Title
Accelerating template generation in resonant anomaly detection searches with optimal transport

Journal
Journal of High Energy Physics, 2025(12)

Authors
Leigh, Matthew
Sengupta, Debajyoti
Nachman, Benjamin
et al.

Publication Date
2025-12-12

DOI
10.1007/jhep12(2025)105

Published for SISSA by Springer

Received: *February 28, 2025*
Revised: *September 21, 2025*
Accepted: *October 17, 2025*
Published: *December 12, 2025*

Accelerating template generation in resonant anomaly detection searches with optimal transport

Matthew Leigh,[a] **Debajyoti Sengupta**,[a] **Benjamin Nachman**[b] **and Tobias Golling**[a]

[a] *University of Geneva,*
24 rue du Général-Dufour, Geneva, Switzerland
[b] *Lawrence Berkeley National Laboratory,*
1 Cyclotron Road, Berkeley, U.S.A.

E-mail: matthew.leigh@unige.ch, debajyoti.sengupta@unige.ch, bpnachman@lbl.gov, tobias.golling@unige.ch

Abstract: We introduce Resonant Anomaly Detection with Optimal Transport (RAD-OT), a method for generating signal templates in resonant anomaly detection searches. RAD-OT leverages the fact that the samples from the conditional probability density of the target features vary approximately linearly along the optimal transport path connecting the resonant feature. This does not assume that the conditional density itself is linear with the resonant feature, allowing RAD-OT to efficiently capture multimodal relationships, changes in resolution, etc. By solving the optimal transport problem, RAD-OT can quickly build a template by interpolating between the background distributions in two sideband regions. We demonstrate the performance of RAD-OT using the LHC Olympics R&D dataset, where we find comparable sensitivity and improved stability with respect to deep learning-based approaches.

Keywords: Automation, Jets and Jet Substructure, Rare Decays, Parton Distributions

ArXiv ePrint: 2407.19818

https://doi.org/10.1007/JHEP12(2025)105

Contents

1 Introduction

The Standard Model (SM) has been very successful in describing the fundamental particles and their interactions, but there are many reasons why it is not the final theory of nature, such as the unexplained dark matter in the universe. One of the main goals of the Large Hadron Collider (LHC) is to search for new physics beyond the Standard Model (BSM) of particle physics. General purpose detectors such as ATLAS [1] and CMS [2] are designed to be sensitive to a wide range of new physics possibilities. As there are limitless possibilities as to what the new physics might be, it is not feasible to individually test each hypothesis. It is thus desirable to have methods that are simultaneously sensitive to numerous possibilities [3–5]. This would complement and could be combined with dedicated searches in specific regions of phase space. A number of data-driven methods [6–23] have been developed and applied [24–29] in the context of resonant anomaly searches.[1] The main assumption is that the new physics is localised in a known (resonant) feature, while the background distribution is featureless. Data away from the resonance in the sideband (SB) are used to build an unbinned template of the background distribution under the resonance in the signal region via interpolation. The most widely-studied approaches use conditional discriminative or generative machine learning models. Once the template is built, it is compared to the data in the SR, often using a classifier to create an anomaly score [6, 7, 32].

Existing methods have shown promising results, but there are a number of motivations for building new techniques. For example, neural network-based methods can be computationally expensive. This is especially the case when ensembling is required for stability to minimize fluctuations from the scholastic nature of training. In this work, we propose a template generation strategy that does not rely on neural networks in order to accelerate the process and enhance stability while also preserving sensitivity. The strategy uses the framework of Optimal Transport (OT), a set of tools for transforming one dataset into another with the least amount of movement. OT has been studied as a method for creating an anomaly score [33–36], but we propose to use it for template generation. Our Resonant Anomaly Detection with Optimal Transport (RAD-OT) leverages the fact that samples from the

[1] We are not counting generic anomaly detection methods applied to the resonant case, see e.g. the recent ATLAS results [30, 31] and method papers they cite.

JHEP12(2025)105

conditional probability density of the target features vary approximately linearly along the OT path connecting the resonant feature. This does not assume that the conditional density itself is linear with the resonant feature, allowing RAD-OT to efficiently capture multimodal relationships, changes in resolution, etc. For modest feature space dimensionality, the OT solution can be approximated without neural networks,[2] leading to high efficiency and stability.

This paper is organized as follows. Section 2 briefly reviews the LHC Olympics (LHCO) R&D dataset [37] that we use to demonstrate RAD-OT. The mathematical aspects of RAD-OT are described in section 3. Numerical results on the LHC are presented in section 4, and the paper ends with conclusions and outlook in section 5.

2 Dataset

The LHCO R&D dataset [37] comprises background events represented by quark/gluon scattering to produce dijets and signal events arising from the all-hadronic decay of a massive particle to two other massive particles $W' \to X(\to q\bar{q})Y(\to q\bar{q})$, with masses $m_{W'} = 3.5\,\text{TeV}$, $m_X = 500\,\text{GeV}$, and $m_Y = 100\,\text{GeV}$. Both processes are simulated with `Pythia` 8.219 [38] and interfaced to `Delphes` 3.4.1 [39] for detector simulation. Jets are reconstructed using the anti-k_T clustering algorithm [40] with a radius parameter $R = 1.0$, using the `FastJet` [41] package. In total there are 1 million dijet events and 100,000 signal events.

