

Reconstructing the invisible fraction of semivisible jets in ISR-boosted events via neural network regression

Yin Li[✉], Jian Bin Wang[✉], Jia Qi Xie[✉], Kai Rong Xu[✉], Rui Han Ye[✉], and Zi Huan Huang[✉]

*School of Science, Shenzhen Campus of Sun Yat-sen University,
66 Gongchang Road, Shenzhen, Guangdong 518107, People's Republic of China*

Bingxuan Liu^{✉*}

*School of Science, Shenzhen Campus of Sun Yat-sen University,
66 Gongchang Road, Shenzhen, Guangdong 518107, People's Republic of China*



(Received 23 April 2026; accepted 27 July 2026; published 17 September 2026)

Semivisible jets (SVJs) provide a characteristic collider signature of strongly interacting dark sectors, in which the key model parameter r_{inv} controls the fraction of dark hadrons decaying to dark matter candidates. In this work, a regression model is developed to reconstruct r_{inv} in SVJ events produced in association with an energetic photon. The model uses information from high-level physics objects only, and the training procedure is optimized to ensure applicability. The performance is found to be robust against varying signal parameters and r_{inv} can be reconstructed at a much higher precision, compared to previously developed analytical method. It offers a new approach to conduct SVJ searches that can potentially unify both s -channel and t -channel productions, enhancing the sensitivities.

DOI: [10.1103/hbxx-5f3k](https://doi.org/10.1103/hbxx-5f3k)

I. INTRODUCTION

The Standard Model of particle physics has achieved great success, but it does not contain a particle that can serve as a dark matter (DM) candidate. It is therefore necessary to introduce new physics, in particular dark matter models beyond the Standard Model (SM), given the strong hints of its existence [1–5]. One of the main candidates for DM is the weakly interacting massive particle (WIMP) [6]. Although searches of this type have covered a wide range of parameter space, they are focused on the scenario where the DM candidate is produced in isolation [7–10]. This picture can be qualitatively modified if the dark sector has its own confining dynamics. In the dark QCD framework, analogous to quantum chromodynamics (QCD) in the SM, the hard interaction produces dark quarks (χ_i), which subsequently shower and hadronize under a new confining force $SU(N)$ [11,12]. The resulting dark hadrons can have different detector signatures, where some remain stable and invisible on detector scales, while others decay back to SM particles. Therefore, the missing transverse momentum is not produced only as an isolated

recoil object, but can be embedded within and geometrically correlated with jetlike visible activity. This gives rise to a striking experimental signature in which a significant fraction of the energy inside the jets is invisible. Such jets are called semivisible jets (SVJs) [13–16].

Previous studies of SVJ used the transverse mass, m_T , constructed from the dijet system and missing transverse energy, E_T^{miss} , as the core discriminating variable to gain sensitivity to the s -channel $Z' \rightarrow \chi\bar{\chi}$ process [13]. The experimental signature is determined by a few effective parameters that dominate the observable behavior, among which r_{inv} controls the ratio of invisible components inside the jets [14]. The community has also attempted to consolidate the parameter choices to ensure smooth cross-talks [17]. Beyond the canonical vector-mediator setup, it has also been studied for more general mediator structures carrying both SM and hidden-sector charges [16]. Both the ATLAS and CMS experiments have probed the s -channel SVJ productions recently [18,19], following the technique proposed in Ref. [13]. Although the t -channel production lacks a clear resonant mass peak, it is possible to take advantage of the event topology since the alignment of jets and E_T^{miss} is unique, as reported by a dedicated ATLAS search in this channel [20]. Efforts have been made to explore the distinguishable features within SVJ using substructure observables [21–24] or constituent-level information [25–27]. Recent studies have further advanced machine-learning approaches for SVJ toward

*Contact author: liubx28@mail.sysu.edu.cn

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

more model-agnostic self-supervised methods and symmetry-aware graph-network architectures [26,28,29].

Along with the identification of SVJ, the reconstruction of key physical quantities has also been developed in parallel. Existing studies investigated the application of machine-learning methods with an information-compression structure to learn the mediator mass from event dynamics [30]. At the same time, initial state radiation (ISR) in SVJ events is found to be beneficial. High- p_T ISR objects help with event triggering and background suppression, provide an additional boost to the SVJ system, and change the geometric relationship between E_T^{miss} and SVJ [31]. Assuming a simple collinear condition that the invisible energy in each SVJ aligns with the jet axis allows us to reconstruct r_{inv} analytically [32]. These attempts are of great interest as they do not only provide discriminating variables, but also enable direct extraction of the model parameters using real data, in case of a discovery. In particular, r_{inv} is the key parameter that governs the phenomenologies in both s -channel and t -channel productions. If one can reconstruct r_{inv} more precisely, it is feasible to consider both channels simultaneously. The positive results from Ref. [32] indicate that it can be achieved using event-level variables already, which is naturally more robust against varying model parameters.

This work tries to establish a workflow to improve the r_{inv} reconstruction performance, in production channels with an energetic ISR photon, as shown in Fig. 1. We demonstrate that by setting up a neural network with sophisticated data preprocessing, it is possible to achieve a much better precision. The performance is robust against various model parameters and effective to both s - and t -channel production channels. The findings in this article reveals the potential of event-topology in SVJ searches, and offer practical strategies that can be applied in future experimental searches.

The article is structured as follows: Sec. II describes the datasets used for this study; Sec. III introduces the event selection criteria and various key variables; Sec. IV lays out the training workflow and algorithm structure, followed by a study to evaluate its impact in Sec. V, and finally Sec. VI concludes this work with some discussion points.

II. DATASETS

All events are generated using Monte Carlo (MC) simulations. In this study, the center-of-mass energy is assumed to be $\sqrt{s} = 13$ TeV, corresponding to the data-taking conditions of the LHC Run 2. The hard scattering processes are generated with MadGraph5_aMC@NLO 3.5.1 [33]. Parton showering and hadronization are performed using Pythia 8.306 [34]. The dark sector dynamics are implemented through the built-in Hidden Valley module in Pythia. The output is subsequently passed to Delphes 3.5.3 [35] for fast detector simulation, using the CMS detector geometry. Jets are clustered using the anti- k_t algorithm [36,37] with a jet radius of $R = 0.4$.

A. Signal processes

Three sets of signal samples are generated to suit different tasks, including the algorithm training, robustness testing and generality testing. Detailed information is given below.

1. Algorithm training samples

The training samples consider a Z' boson produced via an s -channel process, decaying into a pair of dark quarks, χ , i.e., $Z' \rightarrow \chi\bar{\chi}$. The details of the model can be found in Ref. [18]. Since this work studies the case with an energetic ISR object, we require at least one ISR photon at the generation level. Processes with an additional parton are also included. The signal grid spans across Z' mass and r_{inv} , summarized in Table I, together with other important model parameters. Each signal point contains 100k events. The benchmark grid is chosen to cover the main parameter dependence relevant for the photon-associated SVJ topology. The Z' mass range defines the main high-mass training region, while the r_{inv} scan from 0.1 to 0.9 covers the semivisible regime between the mostly visible and mostly invisible limits. The other dark-sector parameters define the nominal hidden valley configuration used for training, with variations studied separately in Sec. V. The hidden valley sector is configured with $N_C^{\text{dark}} = 2$ and $N_F^{\text{dark}} = 2$. The degenerate choice $m_{\pi_D} = m_{\rho_D} = 20$ GeV

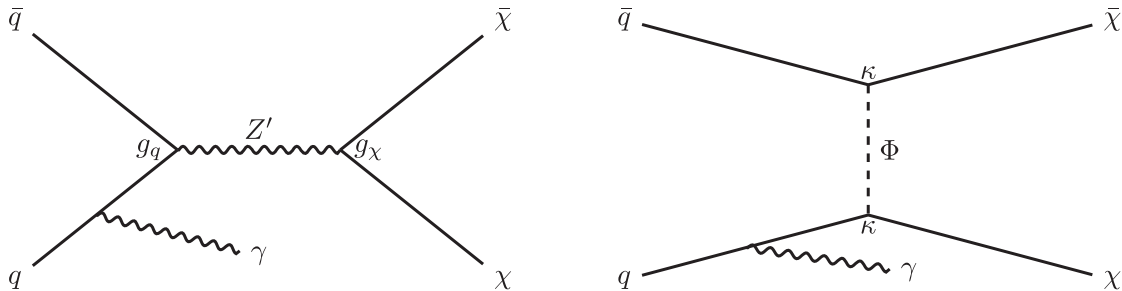


FIG. 1. Representative Feynman diagrams of the signal processes with an ISR photon considered in this work. Left: the s -channel process, $q\bar{q} \rightarrow Z' + \gamma_{\text{ISR}}$, followed by $Z' \rightarrow \chi\bar{\chi}$. Right: the t -channel process, $q\bar{q} \rightarrow \chi\bar{\chi} + \gamma_{\text{ISR}}$, mediated by a scalar Φ .

TABLE I. Main parameters used in the benchmark signal event generation. The corresponding benchmark choices are discussed in the text.

Parameter	Value
Mediator mass $m_{Z'}$	1500–3500 GeV
Center-of-mass energy $\sqrt{s}$	13 TeV
Invisible fraction r_{inv}	0.1–0.9 (step 0.1)
Dark color number N_C^{dark}	2
Dark flavor number N_F^{dark}	2
Dark quark mass m_χ	10 GeV
Dark pion mass m_{π_D}	20 GeV
Dark rho mass m_{ρ_D}	20 GeV
Dark confinement scale Λ_D	35.1539 GeV
Vector meson probability probVector	0.75
Events per signal point	100000

is treated as a simplified generator-level benchmark rather than a full dark-hadron spectroscopy model.

