Jet quenching identification via supervised learning in simulated heavy-ion collisions

Leonardo Lima da Silva [*] and Marcelo Gameiro Munhoz [†]
Instituto de Física, Universidade de São Paulo, Brazil

(Received 24 April 2026; accepted 9 June 2026; published 8 July 2026)

Jet modification in heavy-ion collisions provides microscopic access to the properties of the quark-gluon plasma. However, conventional approaches based on traditional global observables, such as R_{AA}, capture limited information about the complex dynamics of parton-medium interactions during hard scatterings. In this work, we apply sequential machine-learning architectures to the jet declustering history tree, achieving improved classification performance compared with static models that learn only from a single stage of the jet evolution. We find that models trained on different medium implementations exhibit meaningful performance modification under cross-domain validation, indicating that machine learning is sensitive to implementation-specific features that traditional observables may not resolve. These results suggest new opportunities for using machine learning as an analysis tool to overcome some of the limitations of traditional jet-modification studies.

DOI: 10.1103/nhl8-kknn

I. INTRODUCTION

A high-energy collision of two nuclei, whether in a fixed-target or a collider setup, produces numerous new particles. This output exceeds what one might expect from individual nucleon-nucleon collisions. It has been hypothesized that at high temperatures and densities, hadronic matter transitions to a state of free quarks and gluons, known as the quark-gluon plasma (QGP) [1,2].

Jets provide direct experimental access to QGP properties through their modification during medium propagation [3]. Traditionally, these effects are observed using a collection of jets from heavy-ion collisions. One example is the nuclear modification factor R_{AA} that measures the suppression of particle production in the presence of a dense medium [4].

Early studies of inclusive suppression observables showed that the extracted energy-loss coefficient depends on the hydrodynamic medium used to describe the QGP evolution [5]. In the present work, we revisit this medium-dependence from a jet-by-jet perspective using supervised machine-learning classifiers.

Jets are typically defined as a succession of partonic emissions initiated by the scattering of these particles at high energies. Due to confinement, these partons undergo a fragmentation process, forming a collimated set of partons, which subsequently transform into hadrons through a hadronization process and can be experimentally identified.

[*]Contact author: llsilva@if.usp.br
[†]Contact author: munhoz@if.usp.br

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP[3].

The interaction of partons with the medium occurs through energy-loss processes, such as elastic collisions or gluon emission, and is known as jet quenching [3]. As mentioned, the comparison between a collection of jets produced in heavy-ion collisions and those produced in proton-proton collisions is a common way of studying the QCD medium. However, one can wonder about the impact of studying the influence of the QGP using a jet-by-jet approach.

Machine learning is particularly well-suited to learning statistical patterns from individual jets and making predictions based on their corresponding labels. Thus, it is a natural candidate as a tool for studying jet quenching on a jet-by-jet basis.

In recent years, the field of high-energy physics (HEP) has experienced a significant resurgence of interest in the application of Machine Learning (ML) [6]. Previous studies have explored deep-learning strategies for jet-quenching classification, showing that neural-network architectures can learn enough information from jet-level representation to discriminate between medium and vacuum-like jets [7]. In addition, comprehensive reviews have highlighted the growing applications of ML techniques specifically for jet quenching studies [8].

Other approaches have explored energy-flow-network-based representations for jet-quenching classification, combining constituent-level learning with physics-motivated observables [9].

Recent experimental work at ALICE has demonstrated the practical application of ML techniques for differential jet quenching measurements [10]. It is essential to distinguish between jets generated within this medium and those originating from vacuum conditions to interpret the dynamics involved in these collisions accurately.

In this paper, we apply various supervised machine-learning algorithms in the task of jet quenching classification using two hydrodynamical models to study the effect of the medium on quenching effects. We begin with a simple medium description by JEWEL [11,12], which we refer to

as Default. Then, we use JEWEL coupled with a more realistic environment named viscous Ultrarelativistic Smoothed Particle hydrodynamics or simply v-USPhydro [13]. We elaborate more about these two medium frameworks in the following sections. For jet classification, we use five different ML models: Random Forest [14], Multilayer Perceptrons (MLP) [15], Long Short-Term Memory (LSTM) [16], LSTM coupled with an attention mechanism [17], and Transformers [18].

This work is systematically organized as follows: Section II establishes the simulation framework within JEWEL and characterizes the jet substructure observables utilized as machine-learning input features. Section III provides a brief description of the neural network architectures employed in our classification analysis. Section IV presents our results, distinguishing between performance achieved using static (nonsequential) jet substructure features versus sequential jet grooming history as input variables.

Our analysis shows that models trained on sequential jet data significantly outperform those trained on static data. This suggests that understanding jet modification as a temporal process is more effective. Additionally, the models exhibit considerable sensitivity to the details of medium implementation, indicating that machine learning can differentiate between medium frameworks that yield similar results in traditional observables, such as the R_{AA}.

II. SIMULATION SETUP AND DATA PREPARATION

Jet Evolution With Energy Loss, commonly known as JEWEL [11,12], is a Monte Carlo framework for simulating QCD jet evolution in heavy-ion collisions. It models the interplay between QCD radiation and parton-medium interactions using a fully microscopic perturbative approach with minimal assumptions.

JEWEL incorporates PYTHIA [19,20] as the underlying framework for event generation, leveraging its models and formalisms to simulate processes such as parton showers, hadronization, and decays. This allows JEWEL to complement its detailed description of medium-induced effects on jets, including phenomena such as energy loss and radiative coherence.

JEWEL models the medium created in heavy-ion collisions in a simplified manner. The medium consists of a static cloud of thermal partons with momentum distributions following a thermal profile and density evolving according to the Bjorken expansion [12], which assumes longitudinal expansion only. Jet-medium interactions are implemented via elastic scattering and medium-induced radiation, allowing the simulation of parton energy loss while traversing the quark-gluon plasma (QGP). The initial spatial distribution of the medium is determined using Glauber's model [21]. Although JEWEL does not include a realistic hydrodynamic evolution of the QGP, it allows control over key parameters such as the initial temperature, formation time, and the presence of recoiling partons. The energy loss is computed through perturbative processes, combining elastic scatterings and medium-induced gluon radiation. The latter follows the Baier–Dokshitzer–Mueller–Peigné–Schiff–Zakharov (Bjorken-Z) formalism, which incorporates the Landau–Pomeranchuk–Migdal (LPM) effect, and accounts for multiple scatterings within the medium [22,23]. For more details, one can refer to Ref. [24].

To overcome the limitations of this simplified medium, JEWEL can be coupled to more realistic hydrodynamic models. In this work, we employ the v-USPhydro description of hydrodynamics, which solves the relativistic hydrodynamic equations using the Smoothed Particle Hydrodynamics (SPH) method [25]. This model simulates the time evolution of the QGP in $2+1$ dimensions, including viscous effects and pressure gradients, with initial conditions provided by the Reduced Thickness Event-by-event Nuclear Topology ($\mathrm{T_R ENTo}$) model [26]. The hydrodynamic evolution uses an equation of state based on lattice-QCD results at high temperatures, smoothly matched to a hadron-resonance-gas description at lower temperatures. In the present setup, we use the WB21/PDG16+ parametrization [27,28], following the v-USPhydro configuration adopted in previous phenomenological studies [29].

By embedding JEWEL jets into this dynamically evolving medium, one obtains a medium description that includes transverse expansion, local flow, viscous hydrodynamic evolution, and event-by-event fluctuations of the QGP. The JEWEL+v-USPhydro implementation used in this work follows from the work in Ref. [30].

We employ these two complementary medium implementations to probe sensitivity to QGP evolution details. In summary, the simplified thermal description of JEWEL enables controlled studies of basic jet-medium interactions, whereas v-USPhydro's realistic hydrodynamic evolution includes anisotropic expansion and viscous effects characteristic of actual heavy-ion collisions.

The results presented in this paper were obtained using JEWEL 2.4.0, which incorporates a slightly modified version of PYTHIA 6.4.25. The parameters chosen for the simulations follow those used in Ref. [31], except that, in this work, we adopt a broader transverse momentum range of $p_T \in [40,\ 450]$ GeV to investigate the impact of p_T on the quenching effect. All events simulated for this work were performed with recoils on, which means that the medium particles that undergo interactions with the jets are included in the jet reconstruction. The treatment of recoils and background subtraction in JEWEL is known to affect jet substructure observables and has been discussed in detail in Ref. [32]. We do not explore the effects of recoils on observables in this paper. For more information, refer to Ref. [4].

The observables analyzed in this study were computed using the code provided in Ref. [31], with a few modifications to suit our specific goals, and we refer to the implementation in the section Data Availability. We categorize the observables into two main groups, static and sequential. The static, or nonsequential, results are computed from the first splitting that satisfies the Soft Drop condition [33], representing a snapshot of the declustered jet. In contrast, sequential observables are extracted throughout the entire declustering process, associating a value with each step of the jet evolution and thus capturing its dynamical development. We elaborate more about these observables in Sec. III.

Finally, we divide the resulting dataset into several intervals of transverse momentum to study the quenching effect across different regions of the p_T spectrum. In this work, we study three intervals of transverse momentum, namely, [40, 60], [80, 250], and [200, 400] GeV.

We deliberately do not include thermal background contributions in our analysis to enable direct comparison between *pp* and PbPb collision systems, isolating jet modification signatures induced by the medium without thermal complexity. Additionally, our comparative methodology includes a systematic evaluation of how the two different medium descriptions affect machine-learning classification performance by using in-domain and cross-domain evaluations. The former compares the results of models trained and tested within the same medium, and the latter compares models trained in one medium and tested in the other.

III. OBSERVABLES

Particles are clustered into jets via the anti-k_T algorithm [34] with a distance parameter $R = 0.4$. After the initial jet reconstruction, we apply the Cambridge-Aachen algorithm [35] to recluster the jet constituents, building an angular-ordered tree structure that enables a step-by-step analysis of the internal splittings of the jet. The Soft Drop declustering [33] procedure is applied, continuously removing soft branches and keeping the splittings that satisfy the soft drop condition

$$z_g \equiv \frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} \geqslant z_{\text{cut}} \left(\frac{\Delta R}{R_0}\right)^{\beta}, \tag{1}$$

where p_{T1} and p_{T2} are the transverse momenta of the two separated subjets, ΔR is the angular distance between them, and R_0 is the jet radius parameter, fixed in this work to $R_0 = 0.4$, as previously mentioned. The parameters z_{cut} and β set the minimum allowed momentum fraction between the two branches in a splitting, and the angular dependence of the grooming condition, respectively.

In this study, we set the free parameters to $z_{\text{cut}} = 0.1$ [33] and $\beta = 0$ [36]. This choice is motivated by the fact that this configuration provides a theoretically clean grooming procedure. In particular, $\beta = 0$ removes the angular dependence from the grooming condition, making the algorithm sensitive only to the momentum sharing between subjets, while $z_{\text{cut}} = 0.1$ suppresses soft contaminating radiation without significantly affecting the hard core of the jet.

The Soft Drop algorithm works by removing the softer branch in each splitting until it encounters a hard splitting that satisfies the condition given by Eq. (1). Then, the declustering proceeds along the harder subjet, as illustrated in Fig. 1.

