

Explainable equivariant neural networks for particle physics: PELICAN

Alexander Bogatskiy^{},^b Timothy Hoffman^{},^a David W. Miller^{},^{a,c}
Jan T. Offermann^{}^a and Xiaoyang Liu^{}^a

^a*Department of Physics, University of Chicago,
Chicago, IL, U.S.A.*

^b*Center for Computational Mathematics, Flatiron Institute,
New York, NY, U.S.A.*

^c*Enrico Fermi Institute,
Chicago, IL, U.S.A.*

E-mail: abogatskiy@flatironinstitute.org, hoffmant@uchicago.edu,
David.W.Miller@uchicago.edu, jano@uchicago.edu, xiaoyang1@uchicago.edu

ABSTRACT: PELICAN is a novel permutation equivariant and Lorentz invariant or covariant aggregator network designed to overcome common limitations found in architectures applied to particle physics problems. Compared to many approaches that use non-specialized architectures that neglect underlying physics principles and require very large numbers of parameters, PELICAN employs a fundamentally symmetry group-based architecture that demonstrates benefits in terms of reduced complexity, increased interpretability, and raw performance. We present a comprehensive study of the PELICAN algorithm architecture in the context of both tagging (classification) and reconstructing (regression) Lorentz-boosted top quarks, including the difficult task of specifically identifying and measuring the W -boson inside the dense environment of the Lorentz-boosted top-quark hadronic final state. We also extend the application of PELICAN to the tasks of identifying quark-initiated vs. gluon-initiated jets, and a multi-class identification across five separate target categories of jets. When tested on the standard task of Lorentz-boosted top-quark tagging, PELICAN outperforms existing competitors with much lower model complexity and high sample efficiency. On the less common and more complex task of 4-momentum regression, PELICAN also outperforms hand-crafted, non-machine learning algorithms. We discuss the implications of symmetry-restricted architectures for the wider field of machine learning for physics.

KEYWORDS: Jets and Jet Substructure, Specific QCD Phenomenology, Top Quark

ARXIV EPRINT: [2307.16506](https://arxiv.org/abs/2307.16506)

Contents

1	Introduction	2
2	Equivariance and jet physics	5
2.1	Lorentz symmetry and jets	5
2.2	Lorentz invariance	6
2.3	Permutation equivariance	7
2.4	Elementary equivariant aggregators	8
2.5	Equivariance and jet physics	10
3	PELICAN architecture	11
3.1	Inputs and embeddings	11
3.2	Permutation equivariant blocks	12
3.3	Classification and 4-vector regression outputs	13
4	Tagging jets from Lorentz boosted top quarks	15
4.1	Top tagging dataset	15
4.2	Classification training procedure	15
4.3	Top tagging results	17
5	Tagging quark and gluon jets	19
5.1	Quark-gluon jet dataset	19
5.2	Quark-gluon jet tagging results	19
6	Multi-class jet tagging	20
6.1	HLS4ML LHC jet dataset	20
6.2	Multi-class jet tagging results	21
7	W-boson 4-momentum reconstruction	22
7.1	Regression dataset	22
7.2	Regression training procedure	23
7.3	Regression results for p_{true}^W reconstruction	26
7.4	Regression results for p_{cont}^W reconstruction	27
7.5	Discussion	28
8	W-boson mass measurement	28
8.1	Regression results for m_W reconstruction	29
8.2	Model complexity	30
8.3	Discussion	31
9	PELICAN explainability	33
9.1	Distributions of PELICAN weights	33
9.2	Detector effects on PELICAN weights	35

9.3	Event visualization	38
9.4	Explainability	39
10	IRC-safety and PELICAN	40
10.1	Definitions	40
10.2	IRC-safe PELICAN	41
10.3	Testing IRC-safe PELICAN models	43
11	Conclusion	44
A	Additional results and plots	45
A.1	W -boson 4-momentum reconstruction	45
A.2	W -boson mass measurement	45
A.3	Additional event displays	48
B	IRC-safety and Lorentz symmetry	48
B.1	Lorentz meets EFPs	53
B.2	Universality of PELICAN	59
B.3	Quantifying IRC-safety	60

1 Introduction

Identifying, reconstructing, and measuring the properties and dynamics of high-energy, short-distance particle phenomena are inherently inference tasks, since direct access to the fundamental processes is often impossible due to the time and length scales at which they occur. The suite of detection techniques, pattern recognition algorithms, and measurement approaches used to perform such tasks inevitably imposes constraints on both the nature of the information used as well as on the form and structure of the results. Such constraints play a crucial role in the context of jet substructure measurements, in which detailed analysis is performed on the long-distance features of Lorentz-boosted particle decays, parton showering, and radiation patterns found in the collimated sprays of particles that form the jets themselves. We present a comprehensive analysis of a new approach to multiple jet substructure-based inference tasks using a machine learning (ML) architecture that fundamentally respects permutation and Lorentz-group symmetries: PELICAN, the permutation equivariant and Lorentz invariant or covariant aggregator network. Our approach imposes explicit physics-informed symmetry constraints on the architecture and consequently yields new insights and capabilities.

Decades of jet substructure research have yielded a wide range of approaches to performing inference tasks such as: distinguishing quark-initiated from gluon-initiated jets [1–5]; discriminating jets formed from Lorentz-boosted top quarks, Higgs and W -bosons, from the continuum background of jets formed from light-quarks and gluons [6–9]; dissecting and measuring the parton-shower structure of light-quark and gluon jets themselves [10–15]. Many approaches have been adopted to perform these tasks, including the direct use of discriminating high-level observables and multi-variate methods [16–18], as well as a growing number of

ML architectures using a variety of latent-space representations. For a comprehensive overview of jet substructure measurements and approaches, see refs. [19, 20], as well as ref. [21] for a general review of ML methods in high-energy physics (including substructure measurements). As the model complexity has grown, so too have questions regarding the relationship of both the methods and the constraints that they impose on the fundamental physical processes that they are used to model. In particular, the use of observables, architectures, and latent space representations that adhere closely to the structure and dynamics of the physics processes under study have been found to provide not only enhanced performance, but also significant insights and improvements in interpreting the results [18, 22, 23]. Imbuing these models with knowledge of, or even fundamental respect for, the symmetries inherent to the system under study has thus become increasingly impactful in the study of jet substructure, especially in the context of ML models and various neural network (NN) architectures [24–26].

There are several common approaches to enforcing continuous symmetries in NNs. Data augmentation can be used to train a model to have a particular sparsity structure and become approximately symmetric. However, when model complexity and interpretability are of concern, as is the case in many particle physics applications, a different approach is helpful. Similar issues arise when using data preprocessing or normalization, which often come with inherent ambiguities and discontinuities that can be detrimental for more complex tasks.

Traditionally, ML algorithms are evaluated based on basic performance metrics such as accuracy and computational cost. However, in contexts where the trained algorithms are treated not only as predictors or generators, but as actual models for physical processes — which is especially true in scientific applications — other metrics of model quality are valuable. Model complexity (e.g. the number of parameters), explainability and interpretability are also important for using an ML algorithm as a viable physics model. Furthermore, certain problem-specific properties such as symmetries can be critical as well. Symmetries in ML are known to produce less complex models which respect basic geometrical rules and arguably provide more opportunities for interpretability and explainability (e.g. convolutional neural network (CNN) kernels are often interpreted as visual features). Even in realistic settings where the symmetries are merely approximate, symmetry-constrained architectures often outperform more general architectures in terms of pure accuracy (see e.g. section 4), but even in cases when that is not true, symmetric architectures should not be discounted due to their other benefits. For these reasons, as advocated for in ref. [27], we have adopted the approach of building all symmetries directly into the PELICAN network architecture itself, similar to the inherent translational symmetry of CNNs.

Summary of results. In section 2 we discuss equivariance in jet physics and introduce the tools required to build an efficient equivariant architecture. In section 3 we describe the architectures of PELICAN classifiers and regressors. We briefly summarize the main results presented in this work, corresponding to sections 4–10.

Top-tagging with a PELICAN classifier. We train PELICAN top taggers using a public benchmark dataset, to distinguish between top quark jets (signal), and light quark and gluon jets (background). These taggers achieve state-of-the-art performance on the benchmark with fewer learnable parameters than the previous highest-performing network.

PELICAN top taggers with as few as 11k parameters outperform all non-equivariant networks in the benchmark. See section 4 for details.

Quark-vs-gluon-initiated jet tagging with a PELICAN classifier. We extend the study of a PELICAN classifier to the task of distinguishing jets produced by light-quarks (signal) from those produced by gluons (background). As in the task of top-tagging, PELICAN demonstrates state-of-the-art performance on this benchmark task, with fewer learnable parameters than any other similarly performant network architecture. See section 5 for details.

Multi-class jet tagging with a PELICAN classifier. The final classification benchmark task involves identifying 5 different categories of jets (gluon jets, light quark jets, W -boson, Z -boson, and top quark jets). PELICAN achieves state-of-the-art classification performance across each of these five categories and once again demonstrates a more efficient utilization of its parameters than any rival architecture. See section 6 for details.

W -boson 4-momentum reconstruction with PELICAN. We train a PELICAN model using a custom dataset [28] of fully-hadronic top-quark decays to reconstruct the full 4-momentum of the intermediate W -bosons using only the list of 4-momenta of the top quark jet constituents as inputs. PELICAN performs favorably in reconstructing the full W -boson momentum when compared with the Johns Hopkins (JH) top tagger [7], which identifies W -boson candidates for the subset of jets that pass its tagging. PELICAN achieves better transverse momentum (p_T), mass, and angular resolutions on JH top-tagged jets and achieves comparable resolutions to the JH tagger even when evaluated on the full dataset. Additionally, we train a PELICAN model to reconstruct the 4-momentum of only the products of the $W \rightarrow qq'$ decay which are contained within the jet. We discuss differences in performance and effects of this choice in reconstruction targets in section 7.

W -boson mass reconstruction with PELICAN. Particle mass reconstruction is a common particle physics analysis task, and any reconstruction algorithm should be robust and relatively free of bias. In section 8 we discuss the nuances of PELICAN mass reconstruction targeting the W -bosons in the above-mentioned dataset [28] as an example. The results show that eliminating bias in the underlying dataset is required to produce an unbiased final algorithm. In the case of W -boson mass (m_W) reconstruction, this is achieved by training PELICAN on a dataset with a range of values of m_W .

Explaining PELICAN 4-momentum reconstruction. PELICAN’s respect of the particle permutation and Lorentz symmetries inherent to particle datasets provides it with explainability and interpretability rarely found in particle physics ML applications. In section 9 we investigate the rich penultimate layer of PELICAN and its discriminatory and explanatory capabilities. In particular, we discuss interpretations of PELICAN as a form of soft-clustering and detector-unfolding algorithm. By *explainability* of a network we mean the ability to identify physical features that highly correlate with the output. PELICAN’s unique regression architecture, constrained by the simultaneous imposition of Lorentz and permutation symmetries, produces intermediate outputs (“PELICAN weights”) which we directly interpret as jet clustering coefficients and demonstrate the correctness of this interpretation. Despite the task being a simple regression on a single 4-vector, PELICAN

reconstructs a much more complex set of physical features in the form of the labels of all input jet constituents by parent type.

IRC-safety and PELICAN. In particle physics, so-called *IRC-safety* is an algorithmic concern which restricts tools to be robust with respect to soft-particle emissions (infrared — IR) and collinear (C) splittings which proliferate due to divergences in perturbative quantum chromodynamics (QCD). In section 10 we introduce a simple IRC-safe modification of PELICAN and display its state-of-the-art performance.

2 Equivariance and jet physics

This section aims to establish a clear connection between the group theory that underlies the PELICAN architecture and the implementation of this approach for both classification and regression, as described in section 3.

In general, given a symmetry group G and two sets X, Y on which an action of G is defined, a mapping $F : X \rightarrow Y$ is called G -equivariant if $F(g \cdot x) = g \cdot F(x)$ for any $x \in X$ and $g \in G$. In particular, if the action of G on Y happens to be trivial (i.e. $g \cdot y = y$ for all g, y), then F is called *invariant*. In relativistic physics, equivariant maps are typically represented by tensors with equivariant spacetime indices treated via Einstein notation. For instance, the electromagnetic field tensor $F^{\mu\nu}$ can be viewed as a Lorentz-equivariant mapping from covariant vector fields to contravariant ones. In this work we will be interested in tasks from particle physics that can be reduced to learning a Lorentz-equivariant map. In this section we review several aspects of Lorentz symmetry in the context of such tasks.

2.1 Lorentz symmetry and jets

The Lorentz symmetry is one of the fundamental symmetries of the Standard Model of particle physics. The full Lorentz group $O(1, 3)$ can be defined as the set of linear transformations of the 4-dimensional spacetime that preserve the Minkowski metric $\eta = \text{diag}(1, -1, -1, -1)$. However, in this work we will restrict ourselves to the *proper orthochronous* subgroup $SO^+(1, 3)$ that preserves spatial and temporal orientations. Lorentz invariance is the mathematical encapsulation of the fact that the outcomes of physical phenomena don't depend on the inertial frame of the observer. In the context of particle accelerators, this boils down to the observation that all initial and final states of a particle interaction are the same in all inertial frames. This is formally reflected in the fact that the Standard Model of particle physics is Lorentz-invariant, and therefore any model of any physically relevant processes described by the Standard Model can be as well.

Several subtle points are worth addressing before applying Lorentz symmetry to experimental tasks in jet physics. Neither the actual particle detectors nor the software simulating particle decays and their detection are Lorentz-invariant. Reasons for this include: non-invariant corrections to perturbative computations in quantum chromodynamics (QCD); non-invariance of jet clustering algorithms; practical limitations of detectors such as finite spatial and temporal resolutions, as well as energy and momentum thresholds. Nevertheless, it is still valid to learn Lorentz-invariant models from data obtained this way. Firstly, QCD is globally Lorentz-invariant and boosting the *entire* event does not change the outcome

of the decay process. As long as inference is performed on data obtained in conditions similar to the conditions of the perturbative simulation, corrections from effects such as the running of the couplings with varying momentum scales are not a concern either. The same applies to jet clustering algorithms and finite detector resolution: as long as the data used for inference was obtained in the same reference frame as the data used for training, the inference is valid and the outputs are expected to be Lorentz-equivariant. Finally, the fact that the detector itself introduces a fixed reference frame can be fully addressed without breaking the symmetry of the model by including detector geometry among its inputs. This is discussed further in section 3.1.

While the imprecisions in our measurements of 4-momenta in themselves do not undermine the validity of Lorentz-invariant models, further investigation of the relative robustness to systematic measurement biases of different architectures is warranted. One step in this direction, via dataset augmentation, was recently taken in ref. [29].

2.2 Lorentz invariance

The classification tasks considered in this work are exactly Lorentz invariant. The physical implications of this statement are discussed below, but may be stated mathematically in the following way: if the inputs to the network are a collection of 4-vectors (energy-momentum vectors in our case) $p_1, \dots, p_N$, the output is $F(p_1, \dots, p_N)$, and $\Lambda \in \text{SO}^+(1, 3)$ is a Lorentz transformation, then

$$F(\Lambda p_1, \dots, \Lambda p_N) = F(p_1, \dots, p_N). \quad (2.1)$$

There are multiple approaches to constructing an ML model that satisfies such a constraint. The simplest one is to hand-pick a set of invariant observables (such as particle masses, jet masses, relative masses, particle identification labels and charge) and use them as input to a generic NN architecture.

Another approach inspired by convolutional networks is to preserve group-equivariant latent representations in the hidden layers. In this case, the neuron nonlinearity must be a Lorentz-equivariant operation, and examples of this can be found in both the Lorentz Group Network (LGN) [25] and LorentzNet [26] architectures. Equivariance with respect to the part of the Lorentz group that fixes the proton beam axis was also used for regression problems in ref. [30]. As in traditional CNN's used in image processing, equivariant latent representations, as opposed to invariant ones, can regularize the network via efficient weight-sharing and improve training.

The PELICAN design adopts a slightly different approach. Given a set of 4-vector inputs $p_1, \dots, p_N$, we compute a *complete* set of Lorentz invariants on that set. For classical groups, including the Lorentz group, the space of invariants constructed out of a collection of vectors in the fundamental representation consists of functions of only the pairwise invariant dot products (using the appropriate invariant quadratic form for the given symmetry group) and of square determinants (e.g. of 4 column-vectors for the Lorentz group) [31]. Furthermore, if the invariant is required to be symmetric in the vector inputs, then it is *only* a function of the dot products (see also the discussion in ref. [32]). In short, all totally symmetric Lorentz

invariants can be written in the following form:

$$I(p_1, \dots, p_N) = f(\{p_i \cdot p_j\}_{i,j}). \quad (2.2)$$

This is the first key idea used in the PELICAN architecture. The first step performed by the input layer is the computation of the $N \times N$ array of dot products between the input 4-momenta (also known as the Gram matrix). The $N(N-1)/2$ components of the Gram matrix $\{p_i \cdot p_j\}$ cannot be independent, which is apparent from dimension counting. The physical manifold inside this high-dimensional space is defined by the set of constraints $\det M_5 = 0$ for *every* 5-minor M_5 of the Gram matrix (that is, any matrix obtained from the original one by crossing out $N-5$ rows and $N-5$ columns). Moreover, a causally related set of points such as a particle jet will always satisfy $p_i \cdot p_j \geq 0$ for all i, j . Therefore a neural network whose input is an $N \times N$ matrix will learn the task only on this $(4N-6)$ -dimensional submanifold of $\mathbb{R}^{N^2}$. The outputs of the trained model on the rest of the space will be uncontrollable and physically meaningless. As a result, PELICAN has the complexity of $\mathcal{O}(N^2)$, similar to any message passing architecture on a fully connected graph. An explicit set of coordinates for this manifold would allow for complete Lorentz-invariant networks of the superior $\mathcal{O}(N)$ complexity, however such coordinates, as of this time, are unknown for $N > 5$ [32].

