Accelerating Resonance Searches via Signature-Oriented Pre-training

Congqiao Li,* Antonios Agapitos, Dawei Fu, Leyun Gao, and Qiang Li†
School of Physics and State Key Laboratory of Nuclear Physics and Technology, Peking University, 100871 Beijing, China

Jovin Drews, Gregor Kasieczka, and Louis Moureaux
*Institute for Experimental Physics, Universität Hamburg,
Luruper Chaussee 149, 22761 Hamburg, Germany*

Javier Duarte and Raghav Kansal
University of California San Diego, La Jolla, CA 92093, USA

Huilin Qu
CERN, EP Department, CH-1121 Geneva 23, Switzerland

Cristina Mantilla Suarez
Particle Physics Division, Fermi National Accelerator Laboratory, Batavia, IL 60510, USA

(Dated: May 22, 2024)

The search for heavy resonances beyond the Standard Model (BSM) is a key objective at the LHC. While the recent use of advanced deep neural networks for boosted-jet tagging significantly enhances the sensitivity of dedicated searches, it is limited to specific final states, leaving vast potential BSM phase space underexplored. We introduce a novel experimental method, Signature-Oriented Pre-training for Heavy-resonance ObservatioN (Sophon), which leverages deep learning to cover an extensive number of boosted final states. Pre-trained on the comprehensive JetClass-II dataset, the Sophon model learns intricate jet signatures, ensuring the optimal constructions of various jet tagging discriminates and enabling high-performance transfer learning capabilities. We show that the method can not only push widespread model-specific searches to their sensitivity frontier, but also greatly improve model-agnostic approaches, accelerating LHC resonance searches in a broad sense.

I. INTRODUCTION

Discovery of heavy resonances beyond the Standard Model (BSM) is a long-standing goal of the LHC program. Despite tremendous efforts to search for resonances up to the TeV mass scale, no concrete evidence of a BSM resonance has been established [1–7]. To date, besides these extensive experiments focusing on specific theoretical models, model-agnostic search techniques have also seen consistent progress [8–22] and their experimental implementations have been initiated [23–25]. Their common goal is to enhance the sensitivity to new physics as much as possible in potentially unexpected phase space.

The boosted topology is widely explored in BSM searches at the ATLAS and CMS experiments as it focuses on high-momentum phase space where high-mass-scale new physics is likely to appear first. When probing signals with boosted hadronic final states, recent LHC measurements of Higgs boson properties [26–29] reveal that the main driver of the sensitivity is the enhanced performance of the deep neural network (DNN) used for large-radius (large-R) jet tagging [30–32] resulting from rapid progress in deep learning applied to jet tagging [33–36]. Training and deploying state-of-the-art jet networks in all possible boosted-jet final states should bring us to the sensitivity frontier for various BSM signal searches. However, current boosted-jet taggers deployed in experiments cover limited final states as they are developed for specific tagging purposes [30–32, 37–41]. In contrast, unknown BSM processes may produce jets with unpredictable signatures and may be initiated from arbitrary combinations of SM particles [42, 43]. This leaves the majority of such BSM signal phase space underexplored. Therefore, a tool that enables us to push a broad range of final states towards their sensitivity frontier will accelerate our search for heavy new resonances at the LHC.

In this work, we propose novel LHC experimental methodology called Signature-Oriented Pre-training for Heavy-resonance ObservatioN (abbreviated *Sophon*) to achieve the goal. This methodology introduces a boosted-jet DNN model (the *Sophon model*) learned from a comprehensive jet dataset. It is capable of pushing a broad range of hadronic final-state searches toward the sensitivity frontier and also improving model-agnostic approaches. The Sophon model is pre-trained on a large-scale jet dataset, including various resonance decays that span as wide a range of jet signatures as possible. Thus, it is expected to learn a comprehensive latent representation of jets. For the pre-training task, this work implements large-scale classification, using finely catego-

[*] congqiao.li@cern.ch
[†] qliphy0@pku.edu.cn

rized labels indicating which partons, leptons, or combinations thereof initiated the jet. In total, there are 188 classes. When used in LHC experimental searches, as shown in Fig. 1, the model offers the ability to construct various tagging discriminants directly from its output nodes, which are already optimized for dedicated signals. Moreover, one can adopt the transfer learning technique [44, 45], specifically, using its latent representation nodes as input to train a lightweight DNN for dedicated model-specific or model-agnostic tasks. This approach is highly performant in both tagging capabilities and computational efficiency.

The rest of this paper is organized as follows. Section II introduces the new dataset, the Sophon model, and training details. Section III describes a benchmark of its tagging performance. Section IV presents several experiments demonstrating its large potential for LHC resonance searches. Finally, Section V offers our conclusion and outlook.

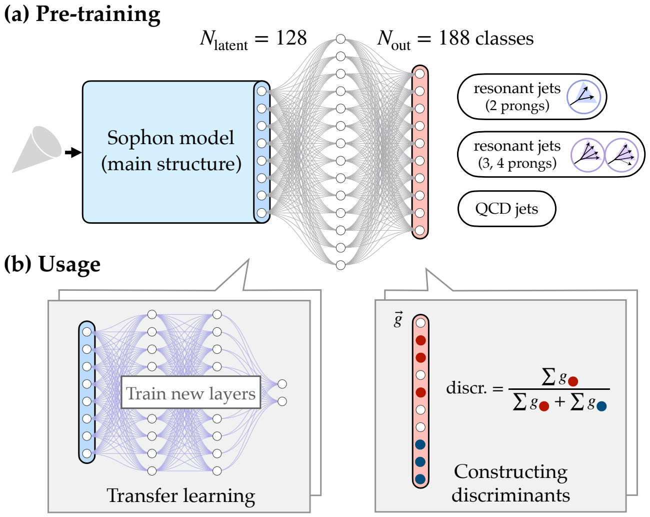
(a) Pre-training
N_{latent} = 128
N_{out} = 188 classes
Sophon model (main structure)
resonant jets (2 prongs)
resonant jets (3, 4 prongs)
QCD jets
(b) Usage
Train new layers
Transfer learning
$\vec{g}$
discr. = $\frac{\sum g_{●}}{\sum g_{●} + \sum g_{●}}$
Constructing discriminants

FIG. 1. Illustration of (a) the Sophon model pre-trained as a large-scale classifier over 188 classes of finely categorized jet signatures, and (b) the usage of the Sophon model by performing transfer learning or constructing discriminants from selected output scores.

II. DATASET AND MODEL

A prerequisite for making the Sophon method effective is having a large-scale and comprehensive dataset. We present the JetClass-II dataset, which includes 188 jet classes to facilitate Sophon model training. Compared with the JetClass dataset [36, 46], JetClass-II contains resonant jets with a broader mass range and an extended set of final states. A resonant jet may contain 2, 3, and 4 prongs, where each prong is initiated by a quark, gluon, or lepton. They are finely categorized with respect to the particle types, quark or lepton flavors, and the tau lepton's decay mode. Specifically, to generate two-prong jets, we consider a generic spin-0 resonance X with mass up to 500 GeV, transverse momentum p_{T} up to 2500 GeV, either charged or neutral, that decays to diparton bb, cc, ss, qq ($q = u/d$), bc, cs, bq, cq, sq, and gg, dilepton ee, $\mu\mu$, $\tau_{\mathrm{h}}\tau_e$, $\tau_{\mathrm{h}}\tau_\mu$, and ditau $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ signatures, where $\tau_{e/\mu/\mathrm{h}}$ represents a tau lepton that subsequently decays into an e, μ, or hadrons, respectively. For jets with 3 or 4 prongs, a decay of $X \to Y^{(*)}Y^{(*)}$ is first performed with a wide m_Y range, $m_Y/m_X \in (0.2, 0.8)$, followed by a similar diparton and dilepton decay of the secondary resonance Y with the additional inclusion of $Y \to e\nu$, $\mu\nu$, $\tau_e\nu$, $\tau_\mu\nu$, and $\tau_{\mathrm{h}}\nu$ signatures. Thus, a complete set of jet final states arising from $Y^{(*)}$ pairs is considered. This results primarily in 4-prong signatures but can also be 3 prongs if an object leaks out of the jet cone or if one of the objects is a neutrino. In addition to the resonant jet, the background jets from quantum chromodynamics (QCD) multijet events are simulated to cover a wide p_{T} and mass range, and they are subdivided into 27 classes based on the number of quarks within the jet and their flavors. A summary of the jet classes can be found in Appendix A.

The simulation of the JetClass-II dataset follows the JetClass simulation workflow [36], while additionally emulating for the effect of pileup (PU) with an average of 50 PU interactions and adopting the PU per particle identification algorithm [47] to remove the PU, with a configuration similar to that used in the CMS experiment [48]. This creates a more realistic dataset that mimics LHC data collected in Run 2. Large-R jets are clustered from the processed E-flow objects in the delphes software [49] using the anti-k_{T} algorithm [50] with $R = 0.8$.