Events are required to have at least one $R = 1.0$ jet J with pseudrapidity $|\eta| < 2.5$, and transverse momentum $p_\text{T}^J > 1.2\,\text{TeV}$. The top two leading p_T jets are selected and ordered by decreasing mass; they are labelled J_1 and J_2. In order to remove the turn on in the m_{JJ} distribution arising from the jet selections, we only consider events with $m_{JJ} > 2.8\,\text{TeV}$. To construct the training datasets, we use varying amounts of signal events mixed in with the dijet events.

To study the performance of RAD-OT, we use the same high-level features employed by many of the existing methods demonstrated on the same dataset. These features are $m_{JJ}, m_{J_1}, \Delta m_J = m_{J_1} - m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}$, and ΔR_{JJ}, where τ_{21}^J is the N-subjettinness [42] ratio of jet J, and ΔR_{JJ} is the angular separation of the two jets in the detector $\eta - \phi$ space.

3 Template building from optimal interpolants

This section describes the procedure for building a data-driven template using RAD-OT. A diagrammatic representation of the method is shown in figure 1.

First, the features $\mathbf{f}$ of a dataset are split into a selected resonant feature m and additional attributes that characterise the template, $\mathbf{x}$. The data is then partitioned on m into a signal region (SR) and two flanking sideband regions, SB1 and SB2. The goal is to sample from values of $\mathbf{f} = (m, \mathbf{x})$ that are representative of the background distribution in the SR. Here, the assumption is that the population of the sidebands is predominantly background. Furthermore, this method assumes that samples from the conditional probability density of the additional features vary linearly along the optimal transport path connecting m. While this assumption may not hold exactly, it is a reasonable first order approximation to apply to narrow SRs. For our numerical tests, we define our SR based on a mass cut of $3300 \leq m_{JJ} < 3700\,\text{GeV}$,

[2] See ref. [15] for OT-related approaches based on neural networks.

JHEP12(2025)105

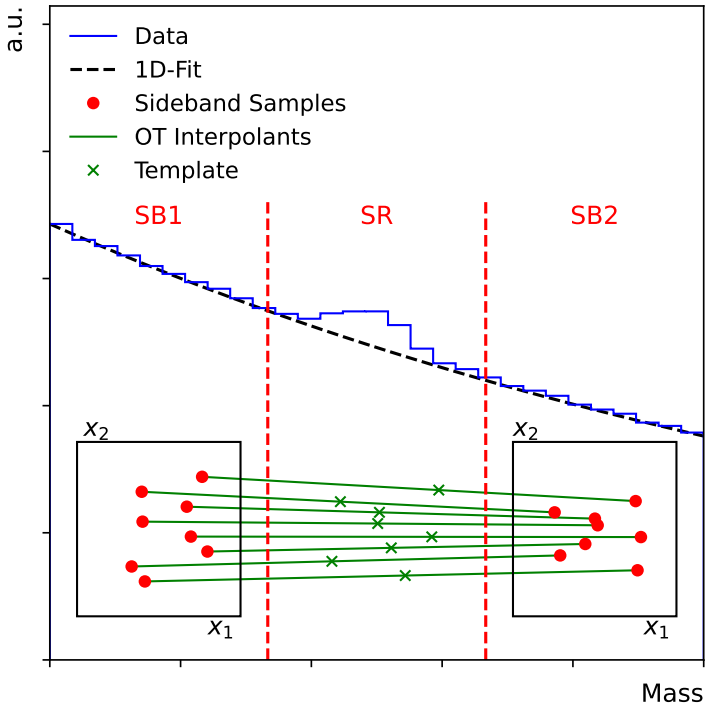

Figure 1. A simple diagrammatic representation of the RAD-OT method. Features $\mathbf{x}$ (red) are sampled from the mass sidebands SB1 and SB2 and paired using the optimal transport map (green). The resonant feature m is sampled from the Kernel Density Estimator (KDE) (dashed black) and the SR template (green cross) is generated using linear interpolation.

and sideband regions $3100 \leq m_{JJ} < 3300\,\text{GeV}$ and $3700 \leq m_{JJ} < 3900\,\text{GeV}$. In practice, these regions would be swept across the spectrum, but this region was chosen, as in previous papers, as it is centered near the LHCO signal mass peak.

Next, a one-dimensional fit is performed on the m distribution to approximate $p(m)$ in the SR [14]. This can be done with a parametric or non-parametric (such as a Kernel Density Estimator, KDE) fit with just the sidebands or on all data. This fit is then used to sample the mass values in the SR.

The next step is to build a linear interpolation paths between samples drawn from each sideband. For $\mathbf{f}_1 = (m_1, \mathbf{x}_1) \in \text{SB1}$ and $\mathbf{f}_2 = (m_2, \mathbf{x}_2) \in \text{SB2}$, this sets up the basic parametric function

$$\mathbf{f}_t = \left(m_t, \mathbf{x}_1 + \frac{m_t - m_1}{m_2 - m_1}(\mathbf{x}_2 - \mathbf{x}_1) \right), \tag{3.1}$$

where m_t is the desired mass value in the SR. We note that the conditional density of $\mathbf{x}$ given m is not known explicitly in terms of the input densities — eq. (3.1) defines an operation on the inputs and thus implicitly modifies the conditional density.

The crucial aspect of this method is to pair the samples f_1 and f_2 such that the interpolation is meaningful. For instance, if the samples (pairs f_1 and f_2) are drawn randomly and independently, the interpolated features will be pushed towards a normal distribution as illustrated in figure 2.

