

Isolating Unisolated Upsilons with Anomaly Detection in CMS Open Data

Rikab Gambhir,^{1,2,*} Radha Mastandrea^{3,4,†} Benjamin Nachman^{5,6,7,‡} and Jesse Thaler^{1,2,§}

¹*Center for Theoretical Physics – a Leinweber Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

²*The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, USA*

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

⁴*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

⁵*Berkeley Institute for Data Science, University of California, Berkeley, California 94720, USA*

⁶*Department of Particle Physics and Astrophysics, Stanford University, Stanford, California 94305, USA*

⁷*Fundamental Physics Directorate, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA*



(Received 28 February 2025; revised 1 May 2025; accepted 3 June 2025; published 8 July 2025)

We present the first study of anti-isolated Upsilon decays to two muons ($\Upsilon \rightarrow \mu^+ \mu^-$) in proton-proton collisions at the Large Hadron Collider. Using a machine learning (ML)-based anomaly detection strategy, we “rediscover” the Υ in 13 TeV CMS Open Data from 2016, despite overwhelming anti-isolated backgrounds. We elevate the signal significance to 6.4σ using these methods, starting from 1.6σ using the dimuon mass spectrum alone. Moreover, we demonstrate improved sensitivity from using an ML-based estimate of the multifeature likelihood compared to traditional “cut-and-count” methods. This is the first ever detection of anti-isolated Upsilons, which can be useful in the study of heavy-flavor fragmentation in quantum chromodynamics. Our Letter demonstrates that it is possible and practical to find real signals in experimental collider data using ML-based anomaly detection, and we distill a readily accessible benchmark dataset from the CMS Open Data to facilitate future anomaly detection developments.

DOI: [10.1103/physrevlett.135.021902](https://doi.org/10.1103/physrevlett.135.021902)

Quarkonia fragmentation within jets is an important test of quantum chromodynamics (QCD), since it probes the transition between perturbative and nonperturbative scales [1–4]. At the Large Hadron Collider (LHC), measurements of charmonium (e.g., J/ψ) production within jets by LHCb [5] and CMS [6] have led to improved fragmentation models [7] compared to baseline parton shower predictions with leading-order fragmentation functions [8]. To date, the only study of bottomonium (e.g., Υ) inside jets is an LHCb thesis that emphasized relatively isolated Υ 's [9]. To our knowledge, no published study has looked for *anti-isolated* Υ resonances, whose decay products are *not* well-separated from other collision activity, thereby offering a complementary probe of QCD fragmentation. Unlike anti-isolated J/ψ 's, whose signature in the dimuon channel is readily visible over the background, anti-isolated Υ 's are rarer and are more difficult to identify through simple event selections. This motivates the use of high-dimensional auxiliary

features to reveal anti-isolated Υ 's in collider data. Such features could be used in a traditional bump-hunt analysis, possibly augmented with machine learning (ML).

In this Letter, we present the first study of anti-isolated $\Upsilon \rightarrow \mu^+ \mu^-$ decays at the LHC. Our analysis is based on 13 TeV proton-proton collision data taken in 2016, as accessed via the CMS Open Data [10]. By applying ML-based *anomaly detection*, we identify a statistically significant sample of anti-isolated Υ 's, elevating a marginal 1.6σ excess to well over the 5σ discovery threshold. The anomaly detection method we use, called CATHODE [11], is capable of detecting resonant features without requiring a simulation of the signal or background. We further show how to boost the signal significance by reweighting the data according to the learned multifeature likelihood, which yields better sensitivity than a simple “cut and count” on the anomaly score. To our knowledge, this likelihood-based reweighting has never before been applied to anomaly detection.

The discovery potential of ML-based anomaly detection has been amply demonstrated on synthetic datasets for the LHC Olympics [12] and the Dark Machine Anomaly Score Challenge [13]; see Ref. [14] for a review of related techniques. At the LHC, CMS and ATLAS have started to apply ML-based anomaly detection to search for new physics in jet data, though no significant excesses have emerged [15–19]. Past work on CMS Open Data used ML-based anomaly detection to rediscover the top quark in