2.3 Permutation equivariance

Particle data are often interpreted as a point cloud since there is no natural ordering on the vectors. For such problems it makes sense to use a permutation-invariant or equivariant architecture. One of the simplest approaches is called *Deep Sets* [33], which has been applied to jet tagging [24] and even heavy-flavor tagging [34]. The fundamental fact used in Deep Sets is that any permutation-invariant continuous mapping of inputs $x_1, \dots, x_N$ can be written in the form $\psi(\sum_i \varphi(x_i))$, where ψ and φ can be approximated by neural networks.

The main limitation of permutation-invariant architectures such as Deep Sets is the difficulty of training. Since aggregation (summation over the particle index) happens only once, the Deep Sets architecture can struggle with modeling complex higher-order interactions between the particles, as shown rigorously in ref. [35]. The network representing ψ is forced to be a relatively wide fully connected network, which presents difficulties in training.

An alternative to permutation-invariant architectures is provided by permutation-*equivariant* ones. Given a symmetry group G (e.g. the group of permutations), a representation (V, ρ) is a tuple where V is a set and $\rho : G \times V \rightarrow V$ is a map that becomes a bijection $\rho_g = \rho(g, \cdot) : V \rightarrow V$ for any fixed value of the first argument, $\rho_e = \text{id}$, and $\rho_{g^{-1}} = \rho_g^{-1}$. Given two representations (V, ρ) and (V', ρ') of a group G , a map $F : V \rightarrow V'$ is called equivariant if it *intertwines* the two representations, that is:

$$F(\rho_g(v)) = \rho'_g(F(v)), \quad v \in V, g \in G. \quad (2.3)$$

Equivariance is a key property of all convolutional networks — for example, in CNN's the convolution operation is inherently equivariant with respect to translations (up to edge effects).

Similarly, Graph Neural Networks (GNNs) use permutation equivariance with respect to the reordering of the rows and columns of the adjacency matrix for problems where the inputs can be naturally represented by a graph data structure. In this context, we review

the standard definition of a message passing layer where the particles are treated as nodes in a graph (e.g. a fully connected graph), and every layer of the network only updates the activation at every node. If we denote by f_i the data assigned to node i , then the message passing layer will typically construct “messages” $m_{ij} = m(f_i, f_j)$ and update each node by aggregating the messages from all neighbors of that node and combining the result with the original state of the node: $f'_i = \psi(f_i, \sum_j m_{ji})$. Sometimes the graph also possesses “edge data” D_{ij} that can be incorporated into the message-forming stage.

Message passing architectures have been successfully applied to jet tagging, most prominently in refs. [25, 26]. Closely related permutation-equivariant transformer architectures were also applied to particle physics in [30, 36]. However, attempts to combine message passing with Lorentz invariance reduced at the input stage as described above run into a major obstacle: the new inputs to the network consist of *nothing but* edge data $d_{ij} = p_i \cdot p_j$. Traditional message passing would require a reduction of this set of inputs to a point cloud (with only one particle index), potentially restricting the set of possible higher-order interactions between the points or adding redundancy to the network. To avoid making these unnecessary choices, we employ the general permutation-equivariant layers suggested in refs. [37, 38].

In the general setting, permutation equivariance is a constraint on mappings F between arrays $T_{i_1 i_2 \dots i_r}$ of any rank r , every index $i_k \in \{1, \dots, N\}$ referring to a particle label, whereby permutations of the particles commute with the map:

$$F(\pi \circ T_{i_1 i_2 \dots i_r}) = \pi \circ F(T_{i_1 i_2 \dots i_r}), \quad \pi \in S_N. \quad (2.4)$$

Here, the action of permutations is effectively diagonal: $\pi \circ T_{i_1 i_2 \dots i_r} = T_{\pi(i_1) \dots \pi(i_r)}$. GNNs explicitly implement this constraint for rank 1 arrays (node information). A higher-order generalization of the message passing layer can be defined as

$$\textbf{Equivariant Layer: } T^{(\ell+1)} = \text{AGG} \circ \text{MSG} \left(T^{(\ell)} \right). \quad (2.5)$$

In this nomenclature, MSG is a node-wise nonlinear map (*message forming*) shared between all nodes, and AGG is a general permutation-equivariant linear mapping (*aggregation*) acting on the particle indices of T . Note that whether MSG is node-wise and whether AGG is linear is somewhat ambiguous based on how one separates the mappings into their components, which is why, in particular, the traditional formulation of message passing allows messages to be functions of pairs of nodes. In practice, our aggregation block will permit nonlinear aggregation functions such as max-pooling and in fact contain an additional nonlinear activation function.

2.4 Elementary equivariant aggregators

The exact structure of the equivariant aggregation layers defined above must still be specified. Since the general case is presented in refs. [38, 39], here we will only present the layers that are required for jet physics tasks. Since the input is an array of rank 2, the main equivariant layer for us is one that transforms arrays of rank 2 to other arrays of the same rank: $T_{ij} \mapsto T'_{ij}$. The space of all linear maps of this type turns out to be 15-dimensional. The basis elements of this space can be conveniently illustrated using binary arrays of rank 4. There are 15 such

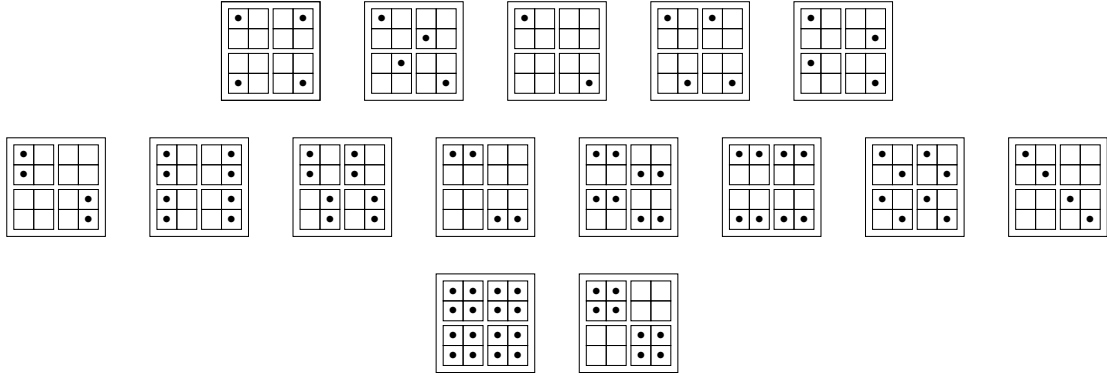


Figure 1. The 15 binary arrays of rank 4 that represent the basis elements of the permutation equivariant aggregators of PELICAN.

arrays B_{ijkl}^a , $a = 1, \dots, 15$, and the action of the equivariant layer can be written as

$$T_{ij}^{\prime a} = \sum_{k,l=1}^N B_{ijkl}^a T_{kl}. \quad (2.6)$$

The 15 aggregators B^a may be visualized as is done in figure 1 for $N = 2$. The smallest squares represent components of the input 2×2 array, and the larger 2×2 squares represent components of the output array. Dots represent the non-zero components of the binary tensors B^a , and every component of the output tensor is the result of aggregation over all inputs marked by the dots. Output components that lack any dots are set to be a fixed constant, by default zero (the affine versions of these mappings include two such parameters: one constant for the diagonal and another for the remaining components). By “aggregation” we mean, in general, any symmetric function, but in practice it is usually a sum or mean. For example, the first aggregator is simply the identity map on matrices: the ij ’th component of the output array is the result of aggregation over only the ij ’th component of the input. The second aggregator realizes the transposition of arrays $T_{ij}' = T_{ji}$. The following three aggregators represent various ways of embedding the diagonal of the input array in an equivariant way. It is easy to see that simultaneously swapping the two rows and the two columns of the input is equivalent to doing the same to the output, which confirms equivariance. These first 5 aggregators are “order zero” in N because they do not actually perform any aggregation. Instead, they can be thought of as permutation-equivariant skip-connections.

The second group of 8 “order one” aggregators aggregate over N components of the input by aggregating either over rows, columns, or the diagonal, and then embedding the result into the output array in all possible equivariant ways. Finally, the last 2 aggregators are the “order two” aggregators that aggregate over all N^2 components of the input.

If we allow aggregators to be nonlinear, then they can take the following form: the binary array B^a selects a subset of the components of the input array, and then a general symmetric function S^a is applied to that subset:

$$T_{ij}^{\prime a} = S^a \left(\{T_{kl} \mid k, l : B_{ijkl}^a \neq 0\} \right). \quad (2.7)$$

In practice we define S^a as the mean of its inputs followed by an additional scaling by a factor of $N^{\alpha_a}/\bar{N}^{\alpha_a}$ with learnable exponents α_a , where $\bar{N}$ is a constant representing the typical number of input vectors expected in the dataset, provided to the model as a hyperparameter.

These equivariant layers and their generalizations have been thoroughly studied since their introduction (see e.g. refs. [40–43]) and applied to problems in graph learning and chemistry, e.g. in refs. [38, 44]. However, many important questions about the universality and expressivity of these networks remain open. For example, networks that can use any $\text{Eq}_{r \rightarrow s}$ layers were proven to be universal in refs. [40, 42], but the exact value of the maximum rank (which is 2 in PELICAN) required for universality in a given problem is not known.

2.5 Equivariance and jet physics

There are several reasons for enforcing the full Lorentz symmetry in our ML models. First and foremost, it is a fundamental symmetry of the space to which the inputs belong. Lorentz transformations represent the effect of switching between different inertial frames, and most fundamental processes in physics are independent of the choice of the observer’s inertial frame: if a given collection of particles consists of the products of a decay of a top quark for one observer, then the same is true for all other observers.

Nevertheless, some processes involved in generating and observing high-energy collision events break the Lorentz symmetry in some subtle ways. At the fundamental level, the running of the couplings in QCD can cause Lorentz symmetry breaking in the parton shower distribution functions. Even the amount of final decay products depends on the transversal boost of the initial parton-level particles. However, there is no question that both the original protons and the final (asymptotic) decay products are accurately represented by a collection of 4-vectors subject to the spacetime Lorentz symmetry: the asymptotic outcome of a collision event is independent of the observer’s reference frame.

Another reason for symmetry-restricted modeling is that, from the geometric perspective, only some mathematical operations are permissible when working with objects that transform in a certain way under a symmetry group. A non-equivariant neural network effectively neglects the vector nature of the inputs by treating individual components of the input vectors as scalars. While improving network expressivity, non-equivariance fails to deliver physically interpretable models. Ultimately, a statement about equivariance is a statement about what the basic *features* of the data are — e.g. vectors are features, but the individual components of those vectors are not.

More relevant to the applications is the fact that both the simulation and the observation of collisions inevitably involves some degree of *clustering*. A particle detector is made of cells (e.g. calorimeters) of finite size and as such is unable to distinguish between some particles that are collinear or very close to collinear. Similarly, the standard algorithms for collision simulation typically perform *jet clustering* to closely reproduce the detector behavior. Clustering of course is not a Lorentz-invariant procedure: particle tracks that diverge by a small angle in one frame will diverge by a large angle in another highly boosted frame. However, this limitation of Lorentz-invariant architectures is fairly minor. Since clustering is always done in a fixed laboratory frame, it is still reasonable to impose the full Lorentz symmetry on the resulting 4-vector data. So unless the pre-clustering data itself

is coming from multiple significantly different inertial frames, clustering is not interfering with the fundamental symmetry. Simply put, however a given set of 4-vectors is obtained and represented in a specific inertial frame, those vectors will respect the Lorentz symmetry. Finally, one important way to address the issue of network sensitivity to the clustering of collinear particles is through enforcing so called IRC-safety, which will be discussed separately in section 10.

3 PELICAN architecture

The PELICAN architecture is simpler than many previous architectures due to its use of a complete set of Lorentz-invariants at the input stage, namely the set of pairwise dot products between the input 4-momenta (see section 2), and this has significant implications for both the overall architecture as well as the ease of training and interpretability of the network. This section discusses each of the primary components of the network, including the inputs and their embedding, the permutation-equivariant blocks, and the output layers that determine the nature of the task, namely classification or 4-vector regression.

3.1 Inputs and embeddings

Dot products and beams. On the input side of the architecture, the first step is to compute all pairwise dot products of the input 4-momenta. Appended to the list of these 4-momenta are two auxiliary beam particles with 4-momenta $(1, 0, 0, \pm 1)$. This is helpful since the datasets we are using are all simulated in a fixed laboratory frame where the original proton-proton collision happens along the z -axis, and the auxiliary inputs restore this orientational knowledge. In particular, the dot products between constituents and beams give PELICAN access to the energies and transverse momenta of all constituents.

It is worth emphasizing that introducing beams in this manner allows us to fix a particular spatial orientation of the events without restricting or violating the global Lorentz symmetry inherent in the architecture. Indeed, if one were to treat the auxiliary beams as constant vectors of hyperparameters, then this action would reduce the full Lorentz symmetry to merely rotations in the xy -plane and z -boosts. However, due to the fact that the beams are fed into the network on equal footing with all other inputs, they are properly treated as full-fledged 4-vectors that should also transform under the global Lorentz symmetry. Thus, counter-intuitively, we let the network access individual energies, transverse momenta and z -momenta while still preserving the full Lorentz symmetry and all the computational benefits that come with it (e.g. the number of learnable parameters remains unchanged, whereas an architecture with a smaller symmetry group would have more degrees of freedom).

Embedding of dot products. Next there is an embedding layer that applies the function $f_\alpha(x) = ((1+x)^{\alpha^2} - 1)/\alpha^2$ to each dot product with several values of the trainable parameter α (initialized to span the interval $[0.05, 0.5]$). This is roughly inspired by the logarithmic embedding in ref. [26] and serves to mollify the extremely heavy-tailed (approximately Pareto) distribution of the input momenta. The addition of the learnable α parameter allows PELICAN to resolve the inputs at multiple momentum scales, providing a slight boost to

performance. Such an embedding may be generally useful for inputs which are naturally Pareto distributed (i.e. positive with density of the form $p(x) \propto x^{-\gamma}, \gamma > 0$).

Embedding of scalar constituent data In addition to dot products, we introduce a binary scalar label to distinguish jet constituents from the custom beam vectors. More generally, jet datasets can involve scalar particle data s_i such as particle IDs (PID), charge, color, spin, etc. Let's say each s_i is a column of dimension C_{scalar} . To be able to process such inputs alongside the dot products, they need to be promoted to arrays with two particle indices. One way to do this is to double the number of channels and define the array $s_i \oplus s_j$ of shape $[B, N_{\text{max}}, N_{\text{max}}, 2C_{\text{scalar}}]$. This array can then be concatenated with the embedded dot products.

However, there is a more natural way of doing this within the PELICAN framework. In the next section we will introduce permutation-equivariant $\text{Eq}_{2 \rightarrow 2}$ blocks that act on arrays with two constituent indices. To embed scalar particle data (“node features”) into the space of such arrays (“edge features”), it is convenient to utilize an analogous equivariant $\text{Eq}_{1 \rightarrow 2}$ block and apply it directly to s , the array of constituent scalars. This way we can produce a flexible number C'_{scalar} of scalar channels. By choosing the dimensionality of the dot product embedding to be $C^0 - C'_{\text{scalar}}$ and then concatenating it with the array of promoted scalars, we get a tensor of shape $[B, N_{\text{max}}, N_{\text{max}}, C^0]$, where the feature vector for each particle pair is C^0 -dimensional and has the form $\left(f_{\alpha_1}(d_{ij}), \dots, f_{\alpha_{C^0 - C'_{\text{scalar}}}}(d_{ij})\right) \oplus \text{Eq}_{1 \rightarrow 2}(s)_{ij}$. This approach allows for any number of Lorentz invariants per constituent, but in all of the tasks considered in this paper we have $C_{\text{scalar}} = 1$.

3.2 Permutation equivariant blocks

The main element of the equivariant architecture is the permutation-equivariant block transforming arrays of rank 2, represented schematically in figure 2. Namely, we assume that the input tensor to the block has shape $[B, N_{\text{max}}, N_{\text{max}}, C^l]$, where B is the batch size, N_{max} is the maximum number of jet constituents per event (with zero padding for events with fewer constituents), and C^l is the number of input channels. We also use a binary mask of shape $[B, N_{\text{max}}, N_{\text{max}}]$ to appropriately exclude the zero padding from operations like **BatchNorm** and aggregation. The output of the block will be a similar tensor of shape $[B, N_{\text{max}}, N_{\text{max}}, C^{l+1}]$ with the same mask.