JHEP12(2025)105

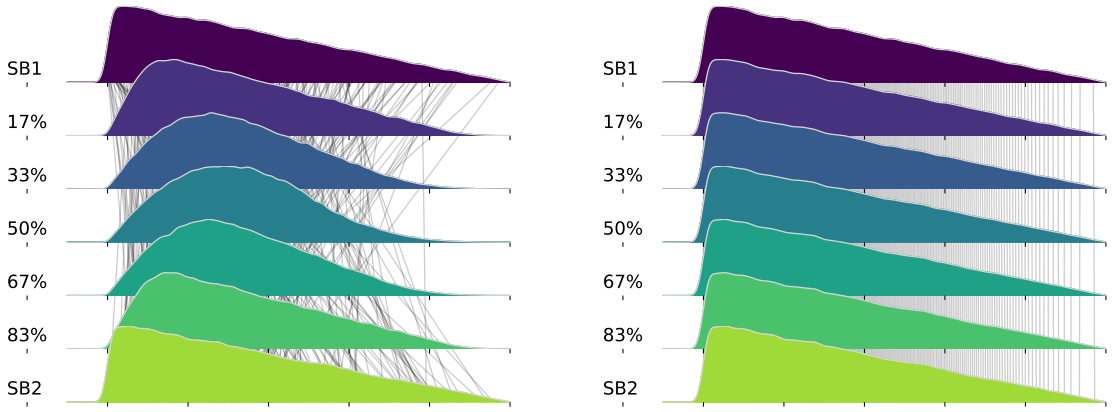

Figure 2. An illustration of the interpolated distributions with and without optimal transport matching on a one-dimensional toy dataset. The y-axis shows the slices of the interpolation between SB1 (0%) and SB2 (100%). Several interpolation paths are shown as black lines. In this toy sample the underlying distributions in each sideband is the same, and therefore the ideal interpolations should be constant. Without OT-matching (left), the interpolations result in a basic convolution and the intermediate stages no longer match the endpoints. The OT-matched interpolation paths (right) correctly keep the distribution constant from sideband to sideband.

We propose to select pairs of samples (f_1, f_2) such that they solve an optimal transport problem. More specifically, given N samples drawn from each sideband, we can construct a cost matrix C where $C_{ij} = c(\mathbf{x}_1^i, \mathbf{x}_2^j)$ is the cost of transporting $\mathbf{x}_1^i$ to $\mathbf{x}_2^j$, where $i, j = 1, \ldots, N$. We can then learn the map Γ^* that minimises the cost of transporting the samples from SB1 to SB2, by solving the following optimization problem:

$$\begin{aligned} \Gamma^* = \min_{\Gamma} & \sum_{i=1}^{N}\sum_{j=1}^{N} \Gamma_{ij} c(\mathbf{x}_1^i, \mathbf{x}_2^j) \\ \text{subject to} & \sum_{j=1}^{N} \Gamma_{ij} = p_i, \quad \forall i = 1, \ldots, n, \\ & \sum_{i=1}^{N} \Gamma_{ij} = q_j, \quad \forall j = 1, \ldots, m, \\ & \Gamma_{ij} \geq 0, \quad \forall i = 1, \ldots, N, \forall j = 1, \ldots, N, \end{aligned} \tag{3.2}$$

where p_i and q_j are the marginal distributions of the samples from SB1 and SB2 respectively. The optimal pair of points is then constructed by first selecting the samples $\mathbf{x}_1$ and then finding the corresponding $\mathbf{x}_2$ using the optimal transport map.

A point to consider is the form of the cost function c. While we have found that the simply using the Euclidean distance works well, it requires transforming the features to have the same scale. This is because the cost of transforming one feature into another is often ill-defined. We have found that using quantile-based scaler which maps each feature to a normal distribution works well, especially when there is no prior knowledge. It should be noted that the transformed features are only used for the cost matrix and the original features are used for the interpolation.

JHEP12(2025)105

Input: Batch size B, Number of batches N, Fitted resonant likelihood $p(m)$, Feature scaler S

$T \leftarrow \varnothing$; ▷ Start with empty template

for *i in range* N **do** ▷ Loop through the batches

  $m_1, x_1 \sim \text{SB1}$; ▷ Sample batch from SB1

  $m_2, x_2 \sim \text{SB2}$; ▷ Sample batch from SB2

  $C \leftarrow \text{compute\_cost}\big(S(x_1), S(x_2)\big)$; ▷ Compute the cost matrix using scaled features

  $\Gamma^* \leftarrow \text{solve\_optimal\_transport}(C)$; ▷ Solve the optimal transport problem

  $P \leftarrow \arg\_\max(\Gamma^*)$; ▷ Find the permutation for SB2

  $x_2 \leftarrow x_2[P],\ m_2 \leftarrow m_2[P]$; ▷ Reorder SB2

  $m_t \sim p(m)$; ▷ Sample batch of mass values in the SR

  $x_t \leftarrow x_1 + \left(\frac{m_t - m_1}{m_2 - m_1}(x_2 - x_1)\right)$; ▷ Sample features using linear interpolation

  $T \leftarrow T \cup (m_t, x_t)$; ▷ Add the batch to the template

Algorithm 1. Pseudocode for the RAD-OT method.