To train the Sophon model, we use the Particle Transformer (ParT) as the backbone with the same model configuration as in Ref. [36], except that the fully connected multilayer perceptron (MLP) is now expanded to two layers, first increasing the dimension to 512, then back to $N_{\mathrm{out}} = 188$ output nodes. The neuron values before passing through the MLP have a dimension of $N_{\mathrm{latent}} = 128$. They are treated as the Sophon model's latent features used for transfer learning. Notably, we adopt two special training techniques in addition to the original ParT training [36]. First, jet samples are selected with a predefined probability during training to ensure a smooth p_{T} and soft-drop mass (m_{SD}) [51, 52] spectrum.This essentially performs a reweighting on p_{T} and m_{SD} of the training dataset, a technique previously explored to minimize the tagger's correlation with jet p_{T} and masses in LHC experiments [32, 34, 53]. The decorrelation with jet mass is especially important for resonance searches to avoid sculpting the background mass distribution when applying a selection on the tagger score. Secondly, the four-momentum of the jet and its constituents are all scaled by a coefficient to satisfy jet $p_{\mathrm{T}} = 500$ GeV before gathering the jet inputs. This improves the scale-invariance of the model and generalizes the tagging performance over a wide mass range. More technical details can be found in Appendix B.

III. PERFORMANCE BENCHMARK

We first compare the Sophon model with the best jet taggers achievable in current experiments as a performance benchmark. Our experiment is delivered on a dataset dedicated to the SM processes, simulated in LHC pp collision at $\sqrt{s} = 13\,\text{TeV}$ and corresponding to the $100\,\text{fb}^{-1}$ of data. This dataset is produced in addition to JetClass-II, using the same configuration for delphes simulation. It imposes a trigger requirement dedicated to the boosted topology study, i.e., the scalar p_T sum of all $R = 0.8$ jets is larger than $800\,\text{GeV}$, and one of the jets should have a trimmed mass [54] larger than $50\,\text{GeV}$. Based on the leading-order cross section calculated from MadGraph5_aMC@nlo [55], this SM dataset includes 5×10^7 events from QCD multijet process, 9×10^5 $V(= W/Z)$+jets events, 3×10^5 events from top quark-antiquark pair ($t\bar{t}$) and single top (ST) quark processes, and other processes including the diboson (VV) and Higgs production.

The performance of tagging resonant $X \to bb$ and $X \to bs$ jets is used for benchmarking. Here, the $X \to bb$ tagging task evaluates the Sophon model's direct tagging capability by constructing discriminants from its output nodes, while $X \to bs$ tagging examines its transfer learning ability since there is no direct correspondence in the model's training classes to the bs signature. The BSM signal process originates from a hypothetical heavy spin-0 (Higgs-boson-like) resonance X_0 with a mass equal to $200\,\text{GeV}$, decaying to bb or bs. For both signal and SM processes, the leading $R = 0.8$ jet satisfying the trigger requirement is used for evaluating various algorithms.

For $X \to bb$ tagging, we begin by discussing the optimal way to construct discriminants from the Sophon model output. A trained multi-class classifier with minimum cross-entropy loss estimates the likelihood ratios of the input classes through the so-called "likelihood-ratio trick" [56]. Specifically, the ith ($i = 1, \cdots, N_\text{out}$) classifier output score $g_i(\mathbf{x})$ given input $\mathbf{x}$ satisfies

$$g_i(\mathbf{x}) = \frac{p(\text{class} = i|\mathbf{x})}{\sum_{j=1}^{N_\text{out}} p(\text{class} = j|\mathbf{x})}, \tag{1}$$

under the ideal DNN assumption, i.e., with sufficient model capacity and data such that the loss reaches the theoretical minimum. Note that the binary classifier form ($N_\text{out} = 2$) of this property has been widely explored in high energy physics [56–58], but its extension to multiple classes form has been investigated less. Here, we show two important properties that can be derived from Eq. (1). They will guide the construction of Sophon's tagging discriminants throughout this work.

Property 1 *Class division property: Consider a classifier with an input class c that is subdivided into multiple exclusive subclasses $\{c_1, \ldots, c_N\}$ to form a new classifier. Let $g_i(\mathbf{x})$ and $g_i'(\mathbf{x})$ denote the output scores of the original and new classifiers, respectively. The output scores of the two classifiers are related as follows.*

$$g_c(\mathbf{x}) = \sum_{l=1}^{N} g_{c_l}'(\mathbf{x}), \quad \textit{and} \;\; g_i(\mathbf{x}) = g_i'(\mathbf{x}) \;\textit{for}\; i \neq c. \tag{2}$$

Property 2 *Extraneous classes property: Consider a classifier that is augmented with additional new input classes $\{e_1, \cdots, e_N\}$ to form a new classifier. The ratios of the output scores for the original classes should remain unchanged, i.e.,*

$$\frac{g_i(\mathbf{x})}{g_j(\mathbf{x})} = \frac{g_i'(\mathbf{x})}{g_j'(\mathbf{x})}, \quad \textit{for}\; i,\, j \notin \{e_1, \cdots, e_N\}. \tag{3}$$

Built on the above properties, the optimal discriminant for distinguishing $X \to bb$ from QCD jets is constructed as

$$\text{discr}\ (X \to bb\ \text{vs. QCD}) = \frac{g_{X\to bb}}{g_{X\to bb} + \sum_{l=1}^{27} g_{\text{QCD}_l}}, \tag{4}$$

where $g_{X\to bb}$ corresponds the $X \to bb$ output score and g_{QCD_l} corresponds to the scores of 27 QCD classes. Ideally, this should be equivalent to training a binary classifier DNN to classify the same $X \to bb$ jets and the undivided QCD jets, then using the $X \to bb$ score as the discriminant. According to the Neyman–Pearson lemma [59], this serves as the strongest discriminant to distinguish the $X \to bb$ and QCD jets.

For $X \to bs$ tagging, transfer learning is applied to the Sophon model from its latent features with a dimension $N_\text{latent} = 128$, using a two-layer MLP with $(512, 2)$ nodes. The parameters of the first linear layer are preloaded from the corresponding part in the Sophon model to ease the learning. It only has two output nodes for classifying $X \to bs$ jets and QCD jets. Here, $X \to bs$ jets are again produced with variable m_X in the same kinematics as jets in JetClass-II. The same mass-decorrelation technique is applied during training. The transfer learning training is much simpler and faster than the original Sophon model training. Only a small fraction (1/320) of the total Sophon training dataset is needed for the transfer learning and the total computational cost (in terms of floating point operations per second) is 1/1 000 000 of the original training.

Figure 2 shows the tagging performance in terms of the discovery significance Z [60] as a function of the SM background selection efficiency, using $40\,\text{fb}^{-1}$ of data and considering events within the mass window $150 < m_\text{SD} < 230\,\text{GeV}$. Several tagging models are compared at a given number of signal injections. To illustrate the current tagging performance achievable at LHC experiments, we train two dedicated tagging models for both tasks: one using state-of-the-art ParT [36] architecture, and one using the ParticleNet model [61]. These are representative of the current tagging capability within the CMS experiment [27, 29, 62]. These models are trained as binary classifiers to distinguish $X \to bb\,(bs)$ jets against the QCD jets, applying similar training settings. The performance of the Sophon model in $X \to bb$ tagging and its transfer learning version in $X \to bs$ tagging already surpasses that of dedicated ParT or ParticleNet trainings. This demonstrates the ability of the method to adapt to various model-specific jet tagging tasks [1].

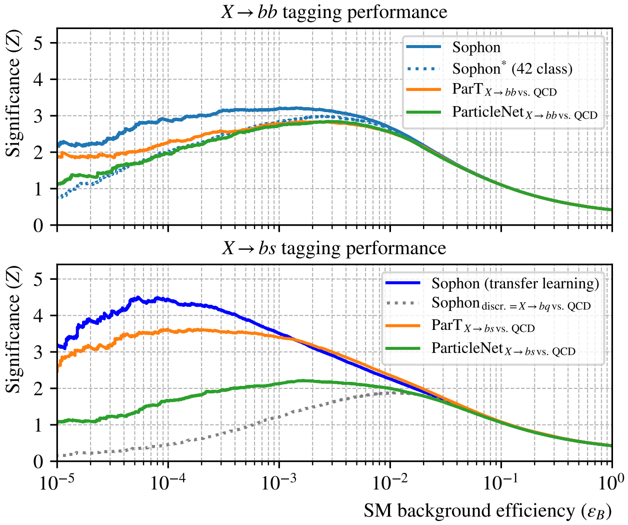

FIG. 2. Benchmark of Sophon model's performance in the $X \to bb$ and $X \to bs$ jet tagging tasks, with the signal originated from a BSM resonance X_0 with $m_{X_0} = 200\,\mathrm{GeV}$ that decays to bb and bs, and backgrounds corresponding to the $40\,\mathrm{fb}^{-1}$ of the full SM processes. The performance of various DNN models is compared in terms of the search's discovery significance versus SM background efficiency calculated within the mass window $150 < m_{\mathrm{SD}} < 230\,\mathrm{GeV}$. A major conclusion is that the Sophon model's direct tagging discriminant (for $X \to bb$ tagging) and its transfer learning version (for $X \to bs$ tagging) both outperform the current best results achievable in the LHC experiment using a ParT or ParticleNet tagger. It also confirms that the model improves performance when trained by large-scale classification.

