

Quantifying vacuum-like jets in heavy-ion collisions: a Machine Learning study

Miguel Crispim Romão^{a,1,2}, João A. Gonçalves^{b,2,3}, José Guilherme Milhano^{c,2,3}

¹Institute for Particle Physics Phenomenology, Durham University, Durham DH1 3LE, United Kingdom

²LIP, Av. Prof. Gama Pinto, 2, P-1649-003 Lisboa, Portugal

³Instituto Superior Técnico (IST), Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Lisbon, Portugal

Received: date / Accepted: date

Abstract The modification of jets by interaction with the Quark Gluon Plasma has been extensively established through the comparison of observables computed for samples of jets produced in nucleus-nucleus collisions and proton-proton collisions. The presence of vacuum-like jets, jets that experienced little interaction with the Quark Gluon Plasma, in the nucleus-nucleus samples dilutes the overall observed modification hindering the detailed study of the underlying physical mechanisms. The ability to ascertain on a jet-by-jet basis the degree of modification of a jet would be an invaluable step in overcoming this limitation. We consider a Transformer classifier, trained on a low-level representation of jets given by the 4-momenta of all its constituents. We show that the Transformer is able to capture discriminating information not accessible to other architectures which use high-level physical observables as input. The Transformer allows us to identify, in the experimentally relevant case where both medium response and underlying event contamination are accounted for, a class of jets that have been unequivocally modified. Further, we perform a robust estimate of the upper bound for the fraction of jets in nucleus-nucleus collisions that are, for all purposes, indistinguishable from those produced in proton-proton collisions.

1 Introduction

Jets result from the reconstruction of the hadronized end products of the branching of highly virtual energetic partons produced in ultra-relativistic particle collisions. In heavy-ion collisions, the branching occurs while traversing and interacting with the Quark Gluon Plasma (QGP) created in

the same collisions. As a result of this interaction, properties of jets reconstructed in heavy-ion collisions differ from those in the proton-proton case. These modifications, globally referred to as jet quenching [1–5], have been hailed as providing a promising pathway towards characterization of the spacetime structure of the QGP [6–20]. Jet quenching has been unequivocally established from a variety of global and substructure jet observables, whose experimentally measured distributions for jet samples reconstructed in heavy ion collisions differ from those measured in proton-proton collisions [4, 5, 21, 22].

The extent to which each individual jet is modified is known to depend on several factors, including the multiplicity of its vacuum-like fragmentation [23–26], the amount of matter traversed, and the gradients of that matter [9–15]. A standard measure of the modification of a jet is the comparison of its reconstructed total transverse momentum with a proxy for the transverse momentum of the parton from which the jet originated, be it the transverse momentum of an electroweak boson in events where it is produced back-to-back with a jet, or in the more abundant dijet case through an estimate of average energy loss [27–30]. This measure disregards the possibility of jets being modified by interaction with QGP without significant energy loss [31].

To devise a novel measure that is sensitive to the modification of properties of a jet beyond its total energy amounts to establishing whether a modified jet contains enough information, without reference to any external proxy, to be distinguishable from a jet produced in the absence of QGP. In the context of Machine Learning, one can ask whether a machine can learn to make such a distinction. Several studies [32–42], including by some of us [34, 37, 39, 41], have addressed this point. In particular, it was established [37] that such discrimination is possible on the basis of combinations of standard observables, with pairs of observables saturat-

^ae-mail: miguel.romao@durham.ac.uk

^be-mail: jgoncalves@lip.pt

^ce-mail: gmlhano@lip.pt

ing the discrimination power of the large set of observables considered in that work.

Here, we address this task by considering a low-level representation of jets, i.e. the list of the 4-momenta of all its constituents, to train a Transformer classifier [43]. This allows us to ask and answer whether there is discriminating information within the jet that is not captured by the standard high-level observables considered in [37]. Transformers arise as the natural architecture to seek such answers. Their versatility and unparalleled performance for sequence-type data, which low-level jet data can be represented as, has taken the High Energy Physics (HEP) community by storm [44–56]. They consistently match or exceed the performance of established architectures across a range of tasks using only low-level data, from jet tagging to generative modelling.

In turn, this allows us to establish an upper bound on the fraction of jets produced in PbPb collisions that are, for all purposes, indistinguishable from their vacuum counterparts produced in proton proton collisions.

The paper is organized as follows: Section 2 describes the generated datasets; Section 3 introduces the transformer classifier; Section 4 shows and discusses our classification results; Section 5 provides an estimate of the fraction of vacuum-like jets in jets produced in PbPb collisions; and we conclude in Section 6. Appendix A and Appendix B detail, respectively, the high-level observables used in [37] and the optimisation of the transformer hyperparameters.

2 Data Simulation

This study is based on di-jet samples produced by the Monte-Carlo event generator JEWEL 2.3 [57]. Two datasets, each including both the proton-proton (vacuum) and Pb-Pb (medium) cases, were prepared. The dataset referred to as *signal only* (SO) does not include any effects due to QGP response, nor effects due to contamination from imperfectly subtracted underlying-event. The *medium response and underlying event* (MR-UE) dataset was prepared following the procedure introduced in [39], where full details can be found. In this case, the PbPb sample includes effects due to QGP response as modelled by JEWEL – including the JEWEL specific subtraction [58] which avoids double counting of underlying event contributions – is embedded in a realistic heavy-ion underlying event which is subsequently subtracted using standard methods. The proton-proton sample is also embedded in a realistic heavy-ion underlying event and subtraction is carried out analogously to the PbPb case. This proton-proton baseline corresponds to the physically relevant case where jets are not modified by interaction with QGP, but display modifications entirely due to underlying-event contamination.

All samples were generated at hadron level for $\sqrt{s} = 5.02$ TeV with a hard matrix element lower transverse momentum cut-off of 50 GeV and the generation spectrum was re-weighted by p_T^5 as to oversample the large p_T region. The resulting event weights were used throughout the study. For the PbPb samples, the QGP was generated according to the simple parametrization described in [59] with $\tau_i = 0.4$ fm/c, $T_i = 590$ MeV, $T_c = 170$ MeV, centrality 0 – 10%, and covering $|\eta| < 4$.

Jets were reconstructed from all particles within $\eta_{\text{part}} < 2.5$ using the anti- k_T clustering algorithm [60], with $R = 0.4$, as implemented in *fastjet* [61]. Only jets with $p_T \in [80, 230]$ GeV, $\eta_{\text{jet}} < 2.0$, and with at least two constituents were retained in the analysis. The approximate number of reconstructed jets is $\mathcal{O}(1.7 \times 10^6)$ for SO ($\mathcal{O}(10^6)$ vacuum, $\mathcal{O}(7 \times 10^5)$ medium) and $\mathcal{O}(1.9 \times 10^6)$ for MR-UE ($\mathcal{O}(10^6)$ vacuum, $\mathcal{O}(9 \times 10^5)$ medium).

Once reconstructed, a jet is a collection of the 4-momenta of its constituents. For this work, we chose to represent the 4-momentum of each constituent by its rapidity and azimuth ($\Delta\eta, \Delta\phi$) relative to the jet axis, its transverse momentum p_T , and its mass m .¹ A single jet is then the set $\{(\log(p_{T,i}/(1\text{GeV})), m_i, \Delta\eta_i, \Delta\phi_i)\}$ with $i = 1, \dots, M$ where M is the number of jet constituents. The use of the logarithm of the transverse momentum ensures that the values are nominally similar across all constituents, which is an important requisite for the neural network. While this representation cannot be obtained by a Lorentz transformation of a physical 4-momenta, it retains all the same information.²

Each jet is represented as a matrix $\mathbf{X} \in \mathbb{R}^{M \times 4}$ with $\mathbf{X} = X_{i\alpha}$, where $i = 1, \dots, M$ runs over the jet M constituents, and $\alpha = 1, \dots, 4$ runs over the four physical quantities.

For comparison purposes, we also computed the high-level observables, such as jet substructure observables, discussed in our previous work [37]. This comparison will allow us to assess how much information inside the jet is not being captured by observables usually employed in phenomenological and experimental studies. The dataset used for the machine learning training and the analysis presented herein is made available here [62].

¹For the mass we used the on-shell mass associated with the particle ID, not the one that *fastjet* derives, as this can lead to non-physical masses after the subtraction steps. More importantly, we observed that this leads to an artificial and non-physical large discrimination power for the MR-UE sample due to the JEWEL-specific subtraction prescription.

²We leave for future work how different handling of the 4-momenta, for example by using Lorentz algebra-aware architectures [53–55], can impact the analysis presented in this work.

3 Transformer and methodology

In previous work [34, 37, 39], we considered different machine learning classification algorithms which embodied different *inductive biases*, i.e. assumptions on how the data representation encodes the relevant information. For example, for jets represented as images, Convolutional Neural Networks were used, thus assuming a compositional hierarchy from local receptive fields to a global structure arising from their successive overlaps. In this work, we employ the *Transformer* block, initially proposed for Natural Language Processing [43] as a sequence-to-sequence model, which provides a *minimal inductive bias* as it intakes *sets* of vectors, with no other data assumption. We now review this architecture with a notation more customary to physicists.