*Contact author: rikab@mit.edu

†Contact author: rmastand@berkeley.edu

‡Contact author: nachman@stanford.edu

§Contact author: jthaler@mit.edu

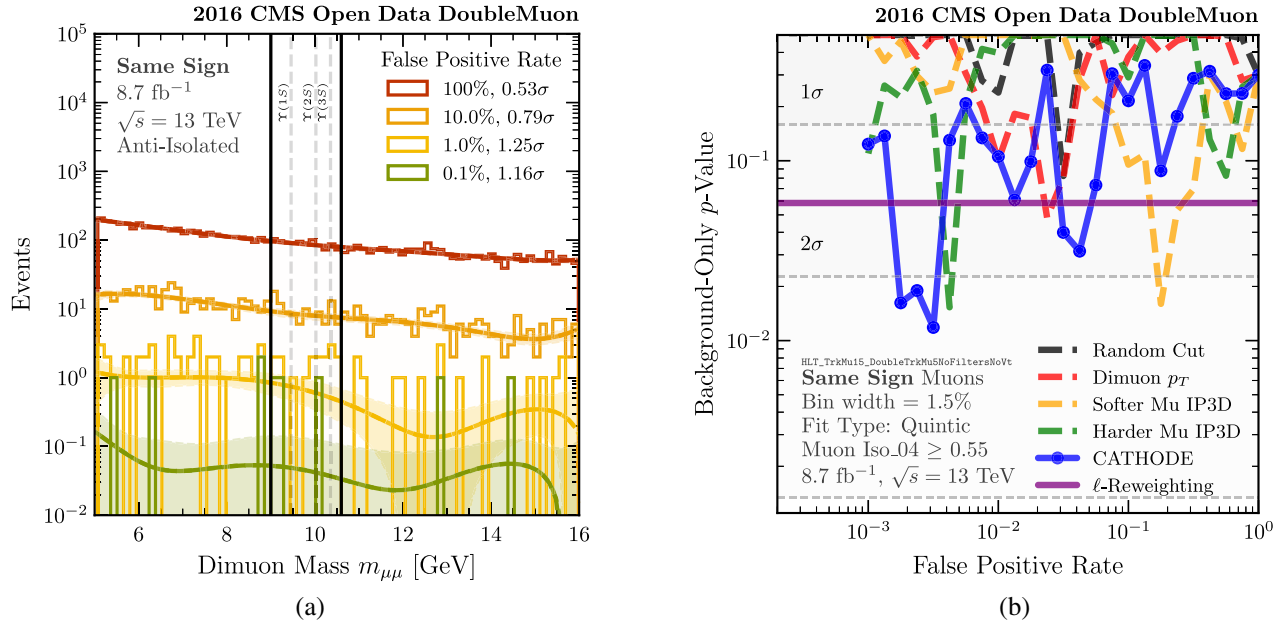


FIG. 1. Results of the SS validation study. (a) Dimuon mass distributions after a series of cuts on the BDT classifier (the second stage of CATHODE), for different FPR working points. For each FPR, a quintic polynomial is fit to the two SB regions and interpolated into the SR, delineated by the vertical black lines. (b) The p value (significance) as a function of FPR for CATHODE (blue), cuts on the dimuon p_T (red) and the muon 3D impact parameters (yellow and green), and a random cut as a baseline (black).

already-known channels [20]. Our Letter of anti-isolated Υ 's builds on this literature by showing that ML-based anomaly detection is capable of finding real signals in real experimental data in regions of phase space not previously studied. To facilitate further ML method development and complement existing synthetic benchmarks, we publish a curated slice of the CMS Open Data used in this Letter along with the code to reproduce our analysis.

Anti-isolated Υ selection—The Υ 's are a series of spin-1 bottom-antibottom ($b\bar{b}$) resonances with masses $m_\Upsilon \geq 2m_b \approx 10$ GeV [21]. There are three main resonances with substantial dimuon branching ratios of $\approx 2\%$: $\Upsilon(1S)$, $\Upsilon(2S)$, and $\Upsilon(3S)$. [The $\Upsilon(4S)$, well-known at B factories, has a thousand-fold smaller dimuon branching ratio.] Our analysis is focused on resonant anti-isolated $\Upsilon \rightarrow \mu^+\mu^-$ decays as seen by CMS in 2016 at $\sqrt{s} = 13$ TeV. This data is made public through the CERN Open Data Portal [22] as the DoubleMu primary dataset [10], corresponding to 8.7 fb⁻¹ of integrated luminosity.

As our primary analysis objects, we select the two highest transverse momentum (p_T) muons that pass the MuonTightID criteria [23]. We also select events that pass the HLT_TrkMu15_DoubleTrkMu5NoFiltersNoVtx trigger, which requires at least 2 tracker muons with p_T 's of at least 15 and 5 GeV, respectively, with no requirement that the muons come from the same primary vertex. To avoid trigger turn-on effects, we enforce further p_T cuts on the muons of 17 and 8 GeV [24]. We refer to the higher (lower) p_T muon as the harder (softer) one. We then divide the data into two channels: *opposite sign* (OS), where

the two hardest muons have opposite electric charges (i.e., $\mu^+\mu^-$), and *same sign* (SS), where they have the same charges (i.e., $\mu^+\mu^+$ or $\mu^-\mu^-$). The SS sample will serve as validation samples for Fig. 1 below.

Many searches for dimuon resonances require the muons to be isolated from the rest of the collision activity. Since we are interested in finding Υ 's produced through QCD fragmentation, we instead impose an *anti-isolation criterion*. Within a rapidity-azimuth (ΔR) distance of 0.4 around each of the two muons, we require the ratio of nonmuon p_T to muon p_T to be *greater* than 0.55. This threshold was chosen to be as stringent as possible while still retaining enough events to apply ML-based techniques—approximately 10k events are needed to ensure our models are well trained. Unlike [9], we do not require any matching to jets. See Ref. [25] for a previous CMS Open Data study that considered prompt dimuons (i.e., coming from the collision vertex), which has a substantial anti-isolated component, and [26] for another CMS Open Data study using ML to study the relation between isolation and promptness.

Anti-isolated Υ 's can only be seen at 1.6σ using the dimuon mass spectrum alone [see the right edge of Fig. 2(b) below]. By comparison, without any cuts, Υ 's are clearly visible at the 28σ level. Anti-isolated Υ 's are typically found within QCD jets and are produced promptly during fragmentation. There are two primary backgrounds to anti-isolated Υ 's. The first is *uncorrelated hadron decays*, where QCD processes produce hadrons (mostly charged pions) that decay in flight to nonprompt muons; this produces approximately equal numbers of SS and OS dimuons. The second is