As outlined above, the equivariant layer consists of a message block and an aggregation block. The message block is chosen to be a dense multilayer perceptron (MLP) acting on the channel dimension with a **LeakyReLU** activation and **BatchNorm2D** (normalization over the first three dimensions of the tensor, for each channel separately, followed by an affine transform with two learnable parameters per channel). Here we use a masked implementation of batch normalization so that the variable particle number is respected. The message block is then followed by **Dropout** that zeroes out each of the $B \times N_{\text{max}}^2 \times C_{\text{eq}}^l$ components independently with a certain probability.

The aggregation block applies 15 linear aggregation functions (**LinEq_{2→2}**) which, for each component of the output tensor, compute the mean over some subset of the components of the input tensor, as explained in section 2.4. Note that this is a non-parametric transformation

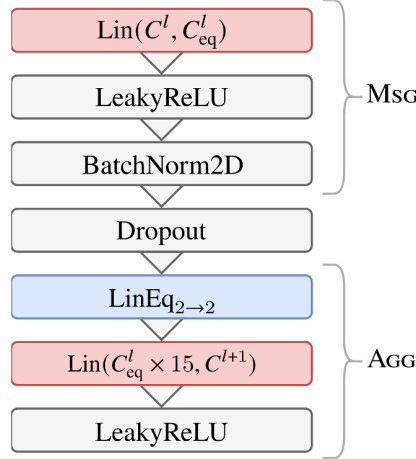


Figure 2. The PELICAN equivariant block updating square arrays.

performed on each channel separately. Each of the $C_{\text{eq}}^l \times 15$ resulting aggregation values is then independently multiplied by $N^\alpha / \bar{N}^\alpha$ with a trainable exponent α (initialized as a random float in $[0, 1]$), where N is the number of particles in the corresponding event. This allows for some flexibility in the aggregation process, for example $\alpha = 1$ returns the sum aggregation function, and combining multiple aggregators is known to boost accuracy, see e.g. ref. [39].

Aggregation is followed by a dense layer that mixes the $C_{\text{eq}}^l \times 15$ aggregators down to C^{l+1} features. Due to the size of this layer, we employ a simple factorization to reduce the number of parameters. Namely the weight tensor W_{abc} , where a is the input channel index, b is the basis index (1 to 15), and c is the output channel index, can be replaced by the following combination:

$$W_{abc}^{\text{factorized}} = W_{ab}^0 W_{ac}^1 + W_{cb}^2 W_{ac}^3. \quad (3.1)$$

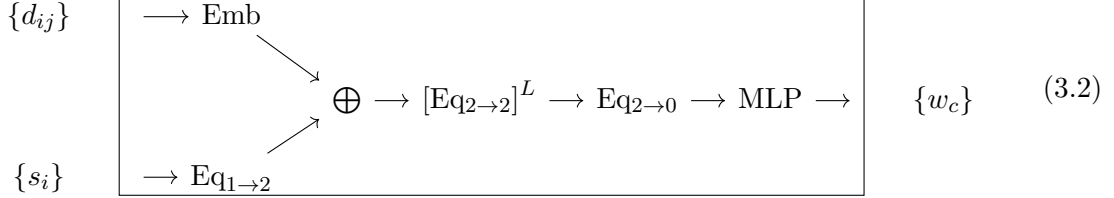
Here, the first term first mixes the 15 aggregators among each other for each output channel, and then mixes the channels. Similarly, the second term first mixes the 15 aggregators for each input channel, and then mixes the channels. It is technically possible to increase the rank of this factorization by adding together multiple tensors of this form, but in practice we find one to be sufficient — in our tasks the factorized network performs as well as the unfactorized one (except at very low network widths, in which case the unfactorized network performs better and may even have fewer parameters). The final result is a tensor of shape $[B, N_{\text{max}}, N_{\text{max}}, C^{l+1}]$, so these equivariant layers can be stacked multiple times.

As already indicated above, a few other equivariant blocks will be useful to us, namely $\text{Eq}_{1 \rightarrow 2}$, $\text{Eq}_{2 \rightarrow 1}$, and $\text{Eq}_{2 \rightarrow 0}$. The only difference in their definition is that they involve just 5, 5, and 2 aggregators, respectively.

3.3 Classification and 4-vector regression outputs

One of the strengths of the PELICAN architecture is the ability to easily switch between serving as a classification tool for jet tagging, to being able to provide 4-vector outputs in tasks such as momentum reconstruction. Here we summarize both architectures.

PELICAN classifier. To build a classifier, aside from the $\text{Eq}_{2 \rightarrow 2}$ equivariant layer one needs a $\text{Eq}_{2 \rightarrow 0}$ layer that reduces the rank 2 array to permutation-invariant scalars, as represented by eq. (3.2). This layer involves just 2 aggregation functions instead of 15 — the trace and the total sum of the input square array, but is otherwise identical to the equivariant layer described in the last section. The input block concatenates the embeddings of the dot products d_{ij} and the scalar data s_i as described above.



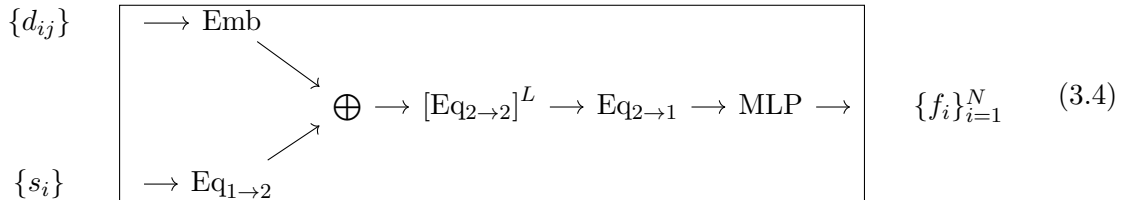
From the input block, the tensor is passed through L equivariant $\text{Eq}_{2 \rightarrow 2}$ layers, and the $\text{Eq}_{2 \rightarrow 0}$ layer with dropout. This produces a tensor of shape $[B, C_{\text{out}}]$. One final MLP mixes this down to just n_{classes} classification weights w_c per event. A cross-entropy loss function is then used for optimization.

PELICAN 4-vector regression. The same architecture can also be easily adapted for 4-vector regression tasks such as momentum reconstruction. Any Lorentz-equivariant map from a collection of 4-vectors $p_1, \dots, p_N$ to one 4-vector (or several) has the form

$$F(p_1, \dots, p_N) = \sum_{i=1}^N f_i(p_1, \dots, p_N) \cdot p_i, \quad (3.3)$$

where f_i 's are Lorentz-invariant functions, see e.g. ref. [45] for a proof. Combining this with permutation invariance, we conclude that the multi-valued map $(p_1, \dots, p_N) \mapsto (f_1, \dots, f_N)$ must also be equivariant with respect to permutations of the inputs.

The only change required to the architecture we've introduced for classification is that $\text{Eq}_{2 \rightarrow 0}$ must be replaced with $\text{Eq}_{2 \rightarrow 1}$ and the final output layer must have only one output channel (assuming we are regressing on a single 4-vector). The $\text{Eq}_{2 \rightarrow 1}$ layer is again identical to $\text{Eq}_{2 \rightarrow 2}$ except that it uses only 5 linear aggregators: taking the diagonal, row sums, column sums, trace, and full sum. The architecture is summarized by the relationship represented in eq. (3.4), where we treat f_i as the outputs, and use eq. (3.3) to recover the final predicted vector.



In summary, the simultaneous imposition of the full Lorentz and permutation symmetries lets us express a Lorentz-equivariant and permutation-invariant vector-valued model via an equivalent Lorentz-invariant and permutation-equivariant scalar-valued one. To recover the

final 4-vector output one needs only aggregate the output using eq. (3.3). As we will see, in particular tasks the coefficients f_i become identifiable physical features, making the model explainable. Crucially, this would not have been possible with only a partial Lorentz symmetry.

4 Tagging jets from Lorentz boosted top quarks

This section presents the dataset, training approach, and results of using PELICAN as a classifier in the context of identifying “Lorentz-boosted” (high transverse momentum) top quarks. Three different versions of PELICAN are discussed, each with a different size in terms both the width of the network and the number of trainable parameters. Lastly, the dependence of the performance on the size of the training dataset is also presented, providing a quantitative relationship between the size of the network, the training dataset efficiency, and the resulting performance.

4.1 Top tagging dataset

We perform top-tagging on the reference dataset [46], which was also used in ref. [8]. This dataset consists of 2M entries, each entry corresponding with a single hadronic top jet or the leading jet from a QCD dijet event. There are 1.2M training entries, 400k validation entries and 400k testing entries. The events were generated with the `PYTHIA8` event generator [47], and the `DELPHES` framework [48] was used for fast detector simulation in order to incorporate detector effects. For each jet, the 4-momentum of the 200 leading constituents are stored in Cartesian coordinates (E, p_x, p_y, p_z) , in order of decreasing p_T . This list is zero-padded, and all jets in the dataset have fewer than 200 constituents. The dataset does not contain any other information on the jet constituents, such as charge or spin.

4.2 Classification training procedure

The top-tagging model contains five $\text{Eq}_{2 \rightarrow 2}$ blocks of identical shapes (followed by a sixth $\text{Eq}_{2 \rightarrow 1}$ block). We train three different versions of the model with different widths. The widest model has 132 input and 78 output channels on every messaging layer (the equivariant layer then produces 132×15 quantities which get mixed down to 78 channels by a fully connected linear layer). The output MLP is just one layer that mixes 132 channels down to 2 classification weights. The number of jet constituents was capped at 80 (no noticeable performance gain was seen beyond that number). The dropout rate was 0.025, and the model was optimized using the `ADAMW` optimizer [49] with weight decay of 0.005. The training on the full dataset went on for 35 epochs with the same learning rate schedule as in ref. [26]: 4 epochs of linear warm-up up to learning rate of 0.001, followed by 28 epochs of `COSINEANNEALINGLR` with T_0 of 4 epochs and $T_{\text{mult}} = 2$, and then 3 epochs of exponentially decaying learning rate with exponent $\gamma = 0.5$ per epoch. We ensure that each minibatch contains an equal number of signal and background events, and shuffle the order of the minibatches at every epoch. The three models were trained on Nvidia H100 GPU’s with batch size of 100, taking 0.43, 0.17, or 0.08 seconds per batch, respectively. Inference took 0.17, 0.07, or 0.04 seconds per batch. Batches were shuffled between epochs.

Architecture	Accuracy	AUC	$1/\epsilon_B$ ($\epsilon_S = 0.3$)	# Params
TopoDNN [50]	0.916	0.972	382 ± 5	59k
EFN [24]	0.927	0.979	729 ± 13	82k
LGN [25]	0.929(1)	0.964(14)	424 ± 82	4.5k
BIP(XGBoost) [51]	0.929	0.978	600 ± 47	312
EFP [18]	0.932	0.980	384	1k
BIP(MLP) [51]	0.931	0.981	853 ± 68	4k
PFN [24]	0.932	0.982	891 ± 18	82k
DisCo-FFS [52]	–	0.982	1249 ± 43	1.4k
ResNeXt [8]	0.936	0.984	1122 ± 47	1.46M
ParticleNet [53]	0.938	0.985	1298 ± 46	498k
ParT [36]	0.940	0.9858	1602 ± 81	2.1M
LorentzNet [26]	0.942	0.9868	2195 ± 173	220k
PELICAN	0.9426(2)	0.9870(1)	2250 ± 75	208k
PELICAN _{IRC}	0.9406(2)	0.9844(11)	1711 ± 208	208k

Table 1. Comparison of different classifiers trained on the full top-tagging dataset. Note that BIP(XGBoost) and EFP are not neural networks. PELICAN’s metrics are averaged and the uncertainties are given by the standard deviation over 5 runs with different values of the random seed. PELICAN_{IRC} is an IRC-safe modification detailed below in section 10. Note that ref. [36] also includes a higher-performing model pre-trained on a large custom dataset. For the direct comparison, this table refers only to the non-pre-trained ParT model.

Depth L	Width	Accuracy	AUC	$1/\epsilon_B$ ($\epsilon_S = 0.3$)	# Params
5	132/78	0.9425(1)	0.9870(1)	2250 ± 75	208k
5	60/35	0.9424(1)	0.9868(1)	2148 ± 125	48k
5	25/15	0.9410(3)	0.9858(4)	1879 ± 103	11k
5	10/6	0.9386(2)	0.9850(1)	1494 ± 43	3k
3	6/4	0.9358(7)	0.9835(2)	1145 ± 74	1k
2	6/3	0.9336(5)	0.9823(3)	901 ± 59	605
1	6/3	0.9291(5)	0.9801(4)	669 ± 41	326
1	-/6	0.9258(8)	0.9780(6)	516 ± 52	248

Table 2. Comparison of PELICAN classifiers of varying shapes trained on the full top-tagging dataset. The depth L is the number of equivariant $\text{Eq}_{2 \rightarrow 2}$ blocks, and the width consists of two numbers. E.g. 132/78 means that each messaging block takes in 132 channels and outputs 78 channels (and the reverse for the aggregation block). The input width of the output MLP matches the messaging blocks. The last model has the fully connected messaging layers disabled altogether, so only the shape of the aggregation blocks is given. See also figure 4.

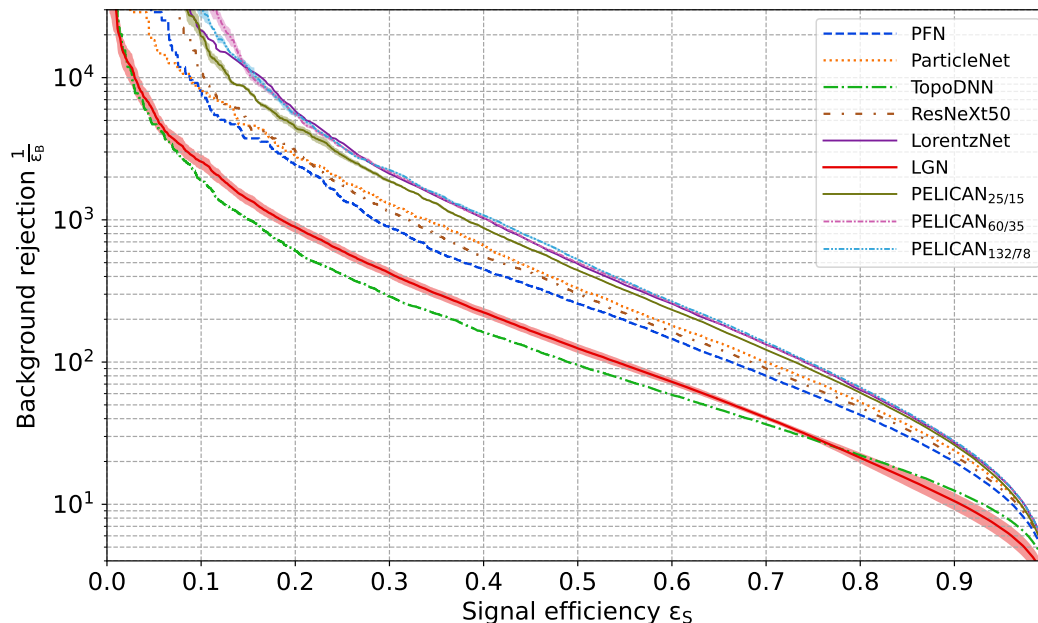


Figure 3. Performance of various ML architectures represented by the background rejection as a function of the signal efficiency.

4.3 Top tagging results

Figure 3 shows the *receiver operating characteristic*, here represented by the background rejection as a function of the signal efficiency, for the classification performance. In table 1 we compare the accuracy, area under the curve (AUC), and background rejection values at 30% signal efficiency between PELICAN and multiple existing ML top-taggers, including the previous state-of-the-art LorentzNet [26]. We also include two non-ML taggers: the Energy Flow Polynomials [18], designed to be IRC-safe; and the Boost Invariant Polynomials [51], designed to be partially Lorentz-invariant and used as inputs to XGBoost [54]. These two taggers stand out due to their low numbers of parameters and efficient utilization of physical constraints. We trained three PELICAN top-taggers with layers of differing widths, with 208k, 48k, and 11k trainable parameters respectively. The results are averaged over 5 random initialization seeds, and the uncertainties are given by the standard deviation. The large PELICAN model improves upon the LorentzNet result with a comparable number of parameters, and the medium model roughly matches LorentzNet despite having 5 times fewer parameters. Perhaps most remarkably, the small model with 11k parameters beats every pre-LorentzNet competitor despite having at least several times fewer, and up to 190 times fewer, parameters than other networks. The metrics for these three and a few even smaller models are presented in table 2, and we visualize this comparison across all models in figure 4.