One drawback of using this method is that while we do not have to train any generative model, the optimal transport problem is itself computationally expensive to solve. Current exact methods use the Hungarian algorithm [43], which has a time complexity of $O(N^3)$. Thus, it is typically infeasible to solve the optimal transport problem for the entirety of the sidebands. We propose therefore to use a batched approach, whereby we iteratively build the template using subsets of the sidebands. We found the variation in the generated template to be small once we used a large enough batch size (~ 2500), which was well within our computational resources.

The full RAD-OT algorithm is summarized in Algorithm 1.

We use the `POT` package [44] to solve the optimal transport problem in a batched manner. We prepare 100 batches of 5000 samples from SB1 and SB2 and solve the optimal transport problem for each batch. All $(m_1, \mathbf{x}_1)$ in the batch sampled from SB1 are paired with $(m_2, \mathbf{x}_2)$ using the optimal map Γ^* computed using that batch. Thereafter, masses m_t are sampled from the KDE and the SR template is formed using eq. (3.1).

4 Anomaly detection on the LHCO

To perform anomaly detection using RAD-OT, we use the widely established method of CWoLa [32]. Here, a classifier is trained to distinguish between the data drawn from the SR and the template. If the template is a good representation of the background, and the SR is a mixture of real background and a some signal samples, the optimal classifier for this task is also the optimal classifier for distinguishing between the background and signal.

While it is not expected that RAD-OT will outperform existing methods with more complex template building schemes, it is far more computationally efficient and robust. Furthermore, the RAD-OT template generation can be run on a single CPU, while efficient training of neural networks typically requires modern GPUs. To benchmark its performance, we compare RAD-OT with other data-driven methods: (1) CurtainsF4F [16], where

JHEP12(2025)105

Method	Device	Time (mins)
RAD-OT	CPU	10
CurtainsF4F	GPU	181

Table 1. Comparison of the time taken for template generation using RAD-OT and CurtainsF4F for one SR. For CurtainsF4F, this also includes the training time.

normalising flows are used to construct the template, and (2) CWoLa [32] where the SB1 and SB2 data are used directly as the template. Note that all the template-based methods use CWoLa to train a classifier between the generated (or sampled) template and data. We label the original sideband-directly-as template approach from ref. [32] as 'Simple CWoLa' to distinguish it from how it is used in the other methods.

While the Simple CWoLa method is the fastest of the three, it is expected to perform the worst as the classifier can pick up on variables which are highly correlated with the mass. We hypothesize that we can improve on this performance using RAD-OT, without requiring the arduous training times of CurtainsF4F.

Boosted Decision Trees (BDTs) have been shown to be very effective in anomaly detection using CWoLa [45, 46]. For all three methods, we use BDTs as opposed to neural networks for the classifier to further reduce the computational cost. We used the `scikit-learn` package [47] to grow an ensemble classifier of 50 Histogram-Gradient BDTs, each grown maximally until early stopping based on a separate 10% hold out validation set. For the template building methods, we used the standard four parameter di-jet fit to produce $p(m_{JJ})$ [31].

Before assessing the quality of the templates, we compare the template generation time of 500000 events using RAD-OT and CurtainsF4F in table 1 (the simple-template generation takes no time). We used a batch size of 5000 for RAD-OT. RAD-OT requires only 10 minutes to generate the template on a single CPU — a factor of 15 times faster than CurtainsF4F, which first needs to train the two flows, then perform the morphing using GPU acceleration with each step of the process.

Next, we evaluate the template quality of RAD-OT qualitatively by comparing the contour plots of the template and the target data in the absence of signal. This is illustrated in figure 3. The marginal distributions of the features in the SR, and the correlations thereof, are captured well, with only slight mismodelling in ΔR. We quantify the RAD-OT accuracy by training an ensemble of BDTs to distinguish between the target and the template data. Figure 4 shows the ROC plot of this test. An Area-Under-the-Curve (AUC) score of 0.5 indicates that the classifier cannot distinguish between these two datasets, which is our goal. The RAD-OT template achieves an AUC score of 0.53 ± 0.01 which is comparable to the AUC score of 0.53 ± 0.01 achieved by the CurtainsF4F method. The values represent the mean and the standard deviation of the AUCs from 5 independent classifiers (initiated with different random seeds) trainings on the same data.

After demonstrating that RAD-OT accurately models the background, we now show how well the method can improve the significance of a signal in the SR. For this we add 3000 signal events to the dataset ($S/\sqrt{B} = 3$). For all methods, we use 5-fold cross-validation to train the classifiers. Figure 5 shows the ROC curve and significance improvement characteristic (SIC