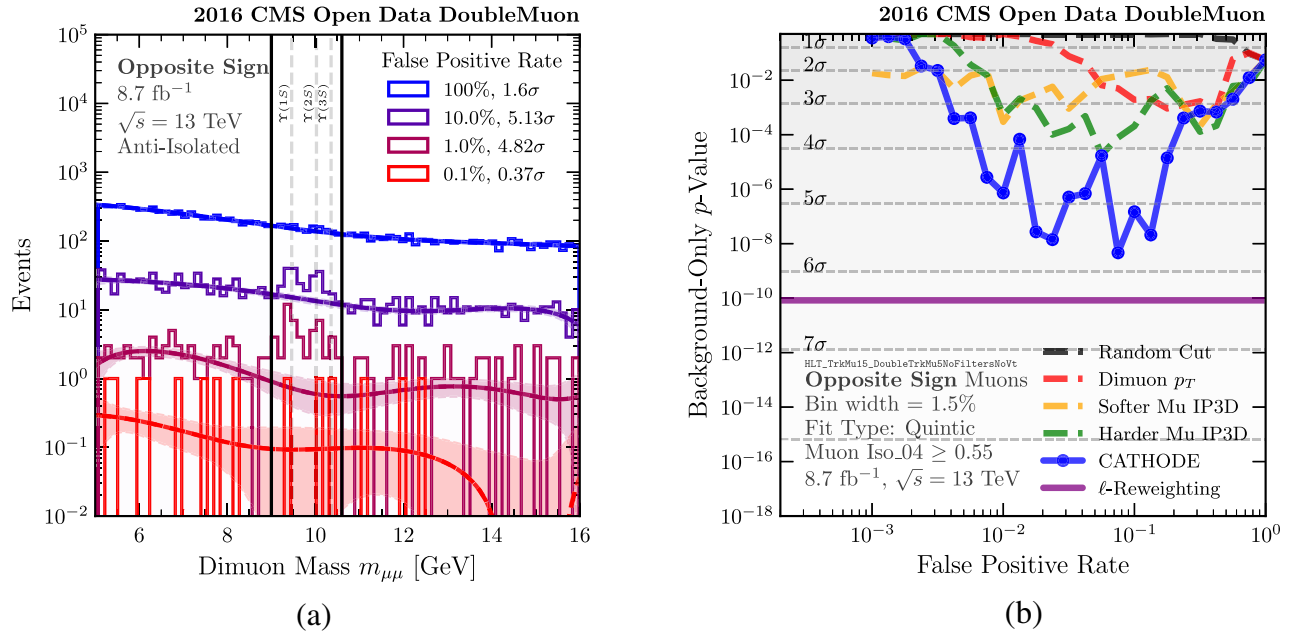


FIG. 2. Results of the OS bump hunt analysis, shown in the same format as Fig. 1. (a) After a cut on the BDT classifier, clear peaks emerge in the SR above the SB-fitted background. (b) Noting the different y-axis scale from Fig. 1(b), the Υ significance grows from 1.6 σ to well over 5 σ with CATHODE, especially with likelihood reweighting.

Drell-Yan production, where virtual photons and Z bosons directly yield prompt dimuons via $\gamma/Z^* \rightarrow \mu^+\mu^-$; this produces only OS dimuons and is essentially an irreducible background. By comparing the OS and SS distributions [Figs. 1(a) and 2(a) below], we estimate that these two backgrounds are roughly the same size in the anti-isolated OS channel.

To enhance the visibility of anti-isolated Υ 's, we consider three auxiliary features—the dimuon transverse momentum $p_{T\mu^+\mu^-}$ and the 3D impact parameters (IP3D) of the harder and softer muons—which we refer to collectively as x . This feature set was chosen both for its performance in reconstructing the Υ and because it does not produce spurious peaks or sculpting artifacts. Other feature sets we considered include various combinations of single- and di-muon kinematic observables; sculpting tended to occur with feature sets too closely correlated with mass, such as those involving the triplet of $p_{T\mu^+\mu^-}$, $\Delta R_{\mu^+\mu^-}$, and $\Delta p_{T\mu^+\mu^-}$. Because of the limited training data, using too many features leads to suboptimal CATHODE performance.

General analysis procedure—The core assumptions of our analysis is that the signal of interest is positive, localized in the dimuon mass $m \equiv m_{\mu^+\mu^-}$, and lives atop a smooth background. Even though we expect three distinct Υ peaks, they count as one resonant feature for a wide enough *signal region* (SR). In the sidebands (SBs), we assume the data is predominantly background, and we distinguish the *low-mass sideband* (SBL) from the *high-mass sideband* (SBH). These regions are separated by vertical black lines in Figs. 1(a) and 2(a) below.

For an ML-based resonance search (first outlined in [27,28]), or *extended bump hunt*, the strategy is to mask the

SR and only use the SB information (and the assumption of smoothness) to construct an estimate of the background distribution $p_{\text{Bkg}}(x, m)$ for a set of auxiliary features x . With these assumptions, we can compare the observed data $p_{\text{Data}}(x, m)$ to the background-only null hypothesis $p_{\text{Bkg}}(x, m)$ in the SR to test for possible resonances. We compare our ML-based results to “classical” cuts on single features, for which only Steps 1 and 4a below are needed.

The full analysis procedure is as follows.

(1) *Choose signal and sideband regions.* To define the bump hunt window, we choose the SR to contain the three dimuon Υ resonances and the SBL and SBH to avoid other known dimuon resonances, corresponding to SBL = [5.0, 9.0] GeV, SR = [9.0, 10.6] GeV, and SBH = [10.6, 16.0] GeV.