In addition to different model sizes, we also explore sample efficiency. Each of the three models above was trained on 0.5%, 1% and 5% of the training data and compared to the original. For these, the training went on for 70 epochs with 60 epochs of COSINEANNEALINGLR instead of 28, and 6 epochs of exponential decay instead of 3. The results can be found in table 3. Notice that at lower amounts of training data the differences in performance between

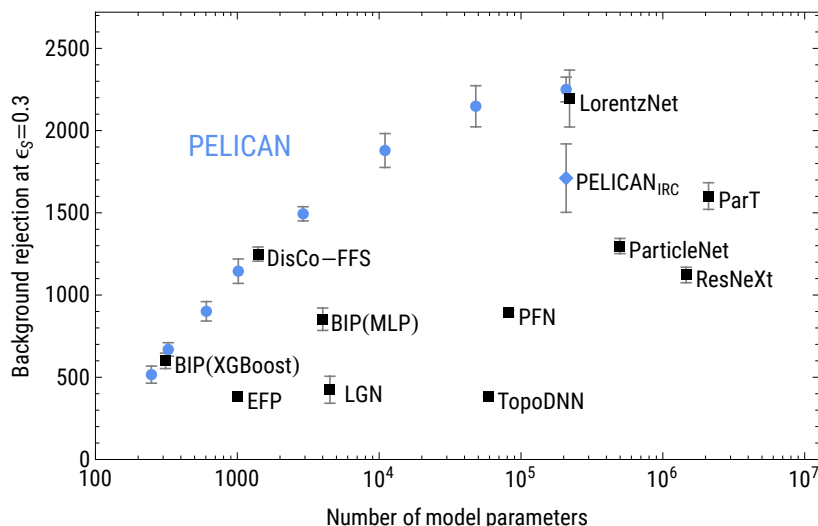


Figure 4. Comparison of top-tagger background rejection performance for fixed signal efficiency ($\epsilon_S = 0.3$) as a function of the number of parameters in each model considered, combining data from table 1 and table 2.

Width	% training data	Accuracy	AUC	$1/\epsilon_B$ ($\epsilon_S = 0.3$)
132/78	100%	0.9425(1)	0.9870(1)	2250 ± 75
	5%	0.9368(3)	0.9843(1)	1270 ± 65
	1%	0.9316(6)	0.9810(5)	789 ± 49
	0.5%	0.9289(11)	0.9800(5)	633 ± 28
60/35	100%	0.9423(1)	0.9868(1)	2133 ± 148
	5%	0.9368(2)	0.9841(1)	1148 ± 49
	1%	0.9323(3)	0.9813(4)	799 ± 52
	0.5%	0.9289(9)	0.9795(5)	637 ± 105
25/15	100%	0.9411(2)	0.9863(1)	1885 ± 109
	5%	0.9360(2)	0.9837(1)	1111 ± 108
	1%	0.9316(1)	0.9810(5)	798 ± 116
	0.5%	0.9286(11)	0.9795(6)	615 ± 133

Table 3. Comparison of PELICAN top-tagging models of different widths (depth is always $L = 5$) trained on different fractions of the training data.

models of different width become much less significant, and at 1% and 0.5% of training data all three models fall within each other's uncertainty ranges. These results suggest that the larger PELICAN networks are likely able to learn a greater range of more subtle features from the training data and thus benefit from seeing a larger training dataset. On the other hand, the primary features are already learnable with just a few percent of the data. In particular, with 5% of the training data and only 11k learnable parameters, the 25/15-channel wide version of the network achieves similar background rejection performance as ResNeXt, which uses 1.46M parameters learning on the full dataset.

Architecture	Accuracy	AUC	$1/\epsilon_B$ ($\epsilon_S = 0.3$)	$1/\epsilon_B$ ($\epsilon_S = 0.5$)	# Params
Not IRC-safe, w/ PID					
PFN-ID [24]	–	0.9052(7)	–	37.4 ± 0.7	82k
ParticleNet-ID [53]	0.840	0.9116	98.6 ± 1.3	39.8 ± 0.2	498k
ABCNet [58]	0.840	0.9126	118.2 ± 1.5	–	230k
LorentzNet [26]	0.844	0.9156	110.2 ± 1.3	42.4 ± 0.4	220k
ParT _{full} [36]	0.849	0.9203	129.5 ± 0.9	47.9 ± 0.5	2.1M
PELICAN _{PID}	0.8551(2)	0.9252(1)	149.8 ± 2.4	52.3 ± 0.3	209k
Not IRC-safe, w/o PID					
PFN [24]	–	0.8911(8)	–	30.8 ± 0.4	82k
ParticleNet [53]	0.828	0.9014	85.4	33.7	498k
PELICAN	0.8349(4)	0.9076(1)	94.5 ± 0.5	37.5 ± 0.1	208k
IRC-safe					
EFN [24]	–	0.8824(5)	–	28.6 ± 0.3	82k
EFP [18]	–	0.8919	–	29.7	1k
EMPN [59]	–	0.8932(6)	–	30.8 ± 0.2	~110k
PELICAN _{IRC}	0.8321(9)	0.9056(7)	91.7 ± 0.4	36.5 ± 0.3	9k

Table 4. Comparison of different quark-gluon classifiers trained on the quark-gluon tagging dataset [57].

5 Tagging quark and gluon jets

In this section we apply PELICAN to another binary classification task: distinguishing jets produced by gluons from those produced by light quarks. We study three variants of PELICAN for this task: both with and without the use of particle ID labels (PID), as well as a third IRC-safe version. We compare the performance of each of these three variants to other published architectures. We also compare PELICAN models of three different widths and find that even a very small PELICAN model achieves state-of-the-art tagging performance.

5.1 Quark-gluon jet dataset

We use the public dataset introduced in ref. [24]. It consists of an equal number of jets produced either by gluons or light quarks (u, d, s). The non-neutrino products are clustered using FASTJET [55] with anti- k_T jet [56] radius $R = 0.4$, and the jets are restricted to $p_T \in [500, 550]$ GeV and $|y| < 2$. There is no detector simulation, and the particle ID is stored. We select the first 200k events for validation, the following 200k for final testing, and the remaining 1.6M for training. The samples can be downloaded from [57].

5.2 Quark-gluon jet tagging results

The training procedure for the PELICAN quark-gluon tagger is identical to that of the top-tagger described above. The hyperparameters were also chosen to be the same, except here we take up to 100 highest- p_T constituents. As before, we train models of three different shapes and compare their performance. In addition, we train these models both with and without PID information supplied as a set of scalar inputs. Finally, we train IRC-safe

PELICAN variant	L	Width	Accuracy	AUC	$1/\epsilon_B$ ($\epsilon_S = 0.3$)	$1/\epsilon_B$ ($\epsilon_S = 0.5$)	# Params
Not IRC-safe, w/ PID	5	132/78	0.8551(2)	0.9252(1)	149.8 $\pm$ 2.4	52.3 $\pm$ 0.3	209k
	5	60/35	0.8544(5)	0.9245(2)	148.7 $\pm$ 2.2	51.5 $\pm$ 0.2	49k
	5	25/15	0.8506(3)	0.9216(2)	134.8 $\pm$ 0.7	48.5 $\pm$ 0.4	12k
Not IRC-safe, w/o PID	5	132/78	0.8349(4)	0.9076(1)	94.5 $\pm$ 0.5	37.5 $\pm$ 0.1	208k
	5	60/35	0.8344(1)	0.9073(1)	93.6 $\pm$ 1.1	37.4 $\pm$ 0.1	48k
	5	25/15	0.8330(5)	0.9063(3)	92.8 $\pm$ 0.5	36.8 $\pm$ 0.3	11k
IRC-safe	5	132/78	0.8294(3)	0.9034(1)	90.5 $\pm$ 1.3	35.7 $\pm$ 0.2	204k
	5	60/35	0.8291(4)	0.9030(1)	90.5 $\pm$ 1.2	35.2 $\pm$ 0.2	46k
	4	25/15	0.8321(9)	0.9056(7)	91.7 $\pm$ 0.4	36.5 $\pm$ 0.3	9k

Table 5. Comparison of PELICAN q/g classifiers of varying widths. The smallest IRC-safe model was found to have a lower optimal depth of $L = 4$.

PELICAN models so that they can be directly compared to other IRC-safe architectures. For the details of the IRC-safe implementation, see section 10.

In table 4 we compare PELICAN to other architectures trained on the quark-gluon dataset. We group the architectures according to whether they use PID inputs, and whether they are IRC-safe, to enable a more direct and fair comparison. Moreover, in table 5 we also compare the performance of PELICAN models of different sizes, analogous to the comparison in section 4. Interestingly, among the IRC-safe models the one with the lowest number of parameters performed the best. This can potentially be explained by the accumulation of floating point errors that we discuss below in section 10. PELICAN achieves impressive state-of-the-art classification performance with as few as 11k parameters, surpassing architectures of up to 190 times larger sizes. Finally, while table 4 includes only the non-pre-trained version of ParT to facilitate a direct and fair comparison, it is worth noting that the PELICAN quark-gluon tagger surpasses even the pre-trained ParT tagger as reported in ref. [36].

6 Multi-class jet tagging

In this section we consider a final, more advanced jet tagging task. It involves identifying jets that are produced by the decays of five different types of particles: gluons, light quarks, W -bosons, Z -bosons, and top quarks. We compare the performance of PELICAN models of three different widths to previously published architectures.

6.1 HLS4ML LHC jet dataset

The HLS4ML dataset consists of jets produced by simulated $\sqrt{s} = 13$ TeV proton-proton collisions using the parameters of a typical LHC detector as detailed in ref. [62]. The jets are clustered using the anti- k_T algorithm [56] with jet radius $R = 0.8$, and the jet transverse momentum is required to be around 1 TeV. There are 5 categories of jets: g, q, W, Z, t , labeled using a one-hot vector (e.g. Z is labeled $(0, 0, 0, 1, 0)$). We use the version of the dataset that includes up to 100 constituents per jet, which can be downloaded from ref. [63]. The original dataset includes 16 features per constituent, 4 of them being the Cartesian components of the 4-momentum, and the rest are various representations of the 4-momentum in cylindrical

Architecture	Gluon	Light quark	W-boson	Z-boson	Top quark	# Params
AUC						
JEDI-net	0.9529	0.9301	0.9739	0.9679	0.9683	34k
PCT	0.9623	0.9414	0.9789	0.9814	0.9757	193k
LorentzNet	0.9681(3)	0.9479(4)	0.9837(2)	0.9813(3)	0.9793(3)	224k
PELICAN	0.9693(1)	0.9493(1)	0.9840(1)	0.9816(1)	0.9803(1)	208k
TPR at FPR=0.10						
JEDI-net	0.878(1)	0.822(1)	0.938(1)	0.910(1)	0.930(1)	34k
PCT	0.891(1)	0.833(1)	0.932(1)	0.946(1)	0.941(1)	193k
LorentzNet	0.912(1)	0.855(1)	0.952(1)	0.939(1)	0.949(1)	224k
PELICAN	0.916(1)	0.860(1)	0.953(1)	0.940(1)	0.951(1)	208k
TPR at FPR=0.01						
JEDI-net	0.485(1)	0.302(1)	0.704(1)	0.769(1)	0.633(1)	34k
PCT	0.513(2)	0.298(2)	0.834(1)	0.781(1)	0.700(3)	193k
LorentzNet	0.557(4)	0.319(2)	0.800(3)	0.850(3)	0.753(3)	224k
PELICAN	0.567(1)	0.320(1)	0.804(1)	0.850(1)	0.761(1)	208k

Table 6. Three metrics of the receiver-operator curves for each jet category in the HLS4ML Jet dataset (in the “one-vs-rest” strategy), compared across architectures: area under the curve and the values of signal efficiencies (true positive rate, or TPR) at background efficiencies (false positive rate, or FPR) of 10% and 1%. LorentzNet and PELICAN results are averaged over 5 random initializations and the uncertainties are given by the standard deviation. The results for JEDI-net and PCT were taken from refs. [60, 61].

coordinates relative either to the beam axis or the jet axis. In addition, it includes a number of jet-level scalar features. For PELICAN, we need only the 4-momentum of each constituent. This simplified version of the dataset appropriate for our dataloader can be found at ref. [64]. To avoid any potential bias, the original training set was split into a fully balanced set of 567k jets and an unbalanced validation set of 63k jets (matching the sizes in ref. [61]). The final testing dataset consists of 240k jets.

6.2 Multi-class jet tagging results

We use the same PELICAN classifier architecture as above, but this time producing 5 classification scores. All other hyperparameters and the training procedure were left unchanged. In addition, we trained LorentzNet [26] on this dataset without modifying the default hyperparameters (other than the number of target classes) or the training procedure. In table 6 we compare the receiver-operator curves of different architectures using three metrics: area under the curve (AUC), and the signal efficiencies (a.k.a. TPR — true positive rate) at background efficiencies (FPR — false positive rate) of 10% and 1%. On net, both PELICAN and LorentzNet (the only two Lorentz-invariant architectures applied to this task) provide a significant improvement over older architectures, yet PELICAN surpasses even LorentzNet despite its similar model size.

Finally, we trained three PELICAN models of different widths and compared their performance in table 7. Even the small PELICAN model with only 11k parameters achieves

Width	Gluon	Light quark	W-boson	Z-boson	Top quark	# Params
AUC						
132/78	0.9693(1)	0.9493(1)	0.9840(1)	0.9816(1)	0.9803(1)	208k
60/35	0.9683(1)	0.9483(1)	0.9835(2)	0.9811(2)	0.9795(2)	48k
25/15	0.9652(2)	0.9441(3)	0.9821(2)	0.9795(3)	0.9769(3)	11k
TPR at FPR=0.10						
132/78	0.916(1)	0.860(1)	0.953(1)	0.940(1)	0.951(1)	208k
60/35	0.912(1)	0.856(1)	0.952(1)	0.938(1)	0.949(1)	48k
25/15	0.902(1)	0.841(1)	0.948(1)	0.934(1)	0.942(1)	11k
TPR at FPR=0.01						
132/78	0.567(1)	0.320(1)	0.804(1)	0.850(1)	0.761(1)	208k
60/35	0.562(1)	0.316(2)	0.801(2)	0.849(1)	0.750(2)	48k
25/15	0.538(2)	0.307(3)	0.794(3)	0.840(2)	0.719(2)	11k

Table 7. Comparison of three PELICAN models of different widths on the multi-class jet tagging task. The width for PELICAN is defined as in table 2.

state-of-the-art performance in 4 out of the 5 jet categories. Meanwhile the medium model with 48k parameters is on par with the much larger LorentzNet. We see similar (small) changes in performance as the model size is changed for this task compared to the other tasks described above.

7 W-boson 4-momentum reconstruction

To test the equivariant regression architecture described in section 3 we chose a task where the aim is to reconstruct (or *predict*) the full 4-momentum of the W -boson within the Lorentz-boosted top-quark decay products. Specifically, we consider the same hadronic top-quark decay process that constitutes the signal in the top-tagging dataset, which uses the $t \rightarrow bW \rightarrow bqq$ two-step decay, followed by hadronization, showering, and detection. Our aim is to reconstruct the true 4-momentum of the W -boson given the full set of observed final state particles of the top-quark decay, as represented by the jet constituents. The work most closely related to this task is the transformer-based reconstruction of the top quark momentum in ref. [30], and we employ similar evaluation criteria for our task.

7.1 Regression dataset

The dataset used for the regression task consists of 1.5M $t\bar{t}$ events simulated with `PYTHIA8`, via the `HEPData4ML` package [65], consisting of 700k events for training, 200k events for validation, and 500k events for testing (with an additional 100k events set aside in a second testing set). From each event, we cluster anti- k_T jets with $R = 0.8$ using `FASTJET` and we select the jet nearest to the truth-level top quark in (η, ϕ) , requiring the distance between the top quark and the jet to satisfy $\Delta R(\text{top quark, jet}) < 0.8$. This jet clustering is done both at truth-level, and using calorimeter tower objects produced by running the event through `DELPHES` fast detector simulation using the ATLAS detector card. Thus, each *event* in the dataset corresponds to a single jet, and includes information for truth-level particles such as

the truth-level top quark — we may therefore use the terms *jet* and *event* interchangeably below with the understanding that each “event” in this dataset has one and only one jet recorded. This dataset is fully reproducible since it contains the full set of parameters used for its own generation via the HEPData4ML package. It is publicly available via Zenodo [28], where a full description of the various data fields is provided. Here we provide only an overview of some key features:

1. There are two versions of the dataset, corresponding with truth- and reconstruction-level (DELPHES) jets. The events are the same between versions, so the two can be compared event-by-event to study the effects of detector reconstruction on network training and performance.
2. The input data for the network are the 4-momenta of the 200 leading jet constituents. For use as possible regression targets and for defining jet containment (explained below), each event contains
 - (a) the truth-level top quark that initiated the jet,
 - (b) the bottom quark from top-quark decay,
 - (c) the W -boson from top-quark decay,
 - (d) the two quarks from subsequent W -boson decay ($W \rightarrow qq'$),

In addition, the event contains the stable W -boson daughter particles. These are the truth-level, final state particles that are traced back to the W -boson by PYTHIA.

3. Each jet is tagged with the Johns Hopkins top tagger [7] (JH), as implemented in FASTJET. This allows us to define a subpopulation of JH-tagged events, which we shall sometimes refer to as *JH events*. For jets that it tags as top-quark jets, JH reconstructs a W -boson candidate from subjects.
4. Each jet is also tagged as whether or not it is *fully-contained* (FC). We define FC events as those where the b -quark, as well as the two quarks from $W \rightarrow qq'$ decay, are within $\Delta R < 0.8$ of the jet centroid (i.e. within the jet radius). In such cases almost all of the W -daughters are contained within the jet and we can expect a good reconstruction of the W momentum. FC events comprise 75% of the dataset.