The Transformer classifier used in this work, $\mathcal{T}$, can be seen as the composition of four different modules: initial input embedding, I ; N_T Transformer blocks, T ; pooling stage, P ; and a classifier, C :

$$\mathcal{T}(\mathbf{X}) = C \circ P \circ T^{N_T} \circ I(\mathbf{X}), \quad (1)$$

where $\mathbf{X}$ is a jet represented as the set of the 4-momenta of its constituents, and each module has its own trainable weights and biases, and $T^{N_T} = T \circ T \circ \dots$ represents a N_T stack of T blocks. The input embedding, I , is an affine projection that maps the momenta of the jet constituents to a higher dimensional space, i.e. $I : \mathbb{R}^4 \rightarrow \mathbb{R}^{d_R}$ acting on the second index of $\mathbf{X}_{i\alpha}$. We will represent this transformation as

$$I(\mathbf{X}) = \mathbf{X}w_I + b_I, \quad (2)$$

with weights $w_I \in \mathbb{R}^{4 \times d_R}$ and bias $b_I \in \mathbb{R}^{M \times d_R}$, where d_R is the *representation* dimension that we will keep the same for every module of Eq. 1 for simplicity. We note that b_I is the same for all jet constituents, i.e. its lines are all the same $b_{I,i\alpha} = \delta_{ii}b_{I,\beta}$ with $b_{I,\alpha} \in \mathbb{R}^{d_R}$ and its representation as a $M \times d_R$ matrix is for convenience.

This step allows the neural network to operate in a higher dimensional space, where it can abstract more complicated representations of the jet. The next module, the Transformer block, is the main ingredient of this architecture and is itself composed of two steps: the multihead self-attention, MHA , and a residual feed-forward, FF , block. The MHA module reads

$$MHA(I) = (\text{Concat}(\text{Att}(I)_1, \dots, \text{Att}(I)_{N_h}))W_O, \quad (3)$$

$$\text{Att}(I)_a = \text{Softmax} \left(\frac{1}{\sqrt{d_R}} W_{Q,a}(I) W_{K,a}(I)^T \right) W_{V,a}(I), \quad (4)$$

where $I = I(\mathbf{X})$, Concat is the concatenation operation along the second index, $\text{Softmax}(A)_{ij} = \exp(A_{ij}) / \sum_k \exp(A_{ik})$ normalises the entries along the second index, Att is a single self-attention head, N_h is the number of heads, $W_{V,Q,K,O}$ represent affine transformations of the form $W(A) = Aw + b$

with weights $w_{V,Q,K} \in \mathbb{R}^{d_R \times d_R}$, $w_O \in \mathbb{R}^{N_h d_R \times d_R}$, and biases $b_{V,Q,K,O} \in \mathbb{R}^{M \times d_R}$ where again the lines of the biases matrices are all the same just as in Eq. 2.³ The subscripts V, Q, K, O have historical names that are not relevant for this discussion, but are kept for bookkeeping. More importantly we note that the MHA module learns a dense adjacency matrix between all constituents of a jet, i.e. $W_Q(I)W_K(I)^T \in \mathbb{R}^{M \times M}$, which is then transformed non-linearly by the Softmax function. The resulting matrix has normalised rows and is then contracted with $W_V(I)$ (often called Scaled Dot-Product), producing a new representation of the jet where each constituent now has the information of the whole jet.⁴ The rest of the Transformer block involves a residual connection and a feed-forward module such as

$$res(I) = \text{LayerNorm}(MHA(I) + I), \quad (5)$$

$$T(I) = \text{LayerNorm}(FF(res(I)) + res(I)), \quad (6)$$

where LayerNorm normalises the inputs over the last index, i.e. the representation, and FF is a multi-layer perceptron with dimension d_R with one hidden layer with the ReLu non-linear activation function. The residual allows for deeper architectures as it prevents vanishing gradient problems during training and intuitively makes the transformer block T act as an incremental mapping around its inputs.

Two features of the Transformer block are worth discussion. The first one, is that it preserves the dimensionality of its inputs, i.e. $T : \mathbb{R}^{M \times d_R} \rightarrow \mathbb{R}^{M \times d_R}$, so one can stack N_T T blocks to form a deeper neural network that learns sequentially more complex relations between the jet constituents. Second, T is covariant under permutations of the jet constituents. If we consider a permutation matrix Π that permutes the constituents of the jet: $\Pi_{\pi(i)i} \mathbf{X}_{i\alpha} = \mathbf{X}_{\pi(i)\alpha}$, where π is some permutation over the index i , then $T(I(\Pi \mathbf{X})) = \Pi T(I(\mathbf{X}))$. This second feature is why Transformers operate optimally over sets of jet constituents.

The fact that the output of the Transformer block has the same dimensionality as its inputs means that the output representation has different dimensions for different jets, as M (the number of constituents) varies on a jet-by-jet basis. In order to perform a classification task over this $M \times d_R$ representation, we need to *pool* along the first index, i.e. over the jet constituents. We do this by defining a pooling module, P , which has the sole purpose of mapping the whole jet to a fixed length representation, $P : \mathbb{R}^{M \times d_R} \rightarrow \mathbb{R}^{d_R}$. We also

³In practice, each attention head, $\text{Att}(I)_a$, acts only on a d_R/N_h slice of the representation, such that the concatenation returns a d_R representation, making W_O unnecessary. Nonetheless, this is how the mechanism was first presented in [43].

⁴The fact that $W_Q(I)W_K(I)^T \in \mathbb{R}^{M \times M}$ means that the new jet representation is affected by the “context” of the rest of the jet. Recently there has been a great deal of research around the idea that Transformer capacity to outperform other architectures for similar tasks is due to this “in-context learning” that Transformers are capable of performing [63, 64].

notice that whereas T is permutation covariant, $P(T)$ is permutation invariant, and consequently the whole Transformer classifier $\mathcal{T}(\mathbf{X})$ will be permutation invariant. In this work we will let our hyperparameter optimisation loop choose between different pooling operations: mean, sum, max, and attention pooling.

Having pooled the whole jet representation into a vector of fixed size, we can perform the classification using C , a feed-forward multilayer perceptron with M_h number of hidden layers of size d_R , which maps the representation $P(T)$ into the probability of a jet being of class one, which we define as being the medium sample, i.e. $C : \mathbb{R}^{d_R} \rightarrow [0, 1]$.

We implemented the Transformer classifier in `pytorch` [65], using `lightning` [66]. Complementarily, a gradient boosted classifier (BDT), implemented with `XGBoost` [67], was trained on the same task using the high level observables considered in [37]. This set of observables includes only observables that return a single value per jet, spanning both global jet properties and jet substructure. In [Appendix A](#) we provide explicit details of the observables that are relevant for the discussion in this work. The inclusion of this classifier will allow us to compare the Transformer classifier performance on low-level jet constituent data against high-level jet data. The details of the hyperparameter optimisation for both algorithms are presented in [Appendix B](#)

4 Classification Results

We now discuss the discrimination performance of the classifiers trained on separating vacuum (proton-proton) from medium (Pb-Pb) samples. We start by discussing the SO case, where neither QGP response nor UE contamination are accounted for. This will allow us to focus the discrimination task on modifications of the jet fragmentation history alone.

In [Fig. 1](#) we show the Receiver Operator Characteristic (ROC) curve and the Area Under the Curve (AUC), for both the Transformer and the BDT classifiers trained on the SO dataset. A very striking observation is the almost complete overlap between the two, with each ROC curve yielding an AUC of 0.70 at two significant digits. This result suggests that our Transformer is not capturing any further information beyond that encoded in the high-level observables studied in [37].

The virtually complete overlap of the ROC curve implies that both discriminants share the same quantiles, suggesting a correlation between them. This is further demonstrated in [Fig 2](#) where we show the distributions of the Transformer and BDT outputs for the vacuum and medium samples of the SO dataset. The correlation is nearly perfect, with both discriminants agreeing to high degree of correlation what is a vacuum and a medium jet. This reinforces the interpretation that there is no novel information inside the jet to be

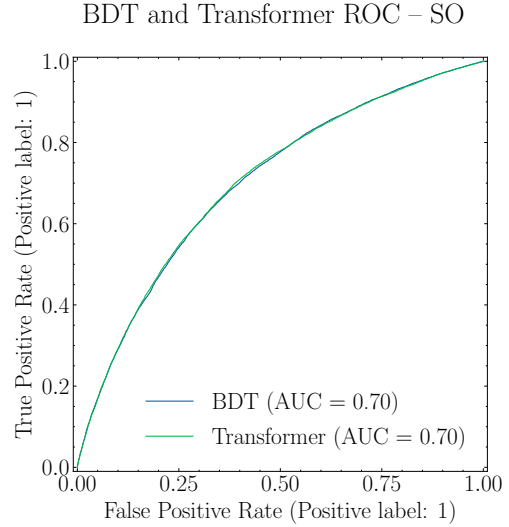


Fig. 1 ROC curves and respective AUC for the BDT (Blue) and Transformer (Green) for SO scenario.

captured by the Transformer that the high-level discriminants have not already used to determine whether the jet has experienced any modification by interaction with the QGP medium.

We now turn to the MR-UE dataset, where the inclusion of both medium response and confounding effects from the imperfectly subtracted underlying event presents a scenario closer to experimental reality. While, on the one hand, the embedding in and subtraction of the underlying event might hide the ‘signal’, making it harder to discriminate between vacuum and medium jets, the presence of medium response within the reconstructed jets should make medium jets less like vacuum (i.e., unmodified) jets [68] and facilitate the discrimination task.

