

Optimal transport event representation for anomaly detection

Tianji Cai^{1,2,*}, Aditya Bhargava³, and Benjamin Nachman^{4,5}

¹*School of Physical Science and Engineering, Tongji University, Shanghai 200092, China*

²*State Key Laboratory of Autonomous Intelligent Unmanned Systems,*

MOE Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 200092, China

³*Department of Physics, University of California, Berkeley, California 94720, USA*

⁴*Department of Particle Physics and Astrophysics, Stanford University, Stanford, California 94305, USA*

⁵*Fundamental Physics Directorate, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA*



(Received 24 March 2026; accepted 24 June 2026; published 10 July 2026)

We introduce optimal transport (OT) as a physics-based intermediate event representation for weakly supervised anomaly detection. With only 0.5% injection of resonant signals in the LHC Olympics benchmark datasets, the OT-augmented feature set achieves nearly twice the significance improvement of the standard high-level observables using an idealized setup, while end-to-end deep learning on low-level four-momenta is less effective in this low-signal regime. The observed gains persist across signal types and classifiers considered in this study, suggesting that structured, physics-informed representations can provide a useful complement to existing approaches for anomaly detection.

DOI: [10.1103/wwzk-25sy](https://doi.org/10.1103/wwzk-25sy)

I. INTRODUCTION

As targeted searches struggle to find new physics beyond the Standard Model, alternative model-agnostic strategies have become increasingly important. Anomaly detection (AD), powered by modern machine learning (ML), offers a promising route to identify rare or unexpected signals without specific model hypotheses [1–4]. A prominent class of AD frameworks follows the *weak supervision* (WS) paradigm [5–8], where classifiers are trained to distinguish between signal-rich data and background-only references. These approaches are particularly effective for resonant anomalies, i.e., localized signals atop smooth background distributions, and have been successfully deployed to experimental data [9–17].

Methods within the WS paradigm differ primarily in background modeling and classifier design [18–23], both of which depend on the underlying event representation. Most studies use a number of standard high-level observables to summarize the event, which may restrict the scope of detectable signals if the chosen observables cannot capture signal characteristics. To broaden sensitivity, recent works [20,24] used as inputs a maximal feature set, i.e., the final-state kinematics of an event. While powerful, methods

based on low-level four-momenta usually require training (or pretraining [24]) a large (foundation) model on substantial amounts of data, and are typically less effective in the (ultra) low-signal regime where anomaly detection is most needed.

We propose an alternative approach that exploits full final-state kinematics while avoiding sophisticated, resource-intensive models and extensive training data. In particular, *optimal transport* (OT) theory is introduced to define a novel intermediate representation, capturing essential event structures in a compact and structured form. OT defines a distance between probability distributions by computing the most efficient way to transform one into another [25]. Its sensitivity to geometric structure has been utilized to equip the space of collider events with physically meaningful metrics [26–33], showing promises across collider applications including jet and event classification [34,35], calibration [36], and more recently, anomaly detection at the event level [37–39].

Nonetheless, the full potential of optimal transport as an *event representation* remains largely untapped. Here we propose the conceptual shift: rather than merely using OT to define metrics between events, we develop an efficient representation through its linearization. When combined with standard high-level observables, features derived from this OT representation yield noticeable improvements for resonant anomaly detection, particularly in the low-signal regime. Notably, only a small number of OT-based features is required to enhance performance, with further gains achieved when augmented by additional higher-order sub-jettiness. These results highlight the value of physically

*Contact author: tianji_cai@tongji.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.

grounded representations as a bridge between engineered features and fully end-to-end machine learning, encouraging further exploration of model latent spaces and alternative intermediate representations beyond the specific constructions considered here.

II. DATASETS AND WEAK SUPERVISION FRAMEWORK

We use datasets from the 2020 LHC Olympics (LHCO) anomaly detection challenge [1,40,41], which includes two signal-injected sets (R&D1 and R&D2) and a background-only set. Each R&D dataset contains one million QCD dijet events and 100 thousand signal events from a resonant decay $W' \rightarrow XY$ with $m_{W'}, m_X, m_Y = 3.5 \text{ TeV}, 500 \text{ GeV}, 100 \text{ GeV}$, respectively. In R&D1, $X, Y \rightarrow qq$, producing two-pronged jets; in R&D2, $X, Y \rightarrow qqq$, yielding three-prong substructure.

Events are simulated in Pythia 8 [42,43] and Delphes 3.4.1 [44], whereas jets are clustered in FastJet [45] via the anti- k_T algorithm [46] with $R = 1$. Events whose leading jet has $p_T > 1.2 \text{ TeV}$ are selected. Only particles in the two leading- p_T jets are retained, which are expected to contain the X/Y decay products. The signal region (SR) is defined as $m_{JJ} \in [3.3, 3.7] \text{ TeV}$, with its complement forming the sideband regions (SB). This is chosen for consistency with prior works [18,20]. Each R&D dataset is further augmented with $\sim 610 \text{ k}$ background events in the SR from Ref. [41], yielding $\sim 731 \text{ k}$ background and $\sim 75 \text{ k}$ signal events in total.

Weak supervision trains on mixed samples without requiring event-level labels or class proportions, and is provably as powerful as full supervision by the Neyman-Pearson lemma in asymptotic limit [47]. This underlies the classification without labels (CWoLa) framework [6], where recent CWoLa-based anomaly detectors model SR background by interpolating SB data via classifiers [48], density estimation [21,22], generative models [18], and invertible networks [49]. These methods are now capable of near optimal matching. We therefore assume perfect background interpolation and adopt an *idealized anomaly detector* (IAD) [18], which enables us to isolate the effect of the proposed representation and to perform a controlled representation-level comparison within the weakly supervised setting.¹ Since background-modeling effects are removed by construction, the resulting performance primarily reflects the discriminating power of the underlying representation itself, rather than the details of a particular background-estimation strategy. The IAD framework is therefore well suited to the central goal of this work, namely understanding

the physical information encoded in the novel OT event representation.

Two training sets (A1 and A2) are constructed in the signal region. The set A1 contains only background events (labeled 0), whereas the set A2 includes background plus a fixed fraction of signal events, i.e., $S/B = 0.5\%, 0.63\%, 0.7\%, 0.8\%, 1\%, 1.2\%, 1.4\%, 1.6\%, 10\%$, covering a large spectrum of possible signal amounts. In total, there are nine A2 sets, one for each S/B level, and all events in A2 are labeled 1. Classifiers are then trained to distinguish events in A1 from those in A2, using only dataset-level labels. The main results use boosted decision trees (BDTs) [50,51] as the classifier, thanks to its robustness to noise and correlated features. Results using multilayer perceptrons (MLPs) are also presented in the Appendix for comparison. To reduce statistical fluctuations, we aggregate over 50 independently trained BDT classifiers (10 for MLPs as in [18]) to obtain the results for one ensemble, and further average over 10 ensembles to estimate the mean and the spread.² Performance is evaluated using event-level labels on two separate test sets—background-only B1 and signal-only B2. On the statistical level, our datasets match those used in Ref. [18], with A1 having 272 k background events, A2 containing 121,352 background events plus injected signal events, B1 having 340 k background events, and B2 having 20 k signal events—all within the SR.