7.2 Regression training procedure

Our model has 4 equivariant $\text{Eq}_{2 \rightarrow 2}$ blocks. Each messaging layer takes in 132 channels and outputs 78 channels. Conversely, each equivariant aggregation layer has 78 input channels and outputs 132 channels. The $\text{Eq}_{2 \rightarrow 1}$ block has the same shape, and the final fully-connected layer has the shape 1×132 . There are 210k parameters in total. Assuming N non-zero input jet constituents, this produces N scalar coefficients c_i with zero-padding, which are the Lorentz invariants as described by eq. (3.3). The reconstructed 4-momentum is then computed via

$$p_{\text{reco}} = \sum_i c_i p_i. \quad (7.1)$$

The training regimen for this task is essentially identical to the one for top-tagging: ADAMW optimizer with weight decay of 0.01, 35 epochs in total with 4 epochs of warm-up and exponential learning rate decay for the last 3 epochs. All matching hyperparameters were copied from the top-tagging model with no extra optimization. The main difference is in the choice of the loss function $L(p_{\text{reco}}, p_{\text{target}})$. Spacetime geometry allows for many choices of this function, which in turn will affect the shape of the landscape near p_{target} and in turn the precision of various reconstructed features of the vector, such as the mass, energy, spatial momentum, transverse momentum, and direction. It is even possible to construct Lorentz-invariant loss functions to make the training process itself equivariant. Nevertheless, for the purpose of simultaneous reconstruction of the direction and the mass m_W of the W -boson, we found

$$\text{Loss}(p_{\text{reco}}, p_{\text{target}}) = 0.01 \|\mathbf{p}_{\text{reco}} - \mathbf{p}_{\text{target}}\| + 0.05 |m_{\text{reco}} - m_{\text{target}}| \quad (7.2)$$

to be very effective. It uses all 4 components of the target vector and strikes a good balance between the precision of the reconstructed mass and spatial momentum. The coefficients are chosen such that each term in the loss function is order unity for our datasets.

Aside from the loss function, another rarely discussed feature of this task is the choice of the target vector p_{target} . Even though our ultimate inference target is the true W momentum p_{true}^W , it is not necessarily the best training target given the nature of the dataset. Detection and jet clustering include multiple energy, momentum, and spatial cuts that exclude some decay products from the final jet. For instance, one of the three quarks in $t \rightarrow bqq$ might fall outside of the $R = 0.8$ radius of the jet clustering algorithm, in which case most of the decay products of that quark are likely to be absent from the event record. If many of the decay products of the W -boson are missing, then we lack the information necessary to make an accurate estimate of its true momentum, or even to identify which of the jet constituents belong to the W -boson. This effect is often referred to as an *acceptance* issue due to the finite purview of the final state reconstruction.

To alleviate this issue and provide better control over the inference stage, we propose an alternative target 4-vector that we call the *contained true W momentum* p_{cont}^W , equal to the total 4-momentum of the *truth-level W decay products* that fall within the radius of the final reconstructed top jet. In the truth-level dataset, this is simply $p_{\text{cont}}^W = \sum_k p_{i_k}$ where i_k are the indices of the constituents whose parent is the W -boson and not the b -quark. In the **DELPHES** dataset, however, there is no simple analytic relationship between p_{cont}^W and the jet constituents p_i . That is to say that the mapping of the truth-level information to the detector-level reconstruction is highly nonlinear. Nonetheless, in either dataset this vector more accurately reflects the available information about the W -boson and allows us to make inferences not only about the W -boson itself, but also about the containment qualities of the event. This will be discussed further in section 7.5 below. For reference, the true mass spectra of both p_{true}^W and p_{cont}^W are shown in figure 5. For fully-contained (FC) events, the mass spectra are similar between the true and the contained W mass as expected. Non-FC events are mostly confined to a clear second peak at 13 GeV corresponding to qb and q jets (where one of the quarks from $W \rightarrow qq$ fell outside the jet), and a minor peak at $m_{\text{cont}}^W = 0$ corresponding to b jets.

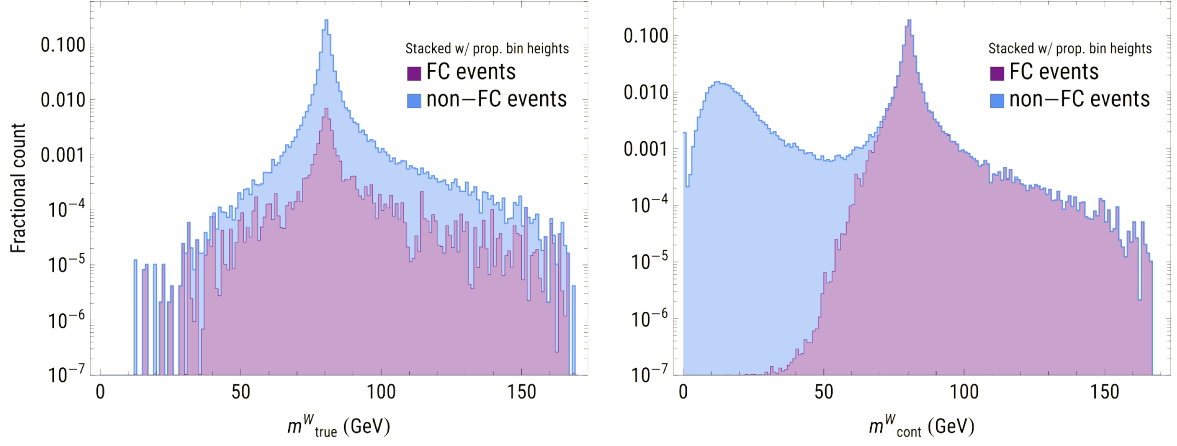


Figure 5. Stacked histogram *with proportional bin heights* showing the mass spectrum of the two targets, the true W momentum p_{true}^W , and the contained true W momentum p_{cont}^W . The left figure shows a stacked histogram of the true W mass spectrum comprised of the FC events (purple) and non-FC events (blue). The right figure shows a stacked histogram of mass spectrum for the contained W mass, similarly comprised of both FC events (purple) and non-FC events (blue). In each case, the bin contents are scaled linearly relative to the total number of events, i.e. the fraction of FC events in a given bin is given by the apparent height of the FC curve divided by the total height of the bin (heights are measured from the x -axis). The two mass spectra of FC events, in fact, match.

	Method	σ_{p_T} (%)	σ_m (%)	$\sigma_{\Delta R}$ (centrad)
Without DELPHES	JH	0.66%	1.26%	0.216
	PELICAN JH	0.26%	0.57%	0.113
	PELICAN FC	0.30%	0.71%	0.139
	PELICAN	0.79%	1.12%	0.473
With DELPHES	JH	9.8 %	8.3 %	9.6
	PELICAN JH	3.5 %	2.6 %	2.8
	PELICAN FC	4.0 %	2.9 %	3.1
	PELICAN	5.1 %	3.0 %	4.7

Table 8. Momentum reconstruction results for JH and PELICAN trained to reconstruct p_{true}^W . We report the relative p_T and mass resolutions, and the interquantile range for the angle ΔR between predicted and true momenta. “PELICAN|JH” refers to PELICAN evaluated only on JH-tagged jets, and “PELICAN|FC” to PELICAN evaluated only on FC events. PELICAN uncertainties are within the last significant digit.

Given the above observations, we prepared two PELICAN models, one trained to reconstruct p_{true}^W , and another trained to reconstruct p_{cont}^W . Otherwise the two models are identical and are trained in the same way and with the same loss function. We then compare the outputs of each model to p_{true}^W and analyze the benefits of the two choices of the target.

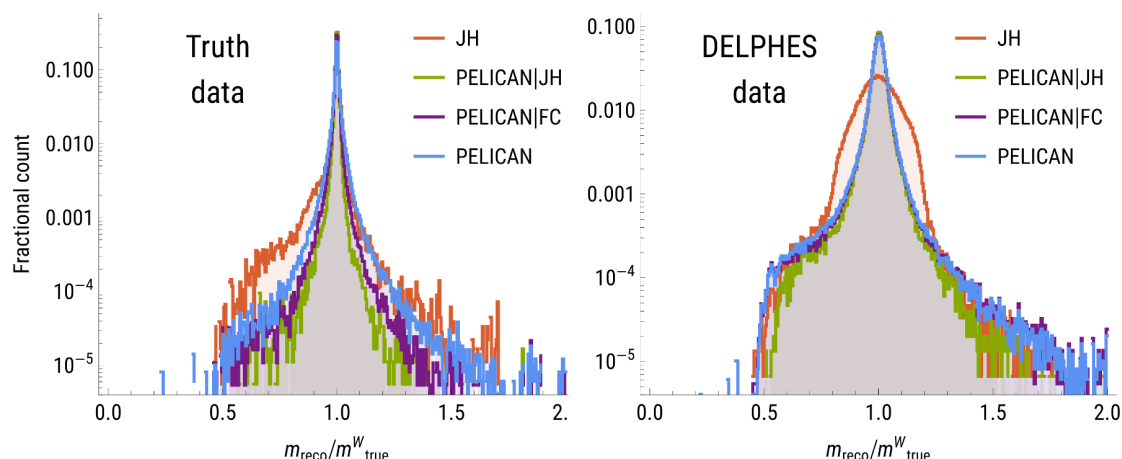


Figure 6. Reconstructed W mass relative to true W mass for the PELICAN model trained on truth (left) or DELPHES (right) data, and targeting p_{true}^W .

7.3 Regression results for p_{true}^W reconstruction

The results are summarized in table 8. We quantify the precision of the reconstruction using the resolution of transverse momentum p_T ¹ and mass as a metric, given by half of the central 68th interquantile range of $(x_{\text{predict}} - x_{\text{true}})/x_{\text{true}}$, where x is m or p_T . In addition we report the lower 68th interquantile range for ΔR , the z -boost-invariant spatial angle between predicted and true momenta.²

Since there are no pre-existing ML-based methods for this task, we use the W -boson identification of JH for the baseline comparison. JH has a 36% efficiency on the truth-level dataset and 31% on the DELPHES one. It can only identify W -boson candidates for jets it tags, so we report PELICAN results both on the JH-tagged jets only (PELICAN|JH) and on the full dataset (PELICAN). Moreover, we evaluate PELICAN on the population of FC events (PELICAN|FC). More than 99.9% of JH-tagged events contain all three true quarks bqq within the jet radius, so this population represents an especially restricted and “ideal” type of event. The results were evaluated over 5 training runs initialized with different random seeds, and the resolutions reported in table 8 are consistent across the runs.

There are significant differences in PELICAN’s performance on the different sub-populations of events. In the direct comparison with the JH tagger, PELICAN|JH is 2–4 times more precise. However, even on the much larger class of FC events, PELICAN produces predictions with almost the same precision. The highest loss of precision happens on non-FC events where many of the W decay products are missing from the jet, leading to lower average precision on the entire dataset. As discussed in section 8, this result can be *explained* by interrogating the PELICAN weights and kinematic information directly.

In figure 6 we show the relative reconstructed W masses for two of the models, one trained on truth data, and one on DELPHES data. The results also include the curve for the JH tagger’s reconstruction, as well as PELICAN|JH and PELICAN|FC. The 68th interquantile

¹ $p_T = \sqrt{p_x^2 + p_y^2}$.

² $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta \ln \tan \theta/2)^2}$.

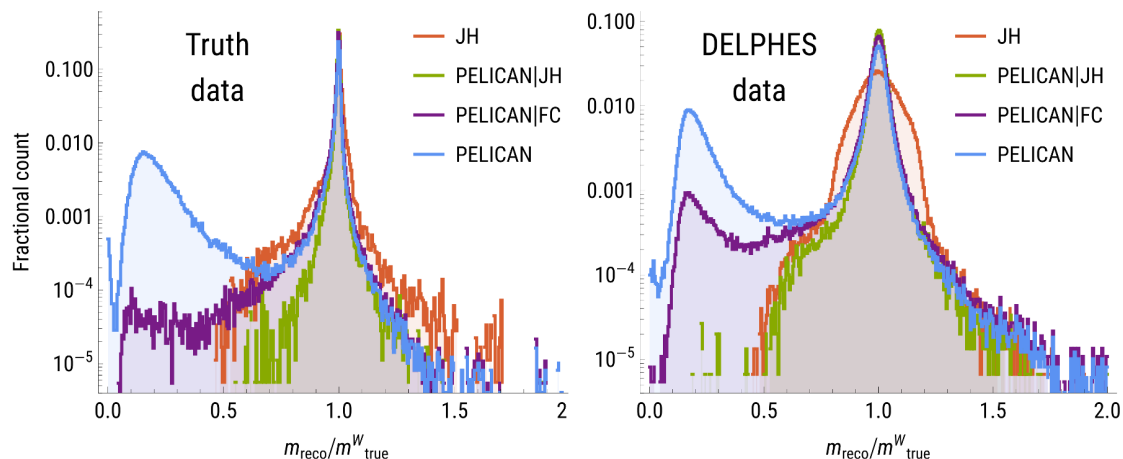


Figure 7. Reconstructed W mass relative to true W mass for the PELICAN model trained (on truth or DELPHES data) targeting p_{true}^W .

	Method	σ_{p_T} (%)	σ_m (%)	$\sigma_{\Delta R}$ (centirad)
Without DELPHES	JH	0.66%	1.26%	0.216
	PELICAN JH	0.27%	0.62%	0.113
	PELICAN FC	0.34%	0.86%	0.142
	PELICAN	2.37%	38.93%	0.681
With DELPHES	JH	9.8 %	8.3 %	9.6
	PELICAN JH	3.6 %	2.8 %	3.1
	PELICAN FC	4.2 %	3.6 %	3.4
	PELICAN	6.2 %	39.6 %	5.6

Table 9. PELICAN resolutions for models trained to reconstruct p_{cont}^W . Resolutions are still obtained by comparing the model predictions to p_{true}^W .

ranges of these curves match the numbers in the σ_m column of table 8. See section 9 for further details on the causes of performance degradation in the DELPHES case. For the complete set of results see appendix A.

7.4 Regression results for p_{cont}^W reconstruction

Now we train new models with the target vector set to the contained true W momentum p_{cont}^W , evaluate their precision by comparing the outputs to the true W momentum p_{true}^W , and compare the results to table 8. As shown in table 9, the resolutions for these models on JH-tagged and FC events are slightly worse than the first set of models, in the DELPHES case by 5–15%. The largest change is in non-FC events, leading to poor average resolutions on the whole dataset. Despite this, as we will now show, these models can in fact be better suited for real-world applications.

7.5 Discussion

To see the main benefit of this model, we present the behavior of the relative reconstructed mass shown in figure 7. PELICAN-reconstructed masses within the range of true W masses are almost as precise on the full dataset as they are on FC events (see figure 7 near the peak at 1). The most prominent feature obvious from these results is that, despite the slightly lower accuracies on FC events (at fixed width and depth of the network), the model trained to reconstruct p_{cont}^W accurately reproduces the mass spectrum of m_{cont}^W in figure 5 and therefore discriminates between FC and non-FC events, allowing us to perform post-inference event selections.

For instance, in the DELPHES case, choosing a 55 GeV cutoff, 97% of all FC events have $m_{\text{reco}} > 55$ GeV, and vice versa, 97% of all events with $m_{\text{reco}} > 55$ GeV are FC. In this manner we can significantly improve the accuracy of the reconstruction without accessing truth-level information that is needed to identify FC events. Notably, this so called “parton labeling” via regression networks was recently studied in ref. [66]. This filtering of events comes at the cost of a modest reduction in signal efficiency — from the ostensible 100% down to 75%. Note that in the DELPHES case, the set of FC events is contaminated with a small number of events with significant losses of W decay products due to detector effects, but it can be refined by reducing the jet radius used in the definition of full containment. Consequently, we propose the following simple routine for real-world applications of these models. First, use the model trained targeting p_{cont}^W as an FC-tagger to refine the data. Then, apply the model targeting p_{true}^W to reconstruct the W -boson.

We conclude that p_{cont}^W is the better target for many common reconstruction tasks where one is willing to sacrifice some signal efficiency — or to only fully measure the 4-momentum on a sub-sample of the identified events — to gain improved accuracy. In the following sections we will not present models trained on both targets, however a complete set of metrics and figures can be found in appendix A.

8 W -boson mass measurement

As we saw above, PELICAN is able to reconstruct the mass of the W -boson, m_W , found within the dense environment of the complete decay products of a top quark jet. For truth-level datasets, the resolution of this reconstruction is below the natural width [67] of the mass spectrum, $\Gamma_W/m_W \approx 2.59\%$. In the DELPHES case, the resolution is too wide to produce any substantial correlation between the true and reconstructed masses (see appendix A for figures that demonstrate this). We would like to eliminate the possibility that the reason that the true masses are highly concentrated around 80 GeV is due in part to the potential for PELICAN to effectively *memorize* a single number: the W mass. In this section we examine a more realistic reconstruction task, where the true mass of the target particle is unknown, and the dataset uniformly covers a wide range of its masses.