In [Fig. 3](#) we show the ROC and respective AUC for the Transformer and the BDT classifiers. Interestingly, while the ROC AUC for the BDT is similar to the SO case, at 0.70, the picture is fundamentally different for the Transformer, yielding a ROC AUC of 0.77. This provides strong evidence that there is extra discriminating information inside the jet arising from the medium response not captured by the high-level observables that the BDT had access to during training.

Since the discriminating power of the Transformer and the BDT on the MR-UE dataset is so different, these discriminants cannot be strongly correlated. In [Fig. 4](#) we show the correlation between the output distributions of these two discriminants for both samples in the MR-UE dataset. The most interesting feature of these plots is the accumulation of medium jets with Transformer prediction very close to 1.0 (i.e., classified as medium), while the BDT prediction is evenly spread out in $[0.5, 1.0]$. The second interesting observation is when we look at the remaining distributions if we ignore this modality. In this case, we observe that the Trans-

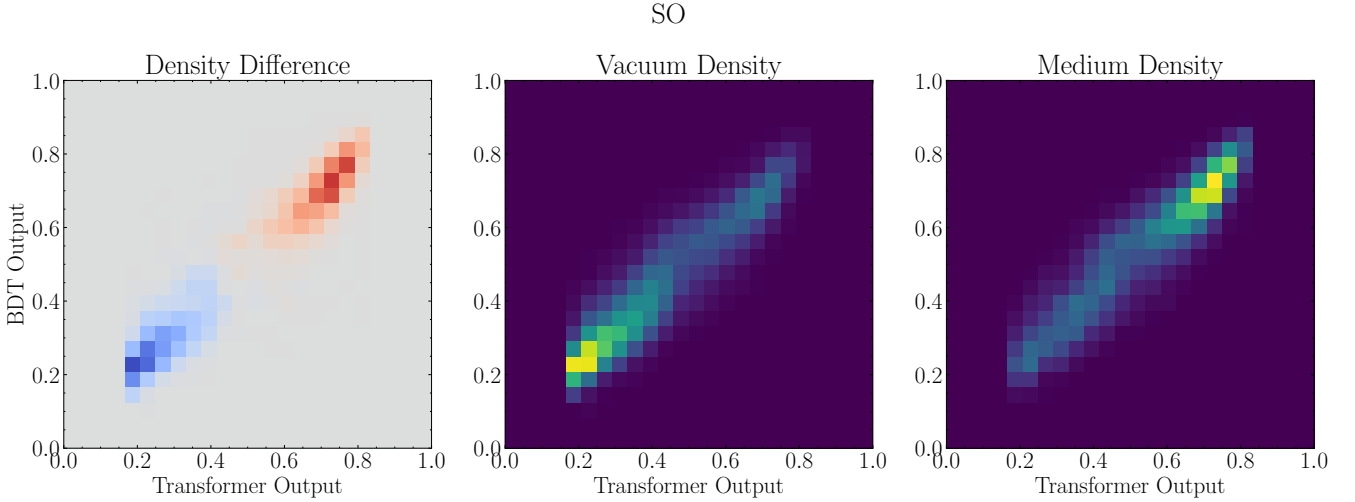


Fig. 2 BDT and Transformer predictions for the SO scenario. Left: Difference of the 2d histogram density between Vacuum and Medium samples, with Blue (Red) showing a greater density of Vacuum (Medium) density. Middle: The 2d histogram density for the Vacuum sample. Right: The 2d histogram density for the Medium sample.

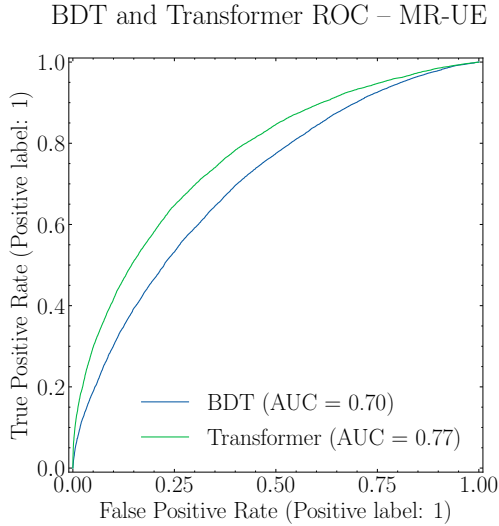


Fig. 3 ROC curves and respective AUC for the BDT (Blue) and Transformer (Green) for MR-UE scenario.

former and the BDT distributions for the vacuum sample and the remainder of the medium sample are highly correlated, similarly to how they were for the SO case. This suggests that in this regime, the medium response is not meaningfully modifying the jets, and we recover fragmentation patterns that are similar to the SO case. Conversely, our methodology suggests that it is possible to isolate with a degree of purity jets that have medium response modifications.

Given the difference in discrimination between the Transformer and the BDT for the MR-UE scenario, it is important to assess what could the Transformer be learning that the BDT could not. While Neural Networks are often perceived as black-box functions allowing little interpretability of their predictions, the explicit expressions for the operations car-

ried out inside the Transformer allow us to gather significant insight.

We start with the crucial ingredient of the Transformer architecture: the dense adjacency matrix produced by the attention mechanism Eq. 4. We see that for the first transformer block we have

$$\begin{aligned} W_{Q,K}(I(\mathbf{X})) &= (\mathbf{X}w_I + b_I)w_{Q,K} + b_{Q,K} \\ &= \mathbf{X}\tilde{w}_{Q,K} + \tilde{b}_{Q,K}, \end{aligned} \quad (7)$$

where $\tilde{w}_{Q,K} = w_I w_{Q,K}$, $\tilde{b}_{Q,K} = b_I w_{Q,K} + b_{Q,K}$, which is nothing else than the fact that the composition of two affine transformations is an affine transformation. This leads to the adjacency matrix

$$\begin{aligned} W_Q(I)W_K(I)^T &= (\mathbf{X}\tilde{w}_Q + \tilde{b}_Q)(\mathbf{X}\tilde{w}_K + \tilde{b}_K)^T \\ &= \mathbf{X}\tilde{w}_Q\tilde{w}_K^T\mathbf{X}^T + \mathbf{X}\tilde{w}_Q\tilde{b}_K^T + \tilde{b}_Q\tilde{w}_K^T\mathbf{X}^T + \tilde{b}_Q\tilde{b}_K^T, \end{aligned} \quad (8)$$

or, in other words, we obtain a second order polynomial in the jet constituents. We also notice that the first term is closely related to the variance matrix between jet constituents and it can be seen as the learnable pair-wise measurement of relation between the jet constituents. Next, we expand the Softmax operation, which for small arguments reads

$$\text{Softmax}(A_{ij}) \simeq (1 + A_{ij}) / \sum_{k=1}^M (1 + A_{ik}) \simeq (1 + A_{ij}) / M.$$

This represents a re-scaling of the entries of this matrix, and therefore the output of the MHA block is schematically

$$MHA(I) \simeq P^2(I)I = P^3(\mathbf{X}), \quad (9)$$

where, to first approximation, P^i is an i^{th} order polynomial of the $\mathbf{X}$ entries. Now we note that the outputs of MHA take

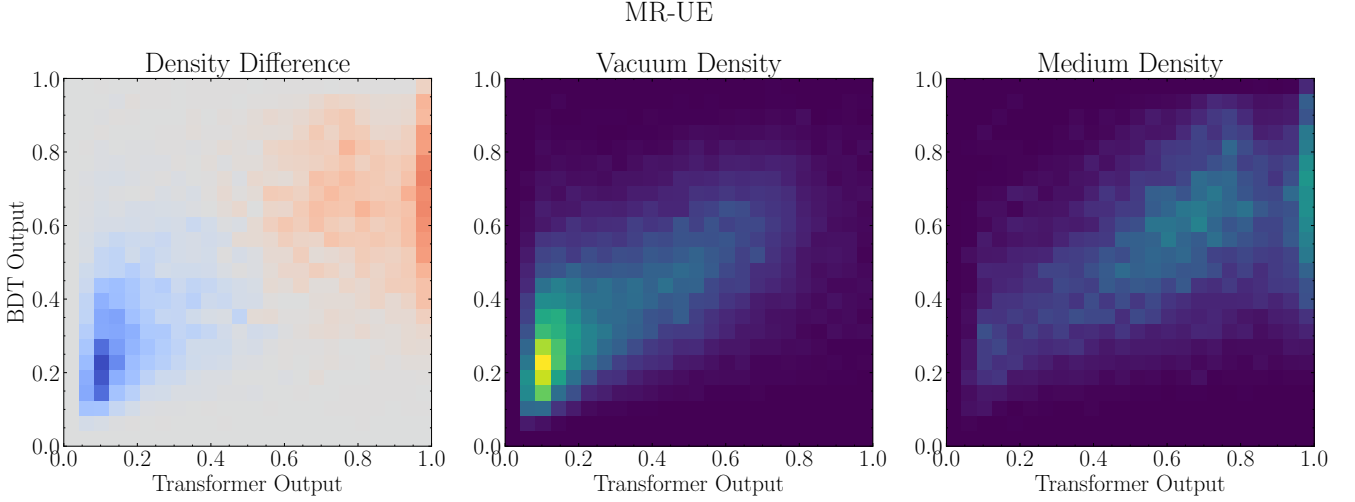


Fig. 4 BDT and Transformer predictions for the MR-UE scenario. Left: Difference of the 2d histogram density between Vacuum and Medium samples, with Blue (Red) showing a greater density of Vacuum (Medium) density. Middle: The 2d histogram density for the Vacuum sample. Right: The 2d histogram density for the Medium sample.

values close to 0 due to the smallness of w and the inputs. Therefore, the LayerNorm operation is another $O(1)$ rescaling, and we can approximate the entire first Transformer block, $T(I(\mathbf{X}))$, as

$$T(I) \simeq I + MHA(I) + FF(I + MHA(I)). \quad (10)$$

Collecting the terms in powers of I dependence, we have

$$T(I(\mathbf{X})) \simeq P^1(I(\mathbf{X})) + P^2(I(\mathbf{X}))I(\mathbf{X}) \simeq P^3(\mathbf{X}), \quad (11)$$

and therefore, at leading order, each Transformer block outputs a cubic polynomial of its inputs. We notice, however, that the true expressive power of the Transformer block is nonetheless greater, as beyond leading order the coefficients of the polynomial are non-linear functions of the entries themselves due to the Softmax and the non-linear activation function of the FF step.