III. OPTIMAL TRANSPORT REPRESENTATION AND DERIVED FEATURES

Standard analyses rely on high-level observables such as jet mass and n -subjettiness ratios [52,53]. The official LHCO R&D1 itself provides $\{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}\}$ for the two leading jets, whereas R&D2 adds $\{\tau_{32}^{J_1}, \tau_{32}^{J_2}\}$. Here we propose an intermediate representation based on particle four-momenta, using a specific OT distance called the 2-Wasserstein metric (W_2). For a pair of events $\mathcal{E}$ and $\tilde{\mathcal{E}}$ containing massless constituent particles at $x_i, \tilde{x}_j \in (y, \phi) \subseteq \mathbb{R}^2$ with normalized transverse momenta $p_T^i, \tilde{p}_T^j$ respectively, the 2-Wasserstein distance is defined as

$$W_2(\mathcal{E}, \tilde{\mathcal{E}}) = \min_{g_{ij} \in \Gamma(\mathcal{E}, \tilde{\mathcal{E}})} \left(\sum_{ij} g_{ij} \|x_i - \tilde{x}_j\|^2 \right)^{1/2},$$

$$\Gamma(\mathcal{E}, \tilde{\mathcal{E}}) = \left\{ g_{ij} : g_{ij} \geq 0, \sum_j g_{ij} = p_T^i, \sum_i g_{ij} = \tilde{p}_T^j \right\}, \quad (1)$$

¹The IAD framework should not necessarily be interpreted as a strict upper bound on anomaly detection performance. In realistic weakly supervised analyses, background mismodeling may either reduce or, in some circumstances, artificially increase the apparent discrimination power of a classifier.

²A tree consists of 200 iterations with up to 31 leaf nodes. An MLP has three hidden layers, each with 64 ReLU units. We found the results to be robust to reasonable architectural variations and therefore did not further optimize model hyperparameters.

where $\Gamma(\mathcal{E}, \tilde{\mathcal{E}})$ contains the transport plans that move g_{ij} amount of p_T from particle i in $\mathcal{E}$ to particle j in $\tilde{\mathcal{E}}$ over a distance $\|x_i - \tilde{x}_j\|$ on the $y - \phi$ plane. Thus, W_2 quantifies the minimal cost to morph one event into another and is by definition infrared and collinear (IRC) safe.

To convert the pairwise distance into an event representation, we employ the linearization method originally introduced to reduce OT's computational cost [28,54]. The linearized W_2 ($\text{Lin}W_2$) embeds each event into the tangent space at a fixed reference event $\mathcal{R}$ containing particles with $p_T = R_i$ at $y_i \in (y, \phi)$. Its i th component is given by

$$\text{Lin}W_2^i(\mathcal{E})|_{\mathcal{R}} := \frac{1}{\sqrt{R_i}} \sum_j r_{ij} x_j \quad \text{for } i \in [1, R], \quad (2)$$

where r_{ij} is an optimal transport plan from $\mathcal{R}$ to $\mathcal{E}$ and the RHS gives the p_T -weighted barycenter of where each reference particle is mapped. Since $x_j = (y_j, \phi_j)$, $\text{Lin}W_2^i$ is itself a vector in $\mathbb{R}^2$, resulting in a $2R$ -dimensional embedding for each event which is likewise IRC-safe. Instead of treating $\text{Lin}W_2$ as merely an efficient approximation to the original W_2 distance, we highlight this novel representation in its own right, marking a major conceptual shift from earlier works.

In practice, the $\text{Lin}W_2$ embedding is applied at the jet level, where each jet is preprocessed by centering its axis and vertically aligning the principal component of its transverse momentum flow. The reference jet $\mathcal{R}$ is constructed as a 10×10 grid over the (y, ϕ) plane, with every particle assigned a uniform $p_T = \frac{1}{100}$.³ The W_2 distance between this 10×10 uniform reference and a given jet serves as a measure of the isotropy of the jet radiation pattern [55], and also approximately gives its 100-subjettiness value, in accordance with the geometric reformulation of n -subjettiness observables in terms of an OT distance [31].

To further obtain the representation for an entire event, we concatenate the $\text{Lin}W_2$ embeddings of the two leading jets, yielding a Euclidean vector in $\mathbb{R}^{2 \times 100 \times 2 = 400}$. Extensions to a more general multiscale OT treatment for an entire collider event [35] are left to future work. Constructing the OT event representation scales at most as $\mathcal{O}(n^3)$ in the number of particles in the reference jet $\mathcal{R}$, which is negligible for $n \sim \mathcal{O}(100)$ and is further parallelizable across events under the linearization scheme.

From the high-dimensional OT representation, we further extract features using a simple principal component analysis (PCA). For a diagnostic sample consisting of 5 k

³The mathematical advantages of using a uniform reference jet were established in [28]. The impact of alternative, including nonuniform, references has been analyzed in [29,54] and found to be minimal provided the reference is neither too concentrated nor too sparse. Accordingly, we adopt a uniform reference with a number of constituents comparable to that of particles in a typical jet.

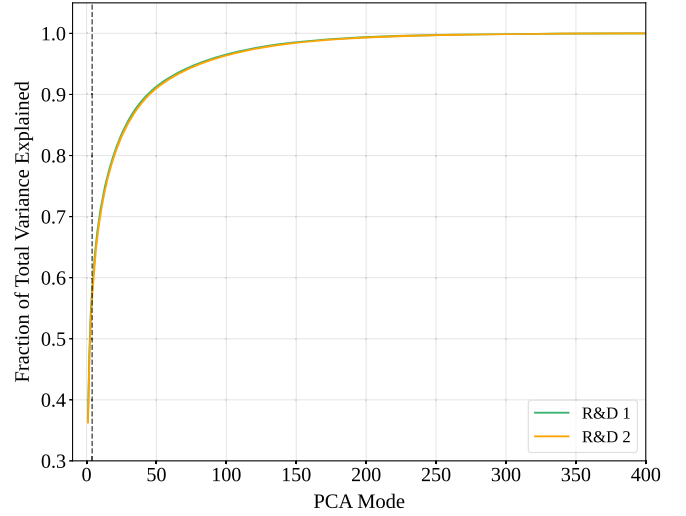


FIG. 1. Total variance explained by increasing numbers of PCA modes of the OT representations for 10 k samples from the R&D1 dataset (green) and the R&D2 dataset (orange).

background and 5 k signal events randomly selected from R&D1 or R&D2, the leading four modes account for almost 60% of the variance, while approximately 100 modes suffice to capture more than 95% of total data variance; see Fig. 1. Note that this diagnostic sample serves only to illustrate the variance structure of the OT representation and plays no role in the anomaly detection analysis. Instead, for each signal injection level in the IAD analysis, the PCA basis is fitted exclusively on the corresponding unlabeled training sample $A_1 + A_2$, without access to event-level signal/background labels. The test sets B_1 and B_2 are subsequently projected onto this training-defined basis, with the first k principal components used as the OT_k feature set.