The reconstruction task is still identical to that of section 7. Even though we could use an invariant scalar-valued version of PELICAN to target the mass of the W -boson, the accuracy of that reconstruction would in fact suffer in comparison with the equivariant 4-vector-valued model. This is simply due to the fact that the 4-momentum contains more

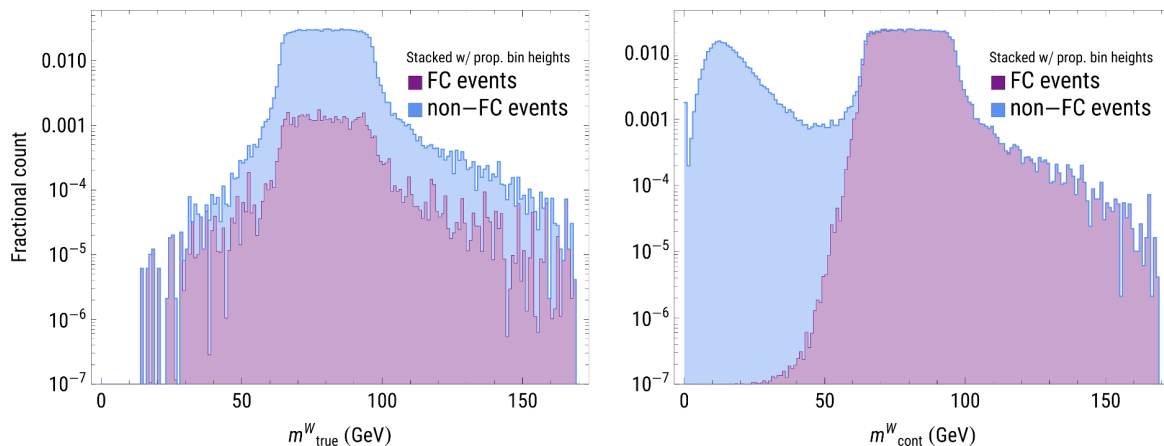


Figure 8. Stacked histogram with proportional bin heights (see description in figure 5) showing the mass spectrum of the two targets, p_{true}^W and p_{cont}^W , in the variable W mass dataset.

	Method	σ_{p^T} (%)	σ_m (%)	$\sigma_{\Delta R}$ (centirad)
Without DELPHES	JH	7.98%	4.75%	22.180
	PELICAN JH	0.27%	0.63%	0.111
	PELICAN FC	0.35%	0.89%	0.143
	PELICAN	2.64%	39.00%	0.744
With DELPHES	JH	16.0 %	12.0 %	25.4
	PELICAN JH	4.2 %	6.5 %	3.4
	PELICAN FC	4.9 %	8.0 %	3.8
	PELICAN	7.3 %	40.7 %	6.7

Table 10. PELICAN resolutions for models trained to reconstruct p_{cont}^W with variable m_W . Resolutions are still obtained by comparing the model predictions to p_{true}^W .

relevant information than the mass alone, since the direction and the energy of the particle are, in part, correlated with the mass. Thus the only new element in this experiment will be the dataset, which will now involve W -bosons of varying masses uniformly covering the range $m_W \in [64, 96]$ GeV. The dataset is also identical to that used in section 7, including the number of events, except that the W mass is set to be variable. This is achieved by combining multiple independently-produced datasets where the generator-level value of m_W was modified from its default value. Figure 8 shows the resulting distribution of W masses, as well as that of the sum of W -daughters contained within each jet.

8.1 Regression results for m_W reconstruction

The hyperparameters and the training regime used here are the same as in section 7. Here we focus on the model trained to reconstruct the contained momentum p_{cont}^W (see appendix A to find the results for the model targeting p_{true}^W). The outputs are then compared to the true W -boson p_{true}^W . The accuracies for the full 4-vector reconstruction are presented in table 10. The largest loss of accuracy relative to section 7 is, unsurprisingly, in the mass column. However, since the true mass now covers a much wider range, while the number of

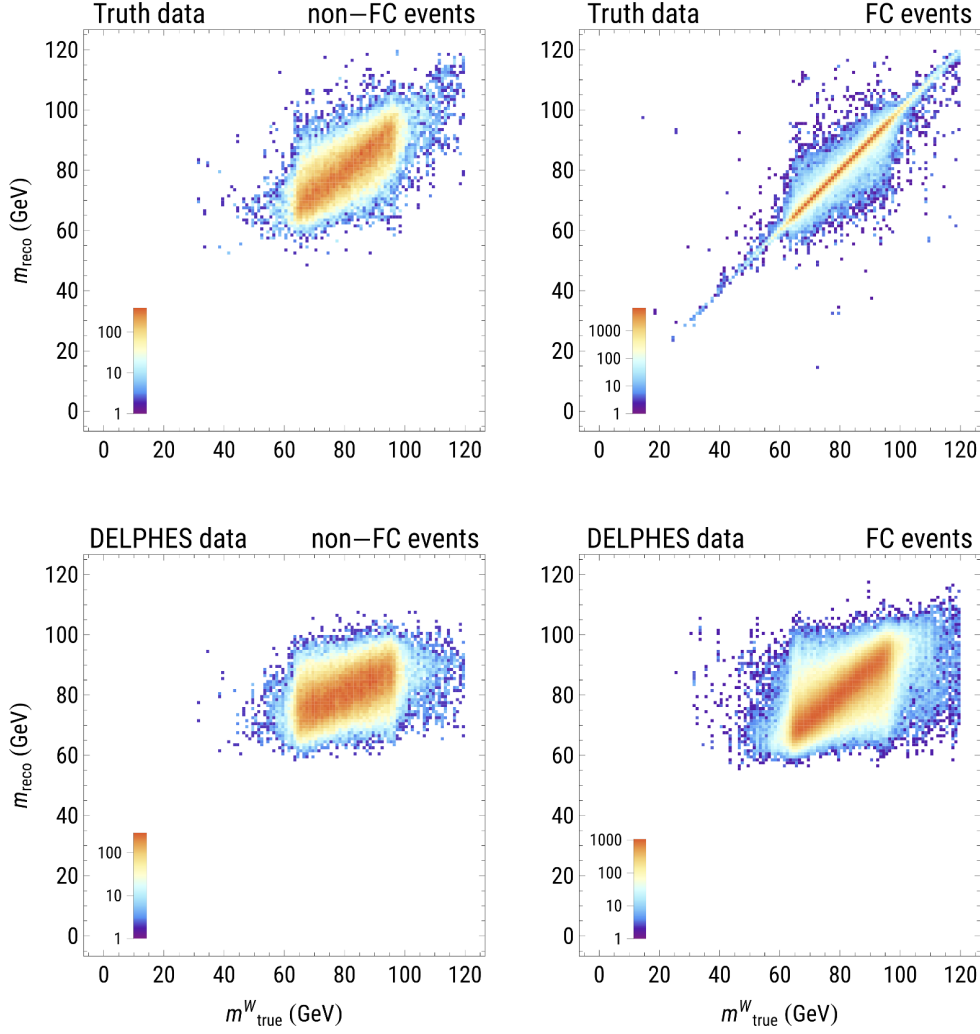


Figure 9. 2D histograms of true vs. reconstructed masses for models trained on the variable mass dataset targeting p_{true}^W (top: truth data; bottom: DELPHES data), broken up into two populations based on jet containment (left: non-FC events; right: FC events). Top right has a correlation of 88%, and bottom right 35%.

training samples remained the same, this still presents a significant improvement in the mass reconstruction capability. To demonstrate this better, we show the 2D correlations between target and reconstructed masses in figures 9 and 10 for the models trained targeting p_{true}^W and p_{cont}^W , respectively. We also differentiate between non-FC (left) and FC (right) events in the two sides of each of the panels in each figure.

8.2 Model complexity

The model examined above has 210k trainable parameters, however even significantly smaller models achieve good accuracy. As an illustration, we compare the resolutions of three PELICAN models trained on the variable mass dataset targeting p_{true}^W . They are obtained from the original model by a proportional rescaling of the widths of all layers. The first

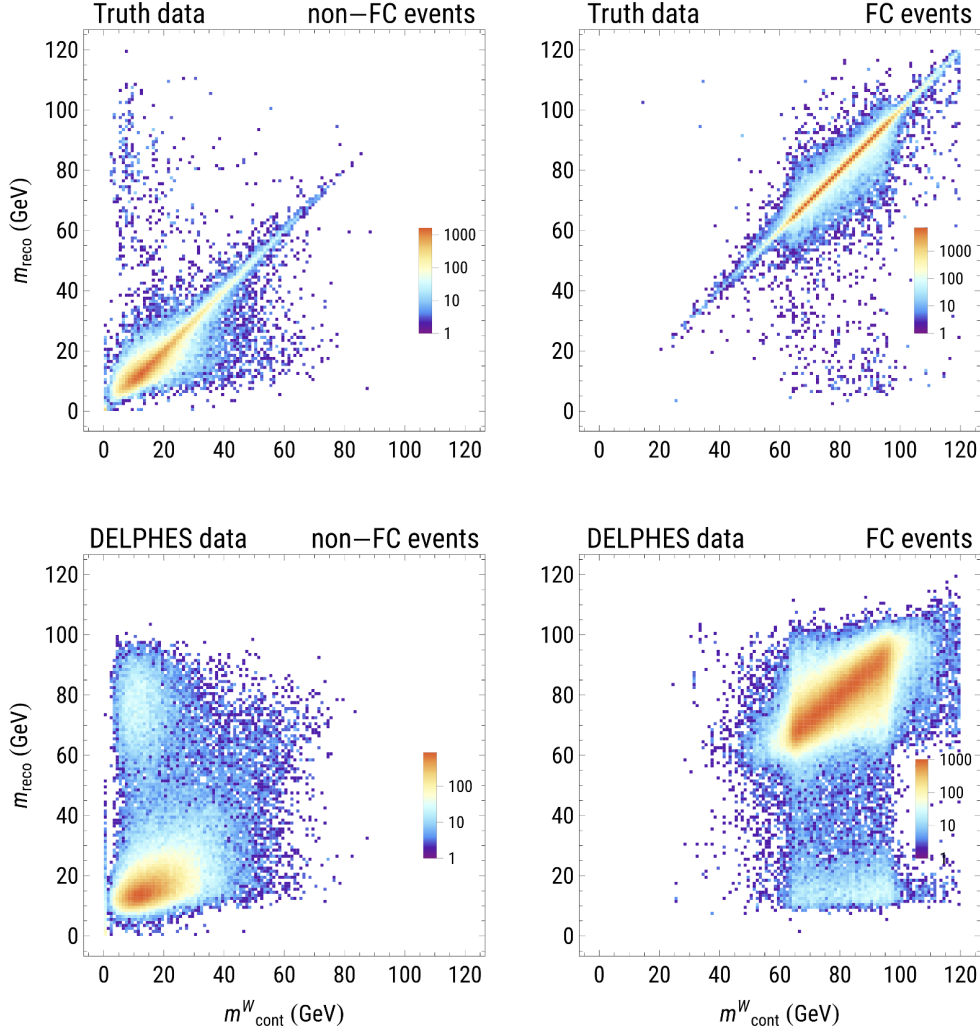


Figure 10. 2D histograms of target vs. reconstructed masses for models trained targeting p_{cont}^W (top: truth data; bottom: DELPHES data), broken up into two populations based on jet containment (left: non-FC events; right: FC events). Top right has a correlation of 95%, and bottom right 65%.

model is the 210k parameter one, with 132/78 channels, i.e. each messaging layer has 132 input and 78 output channels. The second model has 60/35 channels and 49k parameters. The third model has 25/15 channels and 11k parameters. The resolutions over the DELPHES test dataset are reported in table 11, and we observe that even the 11k-parameter model handily beats the JH method.

8.3 Discussion

In the DELPHES dataset, we observe that for non-FC events (bottom left pane of figure 10), the reconstructed contained mass is only weakly correlated with the true contained mass (or with the true W mass, as shown in figure 26 in appendix A). However, in the quadrant where both masses exceed 55 GeV, we find a 65% correlation on FC events in the DELPHES case. The most important type of error PELICAN makes here is when a non-FC event gets assigned

PELICAN width	σ_{p_T} (%)	σ_m (%)	$\sigma_{\Delta R}$ (centirad)	# Params
132/78	6.1%	8.2%	2.8%	210k
60/35	6.5%	8.6%	3.2%	49k
25/15	7.4%	9.5%	3.8%	11k

Table 11. Comparison of PELICAN models of three different widths trained to reconstruct p_{true}^W with variable W mass. Width is defined as in table 2. Trained and tested on DELPHES data.

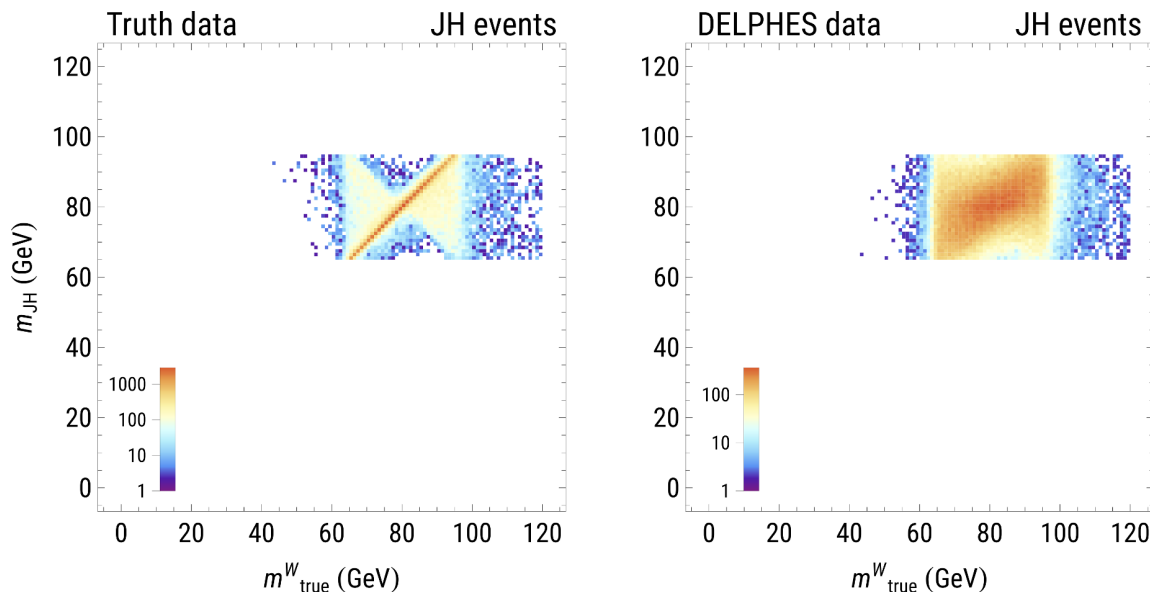


Figure 11. JH tagger’s reconstruction of the W mass on the variable W mass dataset (only JH-tagged events), truth-level and DELPHES versions. The correlation values are 47% and 25%, correspondingly.

a high reconstructed mass, that is a mass near that of the true W -boson was assigned to a jet with few of the W decay products in it. Among all events with $m_{\text{reco}} > 55$ GeV, 3.6% are non-FC, and they bring the correlation among that population down to 51% (p_T , mass, and angular resolutions on this population closely track those of PELICAN|FC above). But since in practice we’re interested in m_{true}^W , the correlation between that and m_{reco} is higher, at 59% among events with $m_{\text{reco}} > 55$ GeV. This is a significant improvement over the model trained on the original $m_{\text{true}}^W \sim 80$ GeV DELPHES dataset, and especially over non-ML methods such as the JH tagger (see figure 11).

Therefore a workflow that guarantees both high background rejection and high reconstruction quality would involve first using a model trained on p_{cont}^W as a classifier to filter out well-contained events, and then using a model trained on p_{true}^W to obtain a high-precision reconstruction on that population. However, even a model trained on DELPHES data to reconstruct p_{true}^W , in fact, achieves a 40% correlation with m_{true}^W on non-FC events (see figure 9), so FC-tagging may not always be necessary. Overall, PELICAN provides a viable method for estimating Lorentz-invariant particle masses.

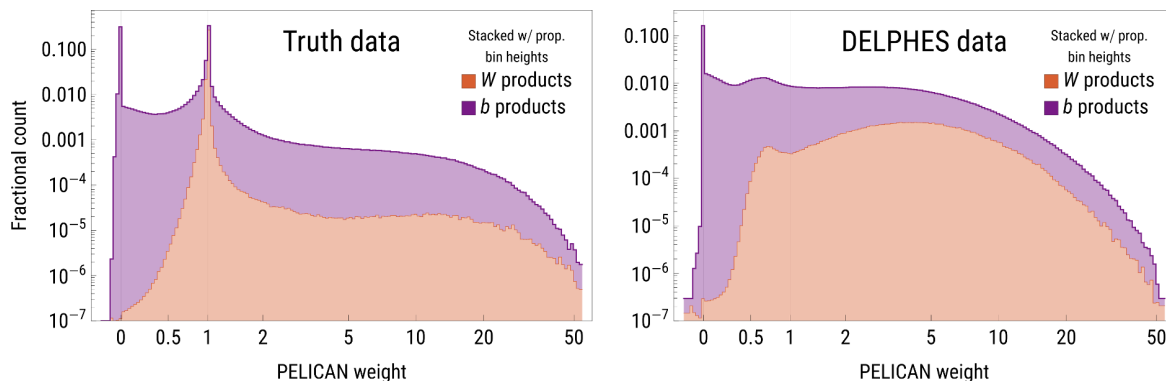


Figure 12. Stacked histograms with proportional bin heights of all PELICAN weights computed over the testing dataset for the 4-vector reconstruction task from section 7 using models trained targeting p_{true}^W . Broken up into two populations — W -boson products and b -quark products. In the DELPHES case, a constituent is considered a W -boson product if the corresponding calorimeter cell detected at least one true W -daughter.