In this work, the best hyperparameters for the Transformer suggested a maximum needed number of stacked Transformer blocks for both datasets at $N_T = 3$. This means that, intuitively, the transformer classifiers studied in this work have operated on a polynomial of approximate degree $3^3 = 27$, which is considerably higher than the order needed to produce the high level observables used to train the BDT as detailed in [37]. This could explain why the BDT was not able to capture the medium response in the MR-UE case and hint at jet substructure observables involving correlations of high order between jet constituents being needed in order to capture medium response. However, for the SO dataset we obtained a high correlation between the Transformer and the BDT, which suggests that it is likely that the optimisation loop for the Transformer ‘exaggerated’ the need for a deeper architecture when a shallower one could have worked as well. This is expected to a certain extent, as

machine learning workflows tend to produce large models that are trained with regularisation that reduce the effective functional representational capacity – in our case dropout and early stop. Therefore, the degree 27 should be read as an approximation to the upper bound on the required degree and not as the necessary degree of a polynomial over the jet constituents to perform this classification task.

5 Estimating the Fraction of Vacuum in Medium

The analysis we carried out in the preceding section highlights the difficulty in isolating jets modified by the QGP as vacuum-like jets still represent an irreducible component of the medium sample. This is evident by the non-existence of a classification regime where the False Positive Rate can be brought to zero while keeping a non-zero True Positive Rate, which would be the required cut to generate a pure sample of modified jets.

A question that arises is how much of the medium sample is effectively vacuum-like, i.e. not modified enough, or at all, by interaction with the QGP. A crude estimate of the fraction of jets that is not modified enough by QGP can be made on the basis on jet suppression alone. Given the steepness of the jet spectrum, jets in PbPb collisions (the medium sample) with a given p_T are roughly those that either did not lose any energy or lost very little [27, 29]. Jets that would have lost a lot of energy, that is that had a substantially higher initial energy, are suppressed by the steepness of the jet spectrum and contribute only marginally to the number of jets found at a given p_T . As such, the value of the jet nuclear modification factor R_{AA} (for the 0 – 10% centrality class $R_{AA} \sim 0.4 \div 0.5$ for all considered p_T) is a rough upper bound for the fraction of jets that are vacuum-like.

More detailed information like that contained both in the observables we considered for BDT training and in the full set of 4-momenta we used as input for the Transformer should lead to an improved bound. It is conceivable that jets are modified in ways, e.g. in their substructure and/or correlations among constituents, without an energy loss large enough to modify the R_{AA} .

Here we will use the discriminants discussed above to derive upper limits on the fraction of vacuum-like jets in the medium sample, or, more colloquially, the fraction of vacuum in the medium. We use a simplified form of topic modelling, similar to that used to discriminate between quark and gluon initiated jets [69–74], towards extraction of the strong coupling [75], for four-top searches [76], and to probe the different interaction of quarks and gluons with QGP [77].

For any discriminant $\mathcal{D}$ we define its distribution for the medium sample as $\mathcal{M}_{\mathcal{D}}(x)$, where x is valued over the domain of $\mathcal{D}$. The shape of this discriminant will change if the amount of modifications to the jet by the medium becomes more or less pronounced, for example by the medium temperature, density, etc. We formally parametrise, but otherwise do not specify, the quenching-inducing quantities as $\{q\}$. Therefore, for a given physical system that produces QGP medium through which the jets traverse and can interact with, we have

$$\mathcal{M}_{\mathcal{D}}(x) = \mathcal{M}_{\mathcal{D}}(x|\{q\}), \quad (12)$$

which is just a way of formally parametrising the ‘amount’ of quenching that the medium can produce.

For a SO-like scenario, jet modifications produce a distributional migration of the different observables along an unidimensional trajectory [37]. Therefore, we take the simplification that there is effectively only one degree of freedom parametrising quenching, i.e. $\{q\} \rightarrow q$. With this framing, we can define the unquenched jets, u , as the contribution to $\mathcal{M}_{\mathcal{D}}$ that has not experienced any modification. In other words, the unquenched jets are those that would have been produced in the limit $q \rightarrow 0$, which leads to

$$\mathcal{M}_{\mathcal{D}}(x|q) = f_q \mathcal{M}_{\mathcal{D}}(x|q \neq 0) + f_u \mathcal{M}_{\mathcal{D}}(x|q = 0), \quad (13)$$

where the relative fractions respect $f_q + f_u = 1$. It is clear now that we should identify the unquenched distribution to the one we would obtain in the absence of medium, i.e. the vacuum, $\mathcal{M}_{\mathcal{D}}(x|q = 0) \simeq \mathcal{V}_{\mathcal{D}}(x)$, and the distribution of the quenched jets as $\mathcal{Q}_{\mathcal{D}}(x) = \mathcal{M}_{\mathcal{D}}(x|q \neq 0)$. When considering experimental data, this only holds when PbPb underlying event contamination is also accounted for in the pp sample [39]. We have then

$$\mathcal{M}_{\mathcal{D}}(x) \simeq f_q \mathcal{Q}_{\mathcal{D}}(x) + f_v \mathcal{V}_{\mathcal{D}}(x), \quad (14)$$

where we renamed $f_u = f_v$ for consistency.

Consider now the histogram approximation of $\mathcal{D}(x)$ using N bins $\mathcal{H}_{\mathcal{D},N} = \{h_{\mathcal{D}}[k]\}$, with $k = 1, \dots, N$ where $h_{\mathcal{D}}[k]$

are the bin heights and with edges over the domain of x . Each bin height, $h_{\mathcal{D}}[k]$, results from two contributions: the vacuum, $h_{\mathcal{D},v}[k]$, and the quenching, $h_{\mathcal{D},q}[k]$, such that

$$h_{\mathcal{D}}[k] = f_q h_{\mathcal{D},q}[k] + f_v h_{\mathcal{D},v}[k]. \quad (15)$$

Given that both $\mathcal{Q}$ and $\mathcal{V}$ are valid distributions generated by different physical phenomena, they need to be non-negative. Immediately, this means that

$$h_{\mathcal{D},q}[k] = \frac{h_{\mathcal{D}}[k] - f_v h_{\mathcal{D},v}[k]}{f_q} = \frac{h_{\mathcal{D}}[k] - f_v h_{\mathcal{D},v}[k]}{1 - f_v} \geq 0 \quad (16)$$

for all bins as long as $f_q \neq 0$, which is a reasonable assumption. Therefore, we can set a maximum bound on f_v , the fraction of vacuum, computed using the discriminant $\mathcal{D}$ by saturating the inequality, i.e.

$$f_{\mathcal{D},v}^{\max} = \min_{k: h_{\mathcal{D},v}[k] > 0} \frac{h_{\mathcal{D}}[k]}{h_{\mathcal{D},v}[k]}. \quad (17)$$

The dependence on the discriminant is crucial and will significantly affect our upper bound estimate. To see this, we notice that if $h_{\mathcal{D},v}$ and $h_{\mathcal{D},q}$ are highly overlapping, then they could both account for the shape of $h_{\mathcal{D}}$. This would make any value of f_v admissible and the minimisation step could overestimate the quantity of f_v as being the unique component. Consequently, the upper bound on the fraction of vacuum in the medium sample is more stringent for discriminants that have more pronounced differences for vacuum and quenched jets. Therefore, the machine learning discriminants developed in the previous section are expected to produce the lower upper bounds for the fraction of vacuum in the medium.

Another important aspect of this estimate is the definition of the bins, specifically regarding both their number and their edges. Naively, one might partition the discriminant into bins of equal width. However, this approach poses the challenge of each bin having a very different number of events, which could result in less populated bins leading to erroneous conclusions due to statistical fluctuations. Moreover, we observe that partitioning the discriminant into too many bins exacerbates this issue; thus, the number of bins cannot be arbitrarily large relative to the given number of simulated (or expected) events.⁵ We address the former problem by constructing bins based on (weighted) quantiles of the medium sample. This ensures that each bin has an equal expected yield, making them relatively robust against statistical fluctuations. With this solution, the latter problem is approached by seeking a regime where the vacuum fraction upper bound remains relatively stable against changes in the number of bins.