To further characterize the physical information encoded in the OT representation, we compare the leading OT PCA modes with standard jet observables using Pearson correlations and mutual information, which respectively probe linear and nonlinear relationships; the results are summarized in Table II in the Appendix. We find that the OT features are partially correlated with jet masses and higher-order N -subjettiness ratios, consistent with the known geometric relation between OT and subjettiness. However, the correlations and mutual information values remain far from saturated, indicating that the OT representation captures overlapping but complementary information beyond that contained in jet mass and subjettiness. This interpretation is further supported by the improved performance observed when OT features are combined with standard observables in the following section.

IV. RESULTS

Figure 2 shows the maximum significance improvement (SI) obtained by scanning the signal-to-background ratio

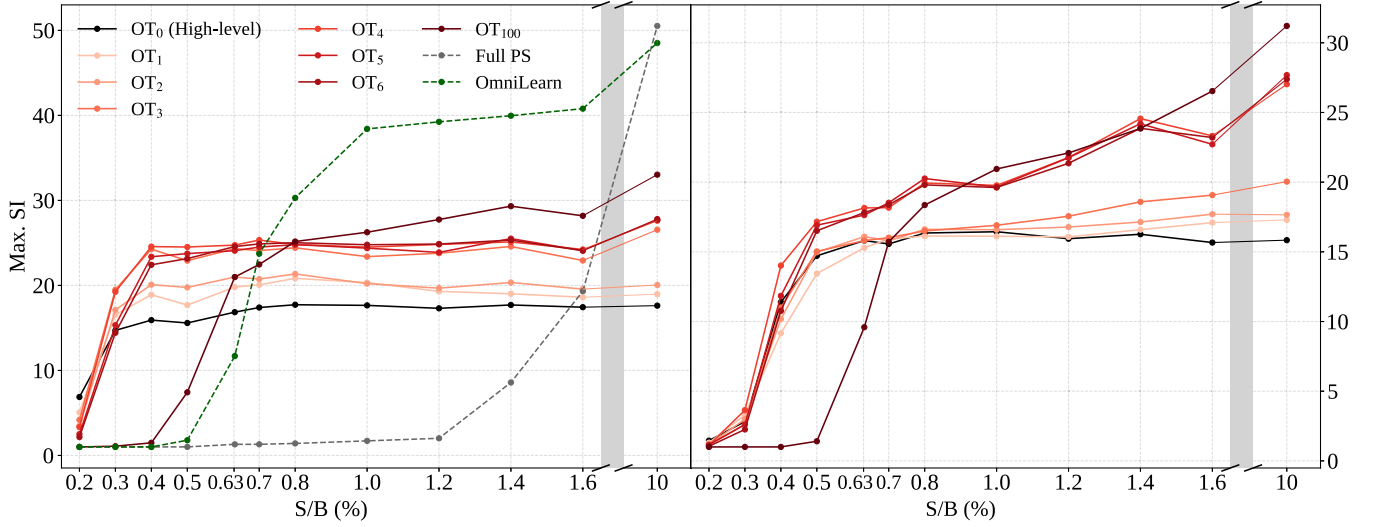


FIG. 2. Maximum significance improvement (SI) for the R&D1 (left) and R&D2 (right) datasets using BDT ensembles as the classifier, with the signal injection level S/B varying from 0.2% to 10%. The group of red curves represent increasing numbers of OT features added to standard high-level observables, i.e., $\{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}\}$ for R&D1, and $\{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}, \tau_{32}^{J_1}, \tau_{32}^{J_2}\}$ for R&D2. The dashed gray line in the left subplot (R&D1) shows the max SI values from Ref. [20] using full phase space as inputs to dedicated models, whereas the dashed green line shows the performance of the pre-trained foundation model OmniLearn from Ref. [24]. Both the full phase space method and foundation models are evaluated in the same IAD setup.

(S/B) from 0.2% to 10% for the R&D1 and R&D2 datasets using BDT ensembles as the classifier. At low signal injection levels ($S/B \lesssim 1.6\%$), feature sets augmented with the first k PCA modes of the OT representation (OT_k) consistently outperform both standard high-level observables provided in the official LHCO datasets and the low-level full phase space (PS) method of Ref. [20]. In the ultralow regime ($S/B \lesssim 0.7\%$), OT_k exceeds the pretrained

foundation model (OmniLearn) of Ref. [24], despite the latter's substantially greater computational cost. Figure 3 further displays the full SI curves at a representative $S/B = 0.63\%$.

On R&D1, the enhancement is most pronounced in the ultralow to low signal regime, where approaches based on the full phase space and foundation models both suffer from limited signal statistics, with the latter alleviating the issue