(2) *Interpolate background into SR.* Following CATHODE [11], we train an ensemble of normalizing flows (NFs) [29–31] on the SBL and SBH data (with the SR masked out) to learn the conditional distribution $p_{\text{Bkg}}(x|m)$. To estimate $p_{\text{Bkg}}(m)$, we fit a quintic polynomial to a histogram of m in the SBs, using a relative bin width of 1.5%. This width was chosen to be above the 1% dimuon mass resolution (verified via J/ψ fits and consistent with similar dimuon studies [25,32]). We can then generate interpolated background events in the SR according to $p_{\text{Bkg}}(x, m) \equiv p_{\text{Bkg}}(x|m)p_{\text{Bkg}}(m)$ by sampling the NFs weighted by the polynomial mass fit.

(3) *Calculate likelihood ratios.* Restricting our focus to the SR, we train an ensemble of boosted decision trees (BDTs) [33,34] to distinguish observed data $p_{\text{Data}}(x)$ from the interpolated background $p_{\text{Bkg}}(x)$. To do this, BDTs

learn optimal sequences of cuts on their inputs, which is the second stage of CATHODE. We choose BDTs over other architectures because they are fast to train, and they were found to be especially robust to noise in previous extended bump hunt studies [35,36]. We do not provide the BDT with mass information. We assume the data distribution can be expressed as $p_{\text{Data}}(x) = \mu p_{\text{Sig}}(x) + (1 - \mu)p_{\text{Bkg}}(x)$, where $\mu = (N_{\text{Sig}}/N_{\text{Sig}} + N_{\text{Bkg}})$ is the signal fraction. In the asymptotic limit, the classifier score $z(x) = [p_{\text{Data}}(x)/p_{\text{Data}}(x) + p_{\text{Bkg}}(x)]$ is monotonically related to the likelihood ratio,

$$\ell(x) \equiv \frac{z(x) - (1 - \mu)(1 - z(x))}{\mu(1 - z(x))} = \frac{p_{\text{Sig}}(x)}{p_{\text{Bkg}}(x)}, \quad (1)$$

which in turn is the most powerful mass-independent test statistic [37,38]. To make the most of the limited SR data, we use five-fold cross-validation [39].

(4) *Estimate significances.* We have two different methods to estimate the signal significance.

(a) *Cut and count.* Here, we cut on a feature of interest, count the number of events that pass the cut, and compare to the background estimate. We place a cut either on the classifier score for the ML-based approach or on individual features for the classical comparisons. Working points are defined by the corresponding false-positive rate (FPR) in the SBs. To estimate the number of background events in the SR after the cut, we fit the SB mass spectrum using another quintic fit. Given the number of observed events in the SR after the cut, we estimate the one-sided p value to reject the background-only hypothesis following [40].

(b) *Reweight by the likelihood.* Instead of a single FPR working point, we can do a weighted analysis across different cut values. Following [41] in adapting the matrix element method [42,43], we assign each event i a weight $w_i = \ell(x_i)$ related to the estimated signal-to-background likelihood ratio from Eq. (1). These weights generalize the FPR cut in Step 4a: there, events have weight 1 if they pass a given threshold and 0 otherwise. Because the estimate of ℓ is imperfect, some weights may be negative, corresponding to the most backgroundlike events. We impose a mild classifier cut to remove these negative weight events (equivalently, set negative weights to zero). We then fit the mass spectrum of the surviving events in the SB to estimate the background-only p value in the SR—this is just like the fits of Step 4a, but now with weighted events. The Poisson binned likelihoods of [40] are replaced with scaled Poisson distributions [44] to account for the event weights.

The likelihood reweighting in Step 4b is expected to be more powerful than the cut-and-count analysis in Step 4a [41]. Step 4b, however, can only be performed for the ML-based method, since obtaining an estimate for the likelihood ratio requires an estimate of $p_{\text{Bkg}}(x, m)$, which is only accessible via the NF. If the classifier-learned ℓ is a

poor estimate of the likelihood ratio, this method is still valid, albeit with degraded statistical power.

To mitigate issues related to ML training, we take ensemble averages of our models (five NFs and 100 BDTs). The NFs are ensembled after Step 2 by drawing samples equally from each NF, and the BDTs are ensembled after training in Step 3 by averaging their scores. For the quintic polynomial fits, we profile over the fit parameters. We briefly discuss alternate analysis choices in the End Matter, in particular the mass fit form and bin width.

Same-sign validation study—To be confident in the statistical validity of our methods, we apply our analysis procedure to a “control” sample of SS dimuons, with roughly half the number of events as the OS sample. A doubly charged resonance decaying to muons ($\Phi^{++} \rightarrow \mu^+\mu^+$) at ~ 10 GeV is strongly constrained by precision electroweak tests, such as the Z -pole width and running of α_{EM} [21]. Such a resonance is also ruled out by model-specific searches for doubly charged scalars [45,46]. Thus, we seek to verify that our procedure *does not* find a signal in the SS channel where we expect there to be none.

In Fig. 1(a), we show the dimuon mass spectrum after various BDT cuts. As we impose stricter criteria on the FPR (i.e., smaller fraction of background events passing the cut), no significant peaks are observed in the SR. We quantify this via the significance curves in Fig. 1(b). The cut-and-count significances (Step 4a) are plotted as a function of the FPR, for both CATHODE and the classical tests, and no method finds evidence of a localized excess above 2.4σ . Note that these are *local* significances, corresponding to different choices of cuts, not a global significance, which would suffer a small look-elsewhere effect penalty when profiling over this choice. The ML-based likelihood reweighted significance (Step 4b) is plotted as a horizontal line at a modest 1.6σ . The likelihood reweighted significance *does not* suffer a look-elsewhere effect, since it circumvents any choice in cuts and is a single search. We conclude that our method successfully avoids sculpting spurious signals. See the End Matter for additional validations.