⁵Of course in the limit of infinite data this would produce a continuous approximation of the true distribution, but such limit is not possible in practice with a finite dataset.

In Fig. 5 (left) we show how the upper bound on the fraction of vacuum changes with the number of bins for the ten discriminants. In addition to the Transformer and the BDT we consider the high-level observables from [37] that produce the lowest upper bounds. We see that from 100 bins onwards, the value of $f_{\mathcal{D},v}^{\max}$ stabilises. On the right panel we fix the number of bins to be 100 and show how $f_{\mathcal{D},v}^{\max}$ gets smaller as the discriminating power of $\mathcal{D}$ increases, as expected.

The results in Fig. 5 indicate that the maximum fraction of vacuum-like jets in the medium sample is just under 0.3. The discriminant that yields the strongest constraint is the BDT output, although it is very similar to that of the Transformer. In Fig. 6 we visually illustrate this by defining the approximate distribution of quenched jets as

$$\hat{Q}_{\mathcal{D}} = \frac{\mathcal{M}_{\mathcal{D}} - f_{\mathcal{D},v}^{\max} \mathcal{V}_{\mathcal{D}}}{1 - f_{\mathcal{D},v}^{\max}}. \quad (18)$$

For the MR-UE case, medium response is a distinctive feature for jets in the medium sample. It is noteworthy that this feature is not captured by the high-level jet observables considered here. In Fig. 7, the left panel shows that both machine learning discriminants produce significantly stricter bounds on $f_{\mathcal{D},v}^{\max}$ that, again, stabilise at around 100 bins. We fix this number of bins for the right panel, where we show the relation between $f_{\mathcal{D},v}^{\max}$ and the discrimination power of different observables. We observe a significant degradation, with respect to the SO case, of the single observable discriminating power with no single observable leading to a ROC AUC above 0.6. This degradation results on significantly poorer bounds for $f_{\mathcal{D},v}^{\max}$ which now are now consistently less stringent than those obtained from both the BDT and the Transformer. While the enhanced sensitivity of the Transformer to medium response leads to the most stringent bound on $f_{\mathcal{D},v}^{\max}$ of just under 0.2, this bound is not significantly better than yielded by the BDT. Both are significantly lower than that obtained in the SO case, confirming the role of medium response as a discriminating feature. In Fig 8 we can observe how the approximate quenching distribution emerges for different number of bins. Of especial relevance, we see the large accumulation of modified jets at the very end of the output, which are the jets that were not so confidently classified by the BDT. We see that this last bin is almost pure in jets from the medium sample, as previously seen in Fig 4. We also note that, if we ignore the right-most bins, the distributions for vacuum and medium are very similar to the ones obtained for SO in Fig. 6.

6 Conclusions

The similar performance of our Transformer and the BDT, when QGP response and underlying event contamination are

neglected, points towards the inability of the transformer to learn any discriminating information beyond that captured by the high-level observables. While it is conceivable that such discriminating information is simply absent, we caution that several factors can limit the performance of the transformer and merit detailed study in future work: (i) transformer scaling laws [78–81] imply that very significant improvement could follow from a larger training data set; and (ii) our choice of parametrization for the 4-momentum of jet constituents, built very similarly to what is used to compute high-level observables, could introduce a bias and limit what the transformer can effectively learn. This limitation could be superseded by considering different representation bases or physics-aware architectures [53–55].

When QGP response and underlying event are included, bringing our simulation results closer to experimental reality, the situation is qualitatively different. In this case, the presence of QGP response provides critical discriminating information, distinguishable from underlying event contamination, that is captured by the transformer and only marginally by the BDT. We argued that this follows from the ability of the transformer to learn particle correlations beyond the 2-particle correlations that dominate the observables used to train the BDT. The discriminating power of multi-particle observables, in the form of Energy Flow Polynomials [82], has been explicitly verified recently [41]. A salient feature of the classification performed by the transformer is the separation of the sample jets in PbPb into two clear classes: one that is unequivocally distinguishable from vacuum jets (with outputs in the near vicinity of 1), and another with outputs with significant overlap with those from pp (vacuum jets). This is directly attributable to the discriminating power of QGP response as the distribution of the transformer output for jets, not clearly identified as modified, is similar for both cases where QGP response is present and absent.

The similarity between the fraction of vacuum-like jets in PbPb extracted from the Transformer and BDT outputs is worth comment. While this was to be expected when neglecting medium response and underlying event contamination as the outputs of the networks are closely correlated, the superior discriminating power of the transformer could be seen as implying a lower bound for the fraction of vacuum-like jets in the medium sample. The procedure we used to determine $f_{\mathcal{D},v}^{\max}$ in Eq. 17 relies on the maximal excess of the vacuum output distribution with respect to the medium one. In other words, it relies on what the discriminants identify as vacuum-like jets. As the outputs of both the transformer and the BDT in the vicinity of 0 are highly correlated even in the presence of QGP response and underlying event, see Fig. 4, it is a natural that the inferred fraction of vacuum-like jets is similar in both cases.

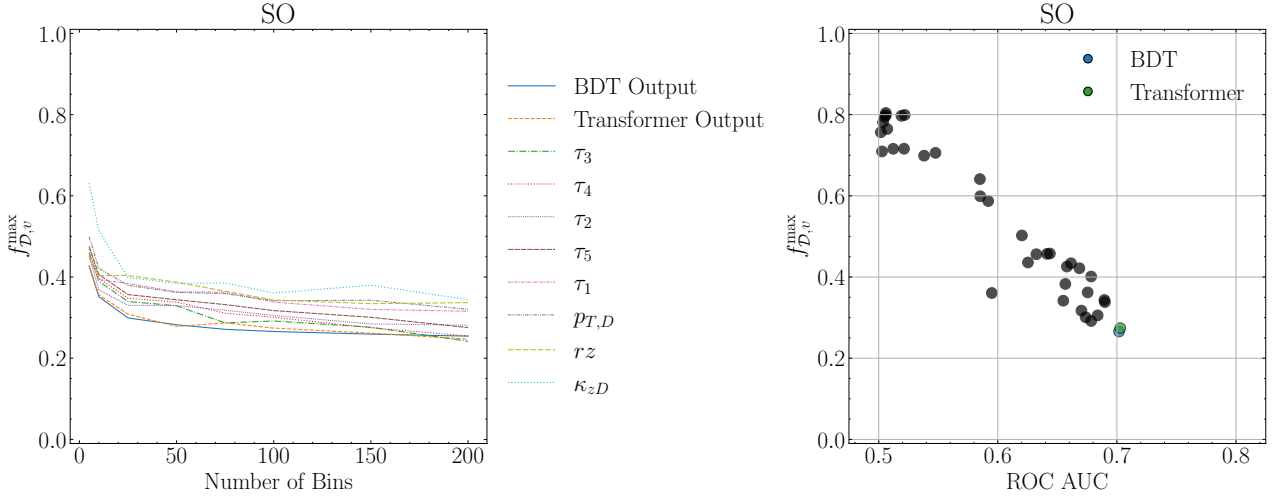


Fig. 5 Left: the evolution of $f_{D,v}^{\max}$ for the ten stringiest discriminants with the number of bins. Right: dependence of $f_{D,v}^{\max}$ on the discriminant discriminating power, as measured by the ROC AUC, for 100 bins for all discriminants. Both plots for the SO scenario.

Acknowledgements This work is part of a project that has received funding from the European Research Council(ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 835105, YoctoLHC). We acknowledge further support from Fundação para a Ciência e a Tecnologia (FCT), under ERC-PT A-Projects ‘Unveiling’, financed by PRR, NextGenerationEU (JGM) and under contract PRT/BD/151554/2021 (JAG). MCR is supported by the STFC under Grant No. ST/T001011/1.

Appendix A: Jet Observables

For convenience, we briefly summarise the jet observables that are discussed in this work. A thorough study of these, and other, observables in the context of Vacuum-Medium discrimination was carried out in [37], which we refer the reader for further details.

Angularities The first set of observables is composed by generalise angularities [83], which are moments of the jet constituent distributions around the jet axis

$$\lambda_{\beta}^{\kappa} = \sum_{i \in \text{jet}} z_i^{\kappa} R_{i,jet}^{\beta}, \quad (\text{A.1})$$

where $z_i = p_{T,i}/p_{T,jet}$ is the fraction of the transverse momentum carried by the constituent i , and $R_{i,jet}$ is the angular distance to the jet axis, i.e.

$$R_{i,j} = \sqrt{(y_i - y_j)^2 + (\phi_i - \phi_j)^2}, \quad (\text{A.2})$$

therefore, for $\beta \neq 0$, the angularities measure the transverse distribution of jet constituents.