2. Robustness testing samples

To validate the robustness of the trained model, additional datasets are generated under different dark sector model assumptions. For s -channel SVJ events, the baseline parameter configuration is chosen to be

$$\begin{aligned} \text{probVector} &= 0.75, & \Lambda &= 35.1539 \text{ GeV}, \\ m_{\pi_D} &= m_{\rho_D} = 20 \text{ GeV}. \end{aligned}$$

Single-parameter variations are then performed for the following quantities:

- (i) the dark vector meson probability (probVector), with values of 0.5 and 0.6;
- (ii) the dark confinement scale (Λ), with values of 10 and 50 GeV;
- (iii) the dark pion and dark rho masses (m_{π_D} and m_{ρ_D}), set to 10 and 40 GeV;
- (iv) the dark pion masses (m_{π_D}), set to 100 and 200 GeV; and
- (v) the dark flavor number (N_F^{dark}), with values of 3 and 4.

In addition, to test the dependence on the mediator mass beyond the main training grid, extra samples with $m_{Z'} = 500$ and 1000 GeV are generated and evaluated separately. To ensure that variations in the machine learning model performance can be attributed to changes in a specific dark sector parameter rather than to correlated multiparameter effects, only one parameter is varied at a time in each trial, while all other parameters are fixed to their baseline values.

3. Generality testing samples

In addition to the resonant s -channel production mode, SVJs can also be produced through a nonresonant t -channel

production. As shown in Fig. 1, the connection between SM and the dark sector is established by a scalar mediator Φ , which couples a quark in SM to a dark quark through a Yukawa-type interaction. The partonic process in proton-proton collisions can produce dark quarks via t -channel mediator exchange, which then undergo dark showering and hadronization to form SVJs. A set of t -channel samples is prepared to examine whether the algorithm trained on the s -channel remains effective. The same event generation, detector simulation, and reconstruction procedure as for the s -channel signal is adopted. For two different Φ mass points at 1.5, 2.5 and 4.0 TeV, samples are generated at three r_{inv} values, $r_{\text{inv}} = 0.1, 0.5$, and 0.9.

B. Background

The background process used for the algorithm training is $\gamma + \text{jets}$. Since its cross section falls rapidly with increasing jet transverse momentum, a slicing strategy based on the minimum transverse momentum of the leading jet is adopted to cover the high-momentum region relevant for the signal without losing statistical precision. Specifically, subsamples are generated with the minimum leading jet transverse momentum $p_{T,\text{jet}_1}^{\text{min}}$ starting from 0 to 2000 GeV, in steps of 500 GeV. Each slice contains 40,000 unweighted events. The background samples are reconstructed and preselected using the same procedures as those applied to the signal samples, ensuring that all datasets share a consistent phase-space definition and kinematic treatment before being fed into the training process.

III. SELECTIONS AND OBSERVABLES

This section discusses the selections applied to mimic actual analyses, and introduces the definition of the training target. A brief summary on the analytical method from Ref. [32] is also given.

A. Event selections

We follow the same strategy as Ref. [32] for preselection. Events are required to contain at least one photon with transverse momentum greater than 150 GeV. After removing the overlapping photons and jets, at least two reconstructed jets are required. The two leading jets are then selected as the SVJ candidates for the construction of the observables used later. The corresponding selection criteria are summarized in Table II. In particular, to retain a broader event sample for the subsequent observable study and regression task, only a loose requirement of $|y^*| < 5$ is imposed in this work.

B. r_{inv} training target

Following a commonly adopted definition in LHC SVJ studies, r_{inv} is defined as

TABLE II. Object definitions and event selection criteria used in this analysis.

Selection requirement	Value
Photon multiplicity	$N_\gamma \geq 1$
Jet multiplicity	$N_{\text{jet}} \geq 2$
ISR transverse momentum	$p_T^{\text{ISR}} > 150 \text{ GeV}$
Jet transverse momentum	$p_T^{j_1}, p_T^{j_2} > 25 \text{ GeV}$
Photon-jet separation	$\Delta R(\gamma, \text{jet}) > 0.4$
Jet azimuthal separation	$ \Delta\phi(\text{jet}_1, \text{jet}_2) > 0.8$
Rapidity separation	$ y^* < 5$

$$r_{\text{inv}} = \frac{N_{\text{invisible}}}{N_{\text{visible}} + N_{\text{invisible}}} \quad (1)$$

[17], where $N_{\text{invisible}}$ (N_{visible}) refers to the number of dark hadrons decaying to detector invisible (visible) particles. For different jets with the same particle-level r_{inv} parameter, the relative fractions of visible and invisible components can still fluctuate following a binomial law. Therefore, we establish a method to obtain the event-by-event r_{inv} value.

It is calculated as

$$r_{\text{inv}} = \frac{N_{\text{DM}}}{2(N_{\pi_D} + N_{\rho_D})} \quad (2)$$

and is used as the target variable in the machine learning training. In the actual implementation, as shown in the formula, the numerator is taken to be the number of final-state dark matter particles, identified by selecting particles with $\text{PID} = 51$ or 53 and $\text{status} = 1$. The denominator is constructed from dark mesons, including both dark pion and dark rho mesons. No distinction is made between diagonal and off-diagonal dark mesons in this definition; all hadronized dark pion and dark rho states recorded by Pythia are included. The former are identified by $|\text{PID}| = 4900111$ or ± 4900211 , while the latter are identified by

$|\text{PID}| = 4900113$ or 4900213 , and in both cases the particles are required to have $\text{status} = 83$ or 84 . The factor two in the denominator comes from the fact that a dark hadron either decays to a pair of DM particles, or a pair of SM quarks in our setup. Figure 2 illustrates the particle-level r_{inv} distributions of a few signal points.

C. Analytical r_{inv} reconstruction

In a realistic experimental environment, generator-level information is not directly accessible, and r_{inv} must be indirectly estimated using reconstruction-level observables. When a high transverse-momentum ISR object (either a jet or a photon) recoils against the SVJ system, the event-topology becomes clearer. Previous studies have attempted to analytically reconstruct the invisible fraction of the jets in this scenario [32].

The analytic method assumes that E_T^{miss} comes from the two invisible components of the SVJs, with each aligned with the corresponding jet axis. Consequently, the transverse momentum conservation of the event can be written as

$$\begin{aligned} p_x^{\text{miss}} &= \alpha_1 p_{x,1} + \alpha_2 p_{x,2}, \\ p_y^{\text{miss}} &= \alpha_1 p_{y,1} + \alpha_2 p_{y,2}, \end{aligned} \quad (3)$$

where $p_{x,i}$ and $p_{y,i}$ denote the transverse momentum components of the i th jet, and α_i represents the ratio between the invisible and visible components within that jet. Solving this linear system yields α_1 and α_2 , from which r_{inv} for a given event can be defined as

$$r_{\text{inv}}^{\text{analytic}} = \frac{1}{2} \left(\frac{\alpha_1}{1 + \alpha_1} + \frac{\alpha_2}{1 + \alpha_2} \right). \quad (4)$$

This analytically reconstructed value, $r_{\text{inv}}^{\text{analytic}}$, has a clear physical interpretation, a simple functional form, and directly exploits the global transverse momentum balance

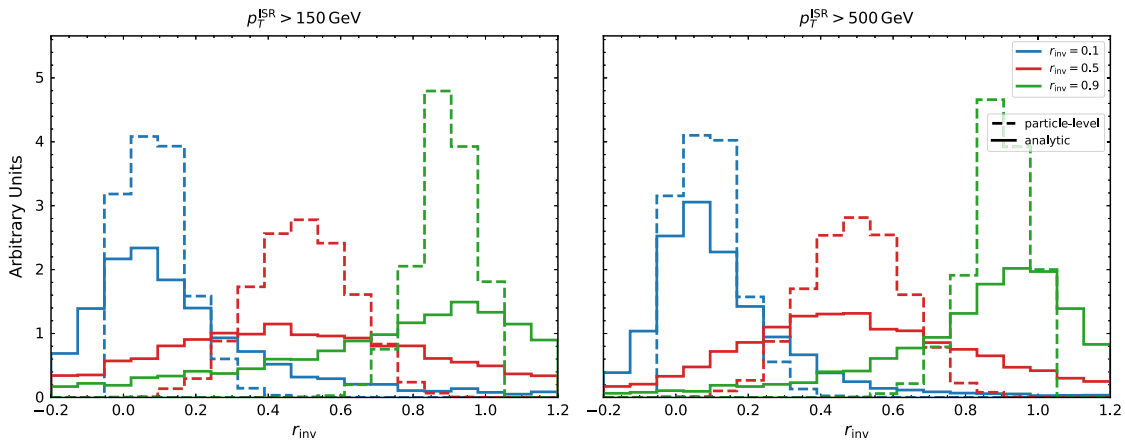


FIG. 2. Distributions of the analytically reconstructed r_{inv} for events with particle-level $r_{\text{inv}} = 0.1, 0.5$, and 0.9 , under the selections $p_T^{\text{ISR}} > 150 \text{ GeV}$ (left) and $p_T^{\text{ISR}} > 500 \text{ GeV}$ (right). The dashed histograms show the corresponding particle-level r_{inv} distributions.

of the event. However, its performance depends strongly on the validity of the underlying assumptions. The geometric interpretation of this decomposition follows Ref. [32]. In this approximation, the two leading jets are treated as the two SVJ candidates, and the missing transverse momentum is assumed to be dominated by the invisible components of these two SVJs. Each invisible component is further assumed to be approximately aligned with the corresponding jet direction. Additional radiation, imperfect SVJ assignment, detector effects, or other sources of missing transverse momentum can therefore spoil this simplified topology and degrade the analytic reconstruction.