Opposite-sign search results—We now present search results in the OS channel for anti-isolated $\Upsilon \rightarrow \mu^+\mu^-$ decays. In Fig. 2(a), we show the dimuon mass distribution after a sequence of BDT cuts. A modest initial excess in the SR is visually amplified by cuts on the classifier. The quantitative gains in signal significance are shown in Fig. 2(b). CATHODE achieves a maximum significance of 5.7σ at a 7.5% FPR working point, which is increased to 6.4σ with likelihood reweighting. By contrast, none of the classical cuts surpass the nominal 5σ discovery threshold—cutting on the harder muon’s IP3D achieves at most 4.1σ significance. This demonstrates substantial gains from using a multidimensional ML-based approach compared to single-feature classical cuts. See the End Matter for a study of multifeature classical cuts.

To better understand how the BDT discriminates Υ -like events from background, we can inspect the impact of

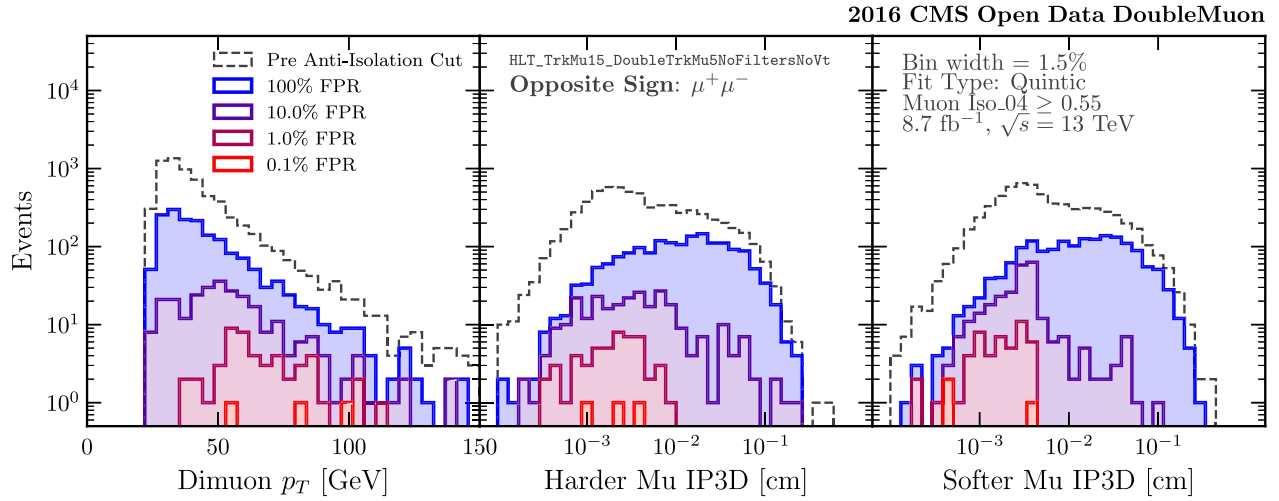


FIG. 3. Distributions in the SR of three auxiliary features—dimuon p_T , harder muon IP3D, and softer muons IP3D—after increasingly stringent cuts on the BDT classifier score. The dashed curves show results before the initial anti-isolation cut.

classifier cuts on the auxiliary features in the SR, shown in Fig. 3. With more stringent FPR cuts, the BDT selects events with moderate dimuon p_T (≈ 60 GeV) and small IP3Ds ($\approx 10^{-3}$ cm), similar to the phase space selection in [25]. In essence, the BDT is attempting to undo the effects of the initial anti-isolation condition to recover the Υ resonances. By focusing on small impact parameters, the BDT is mitigating the background from uncorrelated hadron decays. However, cutting on the IP3D alone does not reproduce the original peak at 10^{-3} cm; if it did, a classical IP3D cut would have matched the performance of CATHODE. By focusing on moderate dimuon p_T , the BDT is additionally rejecting low p_T backgrounds. Both the uncorrelated hadron and Drell-Yan backgrounds fall off at high p_T , although the uncorrelated hadron background does so more sharply, which we can confirm through studies of the SS control sample. Since the Drell-Yan background generates genuine prompt $\mu^+\mu^-$ signatures, the BDT would need access to additional kinematic features to mitigate it further.

Conclusions—Using an ML-based anomaly detection technique, we elevated an anti-isolated $\Upsilon \rightarrow \mu^+\mu^-$ signal to over 5σ significance in CMS Open Data. Our Letter represents the first analysis of anti-isolated Υ production at the LHC. This production channel is particularly useful as a probe for the transition region of QCD, and our Letter is a step toward measuring observables relevant for bottomonium fragmentation. Moreover, we showed that the statistical performance of multi-feature ML-based methods is superior to that of classical single-feature cut and count, with the best performance coming from reweighting events according to the estimated signal-to-background likelihood. This is all done without any simulation of signal or background. Without these novel methods, anti-isolated Υ 's are hard to find: the naive classical cuts we tested yield at best around 4σ significance. With the Υ 's found, it would be interesting to measure their

jet energy fraction to probe bottomonium fragmentation, which would require incorporating detector unfolding into a future analysis.