While κ and β can be any non-negative integer, the relevant angularities for this work are the momentum dispersion, $p_{T,D}$, [84], which reads

$$p_{T,D} = \frac{\sqrt{\sum_{i \in \text{jet}} p_{T,i}^2}}{p_{T,jet}} = \sqrt{\lambda_0^2}, \quad (\text{A.3})$$

and its mean over the number of jet constituents, n_{const} ,

$$\bar{z}^2 = \frac{1}{n_{\text{const}}} \lambda_0^2 = \frac{1}{n_{\text{const}}} \sum_{i \in \text{jet}} z_i^2. \quad (\text{A.4})$$

We notice that for $\kappa \neq 1$ the generalised angularities are not IRC. Two IRC angularities appearing in this work are

$$rz = \lambda_1^1 \quad (\text{A.5})$$

$$r^2 z = \lambda_2^1. \quad (\text{A.6})$$

N -Subjettiness Another set of observables relevant to this work are the N -Subjettiness observables [85]. These also capture the transverse properties of the jet and quantify how dissimilar is the jet from a collection of N subjects. They read

$$\tau_N = \frac{\sum_{i \in \text{jet}} p_T^i \min(R_{1,i}, \dots, R_{N,i})}{R_0 p_{T,jet}}, \quad (\text{A.7})$$

with R_0 the jet clustering radius.

Grooming Derived The final type of observables relevant to this work are those derived from grooming. For example, in SoftDrop, [86], the C/A reclustering jet branching history is recursively declustered, rejecting the softest branch, until the the following condition is met

$$\frac{\min[p_{T,i}, p_{T,j}]}{p_{T,i} + p_{T,j}} > z_{\text{cut}} \left(\frac{R_{i,j}}{R_0} \right)^{\beta}, \quad (\text{A.8})$$

where i, j are the indices of the daughters of the branch, $R_{i,j}$ the angular distance between them, R_0 the jet radius of the initial clustering, and z_{cut} and β parameters governing the grooming. From this, one can derive three observables at the branching where the above condition is met. Let $p_{T,2}$ ($p_{T,1}$) be the softer (harder) branch at the branch splitting where the above condition is met, we then have

Fraction Study – SO – Transformer Output

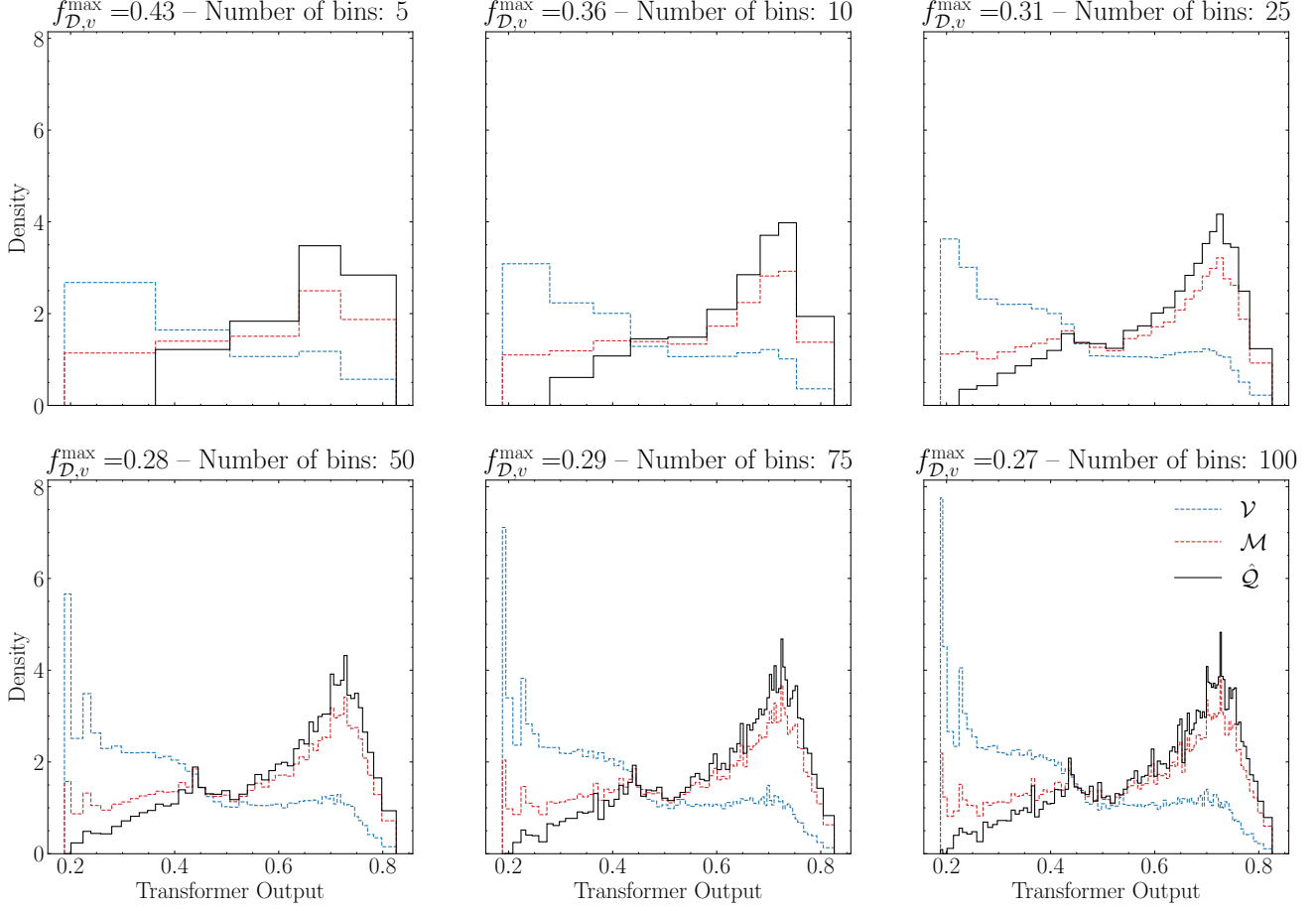


Fig. 6 $f_{\mathcal{D},v}^{\max}$ for the Transformer on the SO scenario for different number of bins. In red (blue) dashed lines we have the medium (vacuum) distributions. In black full line we have the estimated quenched distribution according to Eq. 18.

- $z_g = p_{T,2}/(p_{T,1} + p_{T,2})$, i.e. the fraction of transverse momentum contained in the softer branch,
- $R_g = R_{1,2}$ the angular separation of two branches,
- n_{SD} how many times the condition failed before it was met.

An alternative is to perform dynamical grooming, [87], which selects the first C/A reclustering sequence branch satisfying

$$\kappa^{(a)} = \frac{1}{p_{T,jet}} \max_{i \in \text{C/A seq}} \left[z_i (1 - z_i) p_{T,i} \left(\frac{R_{i,j}}{R_0} \right)^a \right], \quad (\text{A.9})$$

with a a free parameter. Different values of a capture different histories,

- TimeDrop (TD): $a = 2$
- k_T -Drop (ktD): $a = 1$
- z -Drop (zD): $a = 0$

and we refer to [37, 87] for more details. The relevant part for this work is that, for a given choice of a , one can derive

the corresponding R_g , z_g , and $\kappa^{(a)}$ at the first branching that satisfies the grooming condition.

Appendix B: Hyperparameter Optimisation

The values for the hyperparameters were chosen using Bayesian optimisation implemented by Optuna [88], with the prior $N_h = 1, \dots, 8$, $d_R = 4, \dots, 128$, $N_T = 1, \dots, 8$, and $M_h = 0, \dots, 2$, dropout rate in $[0.0.5]$, and maximal learning rate in $\{10^{-5}, 3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$. The remaining hyperparameters were set to default. We used Adam optimizer with a cosine annealing learning rate with warm-up set to the first 10 epochs. The classifier was trained against the binary cross-entropy loss function between the vacuum and the medium classes, with early stop if no improvement was seen after 50 epochs. When stopped, the weights of the best epoch were restored and the Area Under the Curve of the Receiver Operating Characteristic (ROC AUC) was computed on the validation set. After 100 hyperparameter opti-

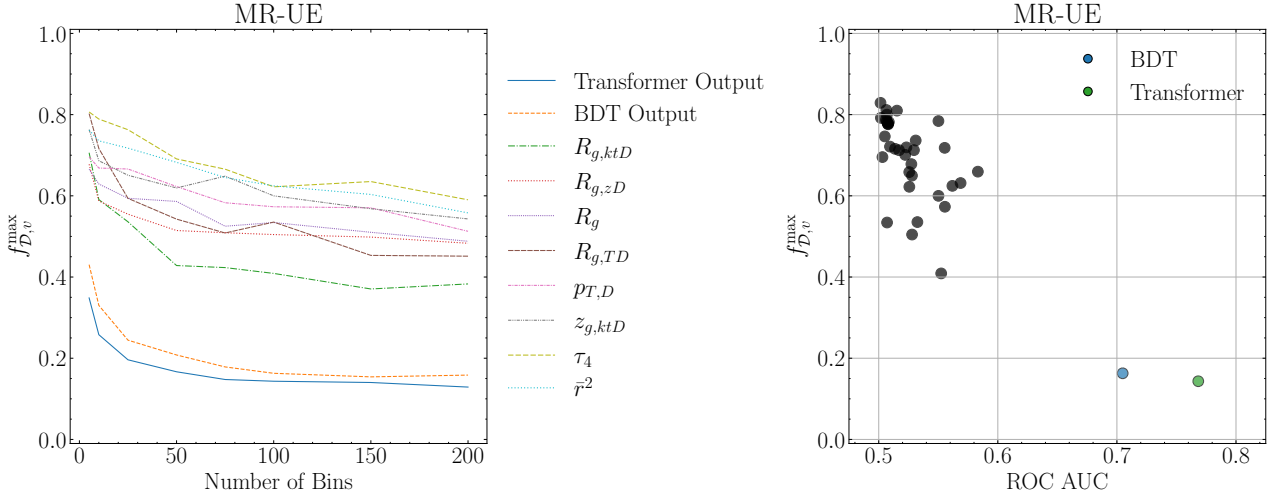


Fig. 7 Left: the evolution of $f_{D,v}^{\max}$ for the ten stringiest discriminants with the number of bins. Right: dependence of $f_{D,v}^{\max}$ on the discriminant discriminating power, as measured by the ROC AUC, for 100 bins for all discriminants. Both plots for the MR-UE scenario.

misation trials, we trained separately for the SO and the MR-UE physics cases. In both cases, the Monte Carlo generation statistical weights returned by JEWEL were used during training to ensure the correct statistics of the loss function, with an overall class reweighing imposed to guarantee the same statistical contribution from both classes during training.