At relatively low ISR transverse momentum thresholds ($p_T > 150$ GeV), the distribution of $r_{\text{inv}}^{\text{analytic}}$ typically remains broad, shows only limited correlation with the generator-level r_{inv} , and contains a non-negligible fraction of negative values or solutions exceeding the physically allowed range, as shown in Fig. 2. This indicates that the geometric approximations in the analytic method are not sufficiently satisfied in this regime. When the ISR transverse momentum is increased ($p_T > 500$ GeV), the analytic reconstruction improves and the $r_{\text{inv}}^{\text{analytic}}$ distribution becomes narrower, reflecting the fact that a more boosted topology better satisfies the collinear condition. Nevertheless, the $r_{\text{inv}}^{\text{analytic}}$ distribution is still wider than that of r_{inv} computed directly from particle-level information. This suggests that a multivariate method trained for various conditions may achieve better performance.

IV. THE REGRESSION MODEL

In this section, we discuss the algorithm structure and the training workflow, followed by a study evaluating the performance.

A. Neural network and training strategy

Since we want to explore the potential of high-level physics objects, such as jets and E_T^{miss} , without any low-level information such as jet substructures, a simple fully connected feed-forward neural network should suit our needs. For the input features, a set of observables that can be robustly obtained at the reconstruction level is selected. These include the four-momentum information of the ISR object, the two leading jets, and E_T^{miss} , as listed in Table III. This feature set provides a sufficient physical basis for the

model to infer r_{inv} . All input features are preprocessed. In the first step, a physics-motivated normalization is applied, in which energy- and momentum-related quantities are divided by p_T of the leading jet, which is therefore not included as input feature. This reduces the dependence of the model performance on the absolute mass scale of Z' and improves the robustness. Standard normalization is then applied to further stabilize the optimization process.

To avoid biases arising from an imbalanced r_{inv} distribution, a sampling procedure is employed to all signal samples listed in Sec. II A 1, after applying the selections in Sec. III A. The entire r_{inv} region starting from zero to one is divided into 20 equal-width bins. The third-lowest bin population is chosen as a common sampling threshold, and the number of events in each bin is randomly reduced to this threshold. This procedure ensures that all bins have approximately equal statistical weights in the training dataset.

The dataset is split into training, validation, and test sets with a ratio of 0.6:0.2:0.2. The regression model is trained separately under different ISR and sample-composition conditions, corresponding to different physical use cases. In particular, the model is trained under two ISR momentum thresholds, $p_T^{\text{ISR}} > 150$ GeV and $p_T^{\text{ISR}} > 500$ GeV, reflecting the difference between moderately and highly boosted regimes. For each ISR selection, the training samples are further divided into two categories: samples containing only signal events, and samples composed of a mixture of signal and dominant background processes. The former is used to study the r_{inv} regression performance under ideal conditions, while the latter more closely resembles the practical analysis scenario in which signal and background coexist. For background events, r_{inv} cannot be computed using Eq. (2). To retain a continuous target variable in the mixed training sample, the background r_{inv} is instead assigned using Eq. (4). In total, four independent regression models are constructed. They share the same fully connected regression framework, while the network depth and training hyperparameters are chosen separately for different training setups. In the implementation used in this work, the models trained with $p_T^{\text{ISR}} > 150$ GeV employ five hidden layers, whereas those trained with $p_T^{\text{ISR}} > 500$ GeV employ four hidden layers. Further technical details of the input preprocessing, network architecture, training setup, final hyperparameters, and public inference workflow are provided in the Appendix.

B. Performance of r_{inv} regression

The regression performance is examined under both ISR selections, $p_T^{\text{ISR}} > 150$ GeV and $p_T^{\text{ISR}} > 500$ GeV. Figs. 3 and 4 show that the trained model gives an overall ordered response to the generator-level r_{inv} in all four training setups. For both ISR thresholds, the reconstructed output follows the change of the true r_{inv} consistently, indicating

TABLE III. Input features used in the r_{inv} regression.

Category	Observables
ISR object	$p_T^{\text{ISR}}, \eta^{\text{ISR}}, \phi^{\text{ISR}}$
Leading jet	$\eta^{j_1}, \phi^{j_1}, m^{j_1}$
Subleading jet	$p_T^{j_2}, \eta^{j_2}, \phi^{j_2}, m^{j_2}$
Missing transverse momentum	$E_T^{\text{miss}}, \phi^{\text{miss}}$

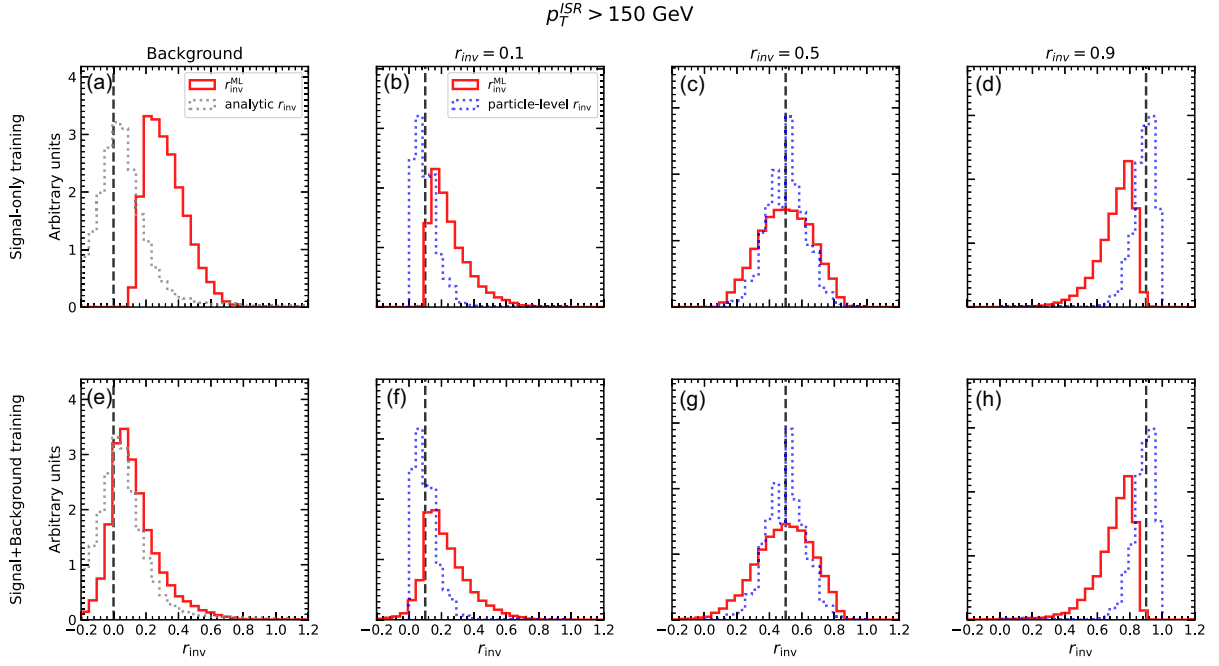


FIG. 3. Distributions of r_{inv}^{ML} for the $p_T^{ISR} > 150 \text{ GeV}$ selection. Panels (a)–(d) show the results for the signal-only training setup, corresponding to background, $r_{inv} = 0.1$, 0.5 , and 0.9 , respectively. Panels (e)–(h) show the corresponding results for the signal-plus-background training setup. In the background panels (a) and (e), the dashed histograms indicate the analytically reconstructed r_{inv} ; in the signal panels, the dashed histograms indicate the particle-level r_{inv} distributions.

that the event-level observables used as inputs retain sufficient information for the regression task.

Figure 5 shows that different r_{inv} hypotheses remain more clearly separated in the comparison with the particle-

level label, while Fig. 6 states that the inclusion of background only mildly weakens the overall behaviors and does not qualitatively change the ordered signal response. Since the two ISR selections are trained independently, the

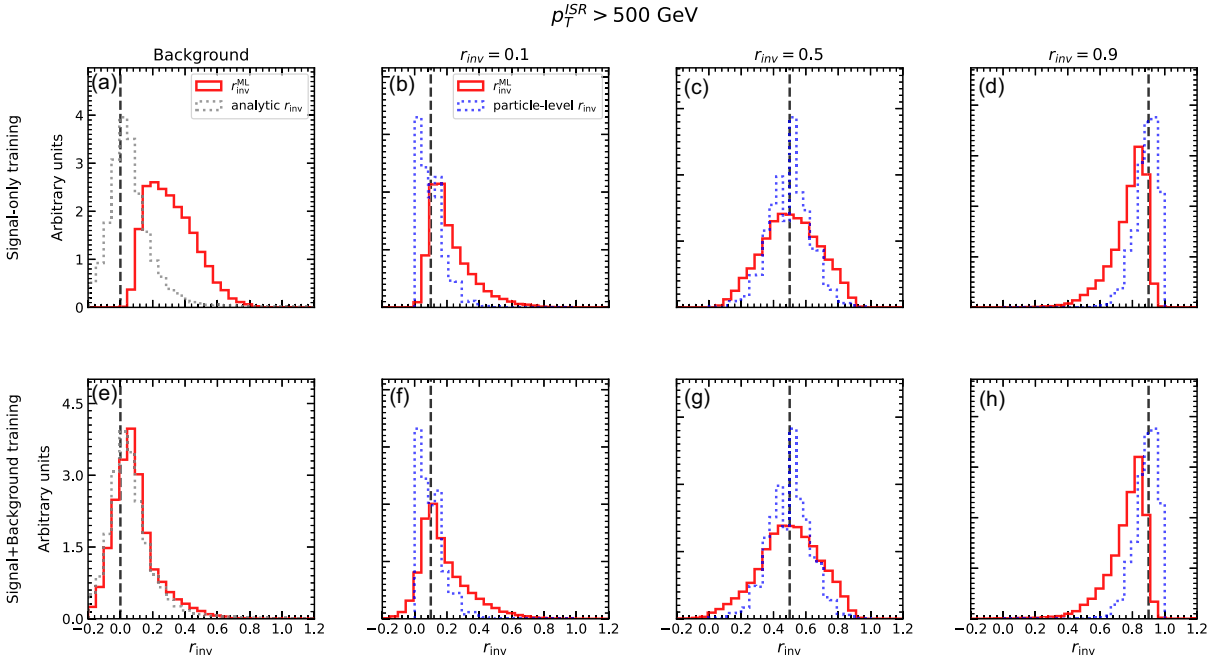


FIG. 4. Distributions of r_{inv}^{ML} for the $p_T^{ISR} > 500 \text{ GeV}$ selection. Panels (a)–(d) show the results for the signal-only training setup, corresponding to background, $r_{inv} = 0.1$, 0.5 , and 0.9 , respectively. Panels (e)–(h) show the corresponding results for the signal-plus-background training setup. In the background panels (a) and (e), the dashed histograms indicate the analytically reconstructed r_{inv} ; in the signal panels, the dashed histograms indicate the particle-level r_{inv} distributions.