Additional comparisons are presented in Fig. 2. First, to study whether the Sophon model gained superior $X \to bb$ vs. QCD jet tagging performance from the large-scale classification task, we conduct an ablation study by training the model only on 42 classes (denoted Sophon*), including all 2-prong resonant-jet classes and the QCD classes. These results show that the Sophon model trained on 188 classes significantly improves the discovery significance at a fixed background efficiency, highlighting the importance of pre-training on a large and comprehensive dataset. Second, to confirm that the high performance in $X \to bs$ vs. QCD jet tagging relies on knowledge transferred from the latent space instead of recycling the tagging ability from existing classification nodes, we identify the output node for $X \to bq$ jets that shares the closest similarity with $X \to bs$ jets and check the performance when using Sophon's $X \to bq$ vs. QCD jet tagging discriminant. The latter significantly underperforms, confirming the important role of transfer learning.

IV. IMPLICATIONS FOR RESONANCE SEARCH

After demonstrating the high performance of the Sophon model, we discuss how this approach, once deployed on LHC experiments, will help to accelerate the search for BSM resonances. We discuss two scenarios to combine the Sophon model with resonance search.

The first method leverages the all-inclusive classification nodes of the Sophon model. Since we are unsure about the exact final state of the resonant, we can use these 188 scores to make certain combinations, building the numerator and denominator as shown in Fig. 1 (b), to create a discriminant for jet selection. A typical bump hunt strategy can then be performed on the mass spectrum to search for potential resonances. This method utilizes the extensive classification ability of the Sophon model to distinguish various jet signatures optimally. The second method embeds Sophon's transfer learning into fast-evolving model-agnostic search strategies. Formally, this only involves replacing the existing method's input jet feature space with the Sophon model's latent feature space. Yet, the extensive knowledge of jet signatures encoded in the feature space is expected to yield improved signal-finding performance for a broad class of signal models.

We evaluate the methods above in the single-jet and the dijet topologies. The first topology aims to identify resonance structure in a single jet m_{SD} spectrum. Utilizing the above techniques, we aim to reveal the existence of SM particles amidst the overwhelming QCD multijet backgrounds. The second experiment performs a standard dijet resonance search to find the resonance peak at the TeV mass scale in the dijet invariant mass m_{JJ}. This serves as a benchmark for the proposed methods by comparing them with established model-agnostic strategies.

First, in the single-jet resonance search, we consider the following discriminant to veto QCD jets while purifying certain signal processes,

$$\mathrm{discr}\ (A\ \mathrm{vs.\ QCD}) = \frac{g_A}{g_A + \sum_{l=1}^{27} g_{\mathrm{QCD}_l}}. \tag{5}$$

[1] Note that the absolute tagging performance does not necessarily match real experiments due to the discrepancies between the DELPHES modeling and real detector conditions. Our purpose is to compare methods and draw conclusions about the capabilities of different models. This will remain valid for real experimental conditions.

This study considers three choices for the signature A for illustration:

$$\begin{aligned} g_{A_1} &= g_{X\to cs}, \qquad g_{A_2} = g_{X\to bb}, \\ g_{A_3} &= g_{X\to bq_{\rm all}q_{\rm all}} \equiv \sum_{\substack{\text{sig}\in\{ccb,ssb,\\ qqb,bcs,bcq,bsq\}}} g_{X\to\text{sig}}, \end{aligned} \tag{6}$$

where the g_{A_3} term aims to select all existing 3-prong signatures composed of three quarks with exactly one b quark included. By selecting jets on the three discriminants, Fig. 3 shows the change in the $m_{\rm SD}$ spectrum of the 40 fb^{-1} of the SM events as the selections become tighter. The stacked histograms are also shown at a selection efficiency of $\epsilon_B = 10^{-4}$. Interestingly, the corresponding resonant signatures from the W, Z bosons, and the t quark are revealed. This example demonstrates the Sophon model's broad ability to construct discriminants and sensitively probe resonances with unknown properties.

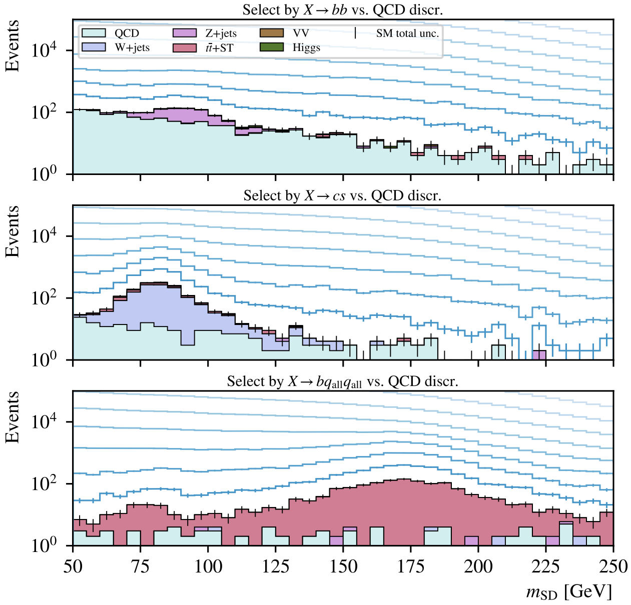

FIG. 3. Distributions of the leading $R = 0.8$ jet $m_{\rm SD}$ for the 40 fb^{-1} of simulated SM events by imposing selections on different Sophon tagging discriminants ($X \to bb$, $X \to cs$, and $X \to bq_{\rm all}q_{\rm all}$ vs. QCD) at various selection efficiencies ϵ. The blue and black curves represents ϵ ranging from 10^0 to 10^{-4}, where the black one corresponds to $\epsilon = 10^{-4}$. The stacked histograms show the contribution of various SM processes at $\epsilon = 10^{-4}$. The $W/Z/t$ peaks can be revealed from the flat QCD multijet background in different cases. The graph demonstrates that constructing various tagging discriminants allows signals with corresponding jet signatures to be purified, revealing distinct signal peaks.

To show the feasibility of using Sophon's learned knowledge in a model-agnostic search, we use the Simulation Assisted Likelihood-free Anomaly Detection (SALAD) method [10] as an illustration. This method utilizes a simulated background dataset to assist in probing the small number of signal events in the data. To search for resonance at $m_{\rm SD} \sim m_0$, we define the signal region (SR) as $(m_0-15,\ m_0+15)$ GeV and the mass sideband (SB) as $(m_0-25,\ m_0-15)\cup(m_0+15,\ m_0+25)$ GeV. Intuitively, this method first learns how to reweight the SB simulation to SB data; with this information, it estimates the background density in SR and trains a classifier to distinguish the estimated SR background from the SR data. The classifier output is theoretically allowed to identify the signal events in SR optimally. We choose the QCD background from 20 fb^{-1} of data as the simulated background. For the rest of the dataset, 40 fb^{-1} samples are used for training, and the other 40 fb^{-1} are used to test the performance. We apply the method with sliding mass windows, changing m_0 from 65 to 295 GeV with a step of 10 GeV. The trained classifier discriminant is applied to the test data in the narrow bin of $(m_0-5,\ m_0+5)$ GeV at a fixed working point to suppress the QCD backgrounds to the 10^{-3} level. Practically, the major difference compared to the original SALAD method [10] is that the input is changed to the jet latent features provided by the Sophon model [2]. Figure 4 shows the $m_{\rm SD}$ spectrum in the test data after the selection is applied, with the $W/Z/t$ related processes being more pronounced. Since no dedicated information on

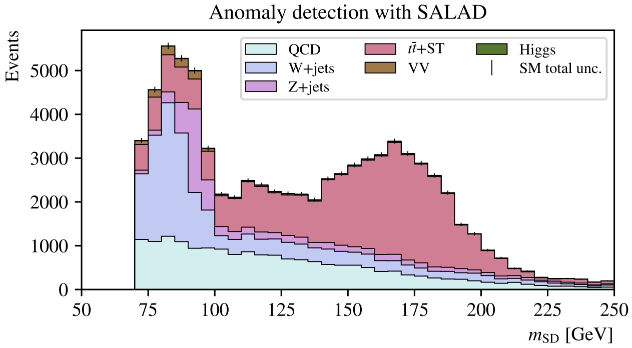

FIG. 4. Distributions of the leading $R = 0.8$ jet $m_{\rm SD}$ for the 40 fb^{-1} of simulated SM events by applying the Simulation Assisted Likelihood-free Anomaly Detection (SALAD) method. The classifier is trained with a sliding mass window, and the selection is applied for each classifier at a QCD multijet efficiency working point of 10^{-3}. The proportion of non-QCD processes is enhanced. The plot shows the feasibility of combining the transfer learning technique with the model-agnostic search strategy.