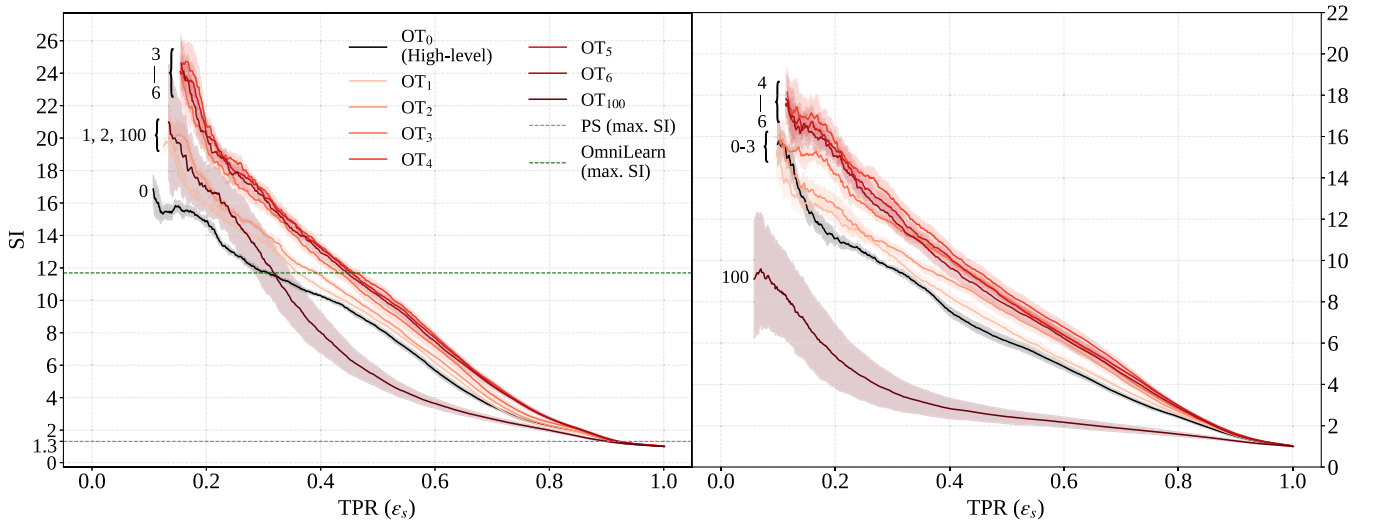


FIG. 3. Significance improvement (SI) curves for the R&D1 (left) and R&D2 (right) datasets at a representative signal injection level of $S/B = 0.63\%$, using BDT ensembles as the classifier. The group of red curves represent increasing numbers of OT features added to standard high-level observables, with shaded bands indicating 1σ variations across the BDT ensembles. For R&D1, the dashed gray line shows the maximum SI values of the full phase space method from Ref. [20], and the dashed green line shows that of the pretrained foundation model OmniLearn from Ref. [24], where both are evaluated in the same IAD setup.

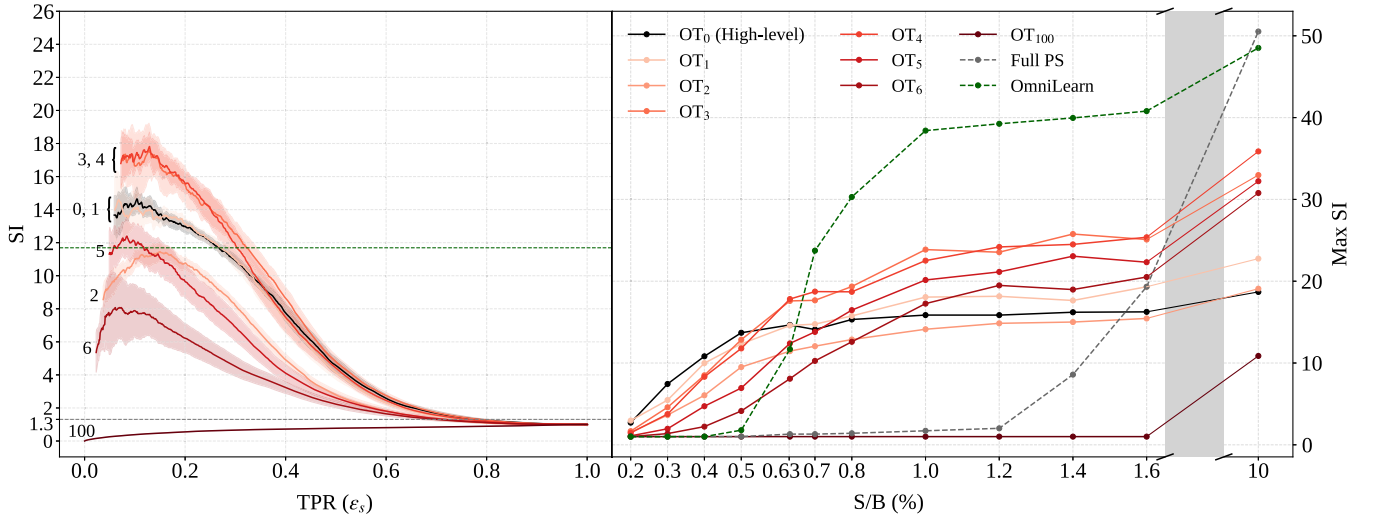


FIG. 4. Anomaly detection for the R&D1 dataset using MLP ensembles as the classifier. Left: the SI curves at the signal injection level $S/B = 0.63\%$. Right: maximum SI at S/B between 0.2% and 10% . The group of red curves represent increasing numbers of OT features added to standard high-level observables $\{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}\}$, with shaded bands denoting 1σ variations across the ensembles of trained models. The dashed gray line shows the performance of the full phase space method from Ref. [20], whereas the dashed green line shows that of the pre-trained foundation model OmniLearn from Ref. [24]. Both the full phase space method and foundation models are evaluated in the same IAD setup.

to some extent through large-scale pre-training. With less than 0.5% signal injection, OT-augmented feature sets raise the maximum SI to $\gtrsim 25$, more than an order of magnitude above the low-level baselines and $\sim 65\%$ above the standard high-level observable set $\{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}\}$. The performance of OT_k (except $k = 100$) remains stable across the entire range of signal fractions. Similar gains are observed in the less-studied R&D2 dataset, albeit with a smaller margin over the standard observables $\{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}, \tau_{32}^{J_1}, \tau_{32}^{J_2}\}$. No full phase space or foundation model benchmark currently exists for R&D2; based on the R&D1 comparison, comparable trends are likely and we encourage future studies to strengthen the low-level benchmarks.

Only a few OT features are needed to achieve the observed gains. The first 3–5 PCA components already saturate the maximum SI, with negligible benefit beyond $k \approx 6$, consistent with the variance profile in Fig. 1. Using very large OT feature sets (e.g., OT_{100}) degrades performance for $S/B \lesssim 0.6\%$, partially reflecting the difficulty of training BDT ensembles with many correlated inputs. The effect is even stronger for MLP classifiers; see Fig. 4 in the Appendix. This illustrates a fundamental trade-off in anomaly detection: expanding the feature set increases the information accessible to more expressive classifiers, but also requires higher signal statistics for effective training. Consequently, well-designed representations and careful feature selection are essential for maximizing discriminative power and attaining optimal sensitivity, especially in the low-signal regime.

In the high-signal regime ($\sim 10\%$), the two low-level approaches perform best, reaching a maximum SI of 50 on

R&D1, consistent with their ability to exploit raw four-momenta when ample signal is present. OT_{100} attains a max SI of 33, $\sim 20\%$ above OT_{3-6} and nearly 80% above the standard observable set $\{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}\}$. An analogous pattern appears in R&D2 at $S/B = 10\%$, where OT_{100} reaches a max SI ≈ 32 , more than doubling the performance of the corresponding jet mass and n -subjettiness ratio inputs. The gap between OT_{100} and the low-level methods likely indicates the presence of information not encoded in the OT representation (see below), possibly associated with IRC-unsafe aspects of the event. Since the OT representation is IRC-safe by construction, its gains are expected to be more robust across signal types and more transferable from simulation to data.