There are a number of ways we could enhance the statistical sensitivity of our analysis. We chose fixed SR and SBs motivated by the known location of the Υ and other QCD dimuon resonances, but one could consider several candidate regions to find the optimal choice. When estimating backgrounds and likelihoods using NFs and BDTs, we used ensembling to mitigate ML training artifacts, but one could account for both statistical and systematic uncertainties in the architecture and training. This would improve sensitivity for the likelihood reweighting method, where there are uncertainties from the estimate of μ in Eq. (1) lead to suboptimal (though still valid) weights. While we do not expect these variations to substantially improve the quoted significances, they could be relevant for revealing more elusive signals.

Now that we have “rediscovered” the Υ , we plan to build upon this Letter and perform a full scan over the dimuon mass range to search for resonances beyond the standard model. Since the NF and BDT architectures are sufficiently expressive, we do not expect significant hyperparameter reoptimization. However, the choice of auxiliary features is analysis dependent and could change with the final states considered (e.g., jets, missing momentum); in addition, we must verify that any chosen feature set avoids mass sculpting.

We encourage the anomaly detection community to use this Letter and this dataset for development and testing, since the Υ peak serves as a “standard candle” whose properties are well known. This dataset is complementary to previously publicized benchmark datasets, since it corresponds to real, noisy, detector-level experimental data. It also presents new challenges; for example, we found that not all auxiliary feature sets avoided sculpting, a hurdle which we had not encountered in previous studies on

synthetic data. We hope this Υ analysis inspires further anomaly detection studies in real collider data.

Acknowledgments—We would like to thank Ed Witkowski for early discussions and contributions to this work, and Cari Cesarotti, Dennis Noll, Matt Strassler, and Daniel Whiteson for detailed feedback and encouragement. R. G. and J. T. are supported by the National Science Foundation (NSF) under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions [47]), and by the U.S. Department of Energy (DOE) Office of High Energy Physics under Grant No. DE-SC0012567. J. T. is additionally supported by the Simons Foundation through Investigator Grant No. 929241, and in part by Grant No. NSF PHY-2309135 to the Kavli Institute for Theoretical Physics (KITP). R. M. and B. N. are supported by the U.S. DOE Office of Science under Contract No. DE-AC02-05CH11231 and Grant No. 63038 from the John Templeton Foundation. 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 using NERSC Award No. HEP-ERCAP0021099.

Data availability—The data that support the findings of this article are openly available [48]. All of the code to reproduce the analyses and figures in this Letter can be found at [49]. The code consists of a series of numbered PYTHON files and Jupyter notebooks, that, when run in order, will completely reproduce our analyses from scratch. The BDTs are trained using the XGBOOST package [50] and the CATHODE NF is built and trained with PYTORCH [51]. The architectures and hyperparameters used in this analysis can be found in the GITHUB repository in the `configs` folder. The datasets used in this Letter are available on Zenodo at [52].

- [1] Per Ernstrom and Leif Lonnblad, Generating heavy quarkonia in a perturbative QCD cascade, *Z. Phys. C* **75**, 51 (1997).
- [2] A. Andronic *et al.*, Heavy-flavour and quarkonium production in the LHC era: From proton–proton to heavy-ion collisions, *Eur. Phys. J. C* **76**, 107 (2016).
- [3] Reggie Bain, Lin Dai, Andrew Hornig, Adam K. Leibovich, Yiannis Makris, and Thomas Mehen, Analytic and Monte Carlo studies of jets with heavy mesons and quarkonia, *J. High Energy Phys.* **06** (2016) 121.
- [4] Francesco Giovanni Celiberto and Michael Fucilla, Diffractive semi-hard production of a J/ψ or a Υ from single-parton fragmentation plus a jet in hybrid factorization, *Eur. Phys. J. C* **82**, 929 (2022).
- [5] Roel Aaij *et al.* (LHCb Collaboration), Study of J/ψ production in jets, *Phys. Rev. Lett.* **118**, 192001 (2017).
- [6] Armen Tumasyan *et al.* (CMS Collaboration), Fragmentation of jets containing a prompt J/ψ meson in PbPb and pp collisions at $\sqrt{s_{NN}} = 5.02$ TeV, *Phys. Lett. B* **825**, 136842 (2022).
- [7] R. Bain, Y. Makris, T. Mehen, L. Dai, and A. K. Leibovich, NRQCD confronts LHCb data on quarkonium production within jets, *Phys. Rev. Lett.* **119**, 032002 (2017).
- [8] Geoffrey T. Bodwin, Eric Braaten, and G. Peter Lepage, Rigorous QCD analysis of inclusive annihilation and production of heavy quarkonium, *Phys. Rev. D* **51**, 1125 (1995); **55**, 5853(E) (1997).
- [9] Naomi Cooke, Measurements of quarkonia and tetraquark production in jets at LHCb, Ph.D. thesis, Glasgow U., 2023.
- [10] CMS Collaboration, DoubleMuon primary dataset in NANO AOD format from RunH of 2016 (/DoubleMuon/Run2016H-UL2016_MiniAODv2_NanoAODv9-v1/NANO-AOD). CERN Open Data Portal (2024), [10.7483/OPENDATA.CMS.UZD7.Z50M](https://opendata.cern.ch/record/10.7483/OPENDATA.CMS.UZD7.Z50M).
- [11] Anna Hallin, Joshua Isaacson, Gregor Kasieczka, Claudius Krause, Benjamin Nachman, Tobias Quadfasel, Matthias Schlaffer, David Shih, and Manuel Sommerhalder, Classifying Anomalies THrough Outer Density Estimation (CATHODE), *Phys. Rev. D* **106**, 055006 (2022).
- [12] Gregor Kasieczka *et al.*, The LHC Olympics 2020: A community challenge for anomaly detection in high energy physics, *Rep. Prog. Phys.* **84**, 124201 (2021).
- [13] 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).
- [14] Living review of machine learning in high energy physics, <https://iml-wg.github.io/HEPML-LivingReview>.
- [15] Georges 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).
- [16] Georges Aad *et al.* (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**, 052009 (2023).
- [17] Vladimir Chekhovsky *et al.* (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).
- [18] Georges Aad *et al.* (ATLAS Collaboration), Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $s = 13$ TeV with the ATLAS detector, *Phys. Rev. Lett.* **132**, 081801 (2024).
- [19] Georges 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, *arXiv:2502.09770*.
- [20] Oliver Knapp, Guenther Dissertori, Olmo Cerri, Thong Q. Nguyen, Jean-Roch Vlimant, and Maurizio Pierini, Adversarially learned anomaly detection on CMS open data: Rediscovering the top quark, *Eur. Phys. J. Plus* **136**, 236 (2021).
- [21] S. Navas *et al.* (Particle Data Group), Review of particle physics, *Phys. Rev. D* **110**, 030001 (2024).