[2] Note that the original SALAD method is applicable in scenarios where the simulation and data have subtle differences in event generation patterns. This study employs a background simulation using the same generator configuration for simplicity. Nonetheless, It can demonstrate the feasibility of the model-agnostic approach, as the fundamental principle of training a weakly-supervised classifier is preserved under our conditions.

the signatures is provided throughout the process, this experiment shows that it is feasible to adopt the Sophon method in model-agnostic searches.

We then experiment with the techniques in a widely explored dijet resonance search, aiming to benchmark the proposed methods. Specifically, we consider a BSM triboson final-state process initiated by $W' \to W\phi \to WWW$, with $m_{W'} = 3$ TeV, $m_\phi = 400\,\mathrm{GeV}$ and all W boson decaying hadronically. The physics model is discussed in Refs. [63, 64] and is adopted by the early model-agnostic study [8]. We simulate this physics process with the same simulation workflow as above. Events passing the trigger must contain at least two $R = 0.8$ jets with $p_\mathrm{T} > 250\,\mathrm{GeV}$ and $|\eta| < 2.5$. To search for the resonance W' on the dijet invariant mass m_{JJ}, we define the SR as $m_{JJ} \in (2500, 3100)\,\mathrm{GeV}$ and SB as $m_{JJ} \in (2200, 2500) \cup (3100, 3400)\,\mathrm{GeV}$. First, we attempt to construct a discriminant with Sophon's output scores to select the expected signature of both jets. Ideally, we expect one jet to be 2-prong while another jet to be 4-prong; however, given that m_ϕ is large, it is probable that the jet only reconstructs three quarks within the cone. Therefore, we select signatures with either 2 prongs initiated by a W boson, or 3 and 4-prong signatures initiated by WW, optimizing an event-selection discriminant as the sum of two jet-tagging discriminates in the form of

$$\text{event discr.} = \sum_{\text{jet}=1,2} \frac{g_{A,\text{jet}}}{g_{A,\text{jet}} + \sum_{l=1}^{27} g_{\text{QCD}_l,\text{jet}}}, \tag{7}$$

where g_A is defined as

$$g_A \equiv 0.3\, g_{W(2)} + 0.1\, g_{WW(4)} + 0.6\, g_{WW(3)}, \tag{8}$$

and

$$\begin{aligned} g_{W(2)} &\equiv g_{X\to cs} + g_{X\to qq}, \\ g_{WW(4)} &\equiv g_{X\to ccss} + g_{X\to qqcs} + g_{X\to qqqq}, \\ g_{WW(3)} &\equiv \sum_{\substack{\text{sig}\in\{ccs,ccq,ssc,\\ ssq,qqc,qqs,qqq\}}} g_{X\to\text{sig}}. \end{aligned} \tag{9}$$

This can be treated as a model-specific tagging discriminant dedicated to the triboson phase space. The model-agnostic search ability based on the weakly-supervised approach is also studied and benchmarked with established strategies. We experiment with the *idealized* case where the classifier is trained to discriminate data in the SR against the SR backgrounds. This assumes the SR background is perfectly modeled and thus provides a simple benchmark to set a performance limit for all relevant anomaly detection methods under the same input feature space. The limit is denoted as the idealized anomaly detection (IAD) limit [16]. We use the Sophon model's latent features as input to train the IAD classifier, comparing it with using high-level jet inputs adopted by various studies [8–12, 16, 17]. We evaluate the maximum significance improvement, defined as $\max\{Z|_{\epsilon_B>10^{-4}}\}/Z|_{\epsilon_B=1}$ when imposing a selection on the IAD classifier discriminant, as a function of the injected signal yield, shown in Fig. 5.

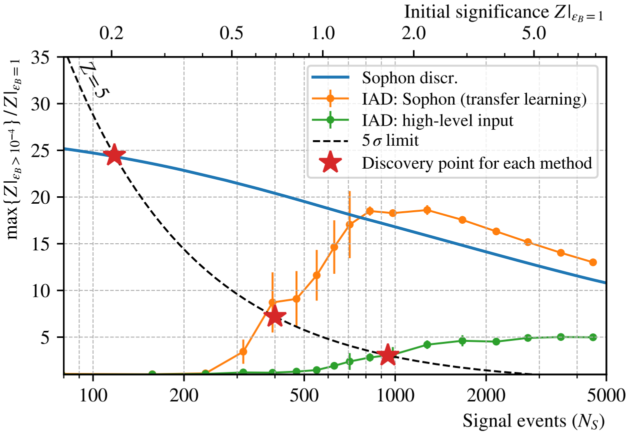

FIG. 5. Benchmark of the model-agnostic dijet anomaly search capability in search of a triboson BSM signal process from $40\,\mathrm{fb}^{-1}$ of simulated SM events. The plot shows the maximum significance improvement defined as $\max\{Z|_{\epsilon_B>10^{-4}}\}/Z|_{\epsilon_B=1}$ with varying signal injection, within the mass window of the dijet invariant mass $2500 < m_{JJ} < 3100\,\mathrm{GeV}$. The idealized anomaly detection (IAD) limit compares the best performance in different input scenarios, including performing transfer learning on the Sophon model and using high-level jet inputs. The error bars correspond to the standard deviation of the maximum significance improvement over 20 trainings. The performance of the constructed Sophon discriminant is also compared. The $5\,\sigma$-limit curve is highlighted to compare the number of signal events required for each method for establishing a first discovery.

Our results show that Sophon's transfer learning combined with IAD enhances not only the maximum significance improvement, but also its sensitivity to the signal at a low injection—the initial significance that the method starts to be aware of the existence of signal is around 0.6–$1.0\,\sigma$. It indicates that by leveraging the learned knowledge of a pre-trained Sophon model, we successfully solve the dilemma that exploring lower-level input features for higher distinguishing capability has to compromise the classifier's sensitivity to low signal injection [65] [3]. Furthermore, we emphasize that a key criterion for evaluating the potential of a method in resonance search is if it can reach the $5\,\sigma$ discovery threshold—widely regarded as the gold standard for new

[3] On this aspect, a recent work [66] in parallel with our study has proposed a similar solution using a pre-trained model for anomaly detection. Our solution differs from this work by proposing the pre-training dedicated to a multitude of jet signatures and utilizing a more lightweight transfer learning to train the weakly-supervised classifier instead of a full model finetuning.

discoveries—in the quickest way. In this regard, Fig. 5 also highlights the $5\,\sigma$-limit curve and the discovery point for each method, showing that Sophon's IAD method can achieve discovery with 2.4 times fewer required signals compared to the traditional methods using high-level inputs, thanks to both improved classifier performance and enhanced sensitivity at lower signal yields. On the other hand, Sophon's signal-targeted discriminant enables a much quicker $5\,\sigma$ discovery, as it requires a further 3.5 times fewer signal events than the much-improved IAD method via Sophon's transfer learning.

This finding implies that if we aim to search for resonance signatures composed of fragments initiated from SM particles, the most efficient strategy is simply to construct discriminants in a multitude of forms and then search for a potential resonant peak in each case. It allows us to push the sensitivity of various resonant searches towards its frontier. On the other hand, for detecting anomalous signals that are totally unlike the known signatures induced by SM particles, the model-agnostic strategy will be advantageous. As physical priors of signals are recognized as important in model-agnostic searches [13, 67], our approach can conceptually enhance transfer capabilities due to the comprehensive jet phase spaces the Sophon model learns from. Overall, by exploring both methods in the broad resonance search program at the LHC, we can expect significant potential to improve search sensitivity and, hopefully, accelerate the next possible discovery.

V. CONCLUSION AND OUTLOOK

We propose the *Sophon* methodology for signature-oriented pre-training over a large-scale dataset and preset JetClass-II, which covers comprehensive boosted jet signatures. Pre-trained on JetClass-II as a large-scale classifier, the *Sophon model* can distinguish over a hundred different jet signatures, showing superior performance in constructed tagging discriminant and transfer learning, outperforming current best results achievable from the LHC experiment. The resonance search studies suggest that it can push a broad range of resonance searches to the sensitivity frontier and also greatly improve model-agnostic searches. It opens a promising direction in conducting future boosted-jet searches at the LHC.

Driven by rapid advancements in deep learning for developing large models, recent LHC phenomenology works have focused on jet model pre-training and fine-tuning applications [66, 68–71]. Compared with established studies, our work demonstrates the importance of building large-scale datasets to train an expressive model and highlights its significance for broadly accelerating the resonance search. Additionally, this methodology can be naturally integrated into the existing LHC analysis workflow. One can store the latent features in the central dataset to facilitate the use of its output scores or the highly efficient transfer learning without the need to revisit the full model. The inference of the Sophon model is also affordable, as it shares a similar computational cost with the default ParT or ParticleNet architecture [36].