In each splitting of the jet, we can compute substructure variables, such as the momentum fraction z, the angular distance ΔR, perpendicular momentum $k_\perp$, and invariant mass m_{inv}. These variables were chosen since they encapsulate key features of the jets, such as the shared energy, the opening angle between two splittings, the quantity of transverse energy from one branch to another, and the combination of the four-momenta of the two splittings. They can be calculated from

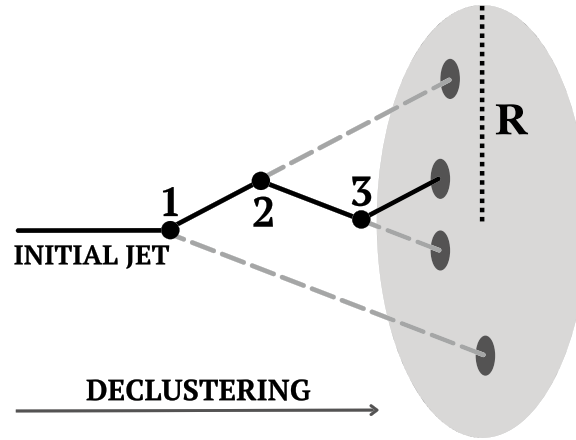


FIG. 1. Schematic representation of the jet declustering history. Starting from the initial jet, the declustering procedure sequentially unfolds the branching structure through successive steps, labeled 1, 2, and 3, until reaching the final constituents inside the jet cone of radius R. The solid black line highlights one possible declustering path, while the dashed gray lines indicate discarded branches of the jet tree by the Soft Drop procedure. Image inspired by Ref. [37].

the corresponding pair (i, j) of declustered subjets,

$$\begin{aligned} z &= \frac{\min(p_{T,i},\ p_{T,j})}{p_{T,i} + p_{T,j}}, \\ \Delta R &= \sqrt{(\phi_i - \phi_j)^2 + (\eta_i - \eta_j)^2}, \\ k_\perp &= \min(p_{T,i},\ p_{T,j})\Delta R, \\ m_{\text{inv}} &= \sqrt{(E_i + E_j)^2 - (\mathbf{p}_i + \mathbf{p}_j)^2}, \end{aligned} \tag{2}$$

where i, j denotes the subjets at declustering step t. We store these variables in a vector x_t, so we have

$$\mathbf{x}_t = [z,\ \Delta R,\ k_\perp,\ m_{\text{inv}}] \tag{3}$$

for each step t. Therefore, each jet can be represented as a sequence $\mathbf{x}$ of vectors x_t that contains the information of its dynamical evolution. For a jet with N declustering steps, we write

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_N]. \tag{4}$$

For each jet, a vector $\mathbf{x}$ in Eq. (4) can be associated with the declustering history, and we refer to this as sequential. However, one can associate each jet only to the first grooming step $\mathbf{x}_1$, which we denote as nonsequential or static data.

The vector presented in Eq. (4) includes important information about the internal structure of jets, highlighting their interaction with the medium. In this paper, we investigate how this information evolves during the jet's progression and examine the influence of different media, the simplified one used in JEWEL and the more elaborated one given by v-USPhydro. Our objective is to understand how these substructures help in classifying quenching phenomena. For this purpose, we utilize Long Short-Term Memory (LSTM) and Transformer neural networks, as they are well-suited for working with sequential data. This sequential approach is strongly inspired by the findings in Ref. [38].

We also investigate the performance of models trained on static representations of the jets, which are used as input to machine-learning models such as Random Forests and Multilayer Perceptrons, since these models do not explicitly capture the temporal dynamics of the jet evolution. As such, they serve as baselines for determining the discriminative power of global jet substructure patterns.

IV. MACHINE-LEARNING MODELS

The investigation in this paper employs a strategically structured set of machine-learning methodologies, encompassing both traditional ensemble techniques and sophisticated sequential architectures, to systematically probe the discriminative signatures of jet quenching phenomena.

We present an architectural progression from nonsequential (Random Forest and MLP) to sequential (LSTM, LSTM+Attention, and Transformer) models that reflects increasing sophistication in temporal feature extraction capabilities. Static (nonsequential) models evaluate discrete snapshots of jet substructure, while sequential architectures integrate complete evolutionary histories. This methodological hierarchy enables a systematic evaluation of whether temporal dynamics provide essential discriminative information beyond static observables.

We also employ cross-domain evaluation to ensure generalization capabilities across different medium descriptions, thereby validating the physical relevance of learned feature representations rather than specific artifacts of simulations.

The hyperparameters of each model are specified in the appendix.

A. Random forest

Random Decision Forests [14], also known as Random Forests, are an ensemble learning method [39,40], generally used for classification and regression problems. To make its predictions, this model combines the outputs of multiple decision trees [41,42] to reach a single result.

The Random Forest algorithm is an extension of the bagging method, as it combines both bootstrap aggregation (bagging) [43] and random feature selection [44] to grow an ensemble of uncorrelated forests of decision trees, improving generalization performance.

This model was selected in this study as a baseline for comparison with subsequent neural network models, primarily because of its ability to capture global information about the features. The model's capacity to capture global feature interactions without temporal dependencies makes it particularly suited for evaluating the discriminative power of static jet substructure observables. Random Forest has been employed in HEP for resolving combinatorial ambiguities in dilepton $\bar{t}t$ event topologies [45].

B. Multilayer perceptron

Multilayer Perceptrons (MLPs) are a type of artificial neural network [46], mathematically modeled to imitate the behavior of the human brain, consisting of multiple layers of neurons [15]. A traditional MLP neuron utilizes nonlinear activation functions [47], such as ReLU [48], which allows the model to capture more complex data patterns. This model has previously been employed in high-energy physics for real-time jet tagging [49].

The performance of MLPs establishes a critical benchmark for assessing whether sequential modeling provides substantive advantages over static feature representations. Its strength lies in its capacity to model global interactions between input features without the need to rely on specific assumptions such as locality or sequential order. In this context, MLPs serve as a middle ground between traditional ensemble models and more advanced architectures, such as recurrent and attention-based networks.

C. Long short-term memory

The Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) [50] designed to model sequential data and capture long-range dependencies over time. This model introduces a memory cell structure regulated by gating mechanisms (input, forget, and output gates) to enable selective information retention across extended sequences. These models have been successfully applied to a wide range of Natural Language Processing (NLP) tasks, such as machine translation [16,51] and constituency parsing [52]. They have also been employed in high-energy physics for top tagging at the CERN Large Hadron Collider (LHC) [53], and for jet quenching identification [38], which inspired part of this work.

The forget gate controls the information that is discarded from the cell state, the input gate determines what must be stored, and the output gate decides what information is propagated to the next hidden state. These operations allow LSTMs to preserve relevant temporal information over sequences, making them an effective classifier for time-series tasks. This capability proves essential for modeling the progressive modifications occurring throughout jet evolution in dense media.

Attention-enhanced LSTM

The attention mechanism [17,54] has transformed the approach to sequence modeling. It allows networks to focus on different parts of the input sequence dynamically when making predictions. Unlike traditional RNNs or LSTMs that fit everything into a single context vector, the attention mechanism takes a more nuanced approach by calculating a weighted sum of all hidden states. The weights are learned based on the importance of each time step for the current prediction.

The integration of an attention mechanism enhances the capability of the model to focus on relevant information within the input sequence after it has been processed. Specifically, for each output prediction generated by the LSTM, the attention layer assigns varying weights to each hidden state, pondering their contribution based on their relevance to the current prediction. This selective focus allows the model to identify and leverage critical substructures associated with the evolution of jets, resulting in greater interpretability and performance metrics.

Long Short-Term Memory coupled with attention has been used in physics, for instance, in accelerated prediction of charge density at the onset condition of dc corona discharge [55]. To the best of our knowledge, studies explicitly exploring attention-enhanced architectures for quenched-jet classification are still limited.

A full mathematical description of this mechanism is beyond the scope of this work. For more details, see Ref. [56].

D. Transformer

The Transformer architecture, introduced by Vaswani *et al.* [18], is a deep-learning model designed to manage sequential data. This architecture utilizes an attention mechanism to effectively compute pairwise interactions between elements in an input sequence. This architecture has been applied in jet physics for tagging [57] and for quark-gluon discrimination and top tagging with a dual attention transformer [58], and, more recently, for the identification of jet quenching [59].

Transformer Architecture represents the most sophisticated approach, and it overcomes the sequential processing limitations of recurrent models, enabling parallel computation of complex feature relationships across the entire jet substructure evolution. The multihead attention framework [18] facilitates the capture of both local splitting characteristics and global evolutionary patterns.

Whereas LSTM with attention employs sequential processing by examining jet substructure elements chronologically, where a time step depends on prior states, the Transformer architecture implements parallel self-attention, allowing simultaneous pairwise interactions across all grooming steps.

V. RESULTS

In this investigation, we use five different performance metrics to ensure rigorous quantitative evaluation across different operational conditions. Accuracy measures the proportion of correctly classified jets relative to the total sample size. Area Under the Curve (AUC) evaluates the discriminative capacity of the model across all classification thresholds, which is particularly valuable for imbalanced datasets. Precision quantifies false positive rates by measuring the specificity of positive predictions and calculating the fraction of correctly identified quenched jets among all jets classified as quenched. Recall evaluates the sensitivity of quenching detection, determining the proportion of actual quenched jets successfully identified by the model, thus measuring false negative rates. F1 Score provides a harmonized metric balancing precision and recall through their harmonic mean, offering a consolidated performance indicator that accounts for both classification errors simultaneously. Please refer to [60] for more information about supervised models' evaluation metrics.

Hyperparameter optimization [61,62] was conducted by training several models with different parameter combinations on the training dataset, followed by an evaluation of their performance using the validation set. The configuration that achieved the best results on the validation set was selected for further use. Once the model was properly tuned, a final evaluation was performed with the test dataset, which remained completely isolated throughout the training and hyperparameter optimization processes. This approach ensures an unbiased evaluation of the predictive capabilities of the model. It also minimizes overfitting and selection bias, as it prevents the model from memorizing patterns that are specific to the training or validation data.

All datasets (training, validation, and test) were generated under the same conditions of JEWEL Default and JEWEL+v-USPhydro and remained strictly separate for each scenario. As discussed in Sec. III, our dataset can be categorized into two main types: static and sequential. The tree-based model, Random Forest, and the fully connected neural network, Multilayer Perceptron (MLP), are trained and evaluated on the static Soft Drop variables computed from the Monte Carlo simulations. Whereas the dynamical evolution of the observables is fed to the recurrent neural network, the LSTM, and to the attention-based model, the Transformer.

To compute uncertainties associated with the performance metrics, we employ the nonparametric bootstrap technique [63,64]. It is widely used in machine learning and statistical modeling for quantifying uncertainty. For more details, see Refs. [65,66].

We use machine-learning methods in an individual jet analysis, but we must address some methodological limitations. First of all, the jet reconstruction is not performed on top of a fully simulated thermal heavy-ion background, as would be the case in realistic PbPb events. Instead, our analysis relies on the recoil partons generated by JEWEL, which encode the medium response associated with the jet-medium interaction, but do not reproduce the full soft environment of an experimental heavy-ion collision.

This distinction is important because, in realistic high-multiplicity environments, jet measurements are affected not only by medium-induced modifications of the hard shower, but also by contamination from uncorrelated soft particles, local background fluctuations, and subtraction ambiguities. Such effects are known to degrade the precision of subjet and grooming observables, especially on a jet-by-jet basis, where the performance of subtraction procedures is substantially less robust than for ensemble-averaged quantities [67–70]. Since our setup does not include a fully thermal PbPb underlying event, the present results should not be understood as a full assessment of experimental background sensitivity, but rather as a controlled study of quenched-jet classification in a simplified medium-response scenario.

The recent work presented in Ref. [71] has investigated the robustness of machine-learning-based quenched-jet classification under underlying-event contamination, highlighting the importance of controlled comparisons between *pp* and PbPb-like samples.