- [22] CERN Open Data Portal. <https://opendata.cern.ch>.
- [23] A. M. Sirunyan *et al.* (CMS Collaboration), Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV, *J. Instrum.* **13**, P06015 (2018).
- [24] Albert M Sirunyan *et al.* (CMS Collaboration), A search for pair production of new light bosons decaying into muons in proton-proton collisions at 13 TeV, *Phys. Lett. B* **796**, 131 (2019).
- [25] Cari Cesarotti, Yotam Soreq, Matthew J. Strassler, Jesse Thaler, and Wei Xue, Searching in CMS open data for dimuon resonances with substantial transverse momentum, *Phys. Rev. D* **100**, 015021 (2019).
- [26] Edmund Witkowski, Benjamin Nachman, and Daniel Whiteson, Learning to isolate muons in data, *Phys. Rev. D* **108**, 092008 (2023).
- [27] Jack H. Collins, Kiel Howe, and Benjamin Nachman, Anomaly detection for resonant new physics with machine learning, *Phys. Rev. Lett.* **121**, 241803 (2018).
- [28] J. H. Collins, Kiel Howe, and Benjamin Nachman, Extending the search for new resonances with machine learning, *Phys. Rev. D* **99**, 014038 (2019).
- [29] Esteban G. Tabak and Eric Vanden-Eijnden, Density estimation by dual ascent of the log-likelihood, *Commun. Math. Sci.* **8**, 217 (2010).
- [30] Ivan Kobyzev, Simon J. D. Prince, and Marcus A. Brubaker, Normalizing flows: An introduction and review of current methods, *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3964 (2021).
- [31] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, *J. Mach. Learn. Res.* **22**, 2617 (2021).
- [32] Search for prompt production of a GeV scale resonance decaying to a pair of muons in proton-proton collisions at $\sqrt{s} = 13$ TeV, Report No. CMS-PAS-EXO-21-005, 2023.
- [33] Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen, Classification and Regression Trees (Chapman and Hall/CRC, 1984).
- [34] Jerome H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Stat.* **29**, 1189 (2000), <https://www.jstor.org/stable/2699986?origin=JSTOR-pdf>.
- [35] Thorben Finke, Marie Hein, Gregor Kasieczka, Michael Krämer, Alexander Mück, Parada Prangchaikul, Tobias Quadfasel, David Shih, and Manuel Sommerhalder, Tree-based algorithms for weakly supervised anomaly detection, *Phys. Rev. D* **109**, 034033 (2024).
- [36] Marat Freytsis, Maxim Perelstein, and Yik Chuen San, Anomaly detection in the presence of irrelevant features, *J. High Energy Phys.* **02** (2024) 220.
- [37] J. Neyman and E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, *Phil. Trans. R. Soc. A* **231**, 289 (1933).
- [38] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler, Classification without labels: Learning from mixed samples in high energy physics, *J. High Energy Phys.* **10** (2017) 174.
- [39] Markus Ojala and Gemma C. Garriga, Permutation tests for studying classifier performance, in *2009 Ninth IEEE International Conference on Data Mining* (2009), pp. 908–913, [10.1109/ICDM.2009.108](https://doi.org/10.1109/ICDM.2009.108).
- [40] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells, Asymptotic formulae for likelihood-based tests of new physics, *Eur. Phys. J. C* **71**, 1554 (2011); **73**, 2501(E) (2013).
- [41] Marat Freytsis, Grigory Ovanessian, and Jesse Thaler, Dark force detection in low energy e-p collisions, *J. High Energy Phys.* **01** (2010) 111.
- [42] K. Kondo, Dynamical likelihood method for reconstruction of events with missing momentum. 1: Method and Toy models, *J. Phys. Soc. Jpn.* **57**, 4126 (1988).
- [43] K. Kondo, Dynamical likelihood method for reconstruction of events with missing momentum. 2: Mass spectra for $2 \rightarrow 2$ processes, *J. Phys. Soc. Jpn.* **60**, 836 (1991).
- [44] G. Bohm and G. Zech, Statistics of weighted Poisson events and its applications, *Nucl. Instrum. Methods Phys. Res., Sect. A* **748**, 1 (2014).
- [45] Morris L. Swartz *et al.*, A search for doubly charged Higgs scalars in Z decay, *Phys. Rev. Lett.* **64**, 2877 (1990).
- [46] S. Atag and K. O. Ozansoy, Realistic constraints on the doubly charged bilepton couplings from Bhabha scattering with LEP data, *Phys. Rev. D* **68**, 093008 (2003).
- [47] <https://iaifi.org/>
- [48] R. Mastandrea, B. Nachman, J. Thaler, and R. Gambhir, DoubleMuon primary dataset from RunH of 2016 | Selected features from validated luminosity runs | Awkward array format, Zenodo (2025), [10.5281/zenodo.14618719](https://doi.org/10.5281/zenodo.14618719).
- [49] Associated Code: Isolating Unisolated Upsilon's with Anomaly Detection in CMS Open Data, <https://github.com/hep-lbdl/dimuonAD/>.
- [50] Tianqi Chen and Carlos Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16* (ACM, 2016), pp. 785–794, [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785).
- [51] J. Ansel *et al.*, PYTORCH2: Faster machine learning through dynamic PYTHON bytecode transformation and graph compilation, in *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)* (ACM, 2024), [10.1145/3620665.3640366](https://doi.org/10.1145/3620665.3640366).
- [52] <https://zenodo.org/records/14618719>