As this work explores a simple classification approach under the pre-training methodology, further studies may extend the use of jet signatures and combine them into novel training targets to improve the model's expressiveness. Additionally, exploring novel applications of the Sophon model within LHC data analyses, other than generic resonance searches, can be an interesting task. Addressing challenges such as calibration can pave the way for its broader application.

The JetClass-II dataset (with the delphes configuration) and the Sophon model will be publicly available.

ACKNOWLEDGEMENT

CL, AA, DF, LG, and QL are supported by National Natural Science Foundation of China (NSFC) under Grants No. 12325504, No. 12061141002, and No. 2075004. JD and RK are supported by the DOE, Office of Science, Office of High Energy Physics Early Career Research program under Grant No. DE-SC0021187, and the U.S. National Science Foundation (NSF) Harnessing the Data Revolution (HDR) Institute for Accelerating AI Algorithms for Data Driven Discovery (A3D3) under Cooperative Agreement No. PHY-2117997. JD, GK, and LM acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC2121 "Quantum Universe" – 390833306. The research is supported in part by the computational resource operated at the Institute of High Energy Physics (IHEP) of the Chinese Academy of Sciences. We appreciate helpful discussions with our colleagues, Anna Benecke, Stephane Cooperstein, Sen Deng, Raffaele Gerosa, Loukas Gouskos, Zichun Hao, Farouk Mokhtar, Spandan Mondal, Alexandre De Moor, Jennifer Ngadiuba, Sitian Qian, Melissa Quinnan, Sebastian Wuchterl, and Yuzhe Zhao in the CMS Collaboration when studying the experimental version of this methodology. CL thanks Zixun Kou, Haonan Lu, and Fanqiang Meng for discussions with them to prepare the manuscript and for the support from Yannan Pan.

[1] G. Aad *et al.* (ATLAS), The quest to discover supersymmetry at the ATLAS experiment, arXiv:2403.02455 [hep-ex] (2024).

[2] ATLAS Collaboration, Exotic Physics Searches (2024), `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults`.

[3] ATLAS Collaboration, Supersymmetry Searches (2024), `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults`.

[4] ATLAS Collaboration, Higgs and Diboson Searches (2024), `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/HDBSPublicResults`.

[5] CMS Collaboration, CMS Exotica Public Physics Results (2024), `https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO`.

[6] CMS Collaboration, CMS Supersymmetry Physics Results (2024), `https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS`.

[7] CMS Collaboration, CMS Beyond-two-generations (B2G) Public Physics Results (2024), `https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G`.

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

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

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

[11] G. Stein, U. Seljak, and B. Dai, Unsupervised in-distribution anomaly detection of new physics through conditional density estimation, in *34th Conference on Neural Information Processing Systems* (2020) arXiv:2012.11638 [cs.LG].

[12] O. Amram and C. M. Suarez, Tag N' Train: a technique to train improved classifiers on unlabeled data, JHEP **01**, 153, arXiv:2002.12376 [hep-ph].

[13] S. E. Park, D. Rankin, S.-M. Udrescu, M. Yunus, and P. Harris, Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge, JHEP **21**, 030, arXiv:2011.03550 [hep-ph].

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

[15] S. Tsan, R. Kansal, A. Aportela, D. Diaz, J. Duarte, S. Krishna, F. Mokhtar, J.-R. Vlimant, and M. Pierini, Particle Graph Autoencoders and Differentiable, Learned Energy Mover's Distance, in *35th Conference on Neural Information Processing Systems* (2021) arXiv:2111.12849 [physics.data-an].

[16] 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), arXiv:2109.00546 [hep-ph].

[17] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, Resonant anomaly detection without background sculpting, Phys. Rev. D **107**, 114012 (2023), arXiv:2210.14924 [hep-ph].

[18] 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**, 899345 (2023), arXiv:2203.09470 [hep-ph].

[19] M. Farina, Y. Nakai, and D. Shih, Searching for New Physics with Deep Autoencoders, Phys. Rev. D **101**, 075021 (2020), arXiv:1808.08992 [hep-ph].

[20] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or What?, SciPost Phys. **6**, 030 (2019), arXiv:1808.08979 [hep-ph].

[21] J. H. Collins, P. Martín-Ramiro, B. Nachman, and D. Shih, Comparing weak- and unsupervised methods for resonant anomaly detection, Eur. Phys. J. C **81**, 617 (2021), arXiv:2104.02092 [hep-ph].

[22] V. Belis, P. Odagiu, and T. K. Aarrestad, Machine learning for anomaly detection in particle physics, Rev. Phys. **12**, 100091 (2024), arXiv:2312.14190 [physics.data-an].

[23] G. Aad *et al.* (ATLAS), Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector, Phys. Rev. Lett. **125**, 131801 (2020), arXiv:2005.02983 [hep-ex].

[24] G. Aad *et al.* (ATLAS), 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), arXiv:2307.01612 [hep-ex].

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

[26] A. M. Sirunyan *et al.* (CMS), Inclusive search for highly boosted Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV, JHEP **12**, 085, arXiv:2006.13251 [hep-ex].

[27] A. Tumasyan *et al.* (CMS), Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. Lett. **131**, 061801 (2023), arXiv:2205.05550 [hep-ex].

[28] A. Tumasyan *et al.* (CMS), Search for Higgs boson and observation of Z boson through their decay into a charm quark-antiquark pair in boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. Lett. **131**, 041801 (2023), arXiv:2211.14181 [hep-ex].

[29] A. Tumasyan *et al.* (CMS), Search for nonresonant pair production of highly energetic Higgs bosons decaying to bottom quarks, Phys. Rev. Lett. **131**, 041803 (2023), arXiv:2205.06667 [hep-ex].

[30] CMS Collaboration, *Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques*, CMS Detector Performance Summary CMS-DP-2020-002 (2020).

[31] CMS Collaboration, *Performance of the mass-decorrelated DeepDoubleX classifier for double-b and double-c large-radius jets with the CMS detector*, CMS Detector Performance Summary CMS-DP-2022-041 (2022).

[32] ATLAS Collaboration, *Transformer Neural Networks for Identifying Boosted Higgs Bosons decaying into* $b\bar{b}$ *and* $c\bar{c}$ *in ATLAS*, Tech. Rep. (CERN, Geneva, 2023).

[33] A. Butter *et al.*, The Machine Learning landscape of top taggers, SciPost Phys. **7**, 014 (2019), arXiv:1902.09914 [hep-ph].

[34] E. A. Moreno, O. Cerri, J. M. Duarte, H. B. Newman, T. Q. Nguyen, A. Periwal, M. Pierini, A. Serikova, M. Spiropulu, and J.-R. Vlimant, JEDI-net: a jet identification algorithm based on interaction networks, Eur. Phys. J. C **80**, 58 (2020), arXiv:1908.05318 [hep-ex].

[35] E. A. Moreno, T. Q. Nguyen, J.-R. Vlimant, O. Cerri, H. B. Newman, A. Periwal, M. Spiropulu, J. M. Duarte, and M. Pierini, Interaction networks for the identification of boosted $H \to b\bar{b}$ decays, Phys. Rev. D **102**, 012010 (2020), arXiv:1909.12285 [hep-ex].

[36] H. Qu, C. Li, and S. Qian, Particle transformer for jet tagging, in *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 162 (PMLR, 2022) p. 18281, arXiv:2202.03772 [hep-ph].

[37] A. M. Sirunyan *et al.* (CMS), Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques, JINST **15**, P06005, arXiv:2004.08262 [hep-ex].

[38] A. M. Sirunyan *et al.* (CMS), Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV, JINST **13**, P05011, arXiv:1712.07158 [physics.ins-det].

[39] G. Aad *et al.* (ATLAS), Identification of boosted Higgs bosons decaying into b-quark pairs with the ATLAS detector at 13 TeV, Eur. Phys. J. C **79**, 836 (2019), arXiv:1906.11005 [hep-ex].

[40] ATLAS Collaboration, *Identification of hadronically-decaying top quarks using UFO jets with ATLAS in Run 2*, ATLAS PUB Note ATL-PHYS-PUB-2021-028 (2021).

[41] ATLAS Collaboration, *Performance of W/Z taggers using UFO jets in ATLAS*, ATLAS PUB Note ATL-PHYS-PUB-2021-029 (2021).

[42] N. Craig, P. Draper, K. Kong, Y. Ng, and D. Whiteson, The unexplored landscape of two-body resonances, Acta Phys. Polon. B **50**, 837 (2019), arXiv:1610.09392 [hep-ph].

[43] J. H. Kim, K. Kong, B. Nachman, and D. Whiteson, The motivation and status of two-body resonance decays after the LHC Run 2 and beyond, JHEP **04**, 030, arXiv:1907.06659 [hep-ph].

[44] S. J. Pan and Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering **22**, 1345 (2010).