A second limitation follows from the intrinsic ambiguity of energy-loss measurements. Jets selected with a given reconstructed transverse momentum do not correspond to a unique physical history because they may originate either from much harder partons that lost a substantial fraction of their energy outside the jet cone, or from moderately energetic partons that experienced only mild modification [72]. This selection bias becomes particularly relevant in steeply falling

TABLE I. Performance metrics for the Random Forest classifier across p_T ranges and media.

p_T [GeV]	Metric	Default	v-USPhydro
	Accuracy	0.7683 (0.0006)	0.7831 (0.0006)
	AUC	0.8588 (0.0005)	0.8737 (0.0005)
40–60	Precision	0.5295 (0.0011)	0.6318 (0.0010)
	Recall	0.8256 (0.0010)	0.8322 (0.0008)
	F1 score	0.6452 (0.0010)	0.7183 (0.0008)
	Accuracy	0.7610 (0.0004)	0.8065 (0.0003)
	AUC	0.8601 (0.0003)	0.8866 (0.0003)
80–250	Precision	0.5105 (0.0007)	0.6646 (0.0006)
	Recall	0.8171 (0.0006)	0.8272 (0.0005)
	F1 score	0.6284 (0.0006)	0.7370 (0.0005)
	Accuracy	0.7776 (0.0006)	0.8173 (0.0005)
	AUC	0.8636 (0.0006)	0.8896 (0.0004)
200–400	Precision	0.5253 (0.0011)	0.6722 (0.0010)
	Recall	0.7886 (0.0011)	0.8285 (0.0008)
	F1 score	0.6306 (0.0010)	0.7422 (0.0007)

TABLE II. Performance metrics for the MLP classifier across p_T ranges and scenarios.

p_T [GeV]	Metric	Default	v-USPhydro
	Accuracy	0.7344 (0.0006)	0.7095 (0.0006)
	AUC	0.8410 (0.0006)	0.7969 (0.0006)
40–60	Precision	0.4881 (0.0010)	0.5452 (0.0010)
	Recall	0.8342 (0.0010)	0.7586 (0.0009)
	F1 score	0.6159 (0.0009)	0.6344 (0.0008)
	Accuracy	0.7286 (0.0004)	0.7482 (0.0003)
	AUC	0.8387 (0.0004)	0.8113 (0.0004)
80–250	Precision	0.4717 (0.0006)	0.6017 (0.0006)
	Recall	0.8124 (0.0007)	0.6850 (0.0007)
	F1 score	0.5969 (0.0006)	0.6407 (0.0005)
	Accuracy	0.7776 (0.0006)	0.7607 (0.0005)
	AUC	0.8438 (0.0006)	0.8129 (0.0006)
200–400	Precision	0.5280 (0.0012)	0.6102 (0.0010)
	Recall	0.7179 (0.0012)	0.6825 (0.0011)
	F1 score	0.6085 (0.0010)	0.6443 (0.0009)

jet spectra, where strongly quenched jets are statistically suppressed [73,74]. As a consequence, even when machine-learning models identify differences between classes, part of the discrimination power may still reflect this kinematic and selection bias rather than a clean isolation of medium-induced substructure effects. In this sense, the classification task should be interpreted with caution, since low-p_T jets may resemble quenched jets not only because of genuine in-medium modification, but also because of the interplay between recoil contamination, jet energy loss, and the biased population selected after reconstruction [38,70].

In light of this issue, a related recent study used Transformer architectures on low-level constituent information to quantify the fraction of vacuum-like jets in heavy-ion samples, including medium response and underlying-event contamination [59].

We fully acknowledge these limitations and interpret results cautiously. Our objective is not absolute jet-level classification, but exploring whether statistical patterns consistent with medium-induced modifications can be systematically learned and interpreted within current simulation and analysis method constraints. This approach recognizes the inherent methodological boundaries while extracting physical insights from data.

For machine-learning applications in jet quenching under realistic experimental conditions, including thermal background, detector effects, and pileup, see Ref. [75].

A. Static data models

1. Performance metrics

Tables I and II show the evaluation metrics measured for the Random Forest (RF) and MLP, respectively. The RF models exhibit a consistent performance across all p_T intervals and for both the Default and v-USPhydro datasets. In particular, AUC scores remain above 0.85 in all cases, with a precision of around 0.5, indicating significant false positive classification.

Although Multilayer Perceptrons (MLPs) are often seen as more flexible and capable due to their ability to represent complex, nonlinear relationships, the findings indicate that the Random Forest model may actually perform better in this context. The MLP achieves slightly lower performance, with AUC values around 0.80 to 0.84. This performance hierarchy suggests that ensemble tree-based methods more effectively capture the jet substructure signatures relevant for quenching classification within the static feature space.

A critical limitation emerges in the precision-recall characteristics of both architectures. While recall values remain consistently high (0.75–0.85), precision scores are substantially lower (0.45–0.7), indicating systematic over-prediction of jet-quenching events. This precision-recall imbalance suggests that static substructure observables contain overlapping signatures between quenched and unquenched jets, preventing clear discrimination based solely on static jet features.

From Tables I and II, one can notice that the v-USPhydro medium consistently yields enhanced classification metrics compared with the Default JEWEL configuration. For the Random Forest models, precision improves from 0.51–0.53 in the Default setup to 0.63–0.67 in v-USPhydro across all p_T intervals. This improvement indicates that v-USPhydro generates more pronounced substructure modifications during jet-medium interactions.

2. Cross-domain validation

To further examine the reliability of these models and eliminate concerns about the dependence on superficial patterns, we conducted cross-domain evaluations (see Tables III and IV). We trained models on jets originating in one medium and tested them on jets from the other medium, i.e., the classifier trained on the JEWEL Default results is evaluated on data from JEWEL + v-USPhydro and vice versa.

Cross-domain validation reveals significant performance degradation when models trained on one medium are evaluated on jets from the other medium description. Random

TABLE III. Random Forest cross-domain performance. Models trained on Default and tested on v-USPhydro, and vice versa, for different p_T intervals.

p_T [GeV]	Metric	Def →v-USP	v-USP →Def
40–60	Accuracy	0.7054 (0.0006)	0.7571 (0.0006)
	AUC	0.7593 (0.0006)	0.8281 (0.0006)
	Precision	0.5504 (0.0011)	0.5165 (0.0012)
	Recall	0.6186 (0.0011)	0.7523 (0.0011)
	F1 score	0.5825 (0.0009)	0.6125 (0.0010)
80–250	Accuracy	0.7263 (0.0004)	0.7766 (0.0004)
	AUC	0.7889 (0.0004)	0.8393 (0.0004)
	Precision	0.5676 (0.0006)	0.5361 (0.0008)
	Recall	0.6929 (0.0006)	0.7160 (0.0008)
	F1 score	0.6240 (0.0005)	0.6131 (0.0007)
200–400	Accuracy	0.7421 (0.0005)	0.7880 (0.0006)
	AUC	0.7958 (0.0006)	0.8424 (0.0006)
	Precision	0.5810 (0.0010)	0.5457 (0.0012)
	Recall	0.6733 (0.0010)	0.7123 (0.0012)
	F1 score	0.6237 (0.0008)	0.6180 (0.0010)

Forest AUC scores decrease to the 0.76–0.84 range, while MLP performance drops to 0.73–0.82. Notably, there is an asymmetric generalization pattern that emerges, since models trained on v-USPhydro and tested on Default configurations (v-USPhydro → Default) outperform the reverse direction (Default → v-USPhydro), as evinced by the increase in accuracy, AUC, and recall scores. However, precision is still about 50%, so the models continue predicting a significant amount of false positives.

This asymmetry suggests that v-USPhydro creates more generalizable jet modification patterns that partially transfer to Default medium descriptions. Conversely, models trained on Default fail to capture the modification signatures present in v-USPhydro, indicating that the latter incorporates additional outcomes not present in the former.

TABLE IV. MLP cross-domain performance. Models trained on Default and tested on v-USPhydro, and vice versa, for different p_T intervals.

p_T [GeV]	Metric	Def →v-USP	v-USP →Def
40–60	Accuracy	0.6512 (0.0006)	0.7172 (0.0006)
	AUC	0.7311 (0.0007)	0.8178 (0.0006)
	Precision	0.4784 (0.0011)	0.4687 (0.0010)
	Recall	0.5525 (0.0011)	0.8110 (0.0011)
	F1 score	0.5128 (0.0010)	0.5941 (0.0009)
80–250	Accuracy	0.7034 (0.0003)	0.7521 (0.0004)
	AUC	0.7732 (0.0004)	0.8188 (0.0004)
	Precision	0.5360 (0.0006)	0.4991 (0.0007)
	Recall	0.7083 (0.0006)	0.6703 (0.0008)
	F1 score	0.6102 (0.0005)	0.5721 (0.0007)
200–400	Accuracy	0.7399 (0.0005)	0.7710 (0.0006)
	AUC	0.7754 (0.0006)	0.8271 (0.0006)
	Precision	0.5856 (0.0011)	0.5183 (0.0012)
	Recall	0.6180 (0.0011)	0.6884 (0.0012)
	F1 score	0.6014 (0.0009)	0.5913 (0.0010)

The analysis of false positives (FPs) across media and cross-domain scenarios shows that in the in-domain context, the Default medium demonstrates a consistently higher rate of false positives than the v-USPhydro medium, as evinced by its lower-precision values (0.47–0.53 compared with 0.54–0.67).

The situation becomes even more pronounced in a cross-domain setting. When models trained on v-USPhydro are tested on Default data (v-USPhydro → Default), there is a notable increase in recall, but this comes at the cost of precision, resulting in a higher occurrence of false positives. This observation indicates that models trained in the more realistic hydrodynamical medium may tend to overinterpret structures in the Default samples as indications of quenching.

Conversely, when training is performed on the Default medium and testing on v-USPhydro (Default → v-USPhydro), the models tend to exhibit higher precision but lower recall. This behavior indicates a more conservative classification regime, in which quenched jets are identified less often, but with greater confidence. Such a pattern suggests that the v-USPhydro medium imprints a broader and more diverse range of modifications on the final observables. As a consequence, a model trained exclusively on the Default scenario may learn a decision boundary that is too restrictive to capture the full variety of quenched jets present in the v-USPhydro case, leading to missed positive instances and therefore reduced recall.

This trade-off between precision and recall underscores a fundamental disparity in the characteristics of the two media. Training on v-USPhydro enhances sensitivity to subtle structures, which can lead to a greater rate of false positives in the v-USPhydro → Default context. Conversely, training on Default confines the classifier to more robust signatures, thereby elevating precision at the risk of overlooking numerous true quenched jets in the hydrodynamical context. These findings emphasize that the choice of training medium plays a critical role in determining the equilibrium between overestimation and underestimation of quenching phenomena, reinforcing the necessity for models capable of effectively disentangling universal physical features from model-specific artifacts.

Another important trend that is observed concerns the recall values across different transverse momentum (p_T) ranges in the MLP model. In the in-domain setting, recall is consistently lower for the v-USPhydro model compared with the Default model, particularly at lower p_T values. This discrepancy suggests that when both models are trained and tested within the hydrodynamical medium, the v-USPhydro model fails to detect a larger fraction of quenched jets. This phenomenon aligns with the expectation that the quenching effects represented in v-USPhydro are more subtle and continuous, making them inherently more challenging for models to identify. In contrast, the Default medium tends to induce stronger and more discrete modifications to the jets, which facilitates easier recognition by the classifiers.