An ablation study on R&D1 at $S/B = 0.63\%$ further isolates the role of OT features; see Fig. 5 in the Appendix. OT features alone, without m_J or τ_{21} , do not yield competitive performance. However, including jet masses and at least the first four OT modes raises the maximum SI to ~ 18 , matching the performance of the standard observable set in Fig. 2. This confirms that OT features encode similar morphological information as n -subjettiness, while the absence of explicit jet-mass information in the current OT construction explains the lower SI (~ 12) obtained when the inputs include only τ_{21} and OT_{3-6} .

On the other hand, to assess whether OT features provide information beyond traditional subjettiness, we repeat the analysis on the R&D1 dataset over $S/B = 0.2\%–10\%$ using BDT ensembles. The inputs now include the *baseline* and three *extended sets* from Ref. [50], consisting of τ_n^β up to $n = 9$ and $\beta = 0.5, 1, 2$. These input sets are further

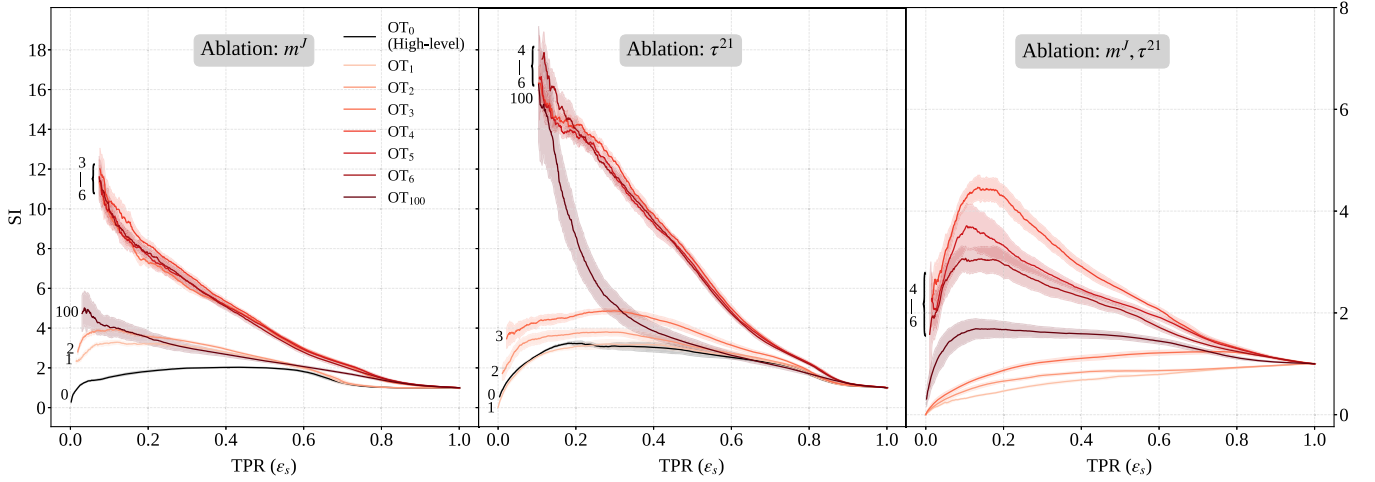


FIG. 5. The SI curves for the R&D1 dataset using BDT ensembles as the classifier at $S/B = 0.63\%$. Left: the input features are $\{\tau_{21}^J, \tau_{21}^{J_2}\}$ and different number of PCA modes for the OT representation (the group of red lines). Middle: the input features are $\{m_{J_1}, m_{J_2}\}$ and different number of PCA modes for the OT representation. Note that the middle subplot shares the same y-axis with the leftmost subplot. Right: the input features only include different number of PCA modes for the OT representation. Note that the rightmost subplot has its own y-axis to the right. Shaded bands denote 1σ variations across the ensembles of trained models.

augmented by the first four PCA modes of the OT representation (OT_4), as more OT features are found to be redundant from earlier results.

As shown in Fig. 6, the inclusion of OT features improves performance across all feature sets, indicating that the OT

representation does capture complementary information beyond traditional subjettness observables. For extended set 3, the improvement becomes marginal, consistent with the similarity between the current OT linearization scheme and high-order subjettness (e.g., τ_{100}). In principle,