[45] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, A survey on deep transfer learning, in *Artificial Neural Networks and Machine Learning – ICANN 2018* (Springer International Publishing, Cham, 2018) pp. 270–279.

[46] H. Qu, C. Li, and S. Qian, JetClass: A Large-Scale Dataset for Deep Learning in Jet Physics, 10.5281/zenodo.6619768 (2022).

[47] D. Bertolini, P. Harris, M. Low, and N. Tran, Pileup Per Particle Identification, JHEP **10**, 059, arXiv:1407.6013 [hep-ph].

[48] A. M. Sirunyan *et al.* (CMS), Pileup mitigation at CMS in 13 TeV data, JINST **15**, P09018, arXiv:2003.00503 [hep-ex].

[49] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, JHEP **02**, 057, arXiv:1307.6346 [hep-ex].

[50] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, JHEP **04**, 063, arXiv:0802.1189 [hep-ph].

[51] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, Towards an understanding of jet substructure, JHEP **09**, 029, arXiv:1307.0007 [hep-ph].

[52] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft Drop, JHEP **05**, 146, arXiv:1402.2657 [hep-ph].

[53] L. Bradshaw, R. K. Mishra, A. Mitridate, and B. Ostdiek, Mass Agnostic Jet Taggers, SciPost Phys. **8**, 011 (2020), arXiv:1908.08959 [hep-ph].

[54] D. Krohn, J. Thaler, and L.-T. Wang, Jet Trimming, JHEP **02**, 084, arXiv:0912.1342 [hep-ph].

[55] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, JHEP **07**, 079, arXiv:1405.0301 [hep-ph].

[56] K. Cranmer, J. Pavez, and G. Louppe, Approximating Likelihood Ratios with Calibrated Discriminative Classifiers, arXiv:1506.02169 [stat.AP] (2015).

[57] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Constraining Effective Field Theories with Machine Learning, Phys. Rev. Lett. **121**, 111801 (2018), arXiv:1805.00013 [hep-ph].

[58] A. Andreassen and B. Nachman, Neural Networks for Full Phase-space Reweighting and Parameter Tuning, Phys. Rev. D **101**, 091901 (2020), arXiv:1907.08209 [hep-ph].

[59] J. Neyman and E. S. Pearson, IX. On the problem of the most efficient tests of statistical hypotheses, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **231**, 289 (1933).

[60] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur. Phys. J. C **71**, 1554 (2011), [Erratum: Eur.Phys.J.C 73, 2501 (2013)], arXiv:1007.1727 [physics.data-an].

[61] H. Qu and L. Gouskos, ParticleNet: Jet Tagging via Particle Clouds, Phys. Rev. D **101**, 056019 (2020), arXiv:1902.08570 [hep-ph].

[62] A. M. Sirunyan *et al.* (CMS), Search for a massive scalar resonance decaying to a light scalar and a Higgs boson in the four b quarks final state with boosted topology, Phys. Lett. B **19**, 10.1016/j.physletb.2022.137392 (2022), arXiv:2204.12413 [hep-ex].

[63] K. Agashe, P. Du, S. Hong, and R. Sundrum, Flavor Universal Resonances and Warped Gravity, JHEP **01**, 016, arXiv:1608.00526 [hep-ph].

[64] K. Agashe, J. H. Collins, P. Du, S. Hong, D. Kim, and R. K. Mishra, Dedicated Strategies for Triboson Signals from Cascade Decays of Vector Resonances, Phys. Rev. D **99**, 075016 (2019), arXiv:1711.09920 [hep-ph].

[65] 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), arXiv:2310.06897 [hep-ph].

[66] V. Mikuni and B. Nachman, OmniLearn: A Method to Simultaneously Facilitate All Jet Physics Tasks, arXiv:2404.16091 [hep-ph] (2024).

[67] C. L. Cheng, G. Singh, and B. Nachman, Incorporating Physical Priors into Weakly-Supervised Anomaly Detection, arXiv:2405.08889 [hep-ph] (2024).

[68] B. M. Dillon, G. Kasieczka, H. Olischlager, T. Plehn, P. Sorrenson, and L. Vogel, Symmetries, safety, and self-supervision, SciPost Phys. **12**, 188 (2022), arXiv:2108.04253 [hep-ph].

[69] M. Vigl, N. Hartman, and L. Heinrich, Finetuning Foundation Models for Joint Analysis Optimization,

arXiv:2401.13536 [hep-ex] (2024).

[70] L. Heinrich, T. Golling, M. Kagan, S. Klein, M. Leigh, M. Osadchy, and J. A. Raine, Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models, arXiv:2401.13537 [hep-ph] (2024).

[71] J. Birk, A. Hallin, and G. Kasieczka, OmniJet-α: The first cross-task foundation model for particle physics, arXiv:2403.05618 [hep-ph] (2024).

[72] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015), arXiv:1410.3012 [hep-ph].

[73] J. Thaler and K. Van Tilburg, Identifying Boosted Objects with N-subjettiness, JHEP **03**, 015, arXiv:1011.2268 [hep-ph].

[74] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, Lookahead optimizer: k steps forward, 1 step back, Advances in Neural Information Processing Systems **32** (2019).

[75] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, On the variance of the adaptive learning rate and beyond, in *International Conference on Learning Representations* (2020).

Appendix A: Supplementary details on JetClass-II

Simulation.— The JetClass-II dataset includes a variety of resonant jets and QCD jets. The resonant jets are initiated from (1) $X \to 2$ prong signatures with neutral resonance X, (2) $X \to 2$ prong signatures with charged resonance X, and (3) $X \to Y^{(*)}Y^{(*)} \to 4$ prong signatures, with an introduction below.

The case (1) is generated by the $p\,p \to H\,H$ process using the MadGraph5_aMC@nlo (MG) 2.9.18 [55] generator at the leading order (LO) with the heft model, with 16 M events in total. To control the resonant jet's p_{T} and mass, the minimum p_{T} of H at the hard-scattering level is sampled in 50 points from (100, 2500) GeV in the logarithm spacing and the H mass is uniformly sampled from (15, 500) GeV with a step of 5 GeV. The decay of the H resonance and parton showering is simulated by pythia 8.3 [72]. The decay modes (branching ratio) are bb $(\frac{1}{8})$, cc $(\frac{1}{8})$, ss $(\frac{1}{8})$, dd $(\frac{1}{16})$, uu $(\frac{1}{16})$, gg $(\frac{1}{8})$, ee $(\frac{1}{16})$, $\mu\mu$ $(\frac{1}{16})$, $\tau\tau$ $(\frac{1}{4})$. The subsequent τ decay follows the SM configuration.

The case (2) is generated by $p\,p \to H^+\,H^-$ process using MG at LO with the 2HDM model, with 12 M events in total. The same configuration on the $H^\pm$ minimum p_{T} and its mass is used. The decay of the $H^\pm$ resonance and parton showering is simulated by pythia 8.3. The decay modes (branching ratio) are du $(\frac{1}{6})$, su $(\frac{1}{6})$, bu $(\frac{1}{6})$, cd $(\frac{1}{6})$, cs $(\frac{1}{6})$, bc $(\frac{1}{6})$.

The case (3) is generated by the $p\,p \to h\,h$ process using MG at LO with the 2HDM model, with 120 M events in total. The same configuration on the h minimum p_{T} and its mass is used. The decay processes are simulated by pythia 8.3, with the resonance h decaying to HH and H^+H^-, each with $\frac{1}{2}$ branching ratio. Here, the h, H, and $H^\pm$ are the two $\mathcal{CP}$-even Higgs and the charged Higgs bosons in the 2HDM model. The mass of H and $H^\pm$ are configured as λm_h, where λ is sampled uniformly within (0.2, 0.8) in MG. H decay is the same with the case (1), while the $H^\pm$ decay is similar to the case (2) but with special inclusion of decay nodes including a neutrino final state. The $H^\pm$ decay modes (branching ratio) are then du $(\frac{5}{48})$, su $(\frac{5}{48})$, bu $(\frac{5}{48})$, cd $(\frac{5}{48})$, cs $(\frac{5}{48})$, bc $(\frac{5}{48})$, $e\nu$ $(\frac{1}{16})$, $\mu\nu$ $(\frac{1}{16})$, $\tau\nu$ $(\frac{1}{4})$. The subsequent τ decay follows the SM configuration.

To initiate QCD jets, a $2 \to 2$ QCD multijet process with the pythia 8.3 generator is simulated with 20 M events. The minimum p_{T} of the hard-scattering process is sampled within (100, 5000) GeV in the logarithm spacing to ensure a wider jet p_{T} and mass coverage.

