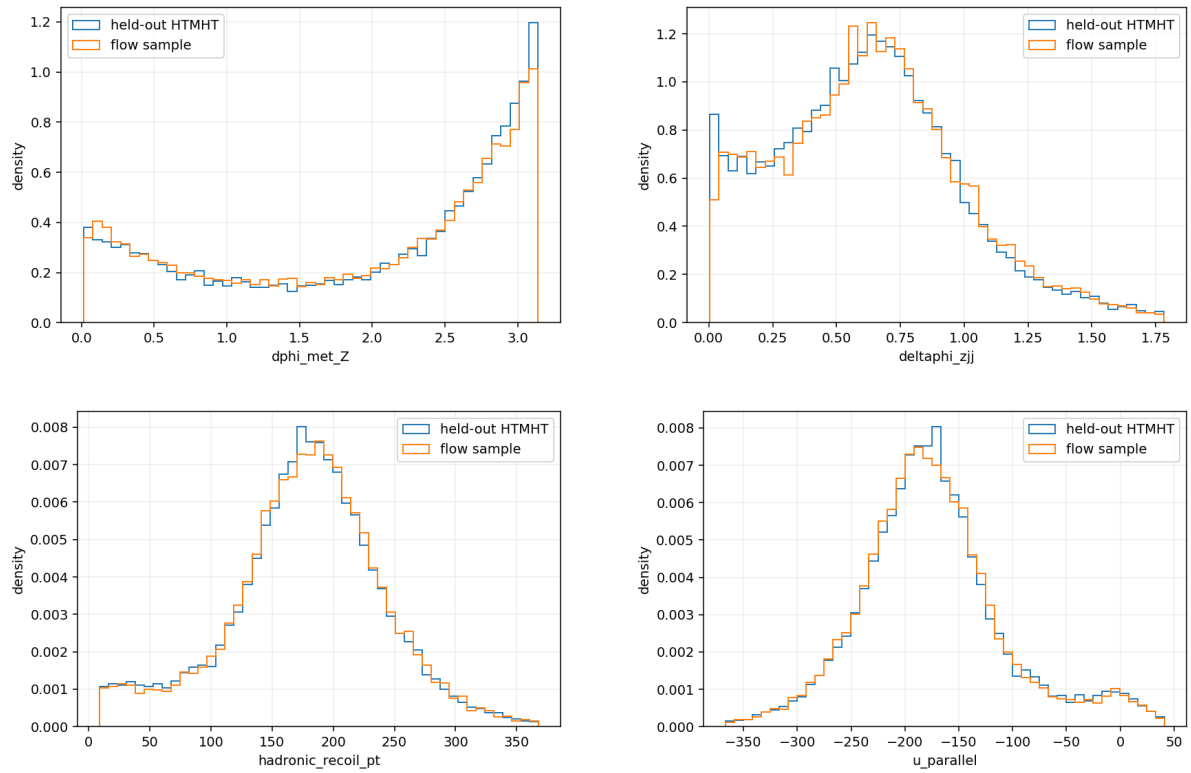


Figure 11: Baseline closure plots for the reconstructed Z direction, dijet geometry, and recoil decomposition.