End Matter

Additional validations—For completeness, we perform two additional validations beyond the SS study in the main text. First, we use a classifier trained on OS data to look for resonances in SS data. We (successfully) do not find a SS signal, which implies that a classifier trained on

(signal-containing) OS data does not sculpt peaks where none exist. Second, we use a classifier trained on SS data to look for resonances in OS data. There is indeed a initial $1.6\sigma\Upsilon$ signal in the OS channel (as in the OS search), but we (successfully) do not elevate the signal significance.

We conclude that in order to reveal the Υ signal, the BDT classifier must be able to learn nontrivial correlations between auxiliary features, which are not present in the SS control sample.

Systematic variations and pseudoexperiments—To test the robustness of our anomaly detection strategy, we want to show its behavior under systematic variations. The baseline results in the main text were based on a quintic polynomial background fit and a dimuon relative bin width of 1.5%. For fit variations, we consider cubic and septic polynomials. For bin width variations, we consider 1.1% and 2.3%.

In Fig. 4, we plot the CATHODE p value as a function of FPR for these different variations, analogous to the solid blue line in Fig. 2(b). All of these variations yield similar results to the baseline approach, and all choices considered surpass the 5σ discovery threshold. This shows that the Υ significances are robust with respect to these choices. If anything, the baseline approach is conservative relative to these variations, likely due to a statistical fluctuation. We note that similar robustness to variations holds for the single-feature classical cuts and for the likelihood reweighting method.

Another potential concern is whether our computed p values are valid even though we are using an ML-based approach. To verify that the standard asymptotic formulae [40] for p -values are still valid, we did a numerical analysis with pseudoexperiments. Specifically, we generated 1,000 pseudoexperiments by sampling from a trained NF, which we take to be our null hypothesis. We then computed the observed distribution of the test statistic, which follows the expected asymptotic distribution out to at least the 3σ level.

Scan over multiple cuts—Given the benefits seen from using multidimensional information, it is interesting to check whether classical cuts on multiple features might mimic the CATHODE performance. Rather than cutting on a single feature, we can scan over simultaneous cuts to see if multi-dimensional auxiliary information can yield a more significant Υ peak. We emphasize that this study was carried out after the anomaly detection analysis, so the direction of the cuts was already informed by Fig. 3.

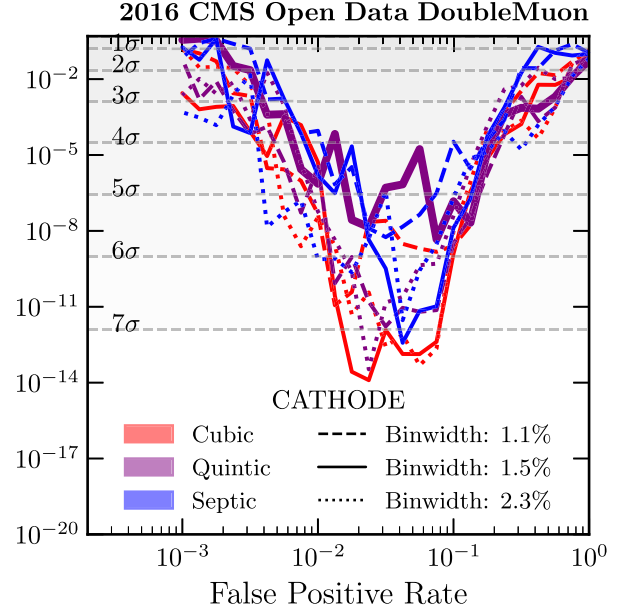


FIG. 4. The CATHODE p values as a function of FPR, for different choices of fit polynomial and bin width. The central fit and binning choice used in Fig. 2(b) is bolded in solid purple.

Nevertheless, it is interesting to see the impact of adding more features without accounting for detailed correlations.

We scan over different cut values for the dimuon p_T and the individual IP3D values, where for simplicity, we use the same cut on both muons. This is a 2D scan, and we choose ten points in each dimension for 100 working points in total. We find a maximum local significance of 5.5σ with $p_T > 45.5$ GeV and IP3D < 0.1 cm, though this does not account for the look-elsewhere effect from considering multiple cut values. Note that this is not the same look-elsewhere effect encountered when scanning over many mass points; rather, each choice of cuts correspond to doing a search in a new signal region. Therefore, the full performance is slightly less than the best CATHODE working point cut (and also less powerful than the likelihood reweighting method), showing that there is a benefit from considering feature correlations.