The simulated events from all the above processes are proceeded with delphes 3 [49] for fast simulation of the detector response and object reconstruction. The delphes simulation card is adapted from the CMS detector configuration but modifies the impact parameter of charged particles to match with the CMS tracker resolution, similar to the JetClass simulation [36] In addition, the pileup (PU) effect with an average of 50 PU interactions are included, adapted from the CMS detector configuration with PU [49]. The PU per particle identification (PUPPI) algorithm [47] is also applied for PU removal, adapted from the CMS detector configuration in Phase-II [49] with additional parameter modifications based on the Phase-I CMS detector condition and taking the CMS experimental configuration as a reference [48]. The PUPPI algorithm assigns a value between 0 and 1 to each E-flow object that signifies the probability the object originates from the genuine interaction. It scales the object's four-momentum with the value. The processed E-flow objects are used to cluster large-R jets using the anti-k_{T} algorithm [50] with $R_0 = 0.8$. Selected jets must satisfy $200 < p_{\mathrm{T}} < 2500\,\mathrm{GeV}$, $20 < m_{\mathrm{SD}} < 500\,\mathrm{GeV}$, and $|\eta| < 2.5$.

Jet labeling.— A large-R jet produced by the diresonant production process is assigned a resonant-jet label if it matches any of the 161 resonant-jet classes or is discarded if it does not match any resonant-jet label. We count the first-generation decay products of the resonance X or $Y^{(*)}Y^{(*)}$. The u and d quarks are merged with and use q to represent them. The tau lepton is further exclusively divided into three subclasses, τ_e, τ_μ, and τ_{h}, if it sequentially decays into an electron, muon, or hadrons. This results in the following truth particle types that may appear in the final state: $b/c/s/q/g/e/\mu/\tau_e/\tau_\mu/\tau_{\mathrm{h}}$ (neutrinos should be excluded). The matching of the jet with a truth label requires the matching rule of all truth particles presented in the label to be satisfied. The matching rule is $\Delta R(\text{particle}, \text{jet}) < R_0$ for particle types expect for τ_e or τ_μ; for the latter, a matching requires their decayed e or μ daughter satisfies $\Delta R < R_0$ with respect to the jet axis.

A large-R jet produced by the QCD multijet process is exclusively assigned a QCD label according to the number of b, c, s quarks, read from the pythia parton list before their hadronization, satisfying $p_{\mathrm{T}} > 10\,\mathrm{GeV}$, and matched with the jet axis by $\Delta R < R_0$. For each number, it is categorized into three cases: 0, 1, ≥ 2. This gives rise to 27 QCD classes in total.

The labels and their indices provided in the JetClass-II dataset are summarised in Table I.

The total number of the labeled jets in the JetClass-II dataset is around 139 M. This includes 22 M resonant jets with 2 prongs, 99 M resonant jets with 3 or 4 prongs, and 18 M QCD jets.

Jet variables.— The jet constituent features are used as the input for neural network training. These features are carried on E-flow objects and include kinematic features, particle identification variables, and impact parameters features, closely following the JetClass dataset. The jet kinematic variables are also provided in the dataset used to construct the neural work input.

Additional variables include the jet N-subjettiness variables [73] up to $N = 4$ as a representative of the high-level jet observables, and several generator-level variables indicating the jet signatures. These include the jet label, assigned by counting for the matched truth particles in-

TABLE I. Summary of the 188 jet labels in the JetClass-II dataset.

Major types	**Index range**	**Label names**
Resonant jets: $X \to 2$ prong	0–14	bb, cc, ss, qq, bc, cs, bq, cq, sq, gg, ee, $\mu\mu$, $\tau_\mathrm{h}\tau_e$, $\tau_\mathrm{h}\tau_\mu$, $\tau_\mathrm{h}\tau_\mathrm{h}$
Resonant jets: $X \to 3$ or 4 prong	15–160	$bbbb$, $bbcc$, $bbss$, $bbqq$, $bbgg$, $bbee$, $bb\mu\mu$, $bb\tau_\mathrm{h}\tau_e$, $bb\tau_\mathrm{h}\tau_\mu$, $bb\tau_\mathrm{h}\tau_\mathrm{h}$, bbb, bbc, bbs, bbq, bbg, bbe, $bb\mu$, $cccc$, $ccss$, $ccqq$, $ccgg$, $ccee$, $cc\mu\mu$, $cc\tau_\mathrm{h}\tau_e$, $cc\tau_\mathrm{h}\tau_\mu$, $cc\tau_\mathrm{h}\tau_\mathrm{h}$, ccb, ccc, ccs, ccq, ccg, cce, $cc\mu$, $ssss$, $ssqq$, $ssgg$, $ssee$, $ss\mu\mu$, $ss\tau_\mathrm{h}\tau_e$, $ss\tau_\mathrm{h}\tau_\mu$, $ss\tau_\mathrm{h}\tau_\mathrm{h}$, ssb, ssc, sss, ssq, ssg, sse, $ss\mu$, $qqqq$, $qqgg$, $qqee$, $qq\mu\mu$, $qq\tau_\mathrm{h}\tau_e$, $qq\tau_\mathrm{h}\tau_\mu$, $qq\tau_\mathrm{h}\tau_\mathrm{h}$, qqb, qqc, qqs, qqq, qqg, qqe, $qq\mu$, $gggg$, $ggee$, $gg\mu\mu$, $gg\tau_\mathrm{h}\tau_e$, $gg\tau_\mathrm{h}\tau_\mu$, $gg\tau_\mathrm{h}\tau_\mathrm{h}$, ggb, ggc, ggs, ggq, ggg, gge, $gg\mu$, bee, cee, see, qee, gee, $b\mu\mu$, $c\mu\mu$, $s\mu\mu$, $q\mu\mu$, $g\mu\mu$, $b\tau_\mathrm{h}\tau_e$, $c\tau_\mathrm{h}\tau_e$, $s\tau_\mathrm{h}\tau_e$, $q\tau_\mathrm{h}\tau_e$, $g\tau_\mathrm{h}\tau_e$, $b\tau_\mathrm{h}\tau_\mu$, $c\tau_\mathrm{h}\tau_\mu$, $s\tau_\mathrm{h}\tau_\mu$, $q\tau_\mathrm{h}\tau_\mu$, $g\tau_\mathrm{h}\tau_\mu$, $b\tau_\mathrm{h}\tau_\mathrm{h}$, $c\tau_\mathrm{h}\tau_\mathrm{h}$, $s\tau_\mathrm{h}\tau_\mathrm{h}$, $q\tau_\mathrm{h}\tau_\mathrm{h}$, $g\tau_\mathrm{h}\tau_\mathrm{h}$, $qqqb$, $qqqc$, $qqqs$, $bbcq$, $ccbs$, $ccbq$, $ccsq$, $sscq$, $qqbc$, $qqbs$, $qqcs$, $bcsq$, bcs, bcq, bsq, csq, $bce\nu$, $cse\nu$, $bqe\nu$, $cqe\nu$, $sqe\nu$, $qqe\nu$, $bc\mu\nu$, $cs\mu\nu$, $bq\mu\nu$, $cq\mu\nu$, $sq\mu\nu$, $qq\mu\nu$, $bc\tau_e\nu$, $cs\tau_e\nu$, $bq\tau_e\nu$, $cq\tau_e\nu$, $sq\tau_e\nu$, $qq\tau_e\nu$, $bc\tau_\mu\nu$, $cs\tau_\mu\nu$, $bq\tau_\mu\nu$, $cq\tau_\mu\nu$, $sq\tau_\mu\nu$, $qq\tau_\mu\nu$, $bc\tau_\mathrm{h}\nu$, $cs\tau_\mathrm{h}\nu$, $bq\tau_\mathrm{h}\nu$, $cq\tau_\mathrm{h}\nu$, $sq\tau_\mathrm{h}\nu$, $qq\tau_\mathrm{h}\nu$
QCD jets	161–187	$bbccss$, $bbccs$, $bbcc$, $bbcss$, $bbcs$, bbc, $bbss$, bbs, bb, $bccss$, $bccs$, bcc, $bcss$, bcs, bc, bss, bs, b, $ccss$, ccs, cc, css, cs, c, ss, s, others

troduced above, and a list of the matched particles with their type and kinematic features included.

Appendix B: Supplementary details on trained models

1. Sophon model

The Sophon model adopts the Particle Transformer architecture following Ref. [36] with the fully connected multilayer perception (MLP) extended to a two layers with dimensions of $(512, 188)$. The main body of the Sophon model includes 6 particle attention blocks and 2 class attention blocks, with an embedding dimension of 128, and the number of heads equals 8. The initial particle features are embedded with a 3-layer MLP with $(128, 512, 128)$ nodes, and the pairwise particle features are embedded with a 4-layer elementwise MLP with $(64, 64, 64, 8)$ nodes. The GELU nonlinearity is used throughout the model. The Sophon model includes 2.3 M parameters in total.

The model takes input from all jet constituents, including the kinematic features, particle identification, and impact parameter features. It adopted the scaled kinematics inputs, where features related to the constituent or jet four-momentum are all scaled by a parameter such that the jet p_T after scaling is 500 GeV.