JHEP12(2025)105

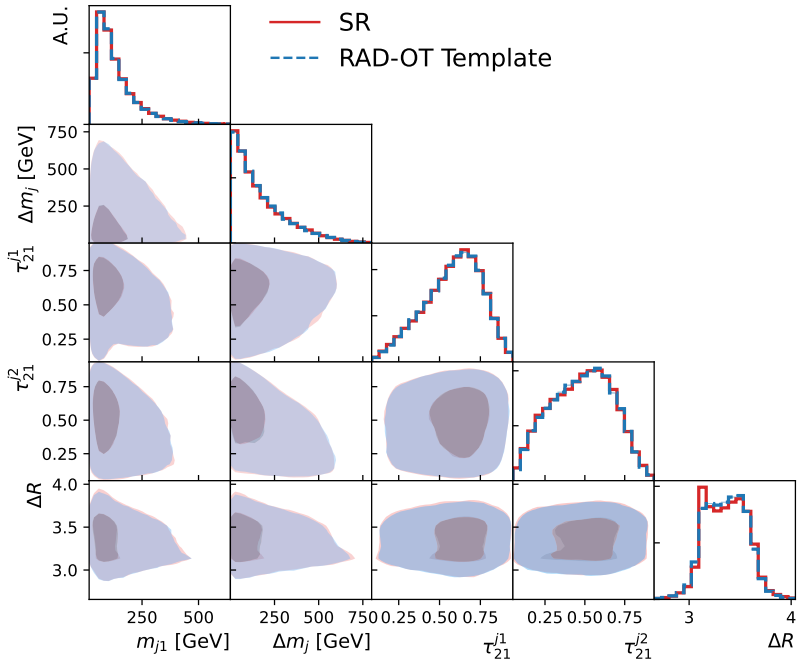

Figure 3. The template is generated using the SR $3300 \leq m_{JJ} < 3700\,\text{GeV}$, and sideband regions $2900 \leq m_{JJ} < 3300\,\text{GeV}$ and $3700 \leq m_{JJ} < 4100\,\text{GeV}$ with no signal injected. The diagonal elements show the marginal distributions of the features in the SR, while the off-diagonal elements show the correlations between the features. The true data is shown in red, while the interpolated template is shown in blue.

$= \text{TPR}/\sqrt{\text{FPR}}$) versus rejection factor. As further benchmarks, we show the performance of a supervised classifier and an idealised classifier, as in previous works. All classifiers are made using the same ensemble of 50 BDTs but differ in the data and labels selected for training. The supervised classifier is trained with true signal and background labels using data from the SR. This provides an upper bound on the achievable classification performance on the dataset. The idealised classifier is also trained using data from the SR but we flip half of the background labels. This sets the limit on the performance that can be achieved with a perfect background template and noisy labels. RAD-OT performs competitively with CurtainsF4F, achieving a SIC of ~ 12 at a rejection factor of 1000. Crucially, RAD-OT outperforms Simple CWoLa, meaning that the simple method of linearly interpolating between the sidebands is an effective strategy for removing the m_{JJ} dependence and enabling more sensitive anomaly detection.

We also track the sensitivity of the method to different levels of signal injection and calculate the SIC at a rejection factor of 1000 which is shown in figure 6. The required SIC for a *discovery* is shown as a dashed line. Here, we use $\frac{S}{\sqrt{B}} = 5$ as a discovery threshold. We see that RAD-OT performs better than standalone CWoLa and is able to *discover* a signal with as few as $\lesssim 700$ signal events in the SR, which corresponds to an initial $\frac{S}{\sqrt{B}} \sim 2$.

The main assumptions of the RAD-OT method is that samples from the conditional probability density of the classifier features varies linearly along the optimal transport path connecting the resonant feature. While this assumption may hold in small regions of the phase (i.e. narrow signal regions and sidebands), it is crucial to investigate the effect of the window

JHEP12(2025)105

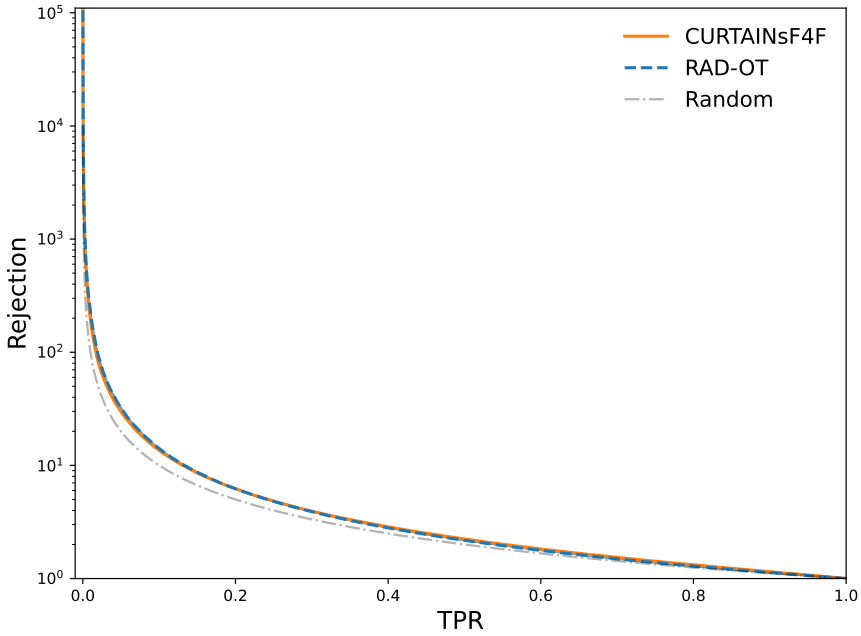

Figure 4. A Receiver Operating Characteristic (ROC) curve showing the trade-off between the true positive rate (TPR) and inverse false positive rate (1/FPR ≡ rejection) for the template generated by RAD-OT (blue) and CURTAINsF4F (orange); the Random line is the case of TPR = FPR. There is no signal here — the TPR is the probability of correctly classifying the data in the SR as such while the FPR is the probability of classifying the template data as target data.