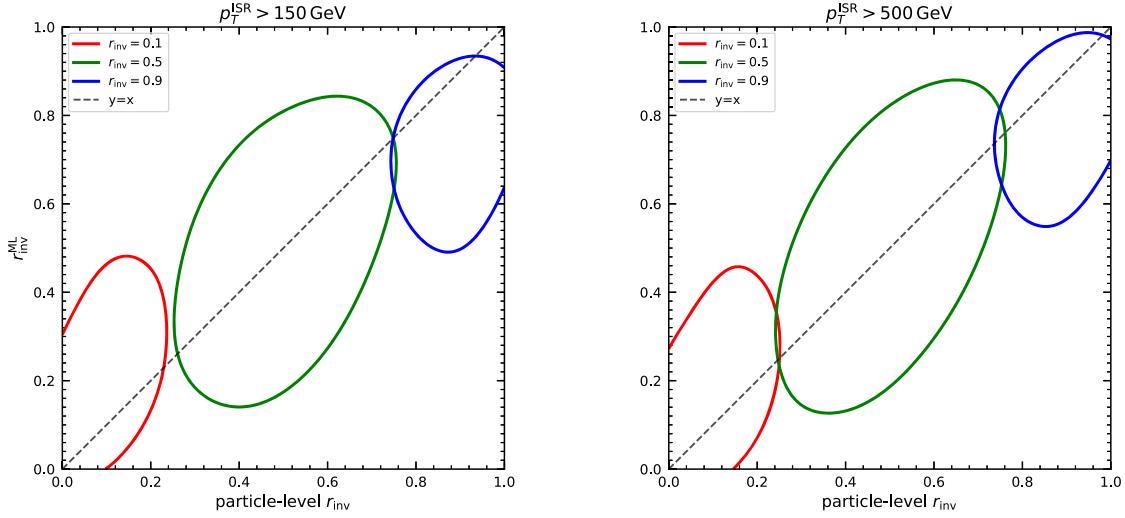


FIG. 5. Two-dimensional comparison between $r_{\text{inv}}^{\text{ML}}$ and the particle-level r_{inv} , shown as 80% contour regions for $r_{\text{inv}} = 0.1, 0.5$, and 0.9 , under the selections $p_T^{\text{ISR}} > 150$ GeV (left) and $p_T^{\text{ISR}} > 500$ GeV (right). The dashed line indicates perfect reconstruction.

change of the ISR threshold has only a limited impact on the regression quality.

As discussed in Sec. III C, $r_{\text{inv}}^{\text{analytic}}$ is strongly constrained by the underlying recoil and collinearity assumptions. Consequently, its reconstructed distributions are substantially broader in the low-ISR regime, as shown in Figs. 7 and 8, which reduces the separation power. On the contrary, the machine learning approach yields narrower overall distributions, and the algorithm can be retrained to adapt for a new ISR condition, offering more flexibility in the analysis usage.

V. IMPACT

We evaluate the potential impact of the regression model in this section. First, the signal parameter dependence is

examined using varied signal samples, and then the model is tested on t -channel production. At last, it is revealed that both s -channel and t -channel may benefit from $r_{\text{inv}}^{\text{ML}}$ in practice.

A. Robustness against dark-sector parameter variations

The physics-motivated scaling method in Sec. IV attempts to remove the model's dependence on the mediator masses. However, given the rich set of signal parameters listed in Table I, it is imperative to make sure the regression model is not overtrained toward a specific setup. Therefore, the stability of the regression results is examined under different dark-sector parameter configurations.

The regression performance is evaluated with only one dark-sector-related parameter varied at a time, while

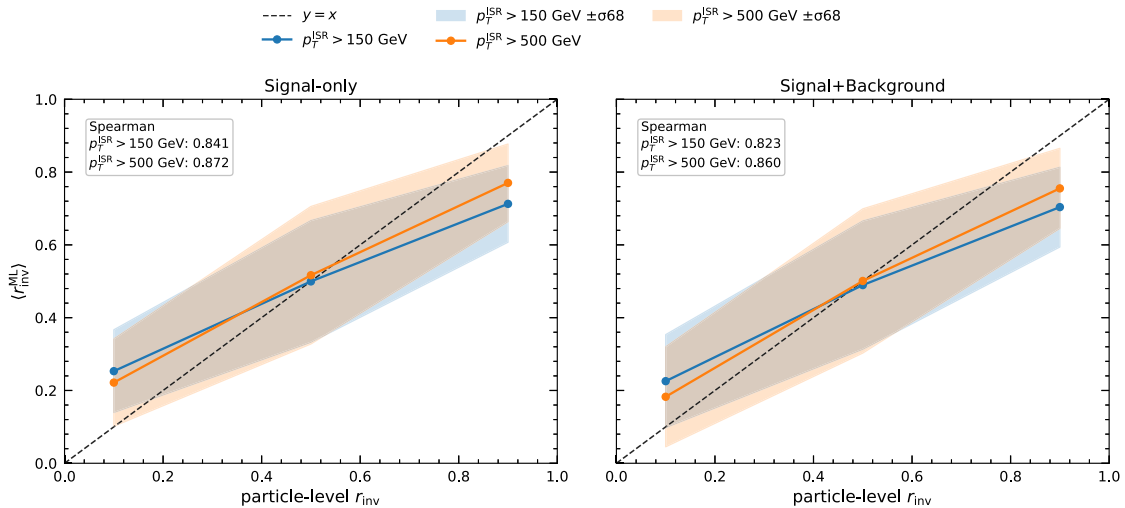


FIG. 6. Mean $r_{\text{inv}}^{\text{ML}}$ response as a function of r_{inv} for the signal-only (left) and signal-plus-background (right) training setups. Results for the two ISR selections are shown together with the corresponding $\pm\sigma$ variations. The dashed line indicates perfect reconstruction.

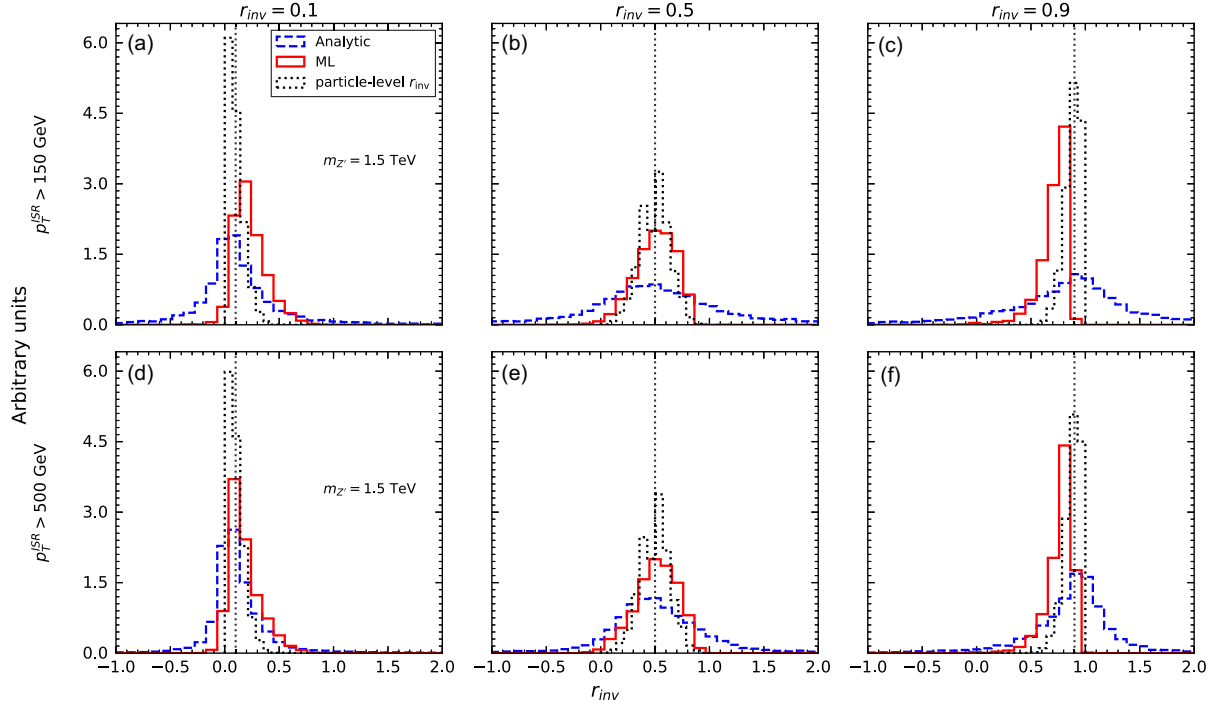


FIG. 7. Comparison between the ML-regressed and analytically reconstructed r_{inv} distributions for $m_{Z'} = 1500$ GeV. Panels (a)–(c) correspond to $p_T^{\text{ISR}} > 150$ GeV with $r_{\text{inv}} = 0.1, 0.5$, and 0.9 , while panels (d)–(f) show the corresponding distributions for $p_T^{\text{ISR}} > 500$ GeV. The dashed histograms indicate the particle-level r_{inv} distributions.