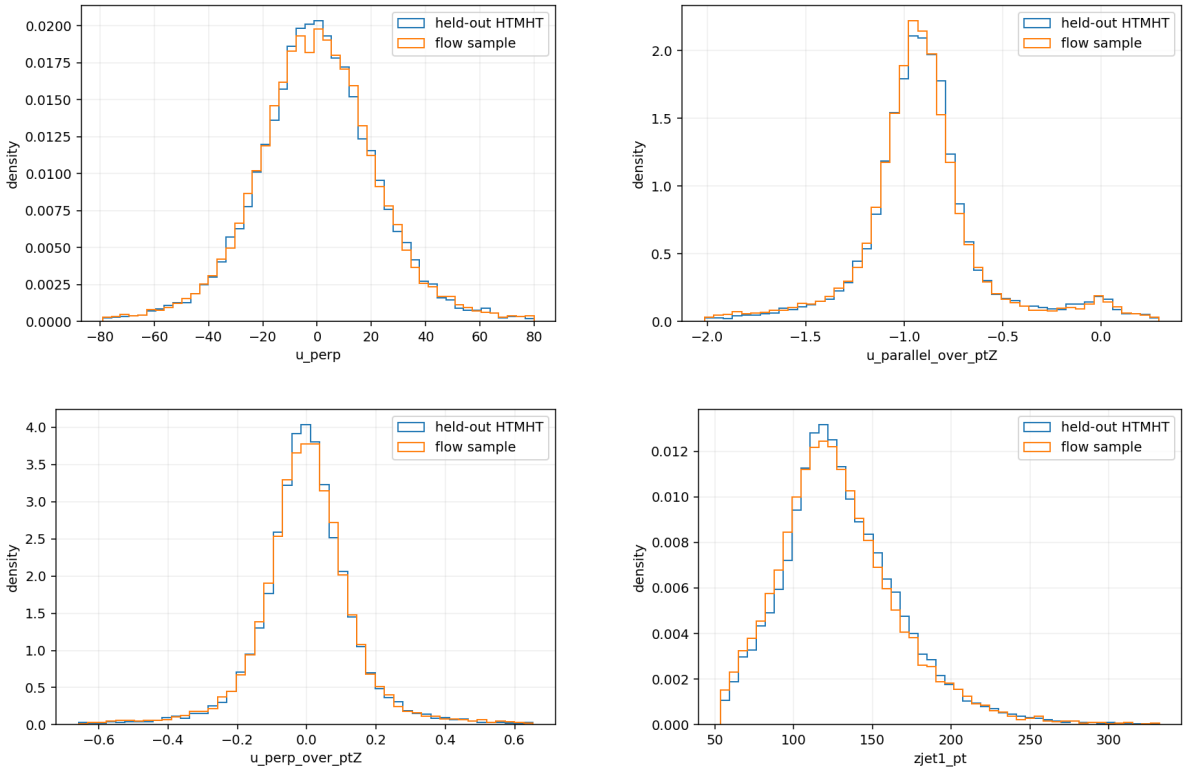


Figure 12: Baseline closure plots for the recoil components and leading Z-daughter jet kinematics.

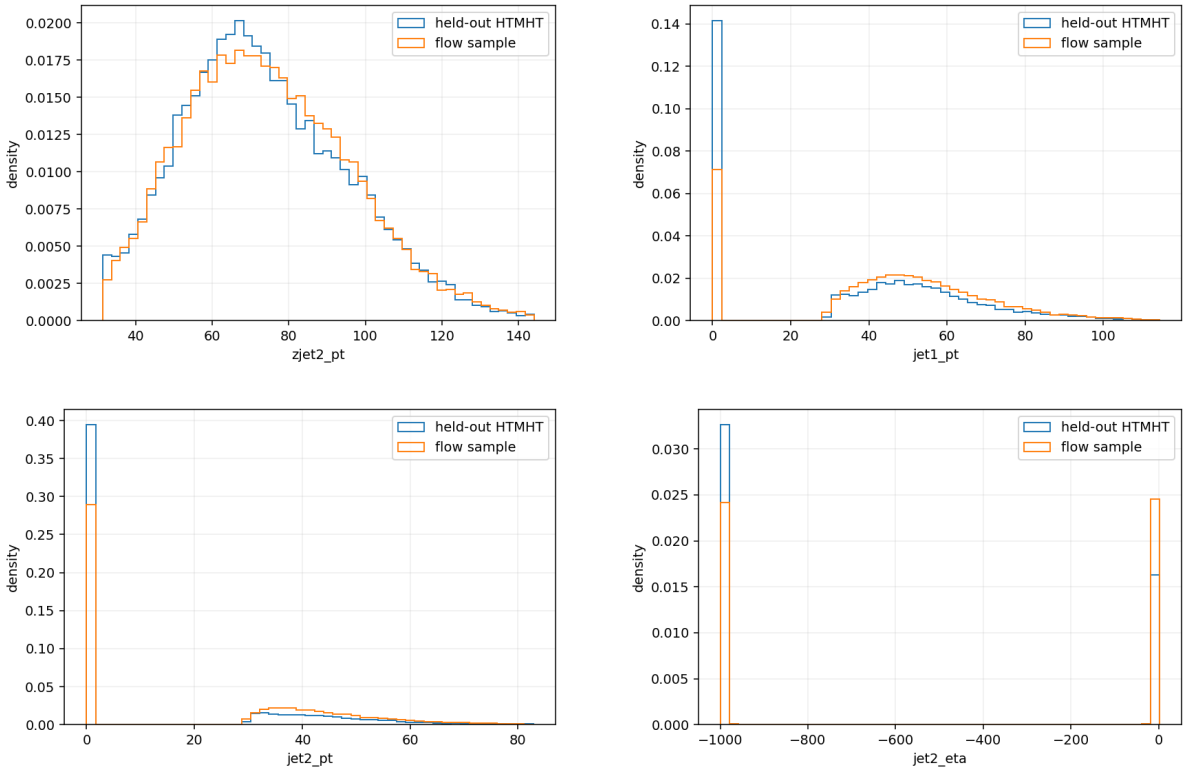


Figure 13: Baseline closure plots for the Z-daughter and extra-jet p_T observables.