Interestingly, as p_T increases, the recall values for both media begin to converge, with the differences becoming

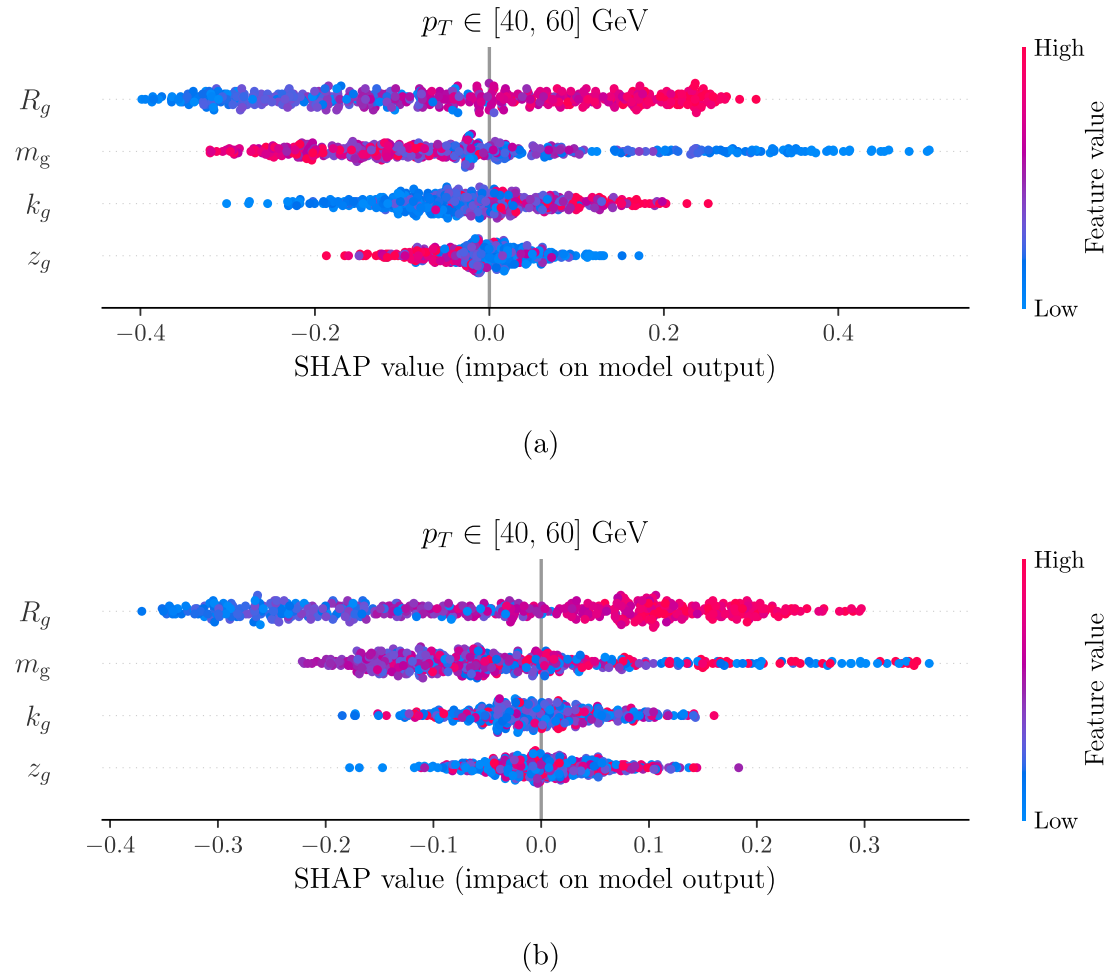


FIG. 2. SHAP values for $40 < p_T < 60$ GeV. R_g and m_g are the most important features in both media. (a) Default and (b) v-USPhydro.

minimal within the 200–400 GeV range. This observation implies that, at higher energies, the quenching signatures that the models capture tend to become more uniform, displaying less sensitivity to the specific description of the underlying medium.

Despite inferior in-domain performance, MLPs demonstrate superior cross-domain generalization compared with Random Forest models. This is evinced by a smaller performance gap when evaluating across different domains for MLPs. Implementing more complex neural networks for sequential data will clarify if the better performance of the MLP in cross-domain scenarios reflects a general advantage over tree-based models, as the Random Forest, in capturing medium dependence.

Overall, these results imply that relying solely on static substructure features for jet quenching classification yields limited discriminative capability, as evinced by high false positive rates across different model architectures and medium configurations. The observed dependence on medium, along with poor cross-domain generalization, underscores the need for more advanced feature representations to effectively encapsulate the physics involved in jet modifications during heavy-ion collisions.

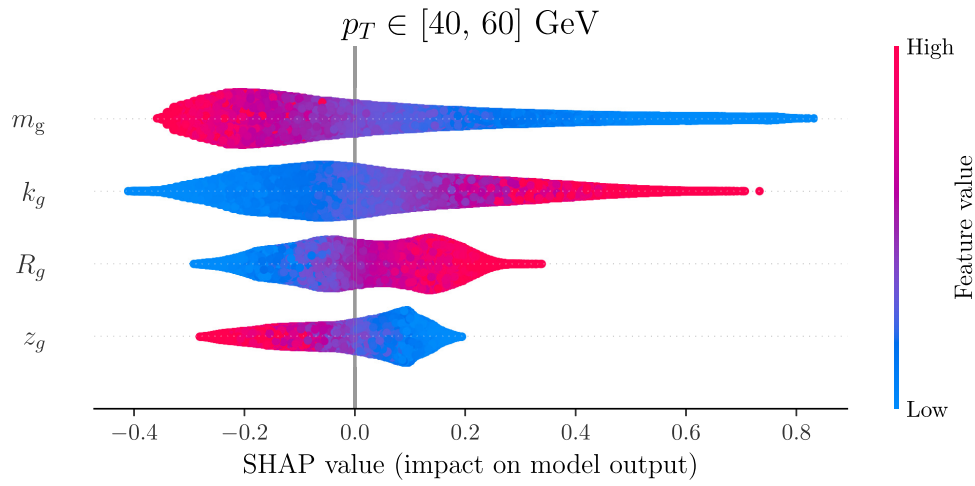
3. Importance analysis

To explore which input features most influence the predictions of the model, we perform a feature importance analysis based on SHAP (SHapley Additive exPlanation) values [76]. This method is grounded in cooperative game theory and provides model-agnostic attributions of feature relevance. Positive SHAP values indicate that a given feature contributed to the prediction of the positive class 1, while negative values reveal that the feature pushes the prediction toward class 0.

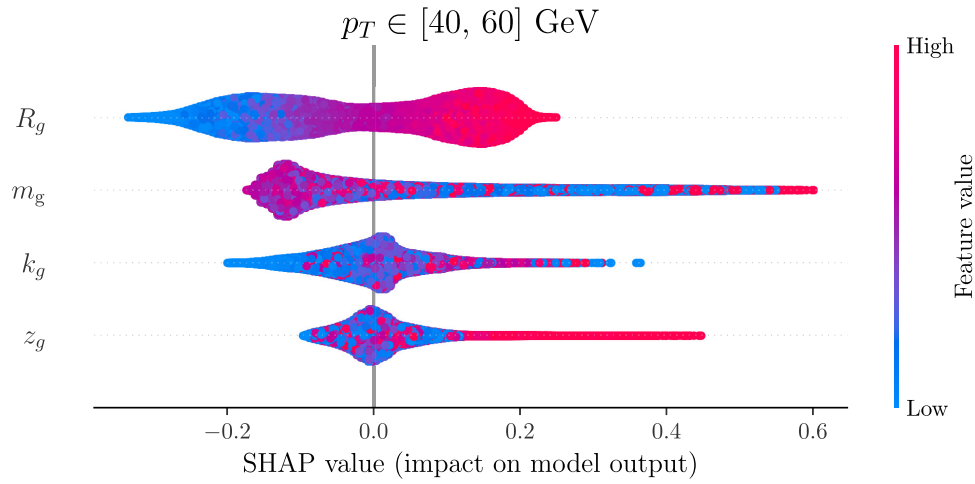
In this work, quenched jets correspond to the positive class, 1, while unquenched jets are typed as 0.

We display the results of the SHAP beeswarm plot in Figs. 2 and 3 for the Random Forest and MLP, respectively. The other two intervals of transverse momentum present the same tendency as the $p_T \in [40, 60]$ GeV results, and for brevity we do not exhibit them in this paper.

From Figs. 2 and 3, one can see that the groomed angular separation, R_g, and the groomed mass, m_g, are generally the two most relevant features in the decision-making of both models. Specifically, higher values of R_g are associated with the classification of jets as quenched, while greater values of m_g indicate that a jet is more likely to be classified as unquenched.



(a)



(b)

FIG. 3. Feature importance (SHAP values) for the MLP model in the $40 < p_T < 60$ GeV interval. (a) Default and (b) v-USPhydro.

The tendencies observed manifest across all examined transverse momentum ranges, medium configurations, and the two ML approaches tested up to this point. The consistency of this pattern across different conditions suggests the presence of a physical signature.

The research presented in Ref. [77], which also utilized the JEWEL framework, illustrates that recoil effects consistently alter jet substructure. These modifications result in an enhancement of angular separation, which aligns with the conclusions drawn from our analysis using SHAP values.

In the Default scenario, we observe a clearer separation in the SHAP values compared with the v-USPhydro scenario. This reinforces our point that the hydrodynamic medium introduces subtle variations in the quenching pattern. As a result, higher feature values become more intermixed with lower ones, although a visible trend still exists.

The difference in granularity observed in the SHAP beeswarm plots between the models is due to the specific explainers used. For the Random Forest, the TreeExplainer leverages the model's discrete tree structure, where multiple observations falling into the same leaf nodes receive identical SHAP values, resulting in distinct vertical clusters. In contrast, the MLP employs the KernelExplainer, which operates on continuous functions and gradients. This leads to a smoother, more organic distribution of points, reflecting the model's continuous nature.

B. Sequential data models

1. Performance metrics

Tables V and VI present the performance metrics for both LSTM and LSTM+Attention models, evaluated on the JEWEL Default and v-USPhydro substructure datasets, across the three p_T intervals that are considered in this study. In all cases, the values reported are averages over independent runs, with the error indicated as well.

TABLE V. Performance metrics for the LSTM model across p_T intervals for Default and v-USPhydro datasets.

p_T [GeV]	Metric	Default	v-USPhydro
40–60	Accuracy	0.9695 (0.0002)	0.9351 (0.0003)
	AUC	0.9896 (0.0002)	0.9727 (0.0002)
	Precision	0.9578 (0.0005)	0.9285 (0.0006)
	Recall	0.9211 (0.0008)	0.8720 (0.0008)
	F1 score	0.9391 (0.0005)	0.8994 (0.0005)
80–250	Accuracy	0.9770 (0.0001)	0.9447 (0.0002)
	AUC	0.9917 (0.0001)	0.9756 (0.0001)
	Precision	0.9700 (0.0003)	0.9339 (0.0003)
	Recall	0.9371 (0.0004)	0.8964 (0.0004)
	F1 score	0.9533 (0.0002)	0.9148 (0.0003)
200–400	Accuracy	0.9778 (0.0002)	0.9481 (0.0003)
	AUC	0.9917 (0.0002)	0.9757 (0.0002)
	Precision	0.9530 (0.0006)	0.9520 (0.0005)
	Recall	0.9579 (0.0005)	0.8858 (0.0007)
	F1 score	0.9554 (0.0004)	0.9177 (0.0004)

Both the LSTM and LSTM+Attention models demonstrate exceptional classification performance. These models achieve accuracy and AUC scores consistently above 95% across all ranges of transverse momentum. The incorporation of the attention mechanism does not increase the predictive capabilities, but rather, it maintains the same levels of performance.

In the in-domain scenario, both the LSTM and LSTM+Attention architectures exhibit a significant advantage when trained and evaluated on the Default medium compared with the v-USPhydro setup. Across all transverse momentum (p_T) intervals, the Default case demonstrates higher precision and recall simultaneously, culminating in superior F_1 scores. For example, within the 40–60 GeV range, the Default-trained LSTM achieves a precision of 0.95 and a recall of 0.92, while the corresponding values for the v-USPhydro configuration decrease to 0.92 and 0.89, respectively. This trend becomes even more pronounced at higher transverse momenta, where Default-trained models maintain recall values approaching 0.95, in stark contrast with the systematically lower values seen in the v-USPhydro setup.