9 PELICAN explainability

Despite the output of the PELICAN regression model ostensibly being a 4-vector (or multiple 4-vectors), the richer and more natural object to treat as the output are the PELICAN weights $\{c_i\}$ introduced in eq. (7.1). Each c_i is attached to its corresponding input constituent p_i due to permutation equivariance and therefore encodes *a scalar feature of that particle within the event*. As we will show in this section, the behavior of these weights is key to the unique explainability and visualization features of the PELICAN architecture.

In essence, PELICAN is able to take a set of N input 4-vectors and assign N scalar features to them (of course there can be several features per input as well) in a Lorentz-invariant way. This can be powerful in a variety of applications, but in the context of particle reconstruction the problem of finding the right values of the weights is similar to a soft clustering problem. Assuming an idealized dataset with perfect information about the decay products, the model should identify the decay products of the W -boson, assign $c_i = 1$ to them, and zero to all other constituents. This is analogous to what taggers like the Johns Hopkins top-tagger aim to do via jet clustering. However, since any five 4-vectors are linearly dependent, there is a continuum family of solutions $\{c_i\}$ and it is not clear that PELICAN will prefer the clustering solution. This section is dedicated to analyzing how well this intuition agrees with the actual behavior of the weights, and to explaining the deviations from this naïve picture caused by detector effects.

9.1 Distributions of PELICAN weights

In figure 12 we display the distributions of all PELICAN weights for models from section 7 trained targeting p_{true}^W . We also mark each constituent as either a W - or a b -daughter. This yields several observations.

Firstly, nearly all weights are either non-negative or very slightly negative (e.g. above -0.1) with a very sharp peak at zero (the peak is entirely to the left of zero to very high precision³).

³The bin $[-10^{-6}, 0)$ contains about 100 times more constituents than the bin $[0, 10^{-6})$.

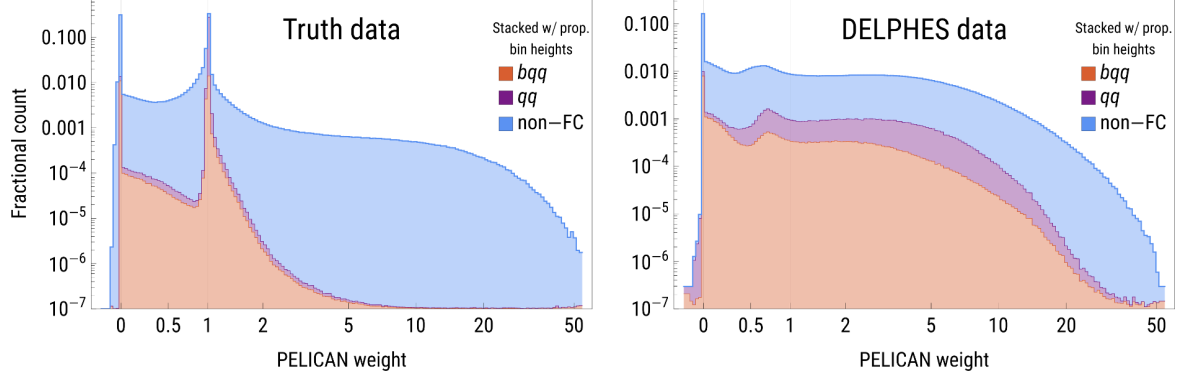


Figure 13. Stacked histograms with proportional bin heights of all PELICAN weights computed over the testing dataset for the 4-vector reconstruction task from section 7 using models trained targeting p_{true}^W . Broken up into three populations by jet containment: bqq events (all three truth-level quarks from the $t \rightarrow bW \rightarrow bqq$ process fall within the jet clustering radius); qq events (only the b -quark fell outside of the jet); and non-FC events, which include bq , b , and q events.

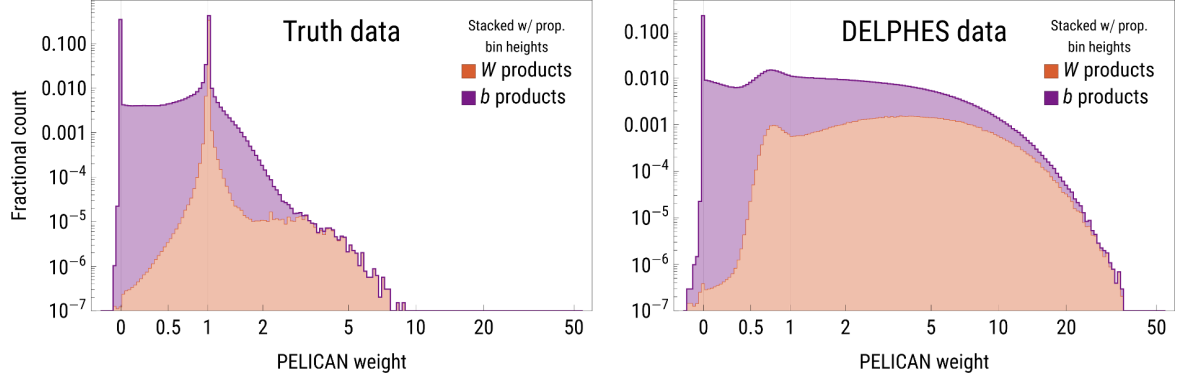


Figure 14. Stacked histograms with proportional bin heights of all PELICAN weights for the 4-vector reconstruction task from section 7 using models trained targeting p_{cont}^W . Broken up into two populations by parent type.

This is the first feature that justifies the interpretation of PELICAN as a *soft clustering* method. Since our inputs represent realistic events, all input 4-vectors in them are timelike, as is the target vector. This implies that no linear combination of these vectors with positive coefficients can produce a zero vector. The distributions, therefore, show that PELICAN weights assigned to b -daughters are not “contaminated” with these degenerate combinations.

Secondly, the truth-level distribution is highly concentrated at 0 and 1 and very closely matches the binary clustering solution. That is, almost all constituents assigned weight 0 are b -daughters, and almost all of those assigned 1 are W -daughters. Nevertheless, 30% of b -daughters are assigned positive weights, prompting further investigation. Moreover, the distribution of W -daughter weights in the DELPHES case is so spread out that it becomes difficult to explain it via a mere analogy with clustering.

We can delve more deeply into the weight distribution by separating the sub-populations of weights based on jet containment. Figure 13 shows the distributions of weights for bqq ,

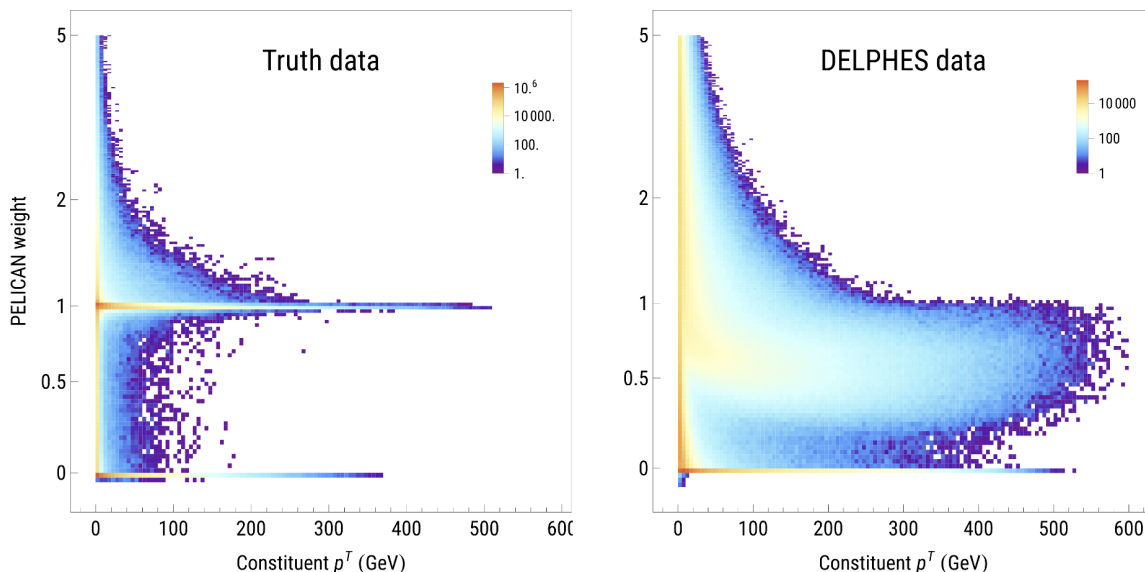


Figure 15. 2D histogram of PELICAN weights vs constituent transverse momentum for the 4-vector reconstruction task from section 7 using models trained targeting p_{cont}^W . Only FC events shown here.

qq , and non-FC events. The majority of constituents at the high end of the weight scale belong to non-FC events. Similarly, the weights produced by the models trained targeting p_{cont}^W , shown in figure 14, are more highly concentrated at 0 and 1, and have much lower and shorter “tails” on the right, especially among b -daughters. This is the first indication that PELICAN tends to upweight some constituents in events where it doesn’t have enough information for an accurate reconstruction.

This approach allows us to characterize the constituents that are being upweighted. Figure 15 shows the constituent weight as a function of the constituent’s p_T . The main observation here is that among high-energy (“hard”) constituents with $p_T > 100$ GeV the weight distribution is bimodal, and the vast majority of constituents with weights away from the two peaks are soft, below 20 GeV. In the DELPHES case PELICAN appears to downweight high-energy W -daughters and upweight soft constituents. Once again, loss of information in the form of detector effects appears to lead to PELICAN upweighting soft constituents.

9.2 Detector effects on PELICAN weights

While the truth-level PELICAN models reliably converge to a binary clustering solution, the weights in the DELPHES case do not permit such a straightforward interpretation. To better understand their behavior, we ran additional experiments using custom datasets that exclude different components of the DELPHES detector simulation one by one. DELPHES performs the following steps: simulate the effect of the magnetic field B_z on charged final-state particles; aggregate truth-level particle energies and apply configured cuts within each electromagnetic calorimeter (ECAL) and hadronic calorimeter (HCAL) detector cell; apply energy smearing by sampling a lognormal distribution; unify the ECAL and HCAL cells; apply spatial smearing by picking a uniformly random point within the detector cell; scale the magnitude of the spatial momentum so that the resulting 4-vector, which represents a detector cell, is massless. We

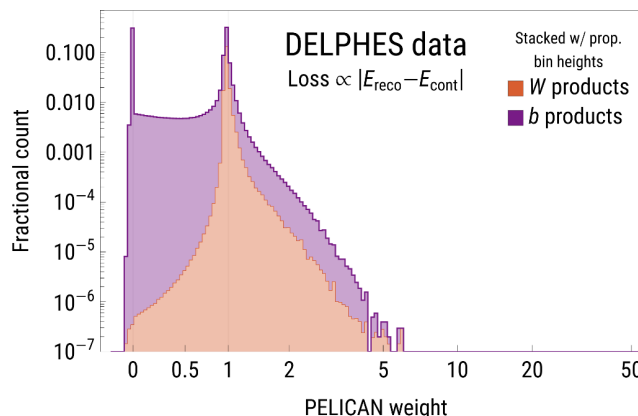


Figure 16. Same as figure 14 but this time the model is trained using a single-term loss function proportional to $|E_{\text{reco}} - E_{\text{cont}}|$. We conclude that the shape of the DELPHES distribution in figure 14 is overwhelmingly due to *spatial* detector effects.

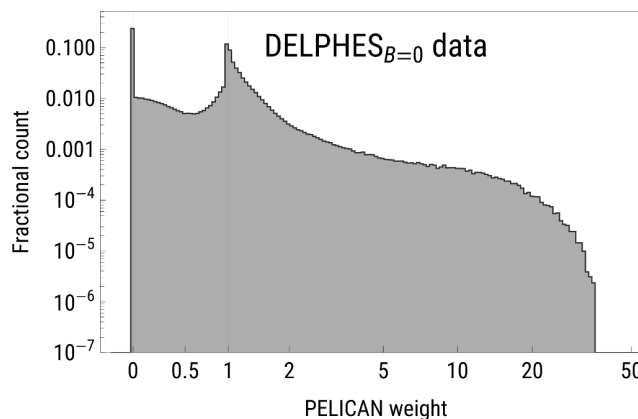


Figure 17. Distribution of PELICAN weights for a model trained on a dataset that differs from a true DELPHES detector simulation only in that the simulated magnetic field is disabled. The model was trained to reconstruct p_{true}^W , so it should be compared to figure 12.

found that while each of these steps contributes to smearing out the truth-level distribution of PELICAN weights and shifting the peak downwards, the magnetic field is responsible for almost all of the differences between truth and DELPHES results.

The simulated magnetic field is able to deflect charged particles very significantly, enough to account for most of the error in PELICAN’s reconstruction relative to the truth-level reconstruction. Our hypothesis for why this leads to lower PELICAN weights for hard constituents is the following: deflected hard particles produce large errors in the direction but not the energy of the reconstruction, and therefore one can downweight them and compensate for the energy deficit using softer constituents. Moreover, by upweighting softer constituents PELICAN can in fact partially correct the error in the direction since the deflections of positively-charged particles can be partially cancelled out by those of negatively-charged particles.

An extra piece of evidence in support of this hypothesis can be found by modifying the loss function. If we re-train the model on the same DELPHES dataset using a loss function

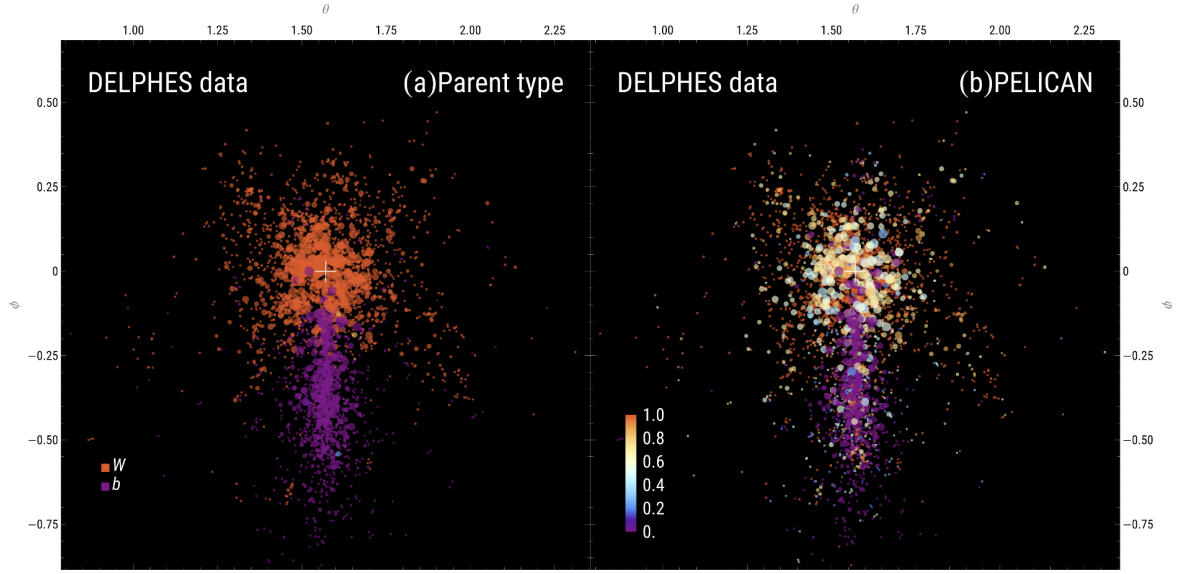


Figure 18. Composite event display of 200 events from the **DELPHES** dataset from section 7. Each event is transformed using a 3D rotation matrix such that the true W -boson ends up at $(\theta, \phi) = (\pi/2, 0)$ (white cross), and the true b -quark is directly below. Each dot is a **DELPHES** constituent and the dot size increases logarithmically with constituent energy. (a) Color reflects parent type: constituents that are fully derived from W -daughters are orange and those from b -daughters are purple; in the rare cases when the fraction of W -derived energy in a given calorimeter cell is between 0 and 1, the corresponding color is taken from the color scale in the right pane. (b) Color reflects the value of the **PELICAN** weight, clipped to the interval $[0, 1]$, as shown in the legend. Note how the hardest W -boson constituents (largest dots) tend to have **PELICAN** weights between 0.5 and 1.