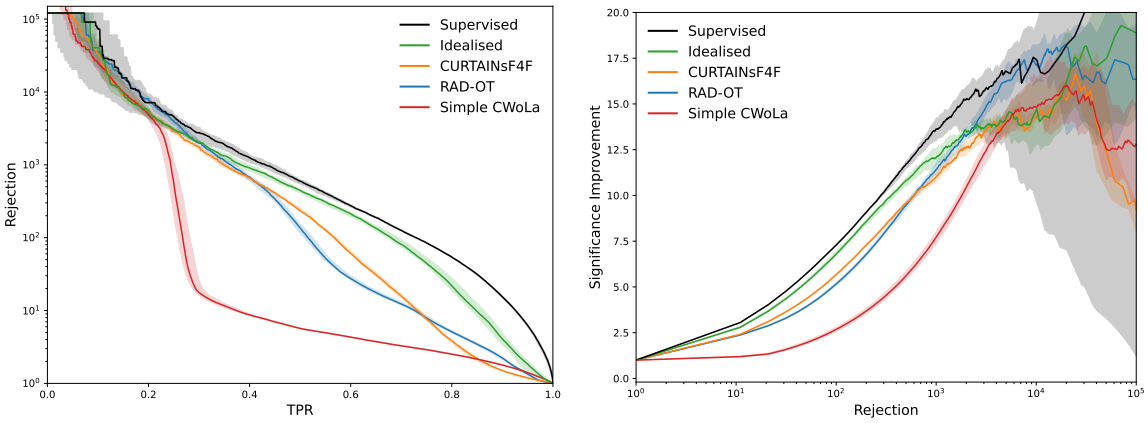

Figure 5. Background rejection as a function of signal efficiency (left) and significance improvement as a function of background rejection (right) for RAD-OT (blue), CWoLa (red), Supervised (black), Idealised (green), and CURTAINsF4F (orange). All classifiers are trained on the sample with 3,000 injected signal events, using a SR $3300 \leq m_{JJ} < 3700\,\text{GeV}$. The lines show the mean value of fifty classifier trainings with different random seeds, with the shaded band covering 68% uncertainty. The same events are used in each training; only the initialization of the machine learning varies. A Supervised classifier and Idealised classifiers are shown for reference.

JHEP12(2025)105

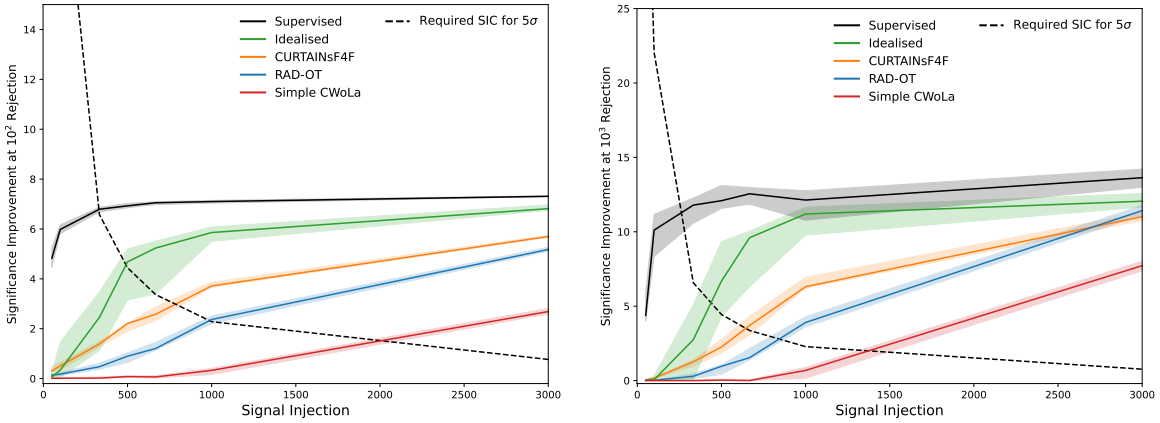

Figure 6. Significance improvement at a background rejection of 10^2 (left) and 10^3 (right) as a function of signal events in the SR $3300 \leq m_{JJ} < 3700\,\text{GeV}$, for RAD-OT (blue), CURTAINsF4F (orange) Idealised (green), and Supervised (black). The lines show the mean value of 5 BDT trainings with different random seeds, with the shaded band covering 68% uncertainty. A Supervised classifier and Idealised classifiers are shown for reference. The required SIC across the doping levels for a discovery is shown as a dashed line.

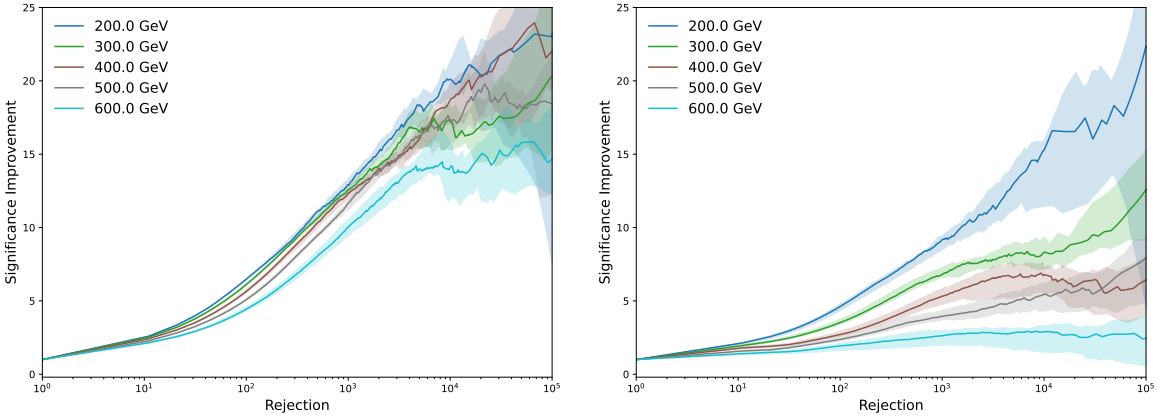

Figure 7. The SIC vs rejection factor for different SR widths using RAD-OT on datasets with 3000 (left) and 1000 (right) injected signal samples. All SRs are centred on 3500 GeV.