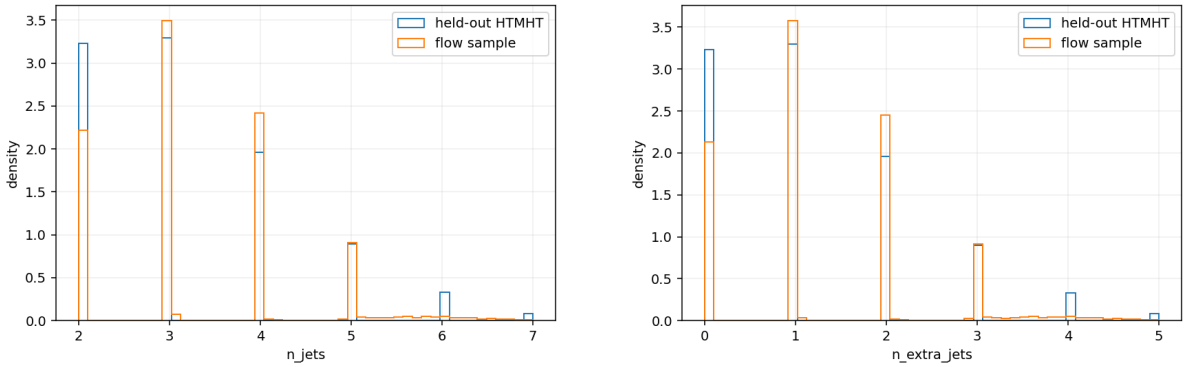


Figure 14: Baseline closure plots for the jet-multiplicity observables.

E Reproducibility and code availability

Full analysis configuration, per-file processing counts, train/validation split indices, training logs, closure diagnostics, and sensitivity outputs are available with the accompanying analysis code. The background extraction pipeline was implemented in Python 3.12 using `uproot` and `awkward` for I/O and array processing and `numpy` for numerical operations; density modelling and sensitivity evaluation used `PyTorch` and `torchdiffeq`. Signal samples were produced with a reproducible MadGraph5_aMC@NLO (3.x branch), Pythia8, and Delphes toolchain, with ROOT, HepMC3, and FastJet installed via `conda-forge`; the `DMsimp_s_spin1` UFO model was used for dark-matter production. The Delphes CMS card (`delphes_card_CMS.tcl`) does not include a pileup overlay module, so `n_vertices` and `rho` are zero placeholders in the signal samples.

Key hyperparameters and environment variables for the flow-matching training are: random seed `SEED=20260730`, batch size `BATCH_SIZE=2048`, learning rate `LEARNING_RATE=2e-4`, weight decay `WEIGHT_DECAY=1e-5`, maximum epochs `MAX_EPOCHS=500`, early-stopping patience `PATIENCE=6`, validation NLL check interval `NLL_EVERY=5`, ODE solver `rk4` with step size 0.05 and tolerances 10^{-3}, divergence mode `hutchinson`, and density batch size `DENSITY_BATCH_SIZE=16`. The sentinel imputation value is `NAN_SENTINEL_VALUE=-999.0`. The ablation mode is activated by setting `DROP_EXTRA_JET_FEATURES=1`.

Per-stage signal-side cutflow counters were added to the signal-generation notebook but the resulting per-stage CSVs have not been reviewed to diagnose why the `axial_mx10_mv20` benchmark has a lower raw selection efficiency than the other two points. The offline post-selection yields quoted in the text (116, 487, and 848 events) reflect the final selected populations prior to trigger-proxy application and NLL scoring.

References

[1] D. Abercrombie et al. Dark Matter Benchmark Models for Early LHC Run-2 Searches: Report of the ATLAS/CMS Dark Matter Forum. *Phys. Dark Univ.*, 27:100371, 2020. doi:10.1016/j.dark.2019.100371.