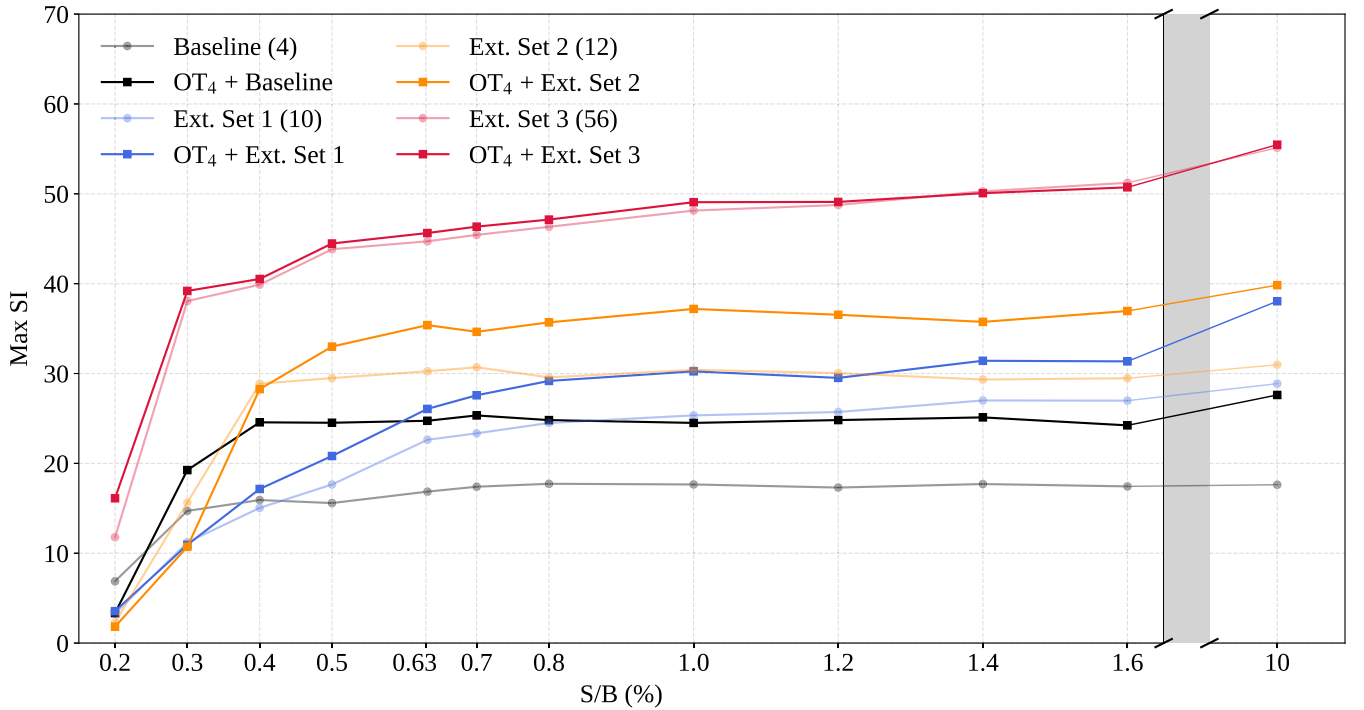


FIG. 6. Maximum SI for the R&D1 dataset using BDT ensembles as the classifier, with the signal injection level S/B varying from 0.2% to 10%. The input features consist of the baseline and three extended sets from Ref. [50], together with the first four PCA modes for the OT representation. Lighter (darker) colors denote feature sets without (with) OT inputs. Numbers in parentheses indicate the dimensionality of each feature set in Ref. [50].

subjettiness observables are encoded in the OT representation itself [31]; extracting this information in practice, however, requires more expressive feature extraction methods than PCA, and is left for future study.

Interestingly, extended set 3 (with OT) achieves a maximum SI of ~ 45 at $S/B = 0.5\%$ and further exceeds 55 at $S/B = 10\%$, surpassing even the two low-level approaches (both ineffective at $S/B = 0.5\%$ and with a max SI ~ 50 at $S/B = 10\%$). This suggests that much of the information learned by the full phase space method and foundation models is likely to have already been encoded in the combined subjettiness and OT features, again underscoring the importance of physics-informed representations while motivating further study of what large models capture.

V. CONCLUSIONS

We presented a physics-aware intermediate representation of collider events based on optimal transport and explored its application to resonant anomaly detection. Within the controlled idealized anomaly detection framework adopted in this work, augmenting standard high-level observables with a small number of OT-derived features leads to substantial improvements in anomaly detection performance, with significance improvements above 25 at signal injection levels of approximately 0.5% in the LHCO R&D1 benchmark and comparable gains in R&D2. Taken together, these results highlight the potential value of physics-informed intermediate representations as a bridge between handcrafted observables and end-to-end machine learning for weakly supervised collider searches.

Several extensions could further enhance performance. While principal component analysis was employed for its simplicity and efficiency, more expressive approaches to feature extraction could be explored. Likewise, advanced classifiers such as neural networks may enable direct use of the complete OT representation without dimensionality reduction. Conversely, the surprising effectiveness of the simple PCA method underscores the importance of carefully balancing data and computational efficiency when introducing sophisticated models, particularly in the low-signal regime.

Beyond boosted dijet signals, our framework is well adapted to more complex event topologies. OT-based observables have demonstrated sensitivity to high-multiplicity, quasi-isotropic radiation patterns typical of hidden-valley scenarios [55–57]. In such new physics settings, OT may be particularly advantageous, as it provides a more flexible and globally structured representation than traditional subjettiness observables. For events featuring distinct total transverse momenta and multiple intrinsic scales, more flexible OT representations with additional degrees of freedom may be required [35,54]. The framework can also be generalized to incorporate information beyond event kinematics, such as particle ID, through multispecies

vector-valued optimal transport [58]. Preliminary studies further suggest applicability of the OT representation to nonresonant anomaly detection, which may open a largely unexplored landscape for new physics discovery.

At the same time, the observed relationship between higher-order subjettiness and OT features underscores the need for a systematic understanding of the physics content encoded in the OT representation. The feature-level diagnostics presented here suggest that OT captures information that is partially overlapping with, yet not fully described by, conventional observables such as jet mass and subjettiness. Since shape-related observables can in principle be derived directly from the OT formulation [31], OT may provide a unified geometric representation of collider events from which a broad class of physics observables can be recovered. Understanding the physical structures encoded in such representations across a wider range of signal and background processes, and determining their predictive capacity, therefore represents an important direction for future study.

In conclusion, the strong performance of both the OT representation and the extended feature sets compared to the available low-level approaches in the IAD setting highlights the importance of physics-aware representations for collider anomaly detection. This advantage appears particularly pronounced in the (ultra) low-signal regime, where inductive biases grounded in physical principles may provide higher sensitivity than purely data-driven approaches.

ACKNOWLEDGMENTS

The authors would like to thank Michael Krämer and Vinicius Mikuni for useful discussions, and Gregor Kasieczka for suggesting performance comparison with the extended feature sets in Ref [50]. The work of T. C. is supported by the Fundamental Research Funds for the Central Universities and further sponsored by Shanghai Pujiang Program. B.N. is supported by the U.S. Department of Energy (DOE), Office of Science under Contract No. DE-AC02-76SF00515. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

DATA AVAILABILITY

The data that support the findings of this article are openly available [1,40,41]. The code for the analysis can be found at [59].

APPENDIX

For completeness, we summarize the standard observables and performance metrics used in this work in Table I.

Table II summarizes Pearson correlations and mutual information values between the leading OT PCA modes

TABLE I. Definitions of the standard observables and performance metrics used throughout the analysis.

Quantity	Definition
m_J	Invariant mass of a reconstructed jet, sensitive to its internal energy distribution.
m_{JJ}	Invariant mass of the two leading- p_T jets, used to define the signal and sideband regions in the LHC0 benchmark.
τ_{21}	N -subjettiness ratio τ_2/τ_1 , commonly used to identify two-prong jet substructure such as boosted W/Z bosons.
τ_{32}	N -subjettiness ratio τ_3/τ_2 , commonly used to identify three-prong jet substructure such as boosted top quarks.
ϵ_s	Signal efficiency: fraction of signal events passing a given classifier-score threshold.
ϵ_b	Background efficiency: fraction of background events passing the same classifier-score threshold.
SI	Significance improvement, defined as $SI \equiv \epsilon_s/\sqrt{\epsilon_b}$, which approximates the improvement in statistical sensitivity relative to an inclusive selection.

and standard jet-substructure observables, including jet masses and higher-order N -subjettiness ratios, computed on the same diagnostic sample used in Fig. 1.