These results suggest that the Default setup offers a cleaner pattern for identifying quenching signatures. This allows the models to differentiate between quenched and unquenched jets with greater sensitivity and reliability. Conversely, the implementation of a realistic hydrodynamical medium, while more representative of actual heavy-ion collisions, introduces additional effects that obscure the quenching signals in the sequential representation, thereby diminishing the efficacy of the models in detecting quenched jets.

The comparison between LSTM and LSTM+ ttention models reveals that neither model consistently outperforms the other across all metrics and p_T ranges. Both exhibit strong performance in the Default and v-USPhydro scenarios, although the results for the latter are generally less favorable. This suggests that while sequential models are more effective for jet classification in a simpler environment, they face greater challenges in a hydrodynamical medium, which introduces a more complex classification landscape.

One possible explanation for the reason why the LSTM+Attention does not outperform the LSTM alone is that the jet substructure sequences used in our analysis are relatively short and may not exhibit strong long-range dependencies. In such cases, the internal gating mechanisms of the LSTM are already sufficient to capture the relevant temporal structure, and the added attention layer does not provide a significant advantage.

Table VII shows the performance of the Transformer model for jet classification in the in-domain medium. As with the previous models, we evaluate accuracy, AUC, precision, recall, and F_1 score.

TABLE VI. Performance metrics for the LSTM+Attention model across p_T intervals for Default and v-USPhydro datasets.

p_T [GeV]	Metric	Default	v-USPhydro
40–60	Accuracy	0.9686 (0.0002)	0.9345 (0.0003)
	AUC	0.9890 (0.0002)	0.9730 (0.0002)
	Precision	0.9549 (0.0006)	0.9098 (0.0007)
	Recall	0.9207 (0.0008)	0.8916 (0.0007)
	F1 score	0.9375 (0.0005)	0.9006 (0.0005)
80–250	Accuracy	0.9766 (0.0001)	0.9458 (0.0002)
	AUC	0.9916 (0.0001)	0.9758 (0.0001)
	Precision	0.9577 (0.0003)	0.9522 (0.0003)
	Recall	0.9483 (0.0004)	0.8804 (0.0004)
	F1 score	0.9530 (0.0002)	0.9149 (0.0003)
200–400	Accuracy	0.9785 (0.0002)	0.9474 (0.0003)
	AUC	0.9919 (0.0001)	0.9757 (0.0002)
	Precision	0.9645 (0.0005)	0.9485 (0.0005)
	Recall	0.9484 (0.0006)	0.8872 (0.0007)
	F1 score	0.9564 (0.0004)	0.9168 (0.0004)

TABLE VII. Performance metrics for the Transformer model on Default and v-USPhydro datasets across different p_T intervals. Values are presented as mean (error).

p_T [GeV]	Metric	Default	v-USPhydro
40–60	Accuracy	0.9725 (0.0002)	0.9327 (0.0003)
	AUC	0.9905 (0.0002)	0.9710 (0.0002)
	Precision	0.9576 (0.0006)	0.9214 (0.0006)
	Recall	0.9336 (0.0007)	0.8720 (0.0008)
	F1 score	0.9454 (0.0005)	0.8960 (0.0005)
80–250	Accuracy	0.9802 (0.0001)	0.9517 (0.0002)
	AUC	0.9931 (0.0001)	0.9792 (0.0001)
	Precision	0.9731 (0.0003)	0.9561 (0.0003)
	Recall	0.9469 (0.0004)	0.8951 (0.0004)
	F1 score	0.9598 (0.0002)	0.9246 (0.0003)
200–400	Accuracy	0.9806 (0.0002)	0.9517 (0.0003)
	AUC	0.9925 (0.0001)	0.9778 (0.0002)
	Precision	0.9708 (0.0005)	0.9476 (0.0005)
	Recall	0.9505 (0.0006)	0.9019 (0.0006)
	F1 score	0.9605 (0.0004)	0.9242 (0.0004)

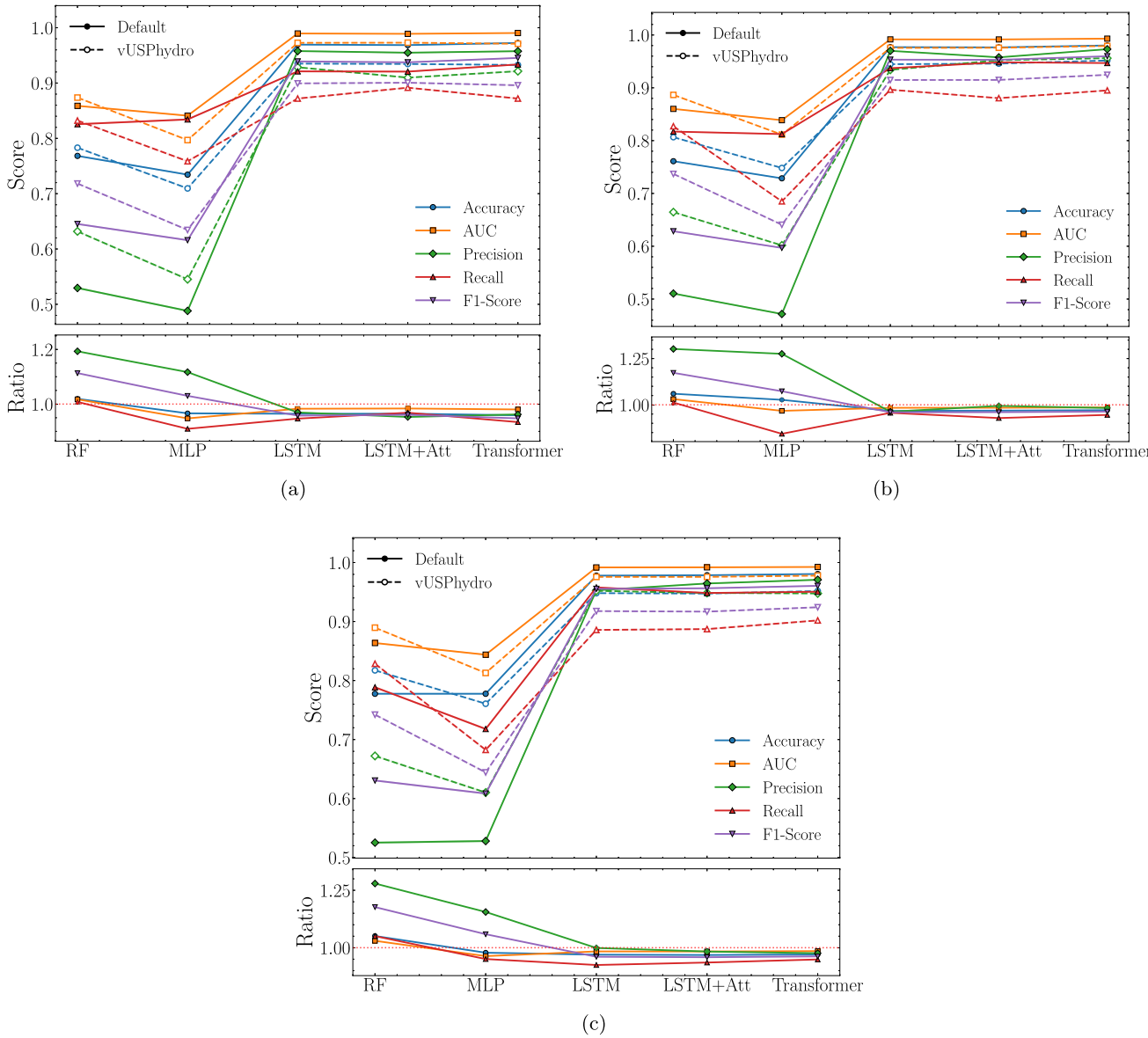


FIG. 4. In-domain performance for each p_T interval. Models are trained and tested on the same medium. All models perform well, with slightly lower performance observed in the v-USPhydro case. Here, the ratio is defined as the value of a given metric obtained with the v-USPhydro model divided by the corresponding value obtained with the Default model. (a) $pt\epsilon$ [40,60] Gev, (b) $pt\epsilon$ [80,250] Gev, and (c) $pt\epsilon$ [200,400] Gev.

When we train and evaluate the model within the same conditions, the Transformer has excellent predictive power, as was the case for the LSTM. Accuracy and AUC score values are above 97%, indicating an almost perfect separation between quenched and unquenched samples. Also, there is a small decrease in recall for the v-USPhydro model, which can be due to the more complex nature and event-by-event variations in a realistic hydrodynamical medium, as explained for the LSTMs. The F_1 score displays significantly high values, indicating that the model does not require a trade-off between precision and recall to achieve strong results.

These results confirm that the Transformer effectively extracts and uses sequential information in jet substructure observables. The inner attention and self-attention mechanisms within the model enable it to dynamically focus on both local features and global patterns that emerge as the jet moves through the medium. The small differences in performance between the JEWEL Default and v-USPhydro scenarios are not necessarily failures of the model but rather a reflection of the underlying dynamics that the realistic medium adds by changing the radiation phase space, which makes it harder to distinguish quenching phenomena, as discussed previously.

Figure 4 displays an intuitive visualization of the difference in performance metrics results for this work.

2. Cross-domain validation

To examine the capacity of LSTM, LSTM+Attention, and Transformer models to generalize the learned data, we con-

TABLE VIII. LSTM cross-domain performance. Models trained on Default and tested on v-USPhydro, and vice versa, for different p_T intervals.

p_T [GeV]	Metric	Def →v-USP	v-USP →Def
40–60	Accuracy	0.9047 (0.0004)	0.9613 (0.0003)
	AUC	0.9621 (0.0003)	0.9887 (0.0002)
	Precision	0.9636 (0.0005)	0.9066 (0.0008)
	Recall	0.7414 (0.0010)	0.9459 (0.0006)
	F1 score	0.8381 (0.0007)	0.9258 (0.0005)
80–250	Accuracy	0.9212 (0.0002)	0.9673 (0.0001)
	AUC	0.9682 (0.0002)	0.9916 (0.0001)
	Precision	0.9756 (0.0002)	0.9110 (0.0005)
	Recall	0.7814 (0.0006)	0.9636 (0.0003)
	F1 score	0.8678 (0.0004)	0.9365 (0.0003)
200–400	Accuracy	0.9262 (0.0003)	0.9751 (0.0002)
	AUC	0.9631 (0.0003)	0.9927 (0.0001)
	Precision	0.9616 (0.0005)	0.9364 (0.0006)
	Recall	0.8064 (0.0009)	0.9654 (0.0005)
	F1 score	0.8772 (0.0006)	0.9507 (0.0004)

TABLE X. Transformer cross-domain performance. Models trained on Default and tested on v-USPhydro, and vice versa, for different p_T intervals. Values are presented as mean (error).

p_T [GeV]	Metric	Def →v-USP	v-USP →Def
40–60	Accuracy	0.9081 (0.0004)	0.9617 (0.0003)
	AUC	0.9647 (0.0003)	0.9897 (0.0002)
	Precision	0.9635 (0.0005)	0.8987 (0.0008)
	Recall	0.7524 (0.0010)	0.9580 (0.0005)
	F1 score	0.8450 (0.0007)	0.9274 (0.0005)
80–250	Accuracy	0.9365 (0.0002)	0.9754 (0.0001)
	AUC	0.9749 (0.0001)	0.9930 (0.0001)
	Precision	0.9791 (0.0002)	0.9407 (0.0004)
	Recall	0.8258 (0.0005)	0.9622 (0.0003)
	F1 score	0.8959 (0.0003)	0.9513 (0.0003)
200–400	Accuracy	0.9297 (0.0003)	0.9737 (0.0002)
	AUC	0.9685 (0.0002)	0.9928 (0.0001)
	Precision	0.9763 (0.0004)	0.9297 (0.0006)
	Recall	0.8044 (0.0009)	0.9674 (0.0005)
	F1 score	0.8820 (0.0005)	0.9481 (0.0004)

duct cross-domain evaluations, similarly to the approach used for the Random Forest and MLP models in Sec. V A 2. In this setup, models are trained on data from one medium, either Default or v-USPhydro, and subsequently evaluated on data from the alternate medium. This testing directly probes the robustness and generalizability of the learned representations, offering insights into the models' ability to adapt across simulation domains.