The best combination of hyperparameters for the Transformer trained on the SO dataset was $N_h = 3$, $d_R = 5$, $N_T = 3$, $M_h = 1$, dropout rate at 0.3, maximal learning rate of 10^{-3} , and attention pooling. For the MR-UE case, the best hyperparameters were $N_h = 1$, $d_R = 109$, $N_T = 3$, $M_h = 1$, dropout rate at 0.2, maximal learning rate of 3×10^{-4} , and attention pooling. The hyperparameters of the BDT were also optimised to guarantee maximal performance, but the details of this step are omitted.

References

1. A. Majumder, M. Van Leeuwen, Prog. Part. Nucl. Phys. **66**, 41 (2011). DOI 10.1016/j.ppnp.2010.09.001
2. Y. Mehtar-Tani, J.G. Milhano, K. Tywoniuk, Int. J. Mod. Phys. A **28**, 1340013 (2013). DOI 10.1142/S0217751X13400137
3. J.P. Blaizot, Y. Mehtar-Tani, Int. J. Mod. Phys. E **24**(11), 1530012 (2015). DOI 10.1142/S021830131530012X
4. L. Apolinário, Y.J. Lee, M. Winn, Prog. Part. Nucl. Phys. **127**, 103990 (2022). DOI 10.1016/j.ppnp.2022.103990
5. L. Apolinário, Y.T. Chien, L. Cunqueiro Mendez, Int. J. Mod. Phys. E **33**(07), 2430003 (2024). DOI 10.1142/9789811294679_0002
6. E. Wang, X.N. Wang, Phys. Rev. Lett. **89**, 162301 (2002). DOI 10.1103/PhysRevLett.89.162301
7. N. Armesto, C.A. Salgado, U.A. Wiedemann, Phys. Rev. Lett. **93**, 242301 (2004). DOI 10.1103/PhysRevLett.93.242301
8. Y. He, L.G. Pang, X.N. Wang, Phys. Rev. Lett. **125**(12), 122301 (2020). DOI 10.1103/PhysRevLett.125.122301
9. J. Barata, A.V. Sadofyev, C.A. Salgado, Phys. Rev. D **105**(11), 114010 (2022). DOI 10.1103/PhysRevD.105.114010
10. J. Barata, X. Mayo López, A.V. Sadofyev, C.A. Salgado, Phys. Rev. D **108**(3), 034018 (2023). DOI 10.1103/PhysRevD.108.034018
11. J. Barata, J.G. Milhano, A.V. Sadofyev, Eur. Phys. J. C **84**(2), 174 (2024). DOI 10.1140/epjc/s10052-024-12514-1
12. J. Barata, C.A. Salgado, J.M. Silva, JHEP **12**, 023 (2024). DOI 10.1007/JHEP12(2024)023
13. J. Bahder, H. Rahman, M.D. Sievert, I. Vitev, (2024)
14. W. Ke, J. Terry, I. Vitev, JHEP **07**, 097 (2025). DOI 10.1007/JHEP07(2025)097
15. J. Barata, C.A. Salgado, J.M. Silva, (2025)
16. L. Apolinário, J.G. Milhano, G.P. Salam, C.A. Salgado, Phys. Rev. Lett. **120**(23), 232301 (2018). DOI 10.1103/PhysRevLett.120.232301
17. L. Apolinário, A. Cordeiro, K. Zapp, Eur. Phys. J. C **81**(6), 561 (2021). DOI 10.1140/epjc/s10052-021-09346-8
18. M. Attems, J. Brewer, G.M. Innocenti, A. Mazeliauskas, S. Park, W. van der Schee, U.A. Wiedemann, JHEP **01**, 080 (2023). DOI 10.1007/JHEP01(2023)080
19. J. Brewer, W. van der Schee, U. Wiedemann, (2025)
20. L. Apolinário, P. Guerrero-Rodríguez, K. Zapp, Eur. Phys. J. C **84**(7), 672 (2024). DOI 10.1140/epjc/s10052-024-13048-2

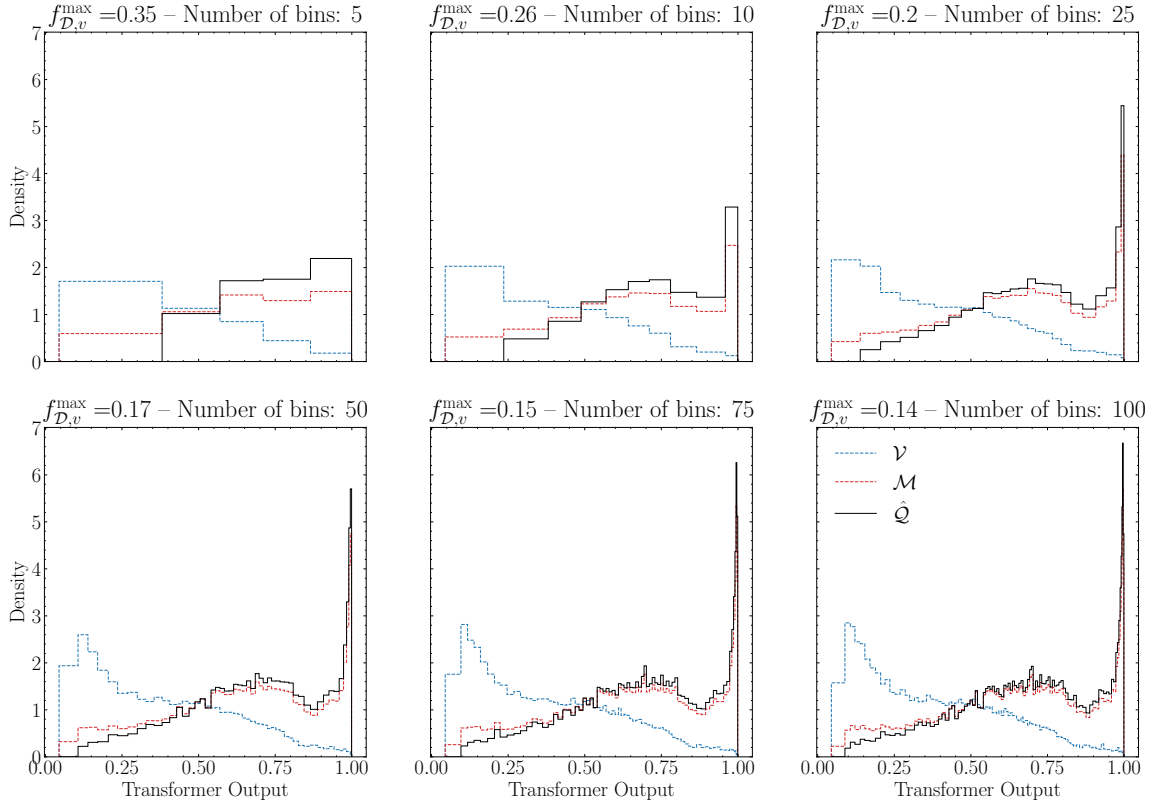


Fig. 8 $f_{\mathcal{D},v}^{\max}$ for the Transformer on the SO scenario for different number of bins. In red (blue) dashed lines we have the medium (vacuum) distributions. In black full line we have the estimated quenched distribution according to Eq. 18.