keeping the other conditions fixed. The parameters considered include the probability of dark hadron production (probVector), the dark confinement scale (Λ), the dark pion and dark rho masses (m_{π_D} and m_{ρ_D}), the dark flavor number (N_F^{dark}), as well as the Z' mass ($m_{Z'}$). Figure 9 shows representative results for variations of the dark-sector parameters at fixed $r_{\text{inv}} = 0.5$, including m_{π_D} , Λ ,

probVector, and N_F^{dark} . Only mild changes are observed, suggesting that the model response is not strongly tied to a specific choice of these dark-sector parameters. The mediator mass dependence is shown separately in Fig. 10. This variation probes the stability of the regression response under changes in the hard-process energy scale. The predicted $r_{\text{inv}}^{\text{ML}}$ distributions remain close to the

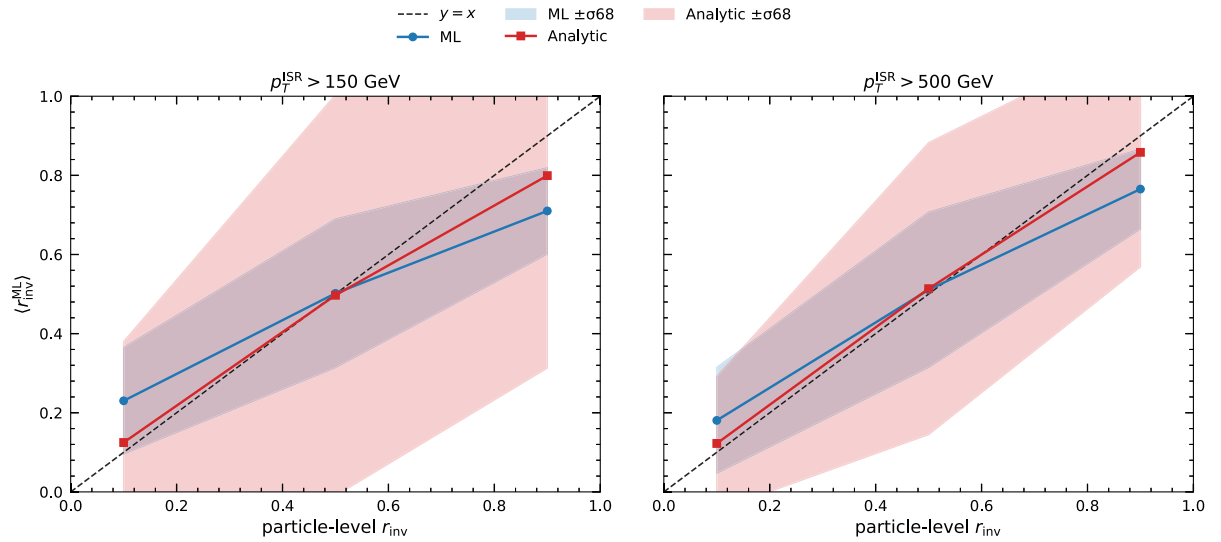


FIG. 8. Mean reconstructed r_{inv} as a function of the particle-level r_{inv} for the ML and analytic methods, shown for $p_T^{\text{ISR}} > 150$ GeV (left) and $p_T^{\text{ISR}} > 500$ GeV (right). The shaded bands indicate the corresponding $\pm\sigma$ variations, and the dashed line corresponds to perfect reconstruction.

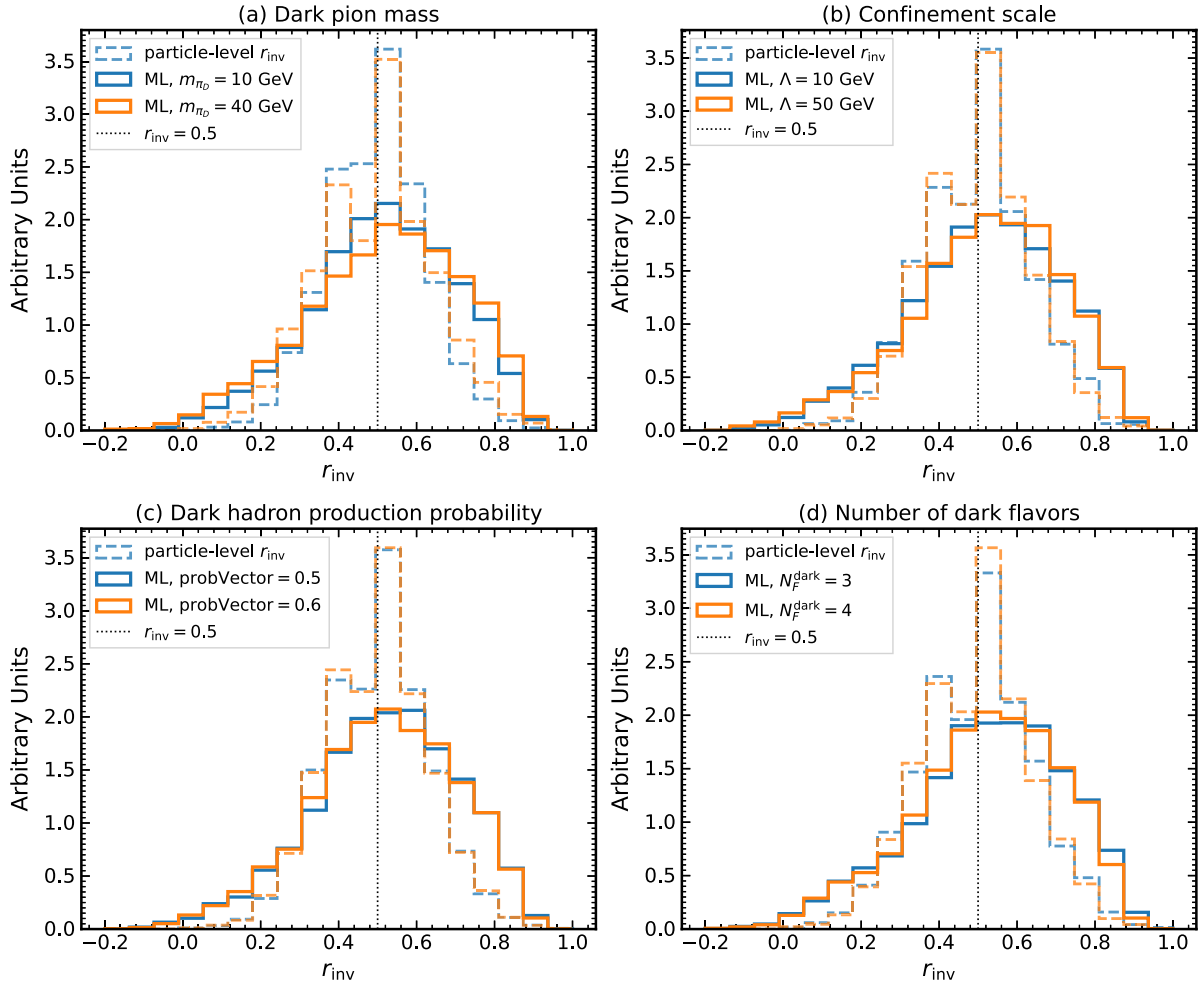


FIG. 9. Regressed $r_{\text{inv}}^{\text{ML}}$ distributions at fixed $r_{\text{inv}} = 0.5$ for variations of the dark-sector parameters: (a) dark pion mass (m_{π_D}), (b) confinement scale (Λ), (c) dark hadron production probability (probVector), and (d) dark flavor number (N_F^{dark}). In each panel, only the parameter indicated is varied, while the others are kept at their baseline values.

corresponding particle-level distributions, indicating that the model response is also stable under the mediator-mass variation considered here.

The results are not surprising. The boosted topology induced by a high- p_T ISR object improves the resolution of the relevant observables and makes the geometric structure of the system clearer [32]. In such a phase-space, those event-level quantities are not impacted by the details of dark shower, unlike jet substructure type variables. Three m_{π_D} points are compared here as illustrative examples, including two extreme values, 100 and 200 GeV. When m_{π_D} is increased to such high values, the dark shower produces fewer dark hadrons, which in turn modifies shower-level observables such as shown in Fig. 11. While the number of charged particles sees a clear shift, the E_T^{miss} distribution changes mildly in the bulk. Figure 12 shows the regression performance remains broadly stable for those conditions. It is reasonable to anticipate an algorithm exploring the correlations between charged particles may be affected to a much greater extent.

B. Applicability to t -channel production

Although, the t -channel involves a scalar mediator Φ that couples an SM quark to a dark quark, very different to the s -channel production, the event-topology is very similar. Both production channels have a back-to-back SVJ system boosted by an ISR object. Given the same r_{inv} , the regression model should achieve similar performance for t -channel as well, even if the model is trained using s -channel samples only, which is confirmed in Fig. 13.