[2] Mihailo Backovic, Michael Kramer, Fabio Maltoni, Antony Martini, Kentarou Mawatari, and Mathieu Pellen. Higher-order QCD predictions for dark matter production at the LHC in simplified models with s-channel mediators. *Eur. Phys. J. C*, 75(10):482, 2015. doi:10.1140/epjc/s10052-015-3700-6.

[3] Matthias Neubert, Jian Wang, and Cen Zhang. Higher-order QCD predictions for dark matter production in mono-Z searches at the LHC. *JHEP*, 02:082, 2016. doi:10.1007/JHEP02(2016)082.

[4] Hitesh Rasineni and Bhavishya Chebrolu. Mono-Z Dark Matter Search with Neural Spline Flows Using CMS Run 2015D Open Data, 2026. doi:10.48550/arXiv.2607.13771.

[5] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural Spline Flows. In *Advances in Neural Information Processing Systems*, volume 32, 2019. doi:10.48550/arXiv.1906.04032.

[6] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021. doi:10.48550/arXiv.1912.02762.

[7] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. *International Conference on Learning Representations*, 2023. doi:10.48550/arXiv.2210.02747.

[8] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. doi:10.48550/arXiv.1806.07366.

[9] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models. In *International Conference on Learning Representations*, 2019. doi:10.48550/arXiv.1810.01367.

[10] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C*, 71:1554, 2011. doi:10.1140/epjc/s10052-011-1554-0.

[11] CMS Collaboration. HTMHT primary dataset in MINIAOD format from RunD of 2015 (/HTMHT/Run2015D-16Dec2015-v1/MINIAOD), 2021. doi:10.7483/OPENDATA.CMS.SRUP.MHPK.

[12] CMS Collaboration. CMS data preservation, re-use and open access policy. CMS Document 6032-v3, 2020. CMS document 6032-v3.

[13] CMS Open Data. Luminosity. CMS Open Data Guide, 2026. CMS Open Data luminosity guide.

[14] Kati Lassila-Perini et al. Using CMS Open Data in research — challenges and directions, 2021. doi:10.48550/arXiv.2106.05726.

[15] Matthew Bellis and Thomas McCauley. CMS Data Preservation and Open Access: status and plans. *Data Preservation in High Energy Physics: Global Report*, 2026. doi:10.48550/arXiv.2607.06775.

[16] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014. doi:10.1007/JHEP07(2014)079.

[17] Torbjorn Sjostrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015. doi:10.1016/j.cpc.2015.01.024.

[18] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 02:057, 2014. doi:10.1007/JHEP02(2014)057.

[19] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k(t) jet clustering algorithm. *JHEP*, 04:063, 2008. doi:10.1088/1126-6708/2008/04/063.

[20] Sabine Kraml, Ulrich Laa, Kentarou Mawatari, and Kenta Yamashita. Simplified dark matter models with a spin-2 mediator at the LHC. *Eur. Phys. J. C*, 77(5):326, 2017. doi:10.1140/epjc/s10052-017-4871-0.

[21] LHC Dark Matter Working Group. Comparing LHC searches for heavy mediators of dark matter production in visible and invisible decay channels, 2017. doi:10.48550/arXiv.1703.05703.

[22] Nicole F. Bell, Giorgio Busoni, and Isaac W. Sanderson. Self-consistent Dark Matter Simplified Models with an s-channel scalar mediator. *JCAP*, 03:015, 2017. doi:10.1088/1475-7516/2017/03/015.

[23] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J. C*, 72:1896, 2012. doi:10.1140/epjc/s10052-012-1896-2.

[24] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015. doi:10.1007/JHEP04(2015)040.

[25] N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.*, 641:A6, 2020. doi:10.1051/0004-6361/201833910.

[26] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017. doi:10.48550/arXiv.1605.08803.

[27] Durk P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018. doi:10.48550/arXiv.1807.03039.

[28] Benjamin Nachman and David Shih. Anomaly Detection with Density Estimation. *Phys. Rev. D*, 101:075042, 2020. doi:10.1103/PhysRevD.101.075042.

[29] Anna Hallin, Gregor Kasieczka, Sabine Kraml, Andre Lessa, and David Shih. Classifying anomalies through outer density estimation. *Phys. Rev. D*, 106:055006, 2022. doi:10.1103/PhysRevD.106.055006.

[30] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Classification without labels: Learning from mixed samples in high energy physics. *JHEP*, 10:174, 2017. doi:10.1007/JHEP10(2017)174.