The results of these evaluations are summarized in Tables VIII–X, which present the performance of the LSTM, LSTM+Attention, and Transformer models, respectively, across different p_T intervals.

Performance reveals a clear asymmetry across all sequential models. When trained on v-USPhydro and tested on Default, the models maintain high performance, showing results comparable to in-domain evaluations. In contrast, models trained on Default and evaluated on v-USPhydro suffer significant performance drops, particularly in recall and F_1 score. This indicates that the hydrodynamical medium enables learning of more generalizable features, while the Default scenario lacks sufficient complexity to train models capable of predicting in a more intricate scenario, as was observed in the nonsequential case.

TABLE IX. LSTM+Attention cross-domain performance. Models trained on Default and tested on v-USPhydro, and vice versa, for different p_T intervals.

p_T [GeV]	Metric	Def →v-USP	v-USP →Def
40–60	Accuracy	0.8998 (0.0004)	0.9577 (0.0003)
	AUC	0.9602 (0.0003)	0.9898 (0.0002)
	Precision	0.9605 (0.0005)	0.8824 (0.0008)
	Recall	0.7290 (0.0010)	0.9627 (0.0005)
	F1 score	0.8289 (0.0007)	0.9208 (0.0005)
80–250	Accuracy	0.9217 (0.0002)	0.9728 (0.0001)
	AUC	0.9677 (0.0002)	0.9919 (0.0001)
	Precision	0.9656 (0.0003)	0.9359 (0.0004)
	Recall	0.7917 (0.0006)	0.9568 (0.0003)
	F1 score	0.8700 (0.0004)	0.9462 (0.0003)
200–400	Accuracy	0.9272 (0.0003)	0.9742 (0.0002)
	AUC	0.9657 (0.0003)	0.9927 (0.0001)
	Precision	0.9712 (0.0004)	0.9318 (0.0007)
	Recall	0.8008 (0.0009)	0.9669 (0.0005)
	F1 score	0.8778 (0.0006)	0.9490 (0.0004)

In the Transformer case, the same asymmetry persists with important nuances. When trained on v-USPhydro and evaluated on Default (v-USP →Def), the model maintains excellent performance, with accuracy, AUC, and F_1 score remaining high, and recall consistently above 0.9 across all p_T intervals. This demonstrates again that training on a realistic medium enables the Transformer with generalized representations that remain informative even in simpler environments.

Conversely, training on Default and testing on v-USPhydro (Def →v-USP) leads to a performance drop, especially in recall. Despite high precision, reflecting a great classification strategy, the model fails to capture a large fraction of quenched jets. This behavior illustrates the presence of a conservative bias in the model that only labels jets as quenched when exposed to strong effects and misses valid cases with more subtle modifications.

These findings demonstrate that even advanced models such as Transformers, designed to capture long-range dependencies, cannot overcome the limitations imposed by insufficient physics in the training data. While their inductive bias allows them to attend to complex correlations across the jet evolution, such correlations must exist in the training domain to be learned effectively. This shows that realistic data are essential to ensure that ML models do not merely interpolate within idealized scenarios, but generalize meaningfully to the full complexity of jet quenching phenomena.

This result becomes especially significant when compared with the behavior of nonsequential models, the Random

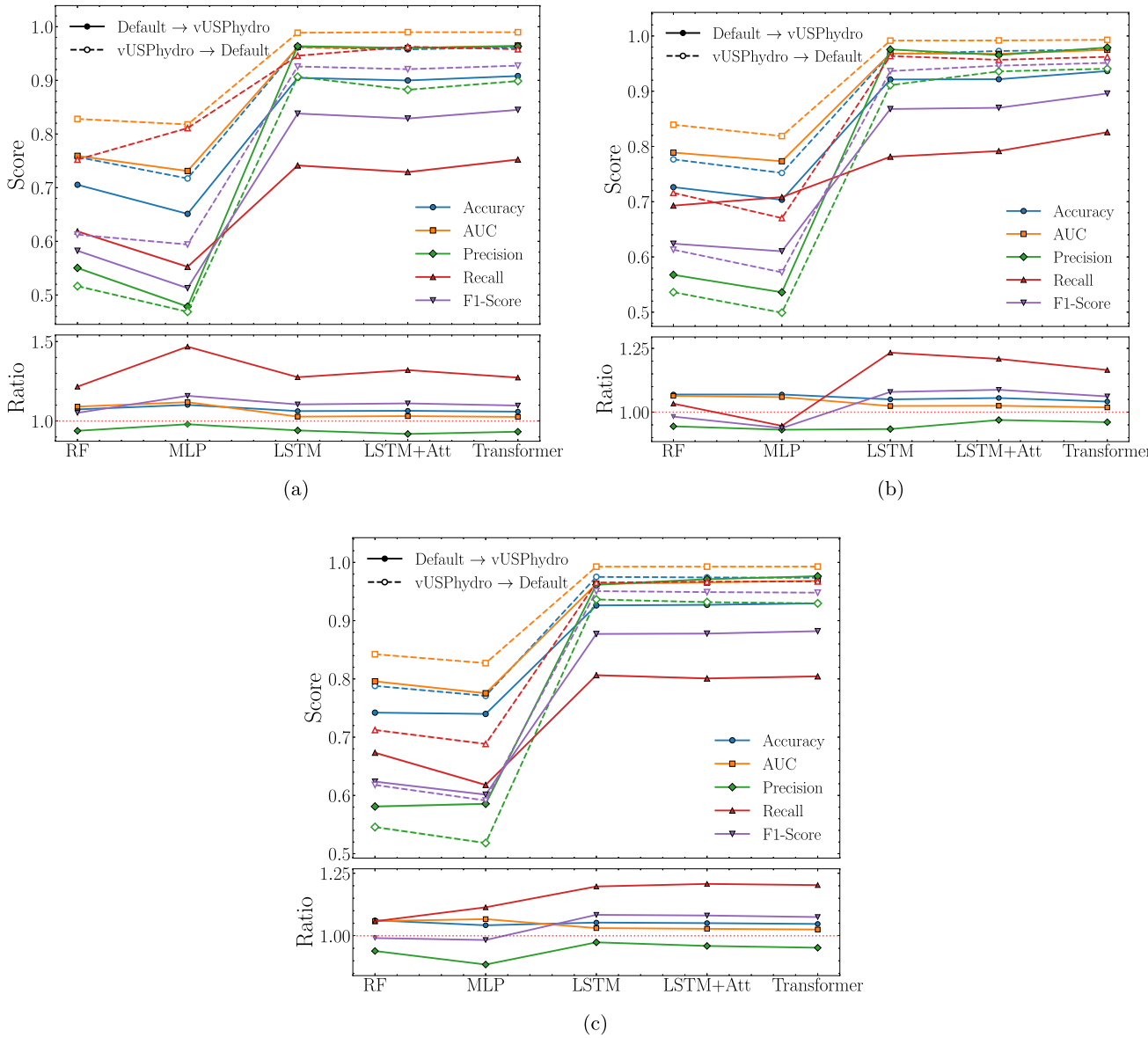


FIG. 5. Cross-domain performance for each p_T interval. Models are trained on one medium (Default or v-USPhydro) and tested on the other. Sequential models retain strong performance when trained on v-USPhydro, but show degradation when trained on Default and applied to the hydrodynamical medium. Here, the ratio is defined as the value of a given metric obtained with v-USPhydro → Default divided by the corresponding value obtained with the Default → v-USPhydro. (a) $pt\epsilon$ [40,60] Gev, (b) $pt\epsilon$ [80,250] Gev, and (c) $pt\epsilon$ [200,400] Gev.

Forest and MLP. In those models, cross-domain performance was inferior, but it remained unclear whether this was due to their architectural limitations or genuine sensitivity to changes in the medium. The sequential models now clarify this point, showing that even with sufficient representational capacity and robust performance, the models remain sensitive to domain shifts arising from different medium descriptions. Therefore, this constitutes an important result, highlighting that machine-learning models, even when powerful and well-regularized, reflect the physical assumptions embedded in the data used for training.

Figures 4 and 5 compare the five metrics results across all the models and medium configurations discussed in this paper.

In conclusion, these results confirm that sequential models, particularly the LSTM and Transformer used in this work, possess greater representational power and generalization capability compared with the chosen nonsequential architectures.

3. Sequence importance analysis

To investigate the relevance of each position in the Soft Drop sequence, we compute the absolute SHAP attribution step by step for every jet and then take the average. In this way, the curves indicate the average importance of each declustering step in the model decision. The distributions of the mean

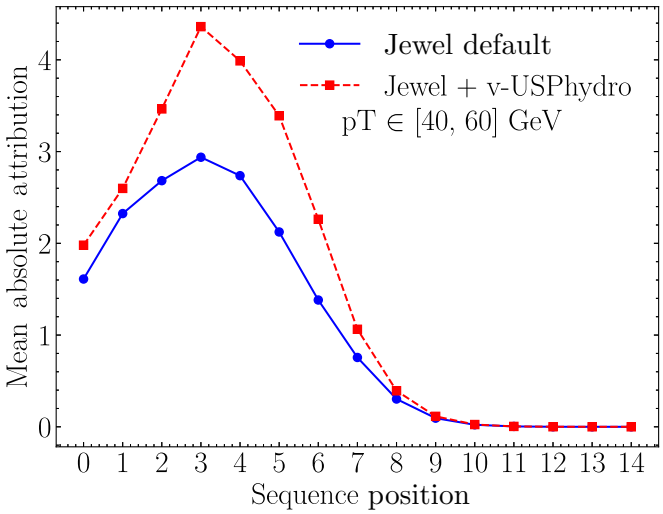


FIG. 6. SHAP mean absolute attribution to each step of the grooming sequence for both JEWEL Default and JEWEL v-USPhydro media in the range $40 \leqslant p_T \leqslant 60$ GeV.

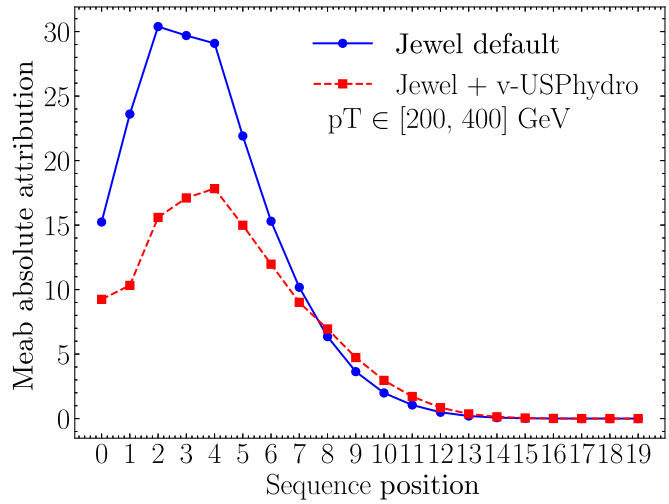


FIG. 8. SHAP mean absolute attribution to each step of the grooming sequence for both JEWEL Default and JEWEL v-USPhydro media in the range $200 \leqslant p_T \leqslant 400$ GeV.

absolute SHAP value along the grooming sequence are shown in Figs. 6–8.