widths on the performance of the method. To do this, we fix the sideband width and vary the SR window width to see how the performance of the method changes. Figure 7 shows and SIC vs rejection factor for different SR window widths for the RAD-OT method. As expected, the performance of RAD-OT generally decreases with increasing SR width, as the linear approximation between the features and the resonant feature no longer holds true, leading to a worse template, and hence a worse classifier performance. With 3000 injected signal events, even with a SR with of 600 GeV, RAD-OT is still able to reach a SIC of ~ 10 at a rejection factor of 1000. However, with only 1000 injected signal events, there is a notable performance drop with even 300 GeV, highlighting the sensitivity of this method to wider signal regions.

JHEP12(2025)105

5 Discussion

In this work, we develop a method for generating templates in resonant anomaly detection using optimal transport and linear interpolation to enhance stability and reduce generation time compared to previous methods. In order to ensure that the interpolation paths are meaningful, we match pairs of samples from each sideband using a mini-batched optimal transport solution. This approach assumes that samples from the conditional probability density of the classifier features vary linearly along the optimal transport path connecting the resonant feature. While this assumption may not be exact, it is a reasonable first order approximation to apply to narrow signal regions. We validated this approach on the LHCO dataset and showed competitive performance with more complex template generation methods, based on neural networks, that take an order of magnitude longer to train. Our new RAD-OT method provides a complementary approach to existing methods and may enable faster sweeps of signal regions when computational challenges are limiting. It would also be interesting to explore how the precision of RAD-OT scales with the number of features and the amount of available data, where it may provide an advantage with limited data as no (neural network) training is required.

Acknowledgments

DS, and TG acknowledge funding through the SNSF Sinergia grant CRSII5_193716 "Robust Deep Density Models for High-Energy Particle Physics and Solar Flare Analysis (RODEM)" and the SNSF project grant 200020_212127 "At the two upgrade frontiers: machine learning and the ITk Pixel detector". ML would like to acknowledge individual funding acquired through the Swiss Government Excellence Scholarships for Foreign Scholars. BN is supported by the U.S. Department of Energy (DOE), Office of Science under contract DE-AC02-05CH11231.

Data Availability Statement. This article has no associated data or the data will not be deposited.

Code Availability Statement. This article has no associated code or the code will not be deposited.

References

[1] ATLAS collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, 2008 *JINST* **3** S08003 [INSPIRE].

[2] CMS collaboration, *The CMS Experiment at the CERN LHC*, 2008 *JINST* **3** S08004 [INSPIRE].

[3] G. Kasieczka et al., *The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics*, *Rept. Prog. Phys.* **84** (2021) 124201 [arXiv:2101.08320] [INSPIRE].

JHEP12(2025)105

[4] 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** (2022) 043 [arXiv:2105.14027] [INSPIRE].

[5] G. Karagiorgi et al., *Machine Learning in the Search for New Fundamental Physics*, arXiv:2112.03769 [INSPIRE].

[6] J.H. Collins, K. Howe and B. Nachman, *Anomaly Detection for Resonant New Physics with Machine Learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

[7] J.H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, *Phys. Rev. D* **99** (2019) 014038 [arXiv:1902.02634] [INSPIRE].

[8] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*, *Phys. Rev. D* **101** (2020) 075042 [arXiv:2001.04990] [INSPIRE].

[9] O. Amram and C.M. Suarez, *Tag N' Train: a technique to train improved classifiers on unlabeled data*, *JHEP* **01** (2021) 153 [arXiv:2002.12376] [INSPIRE].

[10] G. Stein, U. Seljak and B. Dai, *Unsupervised in-distribution anomaly detection of new physics through conditional density estimation*, in the proceedings of the *34th Conference on Neural Information Processing Systems*, Online Conference, Canada, December 06–12 (2020) [arXiv:2012.11638] [INSPIRE].

[11] J.F. Kamenik and M. Szewc, *Null hypothesis test for anomaly detection*, *Phys. Lett. B* **840** (2023) 137836 [arXiv:2210.02226] [INSPIRE].

[12] A. Andreassen, B. Nachman and D. Shih, *Simulation Assisted Likelihood-free Anomaly Detection*, *Phys. Rev. D* **101** (2020) 095004 [arXiv:2001.05001] [INSPIRE].

[13] K. Benkendorfer, L.L. Pottier and B. Nachman, *Simulation-assisted decorrelation for resonant anomaly detection*, *Phys. Rev. D* **104** (2021) 035003 [arXiv:2009.02205] [INSPIRE].

[14] A. Hallin et al., *Classifying anomalies through outer density estimation*, *Phys. Rev. D* **106** (2022) 055006 [arXiv:2109.00546] [INSPIRE].