Conventionally, the searches for s -channel and t -channel production are separated as the former relies on observables trying to reconstruct the resonance peak, which is not feasible for the other. The above finding is encouraging, since the regressed r_{inv} can be a potential powerful discriminant variable that is universally applicable to both production channels. We also want to point it out here that previous searches for t -channel SVJ places stringent selections to reduce the multijet background, resulting in a very different background composition, dominated by electroweak boson processes. $r_{\text{inv}}^{\text{ML}}$ may provide sufficient

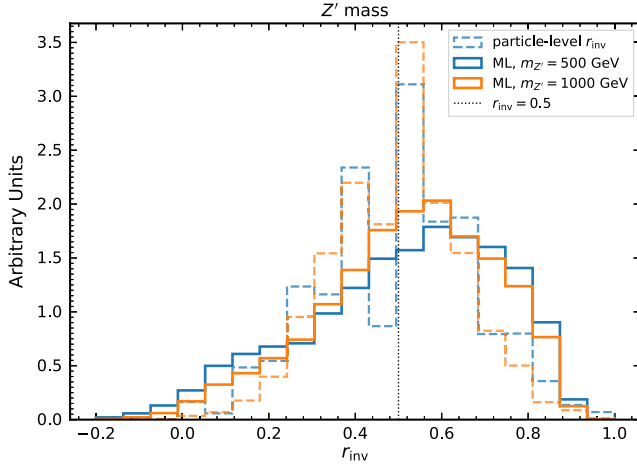


FIG. 10. Regressed $r_{\text{inv}}^{\text{ML}}$ distributions under variations of the mediator mass $m_{Z'}$ at fixed $r_{\text{inv}} = 0.5$. The dashed histograms show the particle-level r_{inv} distributions, while the solid histograms show the corresponding $r_{\text{inv}}^{\text{ML}}$ predictions evaluated with the baseline regression model. The vertical dotted line marks the nominal $r_{\text{inv}} = 0.5$ value. The similar particle-level and ML distributions indicate that the regression response remains stable under the mediator-mass variation considered here.

separation power so that t -channel searches can loosen their selection criteria to have similar background components as the t -channel, which makes it feasible to consider both in a unified approach.

C. Analysis impact

Previous studies have shown that using $r_{\text{inv}}^{\text{analytic}}$ with m_{MAOS} , a mass variable built from the M_{T2} -assisted on-shell reconstruction of the invisible momenta, allows us to

obtain much higher sensitivity [32,38,39] for s -channel. The conclusion applies to $r_{\text{inv}}^{\text{ML}}$ as well, given its improved r_{inv} precision. As shown in Fig. 14, the signal events are more separated from background on the $r_{\text{inv}} - m_{\text{MAOS}}$ plane when $r_{\text{inv}}^{\text{ML}}$ is considered. It is also worth highlighting that both s -channel and t -channel exhibit similar clusters, while the peak is lower for t -channel in m_{MAOS} .

The enhancement can be appreciated more via a quantified approach. We calculate the bin-wise $S/\sqrt{B}$ value in the two-dimensional plane, and Table IV compares the maximum value obtained using analytic r_{inv} and $r_{\text{inv}}^{\text{ML}}$. It is striking that $r_{\text{inv}}^{\text{ML}}$ brings significant gains, especially in the high r_{inv} region.

VI. CONCLUSION

This work proposes a ML-based regression method to reconstruct r_{inv} , the key model parameter in SVJ searches. With the presence of an ISR-object, the model can achieve much better precision compared to the analytical formula developed in Ref. [32]. The training procedure is tailored to reduce the dependence on the signal model parameters, so the performance is robust against various conditions. The underlying physics captures the critical event-topology feature of the ISR + SVJ system, regardless of the production channels. It is found that the regression model trained using s -channel signal events also has similar performance when evaluating on t -channel processes.

Even though we only included photon-ISR processes in this study, the results are still valid for other ISR-objects, as the overall event topology does not change significantly. In the case of jet-ISR, a practical challenge is how to distinguish the ISR-jet from the SVJs. It is shown that simple geometric selections can achieve acceptable

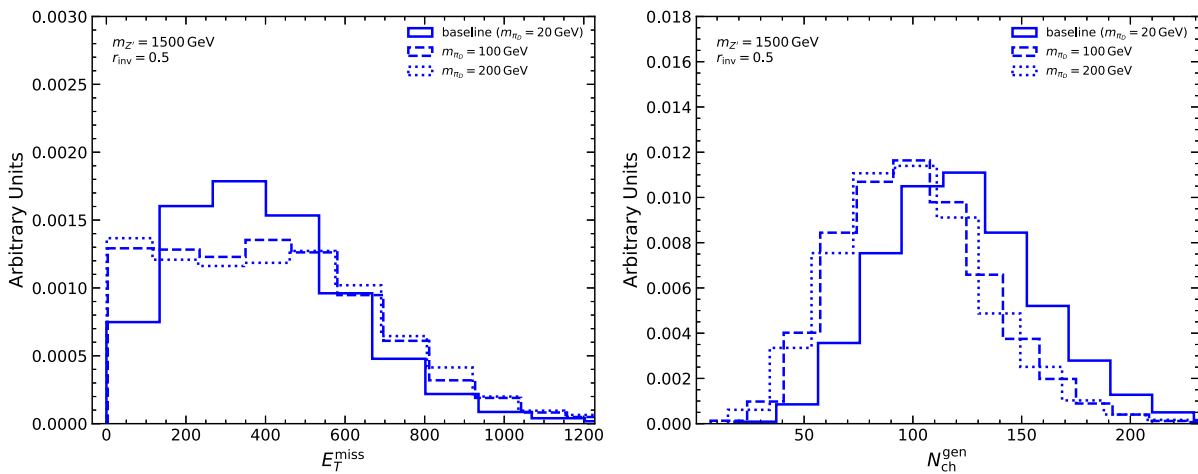


FIG. 11. Left: comparison of the normalized E_T^{miss} distributions for three cases with $m_{\pi_D} = 20$ GeV, 100 GeV, and 200 GeV. When the dark hadron mass is increased to relatively extreme values, the peak of the E_T^{miss} distribution is weakened and the distribution becomes flatter. Right: distribution of the generator-level charged-particle multiplicity $N_{\text{ch}}^{\text{gen}}$, which shifts overall to lower values as the dark meson mass increases.

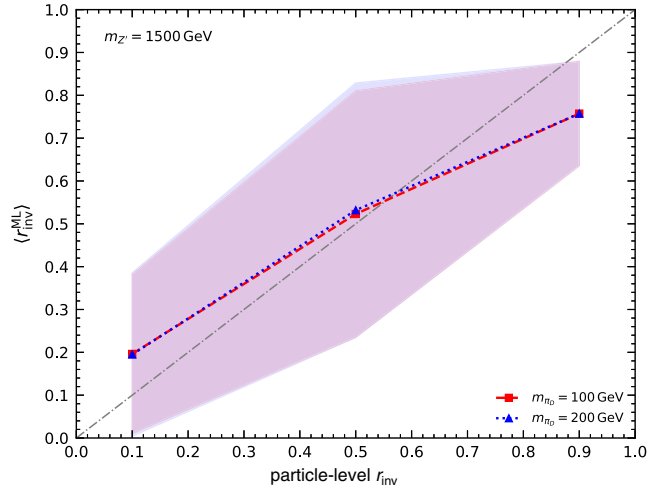


FIG. 12. Mean response of $r_{\text{inv}}^{\text{ML}}$ as a function of the true r_{inv} for the cases with increased dark meson mass. The monotonic and ordered behavior of $r_{\text{inv}}^{\text{ML}}$ is not significantly changed when the dark meson mass is increased, indicating that the regression model remains stable under this parameter variation.

performance [31,32]. Naturally, if one has an excellent SVJ tagging algorithm, this problem is solved automatically [26–28]. It is interesting to further explore the connections between SVJ identification and r_{inv} reconstruction.

This method has great application potential. First, $r_{\text{inv}}^{\text{ML}}$ can serve as a discriminant variable in SVJ searches, given the obvious separation power seen in Fig. 14. Furthermore, it motivates us to rethink about the search strategy as the performance is encouraging for both s -channel and t -channel. It is possible to loosen the selection criteria adopted in t -channel SVJ searches, broadening their coverage. Finally, this regression model is very suitable for reinterpretation [40], since only high-level physics objects are included in the training. It will remain effective as long as the signal models under consideration are characterized by r_{inv} , no matter whether the SVJs are enriched with leptons [41–43], heavy-flavor quarks [44] or long-lived particles [45]. Finally, we want to point out that our method is not tied with how r_{inv} is realized in the generators, as it is calculated using the visible and invisible particles inside the jets directly.

ACKNOWLEDGMENTS

The preliminary result was presented at the LHC Dark Showers Workshop in 2025, where we received several valuable comments. We thank the workshop organizers and participants. In particular, we thank Jackson Burzynski for suggesting t -channel studies. B. X. L. and Y. L. are supported by Shenzhen Campus of the Sun Yat-sen University under project 74140-12240013, and by the Young Scientists Fund (C Class) of the National Natural Science Foundation of China (Grant No. 12405122). B. X. L. is also supported by the Natural Science Foundation of Guangdong Province

(Grant No. 2026A1515012178). J. B. W., J. Q. X., K. R. X., and R. H. Y. are supported by College Students’ Innovative Entrepreneurial Training Plan Program, and they developed the training/validation pipeline. Z. H. H. selected this topic for their undergraduate thesis, and developed the sample production workflow.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

APPENDIX: MODEL AND TRAINING DETAILS

This appendix provides additional technical details of the regression model used to estimate the event-level invisible fraction. The model takes reconstruction-level observables as inputs, including the ISR photon, the two leading jets, the missing transverse momentum, and derived angular and kinematic variables. The input variables are preprocessed using a physics-motivated normalization procedure before training.

The leading-jet transverse momentum is then used as the event-wise normalization scale. The photon transverse momentum, subleading-jet transverse momentum, missing transverse momentum, leading-jet mass, and subleading-jet mass are divided by the leading-jet transverse momentum. The leading-jet transverse momentum itself is used only as the normalization scale and is not retained as an independent input feature. After this physical normalization, the normalized transverse-momentumlike, mass, and missing-momentum variables are standardized using the mean and standard deviation of the training sample. The azimuthal-angle and pseudorapidity variables are kept in their original forms.