Notice that most of the attribution is concentrated in the first few Soft Drop splittings, typically around $t = 2$–5, whereas the later steps contribute very little to the prediction. This pattern is consistently found for both media and across all p_T intervals, indicating that the network primarily relies on information encoded in the earliest and hardest branchings of the declustering sequence.

A more detailed comparison between the two media in the $40 \leqslant p_T \leqslant 60$ GeV interval shows that both attribution profiles reach their maximum at $t = 4$, but with a different internal hierarchy. In the JEWEL+v-USPhydro case, the peak at $t = 4$ is more than twice as large as the attribution of the first splitting, indicating a strong enhancement of the discriminating information after the beginning of the declustering sequence. In contrast, the JEWEL Default profile also peaks at $t = 4$, but the increase relative to the first splitting is milder, rising only from about 2 to 3 in SHAP magnitude rather than doubling. This difference suggests that, while both media encode relevant information beyond the first declustering step, the v-USPhydro medium does so in a much more pronounced way, shifting the model sensitivity toward the next few resolved branchings.

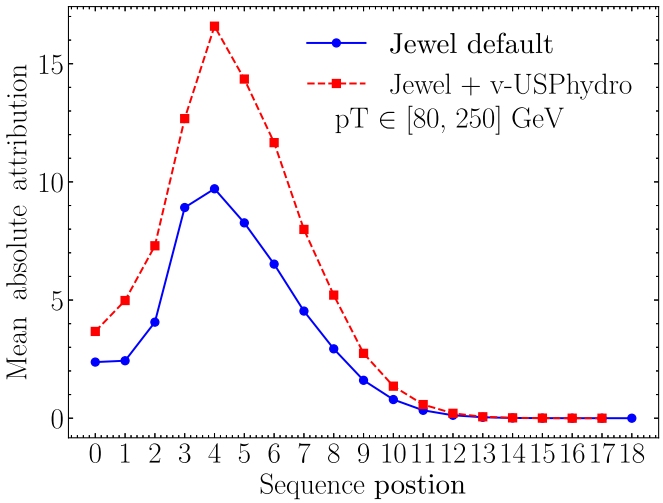


FIG. 7. SHAP mean absolute attribution to each step of the grooming sequence for both JEWEL Default and JEWEL v-USPhydro media in the range $80 \leqslant p_T \leqslant 250$ GeV.

This pattern becomes even clearer in the $80 \leqslant p_T \leqslant 250$ GeV range. In the v-USPhydro case, the peak around $t = 5$ is roughly three to four times larger than the attribution at the first splitting, while in the Default case, the enhancement from the first splitting to the maximum is also strong but noticeably less pronounced. This means that, for v-USPhydro, the model extracts most of the useful information not from the entrance of the sequence, but from a delayed region of the declustering history, after a few branchings have taken place.

The high-p_T interval, $200 \leqslant p_T \leqslant 400$ GeV, shows a qualitatively different behavior. Here, the Default medium displays a much stronger concentration of attribution in the earliest steps, with a very high plateau-like structure between $t = 3$ and $t = 5$. In contrast, the v-USPhydro curve is lower and smoother, with a more modest peak and a less dramatic separation between the leading early steps. This suggests that, at high p_T, the Default medium produces a more sharply localized signature in the first few branchings, whereas v-USPhydro distributes the relevant information more evenly and with less contrast.

These results indicate that the distinction between the Default and v-USPhydro media is p_T-dependent. At low and intermediate p_T, the v-USPhydro medium leads to larger SHAP attributions over the first few declustering steps and to a slightly more extended sequence dependence. At high p_T, however, the Default medium becomes more strongly concentrated in the earliest splittings, while v-USPhydro retains a broader but less sharply peaked attribution profile.

In the present work, this sequential SHAP analysis was carried out only for the LSTM architecture. Extending the same study to the LSTM with Attention and Transformer models will be an important next step, since it will allow us to test whether this dominance of the first splittings can be interpreted as a physical feature of the data or a behavior that depends on the specific bias of the model. Such a comparison may also help clarify whether architectures with explicit attention mechanisms capture the same temporal structure or redistribute the importance across the grooming sequence in a qualitatively different way.

VI. SUMMARY

In this work, we employ the Monte Carlo event generator JEWEL with two different descriptions of the hot and dense medium formed in heavy-ion collisions, such as PbPb, to simulate quenching effects on jets. The main goal of this paper is to determine which of the chosen models of Machine Learning have greater generalization power, using observables computed only at the first splitting of the jet tree history and a sequence of observables computed within the grooming history of a jet.

We demonstrate the existence of an architectural hierarchy in machine-learning approaches for jet quenching identification through systematic cross-domain analysis of medium-modified jets in heavy-ion collisions. Sequential neural networks, LSTM, LSTM+Attention and Transformer, exhibit exceptional classification performance ($\gtrsim$93% accuracy) and remarkable generalization capabilities across distinct medium implementations, while static data-based algorithms, Random Forest and Multilayer Perceptron, display dramatic performance degradation under domain transfer conditions.

The cross-domain validation reveals a critical asymmetry, showing that models trained on the more realistic v-USPhydro hydrodynamic medium maintain robust discrimination when evaluated on the JEWEL Default scenario, reaching accuracies of at least 90%, whereas the reverse configuration leads to a pronounced degradation in performance, with accuracies dropping to around 65% for the nonsequential models. This result indicates that the more complex medium dynamics of v-USPhydro allow the models to learn a richer set of physical patterns, some of which remain identifiable even in the simpler Default setup. In this sense, the medium leaves fingerprints on the final observables, since the information learned in the realistic scenario is sufficiently structured and robust to generalize across media, while the features learned in the Default case are too limited to capture the broader phenomenology present in v-USPhydro.

It is worth noting that the differences caused by the two different media in jet observables are not immediately evident in global statistical measurements, such as R_{AA}, which average over large jet collections [4]. In contrast, they become more visible in a jet-by-jet analysis based on machine learning, highlighting the potential of these methods to uncover patterns that may remain hidden in traditional ensemble observables.

Interpretability analysis of nonsequential models identifies the groomed angular separation and jet mass as the primary discriminative features, with quenched jets exhibiting enhanced substructure broadening consistent with theoretical expectations. In this context, SHAP proves to be a valuable tool for gaining insight into how jets are modified while traversing a hot and dense medium. For the sequential LSTM-based analysis, the attribution patterns further show that the earliest steps of the Soft Drop sequence carry most of the relevant information for the classification task, with the importance strongly concentrated in the first few declustering stages and rapidly decreasing at later steps. Taken together, these results indicate that both the global substructure observables and the early organization of the grooming sequence encode the dominant signatures of medium modification.

However, our simulation framework still lacks embedded thermal background fluctuations, which represents an important limitation and may artificially enhance the classification performance. In addition, one should keep in mind that not every jet reconstructed in a PbPb event is necessarily a strongly quenched jet, since some partons may traverse only a short path in the medium or undergo relatively low modification. This intrinsic physical ambiguity complicates the interpretation of any binary classification task. Future investigations incorporating realistic thermal medium fluctuations, experimental background conditions, and more comprehensive neural-network interpretability frameworks will therefore be essential to assess both the practical applicability and the physical robustness of these sequential machine-learning methodologies for jet-quenching studies in real heavy-ion collision environments.

ACKNOWLEDGMENTS

We acknowledge the Brazilian funding agency FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), Grant No. 2024/02346–6, associated with the Project No. 2020/04867-2, as well as the University of São Paulo (USP) for financial support.

DATA AVAILABILITY

The data and analysis code supporting the findings of this study are available in a public GitHub repository [78].

APPENDIX: MACHINE LEARNING HYPERPARAMETERS

In this Appendix, we present the hyperparameters used for each model in the training and evaluation processes.

The Random Forest model was imported from the scikit-learn [79] library, version 1.5.1. All the supervised neural networks used in this work were built with PYTORCH 2.3.1 [80]. Hyperparameter optimization was conducted with Hyperopt [81].

1. Random forest

The optimized hyperparameters adopted for the Random Forest classifier are summarized in Table XI. These values were used for all transverse-momentum intervals and for both medium descriptions.

TABLE XI. Optimized Random Forest hyperparameters used in all p_T intervals and for both medium descriptions.

Hyperparameter	Value
Number of estimators (`n_estimators`)	400
Maximum depth (`max_depth`)	30
Minimum samples per split (`min_samples_split`)	10
Minimum samples per leaf (`min_samples_leaf`)	2
Bootstrap sampling (`bootstrap`)	False
Maximum features (`max_features`)	`sqrt`
Random seed (`random_state`)	42
Number of jobs (`n_jobs`)	1

2. MLP

For the MLP model, the optimized hyperparameters depend on the transverse-momentum interval and on the medium description. The configuration used for the low-p_T interval is shown in Table XII.

TABLE XII. Optimized MLP hyperparameters for the $p_T \in [40, 60]$ GeV interval. BCE denotes binary cross entropy.

Hyperparameter	Default	v-USPhydro
Decay factor	0.974	0.945
Dropout rate	0.329	0.129
Hidden size 1	64	16
Hidden size 2	12	16
Learning rate	0.003	0.003
Loss function	BCE	BCE
Batch size	64	128
Epochs	40	60

The optimized MLP configuration used for the intermediate- and high-p_T intervals is given in Table XIII.

TABLE XIII. Optimized MLP hyperparameters for $p_T \in [80, 250]$ and $[200, 400]$ GeV. BCE denotes binary cross entropy.

Hyperparameter	Default	v-USPhydro
Decay factor	0.945	0.936
Dropout rate	0.129	0.235
Hidden size 1	16	32
Hidden size 2	16	12
Learning rate	0.003	0.001
Loss function	BCE	BCE
Batch size	128	64
Epochs	60	70

3. LSTM and LSTM+attention

The optimized hyperparameters used for the LSTM and LSTM+attention models are listed in Table XIV. The same configuration was adopted for all transverse-momentum intervals and for both medium descriptions.

TABLE XIV. Optimized LSTM and LSTM+Attention hyperparameters used for all p_T intervals and for both medium descriptions.

Hyperparameter	Value
Decay factor	0.947
Dropout rate	0.456
Hidden size 1	18
Hidden size 2	6
Learning rate	0.016
Loss function	BCE
Batch size	64
Epochs	40
Number of LSTM layers	2

4. Transformer

Finally, the optimized hyperparameters used for the Transformer model are presented in Table XV.

TABLE XV. Optimized Transformer hyperparameters used for all p_T intervals and for both medium descriptions.

Hyperparameter	Value
Batch size	256
Decay factor	0.936
Dropout rate	0.101
Feed-forward dimension	256
Learning rate	0.0013
Model (embedding) dimension	64
Epochs	40
Number of attention heads	4
Number of Transformer layers	5

[1] U. W. Heinz and M. Jacob, Evidence for a new state of matter: An assessment of the results from the CERN lead beam programme, arXiv:nucl-th/0002042.

[2] J. Adams, M. Aggarwal, Z. Ahammed, J. Amonett, B. Anderson, D. Arkhipkin, G. Averichev, S. Badyal, Y. Bai, J. Balewski, *et al.*, Experimental and theoretical challenges in the search for the quark–gluon plasma: The STAR collaboration's critical assessment of the evidence from RHIC collisions, Nucl. Phys. A **757**, 102 (2005).

[3] S. Cao and X.-N. Wang, Jet quenching and medium response in high-energy heavy-ion collisions: A review, Rep. Prog. Phys. **84**, 024301 (2021).

[4] M. M. d. M. Paulino, Study of the medium effects in jet observables in relativistic heavy-ion collisions, Master's thesis, Instituto de Física, Universidade de São Paulo, São Paulo, 2023.

[5] T. Renk, J. Ruppert, C. Nonaka, and S. A. Bass, Jet quenching in a three-dimensional hydrodynamic medium, Phys. Rev. C **75**, 031902(R) (2007).