The procedure of sampling-based reweighting from the training dataset to decorrelate the tagger score with jet p_T and m_SD is introduced as follows. The training samples are selected into the training pool with a predefined probability during the on-the-fly data loading process. These probabilities serve as reweighting factors that reweight the two-dimensional histograms bin by bin, constructed by jet p_T and m_SD within the range of $200 < p_\mathrm{T} < 2500\,\mathrm{GeV}$ and $20 < m_\mathrm{SD} < 500\,\mathrm{GeV}$. The target is to yield the same normalized distributions for several specific reweighting classes. The reweighting classes are formed by merging 188 finely classified categories to some extent: classes with only quark flavor differences have been merged, and all 27 QCD jet classes have been merged into one. This results in 30 reweighting classes. In addition, the relative weights of the 30 reweighting classes are properly chosen to weigh the number of samples in the training pool for each classes.

The model is trained with a batch size of 512 with an initial learning rate (LR) of 5×10^{-4}. The full JetClass-II dataset is split into 80/20% for each file to serve as the training/validation set. It is trained over 80 epochs, with each epoch iterating over 10 M samples. The optimizer and the LR scheduler are the same as the ParT training [36]. We use the Lookahead optimizer [74] with $k = 6$ and $\alpha = 0.5$ and the inner optimizer is RAdam [75] with $\beta_1 = 0.95$, $\beta_2 = 0.999$, and $\epsilon = 10^{-5}$. The LR remains constant for the first 70% of the iterations, and then decays exponentially, changing at the beginning of every following epoch, down to 1% of the initial value at the end of the training. A model checkpoint is saved in every epoch, and the checkpoint with the highest accuracy on the validation set is chosen.

2. Sophon model* (42-class)

This model adopted the same Sophon model configuration except that the classification node dimension is modified to 42. It is trained to classify a subset of jet signatures, which covers all the final states from 2-prong resonant jets and the QCD jets. The training dataset then corresponds to the 2-prong resonant and QCD jets, summed up to 40 M jets.

Compared to the Sophon model training, it is trained over 80 epochs, with each epoch iterating 2.5 M samples. The other training configurations are the same as the Sophon model case.

3. ParT model for $X \to bb$ (bs) vs. QCD

This binary classifier with the ParT architecture [36] is used to benchmark the current state-of-the-art discrimination capability between the resonant $X \to bb$ (bs) jets and QCD jets. It adopts the same ParT model configuration of Ref. [36] with two classification nodes. The input features use the original ParT training input, i.e., no scaling of the jet and constituent four-momenta is applied. This is found to have a negligible difference compared to the scaling case when evaluating the $X \to bb$ (bs) vs. QCD tasks.

The training is performed on the labeled $X \to bb$ (bs) jets and the QCD jets. To prevent too few signal $X \to bb$ (bs) jets in the JetClass-II dataset which might limit the model's capabilities, we produce an extended size of the $X \to bb$ (bs) jets with the same configuration to enlarge this dedicated category, amounting to 16 M jets each, reaching a similar size with the QCD jets. Also, a similar sampling-based reweighting is performed to reach the same 2D histogram on jet p_{T} and m_{SD} for the two classes, with their relative class weights set as 1 : 1.

Compared to the Sophon model training, it is trained over 50 epochs, with each epoch iterating 2.5 M samples. A weight decay of 0.01 is adopted to achieve better performance. The other training configurations are the same as above.

4. ParticleNet model for $X \to bb$ (bs) vs. QCD

The binary classifier with the ParticleNet architecture [61] is used to represent the tagging capability achievable in several present CMS analyses. The ParticleNet model adopts the same configuration as in Ref. [61]. It comprises three edge convolution blocks with increased dimensions of $(64, 128, 256)$ with the number of nearest neighbors for edge computing set as 16. The input to ParticleNet is similar to the experiment on JetClass shown in Ref. [36]. The training is performed on the same extended signal $X \to bb$ (bs) jets, and the QCD jets in JetClass-II. The same sampling-based reweighting on the two classes is adopted as introduced above.

The ParticleNet classifier is trained over 50 epochs, with each epoch iterating 2.5 M samples. A batch size of 512 and initial LR of 5×10^{-3} is adopted. No weight decay is applied in the ParticleNet training. The optimizer and the LR scheduler are the same as above.

5. Transfered Sophon model for $X \to bs$ vs. QCD

To transfer the knowledge of the Sophon model to perform $X \to bs$ vs. QCD jet tagging task, latent space features with a dimension of 128 are used as input to train a 2-layer MLP with $(512, 2)$ nodes with ReLU nonlinearity. Parameters of the first layer `Linear(188, 512)` are preloaded from the corresponding layer in the original Sophon model. The training is performed on the same extended $X \to bs$ resonant jet dataset, and the QCD jet from JetClass-II. The same sampling-based reweighting on the two classes is used.

The training is performed in 1 epoch, which iterates only 2.5 M samples to reach a convergence. A batch size of 1024 and a constant LR of 5×10^{-4} is adopted.

6. SALAD classifiers for single-jet resonance search

In the generic search of resonances on the leading jet m_{SD} spectrum, the Simulation Assisted Likelihood-free Anomaly Detection (SALAD) method [10] is used, involving training two classifiers. The classifier is combined with the Sophon's transfer learning concept; hence, the input of the classifier is the dimensional 128 latent features of the leading jet provided by the Sophon model.

The first classifier is trained to distinguish the QCD background and all events (including QCD jets and other SM processes), in the mass sidebands (SBs) $m_{\mathrm{SD}} \in (m_0 - 25,\, m_0 - 15) \cup (m_0 + 15,\, m_0 + 25)$ GeV at a given m_0. The sampling rate from the two classes is controlled by the data loader to yield the same number of events. The classifier network is a 3-layer MLP with $(512, 64, 2)$ nodes and ReLU nonlinearity, where a preloading of parameters of the first layer is also done. The training is performed in 1 epoch to iterate 5 M samples. It takes a batch size of 50 000 and a constant LR of 1×10^{-3}. The score corresponding to the all-event class $w(\mathbf{x})$ predicted by the network is used in the next step. Notably, An ensemble of 100 networks is trained, and the averaged score $\bar{w}(\mathbf{x})$ is used.

The second classifier is trained to distinguish the QCD background and all events in the signal regions (SRs) $m_{\mathrm{SD}} \in (m_0 - 15,\, m_0 + 15)$ GeV. The training uses the same MLP configuration and the preloading of parameters. It also iterate 5 M samples and takes a batch size of 50 000 and a constant LR of 1×10^{-3}. As proposed in the SALAD method, the per-sample loss function is multiplied by $\bar{w}(\mathbf{x})/(1 - \bar{w}(\mathbf{x}))$ if it belongs to the QCD class. Similarly, an ensemble of 100 networks is trained, and an average of scores that corresponding to the all-event node is used as the final discriminant to suppress the backgrounds.

7. IAD classifiers for dijet resonance search

In the dijet resonance search experiment, we use $40\,\mathrm{fb}^{-1}$ SM data to train the classifier and $40\,\mathrm{fb}^{-1}$ data for test. This amounts to around 330 k SM events in SR $m_{JJ} \in (2500,\, 3100)$ GeV. The former is further split to 80/20% for training/validation.

We train two kinds of classifiers under the idealized anomaly detection (IAD) [16] scheme that can represent the sensitivity boundary of the weakly-supervised methods to detect anomalous resonance. The classifiers differ by their input features. In the IAD scheme, the classifier is trained to distinguish the backgrounds and the data events in SR, where the backgrounds correspond to all SM processes, and the data additionally includes the injected signal events. The experiment evaluates the classifier performance as a function of the injected signal N_s in SR. Practically, for each N_s, we divide the 330 k SM events into two parts and inject $N_s/2$ signal events in one part. The classifier is trained to distinguish the two parts. For each N_s, the experiments are repeated 100 times with different N_s signals chosen for the study. This allows us to calculate the mean and standard deviation of the classifier performance under each N_s.

The first classifier is an application of Sophon's transfer learning. We use a 3-layer MLP with $(512, 64, 2)$ nodes and ReLU nonlinearity. Here, the input is the two latent feature vectors with dimension 128 from the two jets. To proceed with the input with two vectors to the network, the vectors are first passed through the first and second layers of the MLP, respectively. The resulting outputs are then summed and passed through the third layer. An ensemble of 100 networks is trained, and their averaged score corresponding to the data node is used as the discriminant.

The second classifier applies to the high-level input from the dijet system following Ref. [8]. For each jet, we consider the jet m_{SD}, the number of constituent $N_{\mathrm{const.}}$, the N-subjettiness variable [73] τ_1, τ_2, τ_3, and τ_4, taking the logarithm of each variable and standardize it within the range between -2 and 2. The input takes two dimension-6 vectors for both jets. We use a 4-layer MLP with $(512, 128, 128, 2)$ nodes and ReLU nonlinearity. Similarly, the vectors are first passed through the first and second layers of the MLP, respectively, and then the outputs are summed and passed through the third and fourth layers. An ensemble of 100 networks is trained, and their averaged score corresponding to the data node is used as the discriminant.