[15] J.A. Raine, S. Klein, D. Sengupta and T. Golling, *CURTAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals*, *Front. Big Data* **6** (2023) 899345 [arXiv:2203.09470] [INSPIRE].

[16] D. Sengupta, S. Klein, J.A. Raine and T. Golling, *CURTAINs flows for flows: Constructing unobserved regions with maximum likelihood estimation*, *SciPost Phys.* **17** (2024) 046 [arXiv:2305.04646] [INSPIRE].

[17] T. Golling, S. Klein, R. Mastandrea and B. Nachman, *Flow-enhanced transportation for anomaly detection*, *Phys. Rev. D* **107** (2023) 096025 [arXiv:2212.11285] [INSPIRE].

[18] M.F. Chen, B. Nachman and F. Sala, *Resonant anomaly detection with multiple reference datasets*, *JHEP* **07** (2023) 188 [arXiv:2212.10579] [INSPIRE].

[19] D. Sengupta et al., *Improving new physics searches with diffusion models for event observables and jet constituents*, *JHEP* **04** (2024) 109 [arXiv:2312.10130] [INSPIRE].

[20] R. Das, G. Kasieczka and D. Shih, *Residual ANODE*, arXiv:2312.11629 [INSPIRE].

[21] T. Golling et al., *The interplay of machine learning-based resonant anomaly detection methods*, *Eur. Phys. J. C* **84** (2024) 241 [arXiv:2307.11157] [INSPIRE].

[22] E.M. Metodiev, J. Thaler and R. Wynne, *Anomaly detection in collider physics via factorized observables*, *Phys. Rev. D* **110** (2024) 055012 [arXiv:2312.00119] [INSPIRE].

JHEP12(2025)105

[23] E. Buhmann et al., *Full phase space resonant anomaly detection*, *Phys. Rev. D* **109** (2024) 055015 [arXiv:2310.06897] [INSPIRE].

[24] ATLAS collaboration, *Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector*, *Phys. Rev. Lett.* **125** (2020) 131801 [arXiv:2005.02983] [INSPIRE].

[25] D. Shih, M.R. Buckley, L. Necib and J. Tamanas, *via machinae: Searching for stellar streams using unsupervised machine learning*, *Mon. Not. Roy. Astron. Soc.* **509** (2021) 5992 [arXiv:2104.12789] [INSPIRE].

[26] 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. Roy. Astron. Soc.* **529** (2024) 4745 [arXiv:2303.01529] [INSPIRE].

[27] M. Pettee et al., *Weakly-Supervised Anomaly Detection in the Milky Way*, arXiv:2305.03761 [DOI:10.1093/mnras/stad3663] [INSPIRE].

[28] CMS collaboration, *Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV*, CMS-PAS-EXO-22-026, CERN, Geneva (2024).

[29] D. Sengupta et al., *skycurtains: model-agnostic search for stellar streams with Gaia data*, *Mon. Not. Roy. Astron. Soc.* **536** (2024) 1104 [arXiv:2405.12131] [INSPIRE].

[30] ATLAS collaboration, *Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*, *Phys. Rev. D* **108** (2023) 052009 [arXiv:2306.03637] [INSPIRE].

[31] ATLAS collaboration, *Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $\sqrt{s} = 13$ TeV with the ATLAS detector*, ATLAS-CONF-2023-022, CERN, Geneva (2023).

[32] E.M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174 [arXiv:1708.02949] [INSPIRE].

[33] M. Crispim Romão et al., *Use of a generalized energy Mover's distance in the search for rare phenomena at colliders*, *Eur. Phys. J. C* **81** (2021) 192 [arXiv:2004.09360] [INSPIRE].

[34] K. Fraser et al., *Challenges for unsupervised anomaly detection in particle physics*, *JHEP* **03** (2022) 066 [arXiv:2110.06948] [INSPIRE].

[35] S.E. Park, P. Harris and B. Ostdiek, *Neural embedding: learning the embedding of the manifold of physics data*, *JHEP* **07** (2023) 108 [arXiv:2208.05484] [INSPIRE].

[36] N. Craig, J.N. Howard and H. Li, *Exploring Optimal Transport for Event-Level Anomaly Detection at the Large Hadron Collider*, arXiv:2401.15542 [INSPIRE].

[37] G. Kasieczka, B. Nachman and D. Shih, *Official Datasets for LHC Olympics 2020 Anomaly Detection Challenge (Version v6)*, (2019) [DOI:10.5281/zenodo.4536624].

[38] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[39] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].

[40] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[41] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J. C* **72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

JHEP12(2025)105

[42] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **03** (2011) 015 [arXiv:1011.2268] [INSPIRE].

[43] H.W. Kuhn, *The Hungarian method for the assignment problem*, *Nav. Res. Logist. Q.* **2** (1955) 83.

[44] R. Flamary et al., *Pot: Python optimal transport*, *J. Machine Learning Res.* **22** (2021) 1.

[45] T. Finke et al., *Tree-based algorithms for weakly supervised anomaly detection*, *Phys. Rev. D* **109** (2024) 034033 [arXiv:2309.13111] [INSPIRE].

[46] M. Freytsis, M. Perelstein and Y.C. San, *Anomaly detection in the presence of irrelevant features*, *JHEP* **02** (2024) 220 [arXiv:2310.13057] [INSPIRE].

[47] L. Buitinck et al., *API design for machine learning software: experiences from the scikit-learn project*, arXiv:1309.0238 [INSPIRE].

JHEP12(2025)105