[6] Y. Singh, P. K. Bhatia, and O. Sangwan, A review of studies on machine learning techniques, Int. J. Comput. Sci. Secur. **1**, 70 (2007).

[7] L. Apolinário, N. F. Castro, M. Crispim Romão, J. G. Milhano, R. Pedro, and F. C. R. Peres, Deep learning for the classification of quenched jets, J. High Energy Phys. 11 (2021) 219.

[8] Y.-L. Du, Overview: Jet quenching with machine learning, arXiv:2308.10035.

[9] J. A. Gonçalves, All is more: Energy flow networks for jet quenching, arXiv:2510.26879.

[10] H. J. Bossi, Novel uses of machine learning for differential jet quenching measurements at the LHC, Ph.D. thesis, Yale University, 2023.

[11] K. C. Zapp, F. Krauss, and U. A. Wiedemann, A perturbative framework for jet quenching, J. High Energy Phys. 03 (2013) 080.

[12] K. Zapp, JEWEL 2.0.0: Directions for use, Eur. Phys. J. C **74**, 2762 (2014).

[13] J. Noronha-Hostler, G. S. Denicol, J. Noronha, R. P. G. Andrade, and F. Grassi, v-USPhydro: Bulk viscosity effects on event-by-event relativistic hydrodynamics, J. Phys.: Conf. Ser. **458**, 012018 (2013).

[14] L. Breiman, Random forests, Mach. Learn. **45**, 5 (2001).

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, Nature (London) **323**, 533 (1986).

[16] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Comput. **9**, 1735 (1997).

[17] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, in *3rd International Conference on Learning Representations*, Conference Track Proceedings, edited by Y. Bengio and Y. LeCun (ICLR, San Diego, CA, USA, 2015).

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30.

[19] T. Sjöstrand, S. Mrenna, and P. Skands, Pythia 6.4 physics and manual, J. High Energy Phys. 05 (2006) 026.

[20] 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).

[21] M. L. Miller, K. Reygers, S. J. Sanders, and P. Steinberg, Glauber modeling in high-energy nuclear collisions, Annu. Rev. Nucl. Part. Sci. **57**, 205 (2007).

[22] R. Baier, Y. L. Dokshitzer, A. H. Mueller, S. Peigne, and D. Schiff, Radiative energy loss of high energy quarks and gluons in a finite-volume quark-gluon plasma, Nucl. Phys. B **483**, 291 (1997).

[23] B. Zakharov, Radiative energy loss of high-energy quarks in finite-size nuclear matter and quark-gluon plasma, JETP Lett. **65**, 615 (1997).

[24] R. Kunnawalkam Elayavalli and K. C. Zapp, Medium response in JEWEL and its impact on jet shape observables in heavy ion collisions, J. High Energy Phys. 07 (2017) 141.

[25] M. Liu and G. Liu, Smoothed particle hydrodynamics (SPH): An overview and recent developments, Arch. Comput. Method. Eng. **17**, 25 (2010).

[26] J. S. Moreland, J. E. Bernhard, and S. A. Bass, Alternative ansatz to wounded nucleon and binary collision scaling in high-energy nuclear collisions, Phys. Rev. C **92**, 011901 (2015).

[27] P. Alba, R. Bellwied, S. Borsányi, Z. Fodor, J. Günther, S. D. Katz, V. Mantovani Sarti, J. Noronha-Hostler, P. Parotto, A. Pasztor, I. P. Vazquez, and C. Ratti, Constraining the hadronic spectrum through QCD thermodynamics on the lattice, Phys. Rev. D **96**, 034517 (2017).

[28] S. Borsányi, Z. Fodor, J. Guenther, R. Kara, S. D. Katz, P. Parotto, A. Pásztor, C. Ratti, and K. K. Szabó, Lattice QCD equation of state at finite chemical potential from an alternative expansion scheme, Phys. Rev. Lett. **126**, 232001 (2021).

[29] J. E. Bernhard, J. S. Moreland, and S. A. Bass, Bayesian estimation of the specific shear and bulk viscosity of quark–gluon plasma, Nat. Phys. **15**, 1113 (2019).

[30] L. Barreto, F. M. Canedo, M. G. Munhoz, J. Noronha, and J. Noronha-Hostler, Jet cone radius dependence of RAA and v2 at PbPb 5.02 TeV from JEWEL+TRENTo+v-USPhydro, Phys. Lett. B **860**, 139217 (2025).

[31] M. Crispim Romão, J. Guilherme Milhano, and M. van Leeuwen, Jet substructure observables for jet quenching in quark gluon plasma: A machine learning driven analysis, SciPost Phys. **16**, 015 (2024).

[32] J. G. Milhano and K. Zapp, Improved background subtraction and a fresh look at jet sub-structure in JEWEL, Eur. Phys. J. C **82**, 1010 (2022).

[33] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft drop, J. High Energy Phys. 05 (2014) 146.

[34] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.

[35] Y. L. Dokshitzer, G. Leder, S. Moretti, and B. Webber, Better jet clustering algorithms, J. High Energy Phys. 08 (1997) 001.

[36] A. J. Larkoski, S. Marzani, and J. Thaler, Sudakov safety in perturbative QCD, Phys. Rev. D **91**, 111501 (2015).

[37] CMS Collaboration, Measurement of the primary Lund jet plane density in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 05 (2024) 116.

[38] L. Liu, J. Velkovska, Y. Wu, and M. Verweij, Identifying quenched jets in heavy ion collisions with machine learning, Adv. High Energy Phys. (Cairo, Egypt) **2023**, 140 (2023).

[39] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms* (Chapman and Hall/CRC, New York, 2025).

[40] T. G. Dietterich, Ensemble methods in machine learning, in *International Workshop on Multiple Classifier Systems* (Springer, Berlin, 2000), pp. 1–15.

[41] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, 1st ed. (Chapman and Hall/CRC, New York, 1984).

[42] J. R. Quinlan, Induction of decision trees, Mach. Learn. **1**, 81 (1986).

[43] L. Breiman, Bagging predictors, Mach. Learn. **24**, 123 (1996).

[44] T. K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. **20**, 832 (1998).

[45] H. Alhazmi, Z. Dong, L. Huang, J. H. Kim, K. Kong, and D. Shih, Resolving combinatorial ambiguities in dilepton tt event topologies with neural networks, Phys. Rev. D **105**, 115011 (2022).

[46] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychol. Rev. **65**, 386 (1958).

[47] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning* (Springer, New York, 2006), Vol. 4.

[48] V. Nair and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, in *Proceedings of the 27th International Conference on Machine Learning, ICML 2010, Haifa, Israel* (Omnipress, Madison, WI, 2010), pp. 807–814.

[49] C. Sun, J. Ngadiuba, M. Pierini, and M. Spiropulu, Fast jet tagging with MLP-mixers on FPGAs, Mach. Learn.: Sci. Technol. **6**, 035025 (2025).

[50] J. L. Elman, Finding structure in time, Cognit. Sci. **14**, 179 (1990).

[51] I. Sutskever, O. Vinyals, and Q. V. Le, Sequence to sequence learning with neural networks, in *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014), Vol. 27, pp. 3104–3112.

[52] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, Grammar as a foreign language, in *Advances in Neural Information Processing Systems*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc., 2015), Vol. 28, pp. 2773–2781.

[53] S. Egan, W. Fedorko, A. Lister, J. Pearkes, and C. Gay, Long short-term memory (LSTM) networks with jet constituents for boosted top tagging at the LHC, arXiv:1711.09059.

[54] T. Luong, H. Pham, and C. D. Manning, Effective approaches to attention-based neural machine translation, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, edited by L. Márquez, C. Callison-Burch, and J. Su (Association for Computational Linguistics, Lisbon, Portugal, 2015).

[55] Y. Yi, Z. Chen, and R. Li, LSTM neural networks with attention mechanisms for accelerated prediction of charge density at onset condition of DC corona discharge, IEEE Access **10**, 124697 (2022).

[56] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiský, and P. Blunsom, Reasoning about entailment with neural attention, arXiv:1509.06664.

[57] H. Qu, C. Li, and S. Qian, Particle Transformer for jet tagging, in *International Conference on Machine Learning* (PMLR, 2022), pp. 18281–18292.

[58] M. He and D. Wang, Quark/gluon discrimination and top tagging with dual attention Transformer, Eur. Phys. J. C **83**, 1116 (2023).

[59] M. Crispim Romão, J. A. Gonçalves, and J. G. Milhano, Quantifying vacuum-like jets in heavy-ion collisions: A machine learning study, Eur. Phys. J. C **86**, 422 (2026).

[60] L. Ferrer, Analysis and comparison of classification metrics, arXiv:2209.05355.

[61] J. Bergstra and Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. **13**, 281 (2012).

[62] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, Algorithms for hyper-parameter optimization, in *Advances in Neural Information Processing Systems*, edited by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, K. Q. Weinberger (Curran Associates, Inc., 2011), Vol. 24, pp. 2546–2554.

[63] B. Efron, Bootstrap methods: Another look at the jackknife, The Annals of Statistics **7**, 1 (1979).

[64] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall/CRC, New York, 1994).

[65] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge university press, Cambridge 1997).

[66] M. A. Hernán and J. M. Robins, *Causal Inference: What if* (Chapman & Hall/CRC, Boca Raton, 2020).

[67] L. Apolinario, N. Armesto, and L. Cunqueiro, An analysis of the influence of background subtraction and quenching on jet observables in heavy-ion collisions, J. High Energy Phys. 02 (2013) 022.

[68] T. Mengel, P. Steffanic, C. Hughes, A. C. O. D. Silva, and C. Nattrass, Multiplicity based background subtraction for jets in heavy ion collisions, arXiv:2402.10945.

[69] T. Mengel, P. Steffanic, C. Hughes, A. C. O. da Silva, and C. Nattrass, Interpretable machine learning methods applied to jet background subtraction in heavy-ion collisions, Phys. Rev. C **108**, L021901 (2023).

[70] A. Budhraja, M. van Leeuwen, and J. G. Milhano, Jet observables in heavy ion collisions: A white paper, arXiv:2409.03017.

[71] J. A. Gonçalves and J. G. Milhano, Apples to apples in jet quenching: Robustness of machine learning classification of quenched jets to underlying event contamination, J. High Energy Phys. 09 (2025) 028.

[72] M. Connors, C. Nattrass, R. Reed, and S. Salur, Jet measurements in heavy ion physics, Rev. Mod. Phys. **90**, 025005 (2018).

[73] A. Falcão and K. Tywoniuk, Constraining jet quenching in heavy-ion collisions with Bayesian inference, J. High Energy Physics 02 (2026) 69.

[74] C. Andres, J. Holguin, R. K. Elayavalli, and J. Viinikainen, Minimizing selection bias in inclusive jets in heavy-ion collisions with energy correlators, Phys. Rev. Lett. **134**, 082303 (2025).

[75] U. S. Qureshi and R. K. Elayavalli, Model-agnostic tagging of quenched jets in heavy-ion collisions, arXiv:2411.19389.

[76] S. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30, pp. 4765–4774.

[77] G. Milhano, U. A. Wiedemann, and K. C. Zapp, Sensitivity of jet substructure to jet-induced medium response, Phys. Lett. B **779**, 409 (2018).

[78] L. L. d. Silva and M. G. Munhoz, Jet quenching identification via supervised learning in simulated heavy-ion collisions, GitHub repository, 2026, https://github.com/nardols/Jet-quenching-identification-via-supervised-learning-in-simulated-heavy-ion-collisions.

[79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).

[80] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), pp. 8024–8035.

[81] J. Bergstra, D. Yamins, and D. D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research (2013), Vol. 28, pp. 115–123.