Interestingly, the leading OT feature (OT-PC1) shows a strong correlation with m_{J2} in the R&D1 dataset, despite the fact that the OT representation is constructed purely from jet morphology and contains no explicit mass

information. This observation suggests that nontrivial correlations may exist between the geometric structures captured by OT and conventional jet observables, and warrants further investigation.

More generally, neither the correlations nor the mutual information values are saturated, indicating that the OT features are not simply redundant with the standard

TABLE II. Pearson correlation coefficients $\rho(X, Y)$ (left) and mutual information values $I(X; Y)$ (right) between jet-substructure observables X and OT-derived PCA features Y for the R&D1 diagnostic sample.

ρ	OT ₁	OT ₂	OT ₃	OT ₄	OT ₁₀₀
m_{J1}	−0.342	−0.188	−0.628	−0.035	−0.016
m_{J2}	−0.911	0.073	0.279	−0.143	0.002
τ_{21}^{J1}	0.337	−0.061	−0.013	0.206	−0.004
τ_{21}^{J2}	0.677	−0.197	−0.286	0.536	−0.007
$\tau_{4,1}^{J1}$	−0.214	−0.165	−0.516	0.027	−0.005
$\tau_{4,2}^{J1}$	−0.109	−0.167	−0.477	0.083	−0.010
$\tau_{4,5}^{J1}$	−0.262	−0.121	−0.391	−0.035	0.002
$\tau_{4,1}^{J2}$	−0.375	−0.008	−0.017	0.455	−0.013
$\tau_{4,2}^{J2}$	−0.264	−0.060	−0.044	0.459	−0.010
$\tau_{4,5}^{J2}$	−0.396	0.039	0.002	0.369	−0.010
$\tau_{9,1}^{J1}$	−0.184	−0.175	−0.529	0.048	−0.006
$\tau_{9,1}^{J2}$	−0.376	−0.015	−0.026	0.443	−0.021
I	OT ₁	OT ₂	OT ₃	OT ₄	OT ₁₀₀
m_{J1}	0.198	0.100	0.253	0.055	0.025
m_{J2}	0.870	0.333	0.227	0.284	0.048
τ_{21}^{J1}	0.087	0.019	0.022	0.040	0.000
τ_{21}^{J2}	0.339	0.211	0.104	0.309	0.020
$\tau_{4,1}^{J1}$	0.082	0.049	0.160	0.025	0.008
$\tau_{4,2}^{J1}$	0.057	0.019	0.146	0.015	0.000
$\tau_{4,5}^{J1}$	0.086	0.041	0.102	0.008	0.011
$\tau_{4,1}^{J2}$	0.186	0.040	0.037	0.191	0.043
$\tau_{4,2}^{J2}$	0.111	0.007	0.035	0.162	0.028
$\tau_{4,5}^{J2}$	0.164	0.049	0.038	0.147	0.023
$\tau_{9,1}^{J1}$	0.067	0.041	0.185	0.016	0.001
$\tau_{9,1}^{J2}$	0.200	0.043	0.044	0.185	0.033

observables. This supports the interpretation that the OT representation captures overlapping but complementary information, providing a possible explanation for the performance gains observed when OT features are added to extended subjettness feature sets.

-
- [1] G. Kasieczka *et al.*, The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics, *Rep. Prog. Phys.* **84**, 124201 (2021).
 - [2] T. Aarrestad *et al.*, The Dark machines anomaly score challenge: Benchmark data and model independent event classification for the large hadron collider, *SciPost Phys.* **12**, 043 (2022).
 - [3] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, Machine learning in the search for new fundamental physics, [arXiv:2112.03769](https://arxiv.org/abs/2112.03769).
 - [4] V. Belis, P. Odagiu, and T. K. Aarrestad, Machine learning for anomaly detection in particle physics, *Rev. Phys.* **12**, 100091 (2024).
 - [5] J. Hernández-González, I. Inza, and J. A. Lozano, Weak supervision and other non-standard classification problems: A taxonomy, *Pattern Recogn. Lett.* **69**, 49 (2016).
 - [6] E. M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, *J. High Energy Phys.* **10** (2017) 174.
 - [7] J. Collins, K. Howe, and B. Nachman, Anomaly detection for resonant new physics with machine learning, *Phys. Rev. Lett.* **121**, 241803 (2018).
 - [8] R. T. D’Agnolo and A. Wulzer, Learning new physics from a machine, *Phys. Rev. D* **99**, 015014 (2019).
 - [9] G. Aad *et al.* (ATLAS Collaboration), Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the atlas detector, *Phys. Rev. Lett.* **125**, 131801 (2020).
 - [10] G. Aad *et al.* (ATLAS Collaboration), Weakly supervised anomaly detection for resonant new physics in the dijet final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the atlas detector, *Phys. Rev. D* **112**, 072009 (2025).
 - [11] V. Chekhovsky *et al.* (The CMS Collaboration), Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV, *Rep. Prog. Phys.* **88**, 067802 (2025).
 - [12] R. Gambhir, R. Mastandrea, B. Nachman, and J. Thaler, Isolating unisolated upsilons with anomaly detection in cms open data, *Phys. Rev. Lett.* **135**, 021902 (2025).
 - [13] D. Shih, M. R. Buckley, L. Necib, and J. Tamanas, via machinae: Searching for stellar streams using unsupervised machine learning, *Mon. Not. R. Astron. Soc.* **509**, 5992 (2021).
 - [14] D. Shih, M. R. Buckley, and L. Necib, Via machinae 2.0: Full-sky, model-agnostic search for stellar streams in Gaia DR2, *Mon. Not. R. Astron. Soc.* **529**, 4745 (2024).
 - [15] A. Hallin, D. Shih, C. Krause, and M. R. Buckley, Via machinae 3.0: A search for stellar streams in Gaia with the CATHODE algorithm, [arXiv:2509.08064](https://arxiv.org/abs/2509.08064).
 - [16] M. Pettee, S. Thanvantri, B. Nachman, D. Shih, M. R. Buckley, and J. H. Collins, Weakly supervised anomaly detection in the milky way, *Mon. Not. R. Astron. Soc.* **527**, 8459 (2023).
 - [17] D. Sengupta, S. Mulligan, D. Shih, J. A. Raine, and T. Golling, skyCURTAINss: Model-agnostic search for stellar streams with Gaia data, *Mon. Not. R. Astron. Soc.* **536**, 1104 (2024).
 - [18] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, Classifying anomalies through outer density estimation, *Phys. Rev. D* **106**, 055006 (2022).
 - [19] K. Benkendorfer, L. Le Pottier, and B. Nachman, Simulation-assisted decorrelation for resonant anomaly detection, *Phys. Rev. D* **104**, 035003 (2021).
 - [20] E. Buhmann, C. Ewen, G. Kasieczka, V. Mikuni, B. Nachman, and D. Shih, Full phase space resonant anomaly detection, *Phys. Rev. D* **109**, 055015 (2024).
 - [21] B. Nachman and D. Shih, Anomaly detection with density estimation, *Phys. Rev. D* **101**, 075042 (2020).
 - [22] G. Stein, U. Seljak, and B. Dai, Unsupervised in-distribution anomaly detection of new physics through conditional density estimation, [arXiv:2012.11638](https://arxiv.org/abs/2012.11638).
 - [23] C. L. Cheng, G. Singh, and B. Nachman, Incorporating physical priors into weakly supervised anomaly detection, *Phys. Rev. Lett.* **135**, 021801 (2025).
 - [24] V. Mikuni and B. Nachman, Method to simultaneously facilitate all jet physics tasks, *Phys. Rev. D* **111**, L054015 (2025).
 - [25] G. Peyré and M. Cuturi, Computational optimal transport, [arXiv:1803.00567](https://arxiv.org/abs/1803.00567).
 - [26] P. T. Komiske, E. M. Metodiev, and J. Thaler, Metric space of collider events, *Phys. Rev. Lett.* **123**, 041801 (2019).
 - [27] T. Cai, Optimal transport for high energy physics, Ph.D. thesis, UC, Santa Barbara (main), 2023.
 - [28] T. Cai, J. Cheng, N. Craig, and K. Craig, Linearized optimal transport for collider events, *Phys. Rev. D* **102**, 116019 (2020).
 - [29] T. Cai, J. Cheng, K. Craig, and N. Craig, Which metric on the space of collider events?, *Phys. Rev. D* **105**, 076003 (2022).
 - [30] T. Cai, J. Cheng, N. Craig, G. Koszegi, and A. J. Larkoski, The phase space distance between collider events, *J. High Energy Phys.* **09** (2024) 054.
 - [31] P. T. Komiske, E. M. Metodiev, and J. Thaler, The hidden geometry of particle collisions, *J. High Energy Phys.* **07** (2020) 006.
 - [32] A. J. Larkoski and J. Thaler, A spectral metric for collider geometry, *J. High Energy Phys.* **12** (2023) 219.