21. M. Connors, C. Nattrass, R. Reed, S. Salur, *Rev. Mod. Phys.* **90**, 025005 (2018). DOI 10.1103/RevModPhys.90.025005
22. L. Cunqueiro, A.M. Sickles, *Prog. Part. Nucl. Phys.* **124**, 103940 (2022). DOI 10.1016/j.pnpnp.2022.103940
23. J.G. Milhano, K.C. Zapp, *Eur. Phys. J. C* **76**(5), 288 (2016). DOI 10.1140/epjc/s10052-016-4130-9
24. K. Rajagopal, A.V. Sadofyev, W. van der Schee, *Phys. Rev. Lett.* **116**(21), 211603 (2016). DOI 10.1103/PhysRevLett.116.211603
25. J. Casalderrey-Solana, Z. Hulcher, G. Milhano, D. Pablos, K. Rajagopal, *Phys. Rev. C* **99**(5), 051901 (2019). DOI 10.1103/PhysRevC.99.051901
26. M.A. Escobedo, E. Iancu, *JHEP* **05**, 008 (2016). DOI 10.1007/JHEP05(2016)008
27. J. Brewer, J.G. Milhano, J. Thaler, *Phys. Rev. Lett.* **122**(22), 222301 (2019). DOI 10.1103/PhysRevLett.122.222301
28. A. Takacs, K. Tywoniuk, *JHEP* **10**, 038 (2021). DOI 10.1007/JHEP10(2021)038
29. L. Apolinário, L. Luís, J.G. Milhano, J.M. Silva, *JHEP* **01**, 164 (2025). DOI 10.1007/JHEP01(2025)164
30. C. Andres, J. Holguin, R. Kunnawalkam Elayavalli, J. Viinikainen, *Phys. Rev. Lett.* **134**(8), 082303 (2025). DOI 10.1103/PhysRevLett.134.082303
31. L. Apolinário, C.L. Roux, K. Zapp, (2025)
32. Y.L. Du, D. Pablos, K. Tywoniuk, *JHEP* **21**, 206 (2020). DOI 10.1007/JHEP03(2021)206
33. Y.L. Du, D. Pablos, K. Tywoniuk, *Phys. Rev. Lett.* **128**(1), 012301 (2022). DOI 10.1103/PhysRevLett.128.012301
34. L. Apolinário, N.F. Castro, M. Crispim Romão, J.G. Milhano, R. Pedro, F.C.R. Peres, *JHEP* **11**, 219 (2021). DOI 10.1007/JHEP11(2021)219
35. Y.S. Lai, J. Mulligan, M. Płoskoń, F. Ringer, *JHEP* **10**, 011 (2022). DOI 10.1007/JHEP10(2022)011
36. L. Liu, J. Velkovska, Y. Wu, M. Verweij, *JHEP* **04**, 140 (2023). DOI 10.1007/JHEP04(2023)140
37. M. Crispim Romão, J.G. Milhano, M. van Leeuwen, *SciPost Phys.* **16**(1), 015 (2024). DOI 10.21468/SciPostPhys.16.1.015
38. U.S. Qureshi, R. Kunnawalkam Elayavalli, (2024)
39. J. Arruda Gonçalves, J.G. Milhano, *JHEP* **09**, 028 (2025). DOI 10.1007/JHEP09(2025)028
40. R. Li, Y.L. Du, S. Cao, (2025)

41. J.A. Gonçalves, (2025)
42. Y. Wu, Yi, Chen, J. Velkovska, (2025)
43. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, *Advances in neural information processing systems* **30** (2017)
44. V. Mikuni, F. Canelli, *Mach. Learn. Sci. Tech.* **2**(3), 035027 (2021). DOI 10.1088/2632-2153/ac07f6
45. H. Qu, C. Li, S. Qian, (2022)
46. L. Builtjes, S. Caron, P. Moskvitina, C. Nellist, R.R. de Austri, R. Verheyen, Z. Zhang, *SciPost Phys.* **19**(1), 028 (2025). DOI 10.21468/SciPostPhys.19.1.028
47. B. Käch, D. Krücker, I. Melzer-Pellmann, (2022)
48. F.A. Di Bello, et al., *Eur. Phys. J. C* **83**(7), 596 (2023). DOI 10.1140/epjc/s10052-023-11677-7
49. T. Finke, M. Krämer, A. Mück, J. Tönshoff, *JHEP* **06**, 184 (2023). DOI 10.1007/JHEP06(2023)184
50. A. Butter, N. Huetsch, S. Palacios Schweitzer, T. Plehn, P. Sorrenson, J. Spinner, *SciPost Phys. Core* **8**, 026 (2025). DOI 10.21468/SciPostPhysCore.8.1.026
51. M. He, D. Wang, *Eur. Phys. J. C* **83**(12), 1116 (2023). DOI 10.1140/epjc/s10052-023-12293-1
52. A. Hammad, S. Moretti, M. Nojiri, *JHEP* **03**, 144 (2024). DOI 10.1007/JHEP03(2024)144
53. J. Spinner, V. Bresó, P. de Haan, T. Plehn, J. Thaler, J. Brehmer, (2024)
54. J. Brehmer, V. Bresó, P. de Haan, T. Plehn, H. Qu, J. Spinner, J. Thaler, (2024)
55. J. Spinner, L. Favaro, P. Lippmann, S. Pitz, G. Gerhartz, T. Plehn, F.A. Hamprecht, (2025)
56. R. Gambhir, M. LeBlanc, Y. Zhou, in *39th Annual Conference on Neural Information Processing Systems: Includes Machine Learning and the Physical Sciences (MLAPS)* (2025)
57. K.C. Zapp, *Eur. Phys. J. C* **74**(2), 2762 (2014). DOI 10.1140/epjc/s10052-014-2762-1
58. J.G. Milhano, K. Zapp, *Eur. Phys. J. C* **82**(11), 1010 (2022). DOI 10.1140/epjc/s10052-022-10954-1
59. K.C. Zapp, *Phys. Lett. B* **735**, 157 (2014). DOI 10.1016/j.physletb.2014.06.020
60. M. Cacciari, G.P. Salam, G. Soyez, *JHEP* **04**, 063 (2008). DOI 10.1088/1126-6708/2008/04/063
61. M. Cacciari, G.P. Salam, G. Soyez, *Eur. Phys. J. C* **72**, 1896 (2012). DOI 10.1140/epjc/s10052-012-1896-2
62. M. Crispim Romão, J.A. Gonçalves, J.G. Milhano. JEWEL 2.3 simulated pp and PbPb di-jet samples at 5.02 TeV - With and without Medium Response and Underlying Event (2025). DOI 10.5281/zenodo.17591396. <https://doi.org/10.5281/zenodo.17591396>
63. S. Garg, D. Tsipras, P.S. Liang, G. Valiant, *Advances in neural information processing systems* **35**, 30583 (2022)
64. Y. Bai, F. Chen, H. Wang, C. Xiong, S. Mei, *Advances in neural information processing systems* **36**, 57125 (2023)
65. A. Paszke, arXiv preprint arXiv:1912.01703 (2019)
66. W.A. Falcon, GitHub **3** (2019)
67. T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2016), KDD '16, pp. 785–794. DOI 10.1145/2939672.2939785. URL <http://doi.acm.org/10.1145/2939672.2939785>
68. J. Arruda Gonçalves, J.G. Milhano, *JHEP* **09**, 028 (2025). DOI 10.1007/JHEP09(2025)028
69. E.M. Metodiev, J. Thaler, *Phys. Rev. Lett.* **120**(24), 241602 (2018). DOI 10.1103/PhysRevLett.120.241602
70. P.T. Komiske, E.M. Metodiev, J. Thaler, *JHEP* **11**, 059 (2018). DOI 10.1007/JHEP11(2018)059
71. G. Aad, et al., *Phys. Rev. D* **100**(5), 052011 (2019). DOI 10.1103/PhysRevD.100.052011
72. P.T. Komiske, S. Kryhin, J. Thaler, *Phys. Rev. D* **106**(9), 094021 (2022). DOI 10.1103/PhysRevD.106.094021
73. M.J. Dolan, J. Gargalionis, A. Ore, *JHEP* **08**, 024 (2025). DOI 10.1007/JHEP08(2025)024
74. X.P. Duan, L. Chen, G.L. Ma, C.A. Salgado, B. Wu, (2025)
75. M. LeBlanc, B. Nachman, C. Sauer, *JHEP* **02**, 150 (2023). DOI 10.1007/JHEP02(2023)150
76. E. Alvarez, F. Lamagna, M. Szwec, *JHEP* **01**, 049 (2020). DOI 10.1007/JHEP01(2020)049
77. J. Brewer, J. Thaler, A.P. Turner, *Phys. Rev. C* **103**(2), L021901 (2021). DOI 10.1103/PhysRevC.103.L021901
78. J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, arXiv preprint arXiv:2001.08361 (2020)
79. T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T.B. Brown, P. Dhariwal, S. Gray, et al., arXiv preprint arXiv:2010.14701 (2020)
80. X. Zhai, A. Kolesnikov, N. Houlsby, L. Beyer, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 12,104–12,113
81. J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D.d.L. Casas, L.A. Hendricks, J. Welbl, A. Clark, et al., arXiv preprint arXiv:2203.15556 (2022)
82. P.T. Komiske, E.M. Metodiev, J. Thaler, *JHEP* **04**, 013 (2018). DOI 10.1007/JHEP04(2018)013
83. A.J. Larkoski, J. Thaler, W.J. Waalewijn, *JHEP* **11**, 129 (2014). DOI 10.1007/JHEP11(2014)129
84. S. Acharya, et al., *JHEP* **10**, 139 (2018). DOI 10.1007/JHEP10(2018)139
85. J. Thaler, K. Van Tilburg, *JHEP* **03**, 015 (2011). DOI 10.1007/JHEP03(2011)015

-
86. A.J. Larkoski, S. Marzani, G. Soyez, J. Thaler, JHEP **05**, 146 (2014). DOI 10.1007/JHEP05(2014)146
87. Y. Mehtar-Tani, A. Soto-Ontoso, K. Tywoniuk, Phys. Rev. D **101**(3), 034004 (2020). DOI 10.1103/PhysRevD.101.034004
88. T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (2019), pp. 2623–2631