The regression network is a fully connected feed-forward neural network. Each hidden layer is followed by batch normalization and a ReLU activation. The output layer consists of a single neuron corresponding to the predicted value of r_{inv} . The model is trained by minimizing the mean-squared error between the predicted r_{inv} and the particle-level target value defined in Eq. (2). The training uses the Adam optimizer, weight decay, a validation-based learning-rate scheduler, and gradient clipping.

The network architecture and training hyperparameters are selected using an Optuna-based hyperparameter optimization [46]. The validation loss is used as the optimization objective, so the different network depths used for the $p_{\text{T}}^{\text{ISR}} > 150$ GeV and $p_{\text{T}}^{\text{ISR}} > 500$ GeV selections are the result of validation-based hyperparameter selection rather than a manual architectural choice. The final hyperparameters used in the analysis are summarized in Table V.

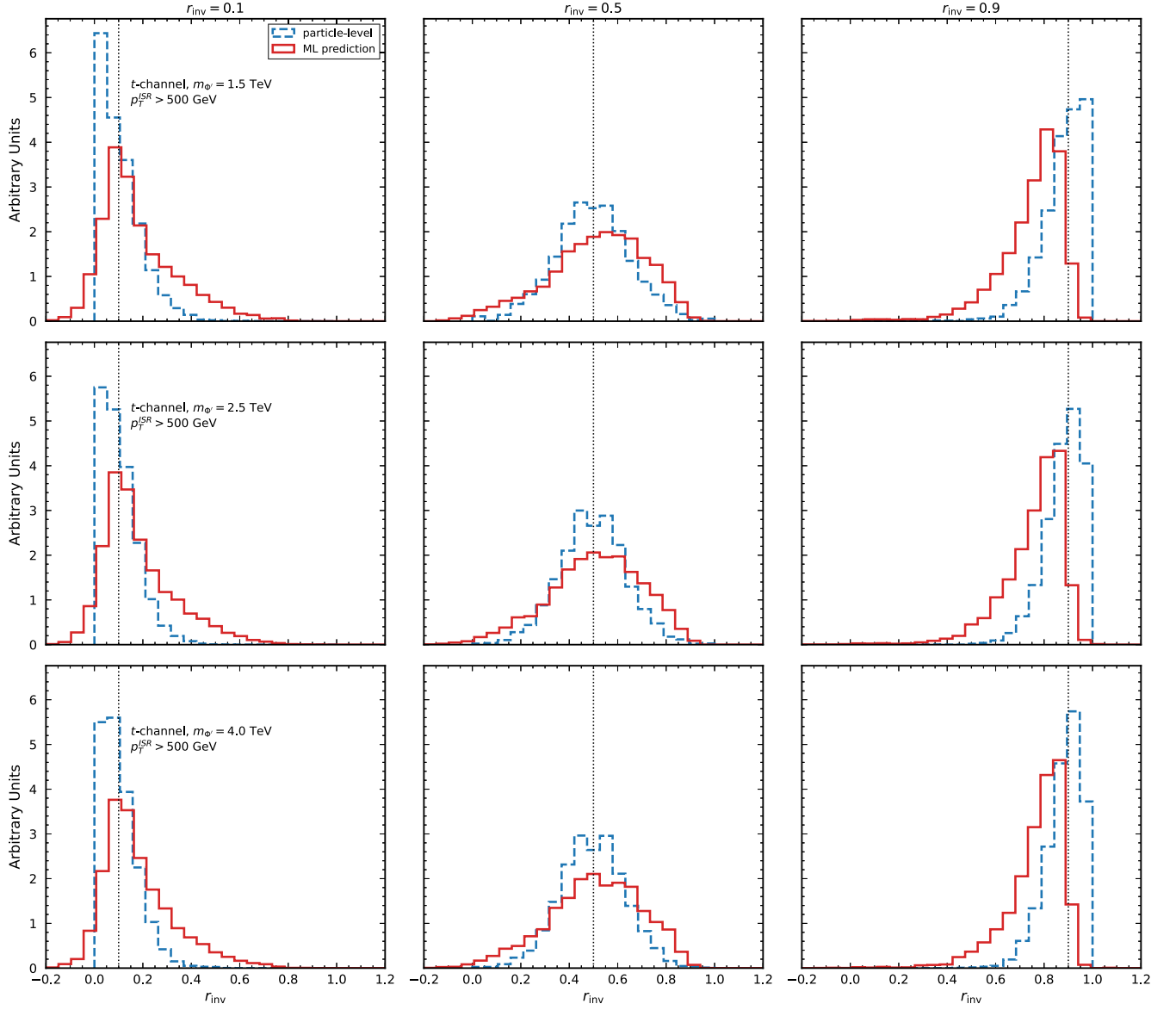


FIG. 13. Distributions of $r_{\text{inv}}^{\text{ML}}$ for t -channel samples with $m_\phi = 1500$ GeV (upper row), 2500 GeV (middle row) and 4000 GeV (lower row), evaluated using the model trained on s -channel samples. From left to right, the panels correspond to $r_{\text{inv}} = 0.1, 0.5$, and 0.9 . The dashed histograms indicate the particle-level r_{inv} distributions.

The code used for the Delphes-to-CSV conversion, r_{inv} prediction, and plotting workflow is publicly available at [47]. The repository contains a lightweight inference workflow, including data preprocessing utilities, trained model loading, r_{inv} prediction, and example plotting scripts.

Starting from Delphes-level inputs, the workflow converts the selected objects into the required tabular format, applies the same preprocessing as in the analysis, evaluates the trained regression model, and produces the corresponding r_{inv} distributions.

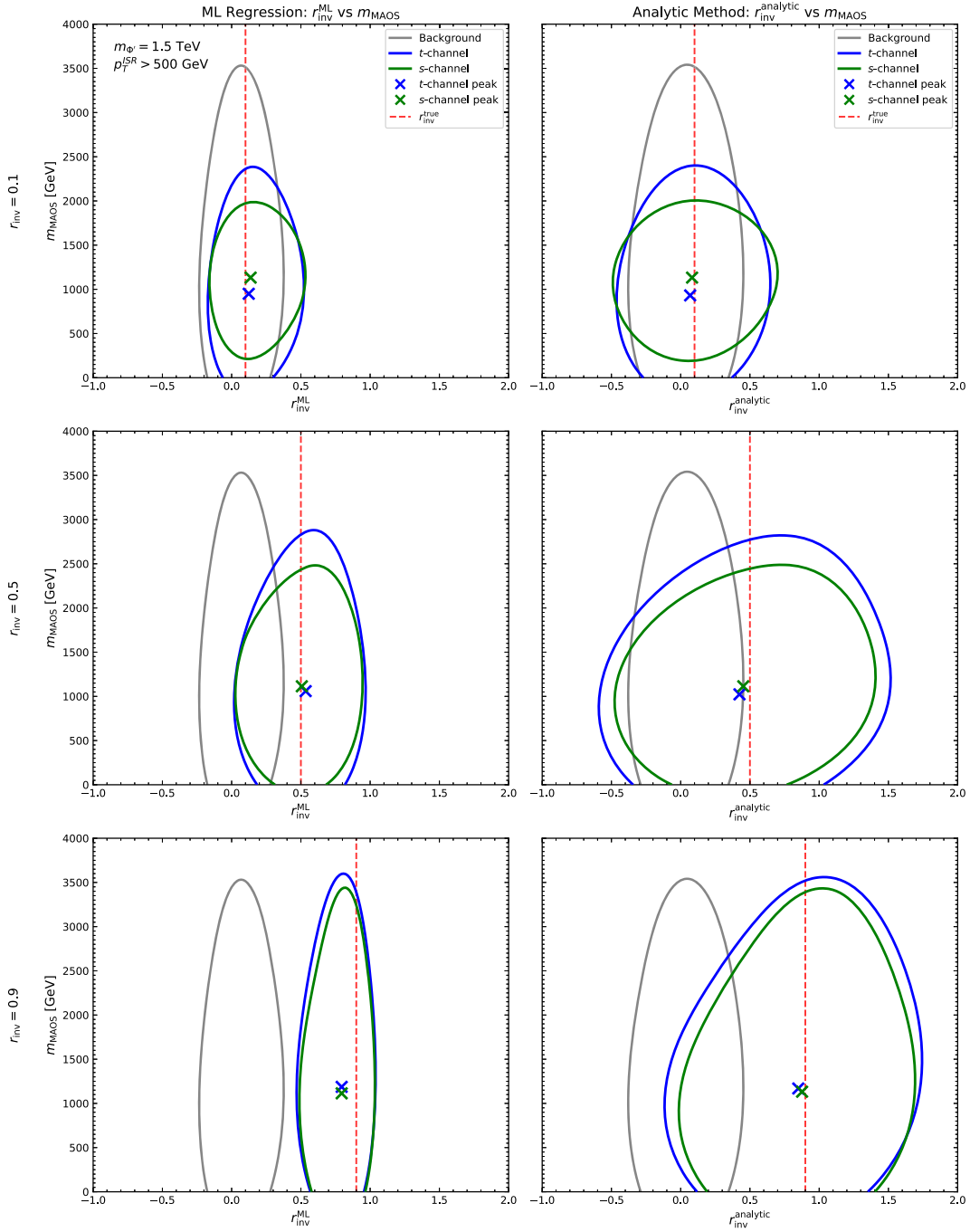


FIG. 14. 80% contour regions in the two-dimensional $(r_{\text{inv}}, m_{\text{MAOS}})$ plane for the t -channel and s -channel samples at $m_\phi = m_{Z'} = 1.5$ TeV and $p_{\text{T}}^{\text{ISR}} > 500$ GeV, shown for $r_{\text{inv}} = 0.1$ (top), 0.5 (middle), and 0.9 (bottom). The left column uses $r_{\text{inv}}^{\text{ML}}$, and the right column uses $r_{\text{inv}}^{\text{analytic}}$. In each panel, the gray contour denotes the background, the blue and green contours denote the t -channel and s -channel signals, respectively, and the markers indicate the corresponding density peaks. The red dashed line marks the particle-level r_{inv} value.