- [33] R. Gambhir, A. J. Larkoski, and J. Thaler, SPECTER: Efficient evaluation of the spectral EMD, *J. High Energy Phys.* **12** (2025) 219.
- [34] P. Onyisi, D. Shen, and J. Thaler, Comparing point cloud strategies for collider event classification, *Phys. Rev. D* **108**, 012001 (2023).
- [35] T. Cai, N. Craig, K. Craig, and X. Lin, Multiscale optimal transport for complete collider events, *Phys. Rev. D* **112**, 036021 (2025).
- [36] G. Aad *et al.* (ATLAS Collaboration), A continuous calibration of the ATLAS flavour-tagging classifiers via optimal transportation maps, *Eur. Phys. J. C* **85**, 1272 (2025).
- [37] N. Craig, J. N. Howard, and H. Li, Exploring optimal transport for event-level anomaly detection at the large hadron collider, [arXiv:2401.15542](https://arxiv.org/abs/2401.15542).
- [38] L. Brennan, T. A. Vami, O. Amram, S. Sekhar, Y. Takahashi, L. Moureaux, M. Sommerhalder, P. Maksimovic, T. Cai, and N. Craig, Weakly supervised anomaly detection with event-level variables, *Phys. Rev. D* **112**, 055040 (2025).
- [39] R. T. D’Agnolo, A. Glioti, G. Rigo, and A. Valenti, The intrinsic dimension of collider events and model-independent searches in 100 dimensions, [arXiv:2511.20760](https://arxiv.org/abs/2511.20760).
- [40] G. Kasieczka, B. Nachman, and D. Shih, R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge, [10.5281/zenodo.6466204](https://doi.org/10.5281/zenodo.6466204) (2022).
- [41] D. Shih, Additional QCD Background Events for LHCO2020 R&D (signal region only). [10.5281/zenodo.8370758](https://doi.org/10.5281/zenodo.8370758) (2023).
- [42] T. Sjostrand, S. Mrenna, and P. Z. Skands, Pythia 6.4 Physics and Manual, *J. High Energy Phys.* **05** (2006) 026.
- [43] T. Sjostrand, S. Mrenna, and P. Z. Skands, A brief introduction to Pythia 8.1, *Comput. Phys. Commun.* **178**, 852 (2008).
- [44] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), Delphes 3, A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [45] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, *Eur. Phys. J. C* **72**, 1896 (2012).
- [46] M. Cacciari, G. P. Salam, and G. Soyez, The anti-kt jet clustering algorithm, *J. High Energy Phys.* **04** (2008) 063.
- [47] J. Neyman, E. S. Pearson, and K. Pearson, IX. on the problem of the most efficient tests of statistical hypotheses, *Phil. Trans. R. Soc. A* **231**, 289 (1933).
- [48] A. Andreassen, B. Nachman, and D. Shih, Simulation assisted likelihood-free anomaly detection, *Phys. Rev. D* **101**, 095004 (2020).
- [49] J. A. Raine, S. Klein, D. Sengupta, and T. Golling, CUR-TAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals, *Front. Big Data* **6**, 899345 (2023).
- [50] T. Finke, M. Hein, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, and M. Sommerhalder, Tree-based algorithms for weakly supervised anomaly detection, *Phys. Rev. D* **109**, 034033 (2024).
- [51] M. Freytsis, M. Perelstein, and Y. C. San, Anomaly detection in the presence of irrelevant features, *J. High Energy Phys.* (2024) 220.
- [52] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, *J. High Energy Phys.* **03** (2010) 015.
- [53] J. Thaler and K. Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness, *J. High Energy Phys.* **02** (2011) 093.
- [54] T. Cai, J. Cheng, B. Schmitzer, and M. Thorpe, The linearized Hellinger–Kantorovich distance, *SIAM J. Imag. Sci.* **15**, 45 (2022).
- [55] C. Cesarotti and J. Thaler, A robust measure of event isotropy at colliders, *J. High Energy Phys.* **08** (2020) 084.
- [56] M. J. Strassler and K. M. Zurek, Echoes of a hidden valley at hadron colliders, *Phys. Lett. B* **651**, 374 (2007).
- [57] 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.
- [58] K. Craig, N. G. Trillos, and D. Nikolić, Vector valued optimal transport: From dynamic to static formulations, [arXiv:2505.03670](https://arxiv.org/abs/2505.03670).
- [59] <https://github.com/TianjiCai/ADwithOT>.