TABLE IV. Maximum $S/\sqrt{B}$ obtained in the two-dimensional $r_{\text{inv}}, m_{\text{MAOS}}$ plane for the machine-learning-regressed and analytically reconstructed r_{inv} . The relative improvement is defined with respect to the analytic method.

$r_{\text{inv}}^{\text{true}}$	ML regression	Analytic reconstruction	Relative improvement
0.1	14.33	11.88	20.7%
0.5	58.67	32.27	81.8%
0.9	218.33	88.47	146.8%

TABLE V. Final hyperparameters selected for the r_{inv} regression models.

Hyperparameter	$p_{\text{T}}^{\text{ISR}} > 150 \text{ GeV}$	$p_{\text{T}}^{\text{ISR}} > 500 \text{ GeV}$
Hidden layers	[256, 128, 64, 32, 16]	[128, 64, 32, 16]
Learning rate	10^{-3}	10^{-5}
Batch size	128	64
Maximum epochs	200	200
Weight decay	10^{-5}	10^{-5}

- [1] V. C. Rubin, N. Thonnard, and W. K. Ford, Jr., Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605/R = 4 kpc/ to UGC 2885/R = 122 kpc/, *Astrophys. J.* **238**, 471 (1980).
- [2] M. Persic, P. Salucci, and F. Stel, The universal rotation curve of spiral galaxies: 1. The dark matter connection, *Mon. Not. R. Astron. Soc.* **281**, 27 (1996).
- [3] D. Clowe, M. Bradac, A. H. Gonzalez, M. Markevitch, S. W. Randall, C. Jones, and D. Zaritsky, A direct empirical proof of the existence of dark matter, *Astrophys. J. Lett.* **648**, L109 (2006).
- [4] DES Collaboration, Dark energy survey year 1 results: Curved-sky weak lensing mass map, *Mon. Not. R. Astron. Soc.* **475**, 3165 (2018).
- [5] Planck Collaboration, Planck 2018 results. VI. Cosmological parameters, *Astron. Astrophys.* **641**, A6 (2020).
- [6] G. Jungman, M. Kamionkowski, and K. Griest, Super-symmetric dark matter, *Phys. Rep.* **267**, 195 (1996).
- [7] ATLAS Collaboration, Search for new phenomena in events with an energetic jet and missing transverse momentum in pp collisions at $\sqrt{s} = 13 \text{ TeV}$ with the ATLAS detector, *Phys. Rev. D* **103**, 112006 (2021).
- [8] CMS Collaboration, Search for new particles in events with energetic jets and large missing transverse momentum in proton-proton collisions at $\sqrt{s} = 13 \text{ TeV}$, *J. High Energy Phys.* **11** (2021) 153.
- [9] ATLAS Collaboration, Exploration at the high-energy frontier: ATLAS run 2 searches investigating the exotic jungle beyond the standard model, *Phys. Rep.* **1116**, 301 (2025).
- [10] CMS Collaboration, Dark sector searches with the CMS experiment, *Phys. Rep.* **1115**, 448 (2025).
- [11] M. J. Strassler and K. M. Zurek, Echoes of a hidden valley at hadron colliders, *Phys. Lett. B* **651**, 374 (2007).
- [12] T. Han, Z. Si, K. M. Zurek, and M. J. Strassler, Phenomenology of hidden valleys at hadron colliders, *J. High Energy Phys.* **07** (2008) 008.
- [13] T. Cohen, M. Lisanti, and H. K. Lou, Semivisible jets: Dark matter undercover at the LHC, *Phys. Rev. Lett.* **115**, 171804 (2015).
- [14] T. Cohen, M. Lisanti, H. K. Lou, and S. Mishra-Sharma, LHC searches for dark sector showers, *J. High Energy Phys.* **11** (2017) 196.
- [15] H. Beaudesne, E. Bertuzzo, and G. Grilli Di Cortona, Dark matter in hidden valley models with stable and unstable light dark mesons, *J. High Energy Phys.* **04** (2019) 118.
- [16] H. Beaudesne, E. Bertuzzo, G. Grilli Di Cortona, and Z. Tabrizi, Collider phenomenology of hidden valley mediators of spin 0 or 1/2 with semivisible jets, *J. High Energy Phys.* **08** (2018) 030.
- [17] G. Albouy *et al.*, Theory, phenomenology, and experimental avenues for dark showers: A snowmass 2021 report, *Eur. Phys. J. C* **82**, 1132 (2022).
- [18] CMS Collaboration, Search for resonant production of strongly coupled dark matter in proton-proton collisions at 13 TeV, *J. High Energy Phys.* **06** (2022) 156.
- [19] ATLAS Collaboration, Search for new physics in final states with semivisible jets or anomalous signatures using the ATLAS detector, *Phys. Rev. D* **112**, 012021 (2025).
- [20] ATLAS Collaboration, Search for non-resonant production of semivisible jets using run 2 data in ATLAS, *Phys. Lett. B* **848**, 138324 (2024).
- [21] D. Kar and S. Sinha, Exploring jet substructure in semivisible jets, *SciPost Phys.* **10**, 084 (2021).
- [22] M. Park and M. Zhang, Tagging a jet from a dark sector with Jet-substructures at colliders, *Phys. Rev. D* **100**, 115009 (2019).
- [23] T. Cohen, J. Doss, and M. Freytsis, Jet substructure from dark sector showers, *J. High Energy Phys.* **09** (2020) 118.
- [24] T. Cohen, J. Roloff, and C. Scherb, Dark sector showers in the Lund jet plane, *Phys. Rev. D* **108**, L031501 (2023).
- [25] T. Faucett, S.-C. Hsu, and D. Whiteson, Learning to identify semi-visible jets, *J. High Energy Phys.* **12** (2022) 132.
- [26] L. Favaro, M. Krämer, T. Modak, T. Plehn, and J. Rüsckamp, Semi-visible jets, energy-based models, and self-supervision, *SciPost Phys.* **18**, 042 (2025).
- [27] E. Bernreuther, T. Finke, F. Kahlhoefer, M. Krämer, and A. Mück, Casting a graph net to catch dark showers, *SciPost Phys.* **10**, 046 (2021).

- [28] F. Canelli, A. de Cosa, L. L. Pottier, J. Niedziela, K. Pedro, and M. Pierini, Autoencoders for semivisible jet detection, *J. High Energy Phys.* **02** (2022) 074.
- [29] A. Bhardwaj, C. Englert, W. Naskar, V. S. Ngairangbam, and M. Spannowsky, Equivariant, safe and sensitive—graph networks for new physics, *J. High Energy Phys.* **07** (2024) 245.
- [30] K. Pedro and P. Shyamsundar, Optimal mass variables for semivisible jets, *SciPost Phys. Core* **6**, 067 (2023).
- [31] ATLAS Collaboration, Search for low-mass resonances decaying into two jets and produced in association with a photon or a jet at $s = 13$ TeV with the ATLAS detector, *Phys. Rev. D* **110**, 032002 (2024).
- [32] B. Liu and K. Pedro, Semi-visible jets + X: Illuminating dark showers with radiation, *J. High Energy Phys.* **12** (2024) 105.
- [33] 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, *J. High Energy Phys.* **07** (2014) 079.
- [34] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and Peter Z. Skands, An introduction to PYTHIA 8.2, *Comput. Phys. Commun.* **191**, 159 (2015).
- [35] DELPHES 3 Collaboration, DELPHES 3, A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [36] M. Cacciari, G. P. Salam, and G. Soyez, The anti- k_t jet clustering algorithm, *J. High Energy Phys.* **04** (2008) 063.
- [37] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, *Eur. Phys. J. C* **72**, 1896 (2012).
- [38] W. S. Cho, K. Choi, Y. G. Kim, and C. B. Park, M(T2)-assisted on-shell reconstruction of missing momenta and its application to spin measurement at the LHC, *Phys. Rev. D* **79**, 031701 (2009).
- [39] K. Choi, J. S. Lee, and C. B. Park, Measuring the Higgs boson mass with transverse mass variables, *Phys. Rev. D* **82**, 113017 (2010).
- [40] A. Buckley, J. Butterworth, L. Corpe, C. Doglioni, D. Kar, C. Prat *et al.*, Probing the sensitivity of semi-visible jets to current LHC measurements using the CONTUR toolkit, [arXiv:2502.11237](https://arxiv.org/abs/2502.11237).
- [41] C. Cazzaniga and A. de Cosa, Leptons lurking in semi-visible jets at the LHC, *Eur. Phys. J. C* **82**, 793 (2022).
- [42] CMS Collaboration, Search for s -channel production of lepton-enriched semivisible jets in proton-proton collisions at 13 TeV, CERN Report No. CMS-PAS-EXO-24-029, 2025.
- [43] CMS Collaboration, Search for s -channel production of tau-enriched semivisible jets in proton-proton collisions at 13 TeV, CERN Report No. CMS-PAS-EXO-26-005, 2025.
- [44] D. Kar, W. Nzuza, and S. Sinha, 2B or not 2B, a study of bottom-quark-philic semi-visible jets, *SciPost Phys. Core* **7**, 071 (2024).
- [45] J. Carrasco, S. Kulkarni, W. Liu, J. Lockyer, and J. Zurita, Semi-visible emerging jets, *J. High Energy Phys.* **07** (2026) 219.
- [46] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, New York, NY, 2019), pp. 2623–2631, [10.1145/3292500.3330701](https://doi.org/10.1145/3292500.3330701).
- [47] https://github.com/YinLi28/svj_rinv_inference_demo.