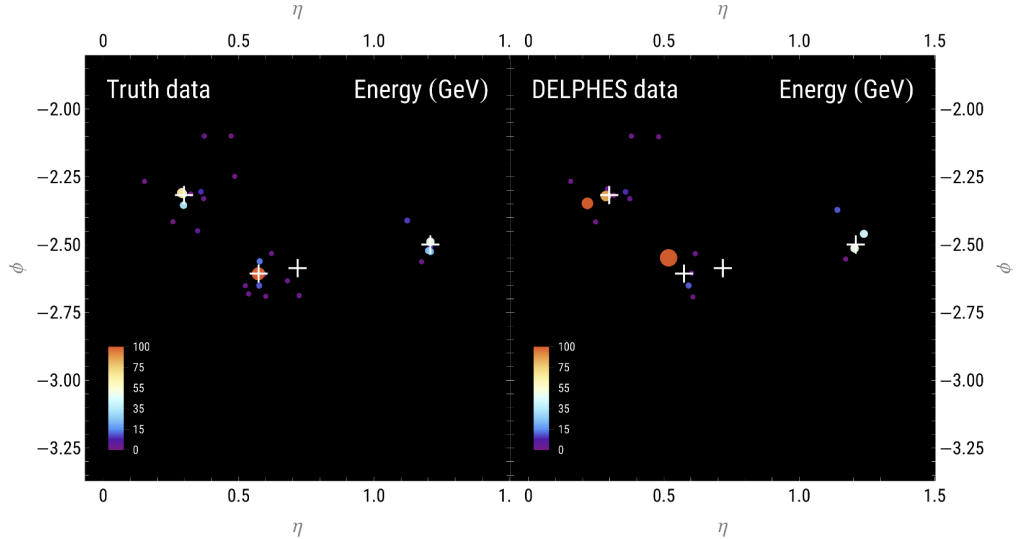


Figure 19. A single event viewed in the η, ϕ plane with color and size dependent on energy. The central cross marks the true W -boson, and the other three crosses mark the three true quarks from the process $t \rightarrow bqq$.

consisting of a single energy term $|E_{\text{reco}} - E_{\text{true}}|$, we find a distribution of weights (see figure 16) nearly as bimodal as the original one trained on truth-level data (see figure 14). This indicates that the source of the error in PELICAN’s reconstruction on DELPHES data is overwhelmingly *spatial*. Out of all the steps that DELPHES performs, only two are purely spatial: momentum smearing within one cell, and the simulated magnetic field. However, the detector cells (approximately 0.02×0.02 in (η, ϕ)) are much smaller than the magnitude of PELICAN’s typical angular error, and thus smearing alone cannot explain the error.

For an ultimate confirmation of the hypothesis that the reduced precision of the DELPHES model is overwhelmingly caused by the magnetic field, we generated another dataset that differs from a true DELPHES dataset only in that the simulated magnetic field is set to zero. Training the same architecture on this dataset to reconstruct p_{true}^W produces a model with overall resolutions $\sigma_{p_T} = 3.1\%$, $\sigma_m = 2.6\%$, and $\sigma_{\Delta R} = 1.6$ centirad. These are significantly better than the DELPHES model in table 8, particularly the angular resolution. However, the largest difference can be seen in the distribution of the PELICAN weights, shown in figure 17. It is much closer to the truth-level model in figure 12 than the DELPHES one, which proves that the vast majority of the smearing of the PELICAN weights is caused by the magnetic field. The effect of the magnetic field can likely be mitigated by adding charge information to the input in the DELPHES’ energy-flow scheme, which we intend to explore in a future study.

9.3 Event visualization

As we discussed above, despite being a single-vector regression model, PELICAN produces one feature *per input constituent* (namely the weight c_i), and these features become interpretable by virtue of eq. (7.1). This gives us a unique opportunity to make event-level visualizations that provide insight into how PELICAN treats jet topology and how it compares to conventional methods such as the JH tagger’s jet clustering.

In figure 18 we show an amalgamation of 200 events from the DELPHES dataset from section 7 projected onto the unit sphere. Each event was spatially rotated so that the position of the true W -boson within the image is fixed and the true b -quark is located in the negative ϕ direction. Due to the rotational invariance of PELICAN weights, this normalization does not affect the inference. In one display the constituents are colored according to their parent being either the W -boson or the b -quark, and in the other they’re colored based on their assigned PELICAN weight. The correlation between the two images is clear: b -daughters tend to be correctly assigned zero weight, whereas W -daughters have positive weights with the hardest constituents having weights between 0.4 and 0.8.

In figure 19 we show a single event in the (η, ϕ) plane, with dot color and size dependent on the constituent energy. Note the reduced number of constituents in the DELPHES display, and how some of the constituents get strongly deflected by the simulated magnetic field. The same event can be visualized in three more helpful ways. In addition to parent type and PELICAN visualizations introduced in figure 18, we can also extract the list of constituents that the JH tagger identifies as belonging to the W -boson and highlight them. Figure 20 displays the same single event in all three ways. In addition, we add a special marker for the direction of the reconstructed W -boson. In the parent type pane, this reconstruction is defined as $\sum_{i=1}^N r_i p_i$ where r_i is the energy of the true W -daughters within that constituent

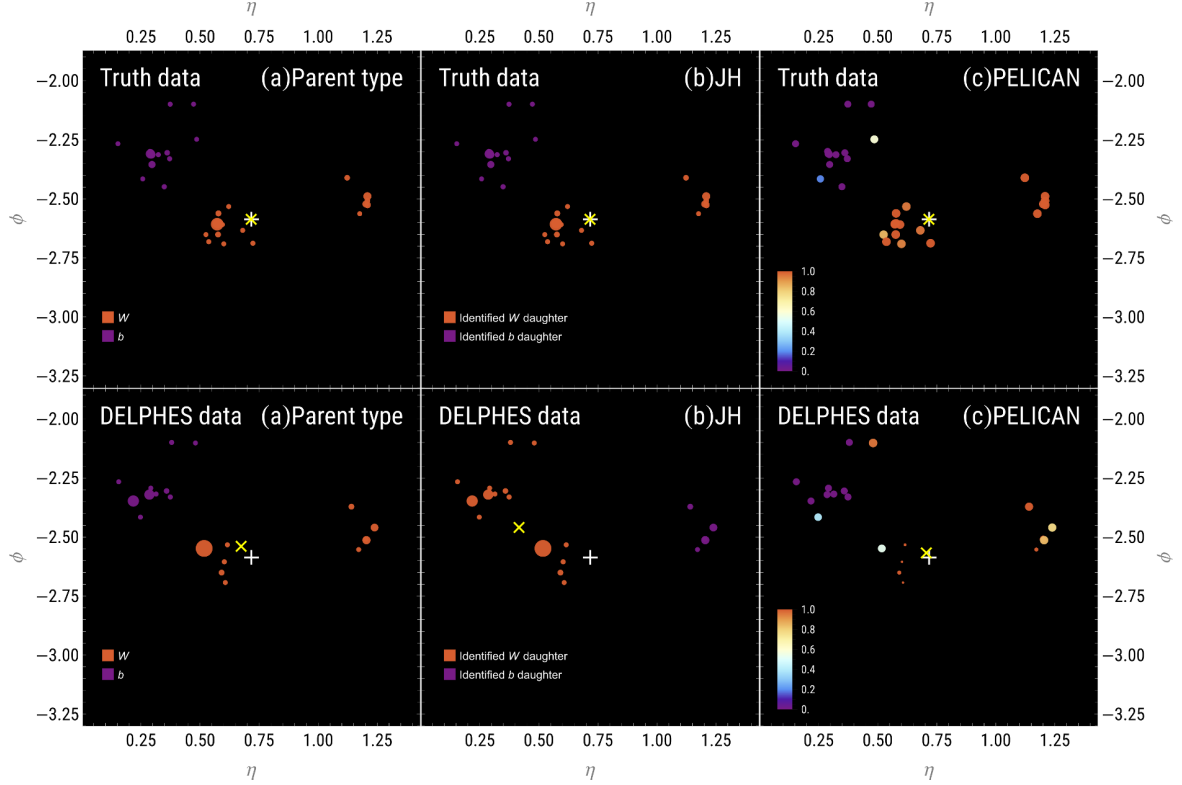


Figure 20. The same event as in figure 19 in the (η, ϕ) plane. (a) Constituents are colored according to the actual parent type; size increases nonlinearly with energy; the yellow cross marks the reconstruction obtained by summing all of the constituents that belong to the W -boson. (b) Constituents are colored according to how they are tagged by the JH-tagger as either W -daughters or not; size increases with energy; the yellow cross marks the JH-reconstructed W -boson. (c) Constituents are colored according to their PELICAN weight clipped to the interval $[0, 1]$; size increases as the weight goes from 0 to 1 and decreases after that. Note how the soft DELPHES W -constituents get assigned high PELICAN weights.

divided by the actual energy of the constituent. In the JH and PELICAN panes, the yellow marker corresponds to the respective reconstructions obtained by those methods.

9.4 Explainability

We claim that the interpretation of the PELICAN weights as clustering coefficients constitutes explainability of the PELICAN 4-vector regression model. We are able to fully understand the way PELICAN produces its 4-vector output in terms of a set of physical features of the input constituents which cannot be inferred from the training target — namely, their identities as W -daughters or b -daughters. This is possible only due to the combination of the *full* Lorentz symmetry and permutation symmetry, which guarantees the universality of eq. (3.3). This approach can be contrasted with that of ref. [66], where unmatched top quarks are assigned to jets by averaging attention weights in a regression network trained to reconstruct kinematic properties of the top quarks across all layers.

10 IRC-safety and PELICAN

10.1 Definitions

Perturbative computations in QCD suffer from a divergence caused by two types of processes: soft emission and collinear splittings. As a consequence, meaningful observables in this theory need to be insensitive to such processes, and this requirement is known as IRC-safety. In this section we provide a precise definition, give a characterization of IRC-safe Lorentz-invariant observables (see details in appendix B), and describe modifications to the PELICAN architecture that make it IR-safe or IRC-safe.

Infrared safety (IR-safety) guarantees insensitivity to soft emissions, i.e. particles with relatively low energies and momenta. A family of continuous symmetric observables $f^{(N)}(p_1, \dots, p_N)$ is said to define an IR-safe observable f if

$$\lim_{\epsilon \rightarrow 0} f^{(N+1)}(p_1, \dots, p_N, \epsilon p) = f^{(N)}(p_1, \dots, p_N) \quad (10.1)$$

for any N and any $p_1, \dots, p_N, p$, where ϵ controls how infinitesimally small the considered soft emission p is.

Collinear safety (C-safety) is a restriction on observables in perturbative QCD that arises from the divergent contributions of collinear emissions of gluons. Positive gluon mass would prevent such divergences, which is why C-safety concerns only massless particles. We can define C-safety formally as follows: an observable $f(p_1, \dots, p_N)$ is C-safe if, whenever two massless 4-momenta p_1 and p_2 become collinear (which happens for massless particles iff $p_1 \cdot p_2 = 0$), the value of f depends only on the total momentum $p_1 + p_2$. Expressed even more explicitly, C-safety says that setting $p_1 = \lambda p$ and $p_2 = (1 - \lambda)p$ with some 4-vector p such that $p^2 = 0$ must lead to the same output regardless of the value of $\lambda \in (0, 1)$, i.e.

$$C_{12}(p)f = \partial_\lambda f(\lambda p, (1 - \lambda)p, p_3, \dots, p_N) = 0, \quad \lambda \in (0, 1). \quad (10.2)$$

In appendix B we characterize a certain class of IRC-safe Lorentz-invariant observables in terms of polynomial bases, but the following summary will suffice for the purpose of designing an IRC-safe version of PELICAN. First, a Lorentz-invariant observable (assumed to be consistently defined for any finite number N of 4-vector inputs) is IR-safe if and only if it has no explicit dependence on the multiplicity N . More precisely, adding the zero 4-vector to the list of inputs should leave the output value invariant. Second, an IRC-safe Lorentz-invariant observable is one that is IR-safe and moreover depends on any of its massless inputs in such a way that when any number of them become collinear, the observable ends up depending only on the sum of their energies. E.g. if p_1, p_2, p_3 are fixed to be massless, then $f(p_1, p_2, p_3, p_4, \dots)$ must depend only on $p_1 + p_2 + p_3$ when $p_1 \parallel p_2 \parallel p_3$. Note that such an observable is still completely unrestricted away from the massless manifold.

It is instructive to reflect on the interplay between IR-safety and C-safety. C-safety allows us to exchange momentum between collinear particles without affecting the values of C-safe observables, whereas IR-safety allows us to omit vanishing 4-momenta from the list of inputs. C-safety in itself doesn't change the number of input 4-vectors — it only requires f to be invariant under a certain transformation that mixes collinear massless inputs. Meanwhile IR-safety effectively requires f to encode an infinite family of observables $f^{(N)}(p_1, \dots, p_N)$

for any possible number of inputs with the compatibility condition $f^{(N)}(p_1, \dots, p_{N-1}, 0) = f^{(N-1)}(p_1, \dots, p_{N-1})$.

The original PELICAN architecture as introduced above is neither IR- nor C-safe. Below we modify the architecture to make it exactly IRC-safe and evaluate the implications.

10.2 IRC-safe PELICAN

To enforce IRC-safety in PELICAN, we follow a strategy almost identical to that of ref. [59], based on the original idea from ref. [24], with the correction that our equivariant layers are a generalization of message passing and that we need to take extra care to preserve Lorentz symmetry. The basic idea is that if F_i are a permutation-equivariant set of IRC-safe observables of our inputs $\{p_i\}$, then so is any observable of the form

$$f\left(\sum_{i=1}^N z_i F_i\right), \quad (10.3)$$

where f is any scalar function and z_i are appropriately picked energy-dependent weights, commonly chosen to be either fractional energy $E_i/\sum_j E_j$ or fractional transverse energy $p_i^T/\sum_j p_j^T$. IR-safety is guaranteed due to the fact that any soft constituent has $z \rightarrow 0$ and the expression above is invariant under the insertion of such terms. C-safety is slightly more involved: if, say, p_1 and p_2 are massless and collinear, then due to the C-safety and permutation-equivariance of F_i , we have $F_1 = F_2$, and all F_j depend on these two particles only through $z_1 + z_2$ and their common spatial direction. Eq. (10.3) then guarantees that the new observable also depends on the magnitudes of these two vectors only through the sum $z_1 + z_2$. All of these arguments obviously also apply to higher numbers of collinear particles. It is also important for us to note that this property applies to permutation-equivariant observables as well.

To apply this idea to PELICAN, we first need a Lorentz-invariant analog of the energy weights z_i and the corresponding normalized vectors $\hat{p}_i$. Since the only permutation-invariant inertial frame that can be defined based on the list of the constituents p_i is the frame of the jet $J = \sum_i p_i$, it is natural to define the jet-frame energies

$$\mathcal{E}_i = \frac{p_i \cdot J}{m_J} = \frac{\sum_{j=1}^N d_{ij}}{\sqrt{\sum_{j,k=1}^N d_{jk}}}, \quad (10.4)$$

where m_J is the invariant jet mass. The energy weights are then simply the fractional jet-frame energies, and the de-dimensionalized 4-momenta are rescaled by these energies:

$$z_i = \frac{\mathcal{E}_i}{\sum_{j=1}^N \mathcal{E}_j}, \quad \hat{p}_i = \frac{p_i}{\mathcal{E}_i}, \quad \hat{d}_{ij} = \hat{p}_i \cdot \hat{p}_j = \frac{d_{ij}}{\mathcal{E}_i \mathcal{E}_j}. \quad (10.5)$$

Note that $\hat{d}_{ij}$ ⁴ are Lorentz-invariant and constitute the natural input to the IRC-safe version of PELICAN. We are finally ready to modify the architecture to make it IRC-safe. Combining the idea from eq. (10.3) with the definition of equivariant aggregators in eq. (2.6), we come to the following prescription for the $\text{Eq}_{2 \rightarrow 2}$ blocks:

⁴If the inputs are massless, $\hat{d}_{ij} = 1 - \cos \Theta_{ij}$, where Θ_{ij} are the pairwise spatial angles in the jet frame.

- The 5 order zero aggregators don't need to be modified since the application of any nonlinear function to an IRC-safe observable preserves its IRC-safety.
- The 8 order one aggregators involve summation over one equivariant input index, $\sum_i \bullet_i$, and all of these need to be modified by including the Lorentz-invariant energy weights: $\sum_i z_i \bullet_i$.
- The 2 order two aggregators are similarly adjusted with two energy weights: $\sum_{i,j} z_i z_j \bullet_{i,j}$.

Note that the aggregation function *has to be* based on summation, since max/min pooling or any other nonlinear aggregation will immediately fail C-safety. The way the outputs of these aggregations are embedded into the output $N \times N$ array is unchanged since it doesn't affect IRC-safety. Since the inputs $\hat{d}_{ij}$ are IRC-safe and since eq. (10.3) guarantees recursive IRC-safety, *every individual component* of the output array of every IRC-safe equivariant block is IRC-safe.

The same kind of prescription can also be used to make the $\text{Eq}_{2 \rightarrow 1}$ (for 4-vector regression), $\text{Eq}_{1 \rightarrow 2}$ (for scalar inputs), and $\text{Eq}_{2 \rightarrow 0}$ (for classification) layers IRC-safe. As a consequence, the PELICAN weights c_i become IRC-safe. The regression case deserves special attention, since strictly speaking the PELICAN weights c_i themselves don't have to be IRC-safe for the combination $p_{\text{reco}} = \sum_i c_i p_i$ to be IRC-safe. In particular, whereas the IR-safety of p_{reco} necessarily makes it possible to find IR-safe c_i , there is no such criterion for C-safety.

However, under the interpretation of c_i 's as “soft clustering” coefficients (as discussed in section 9) it does make sense to require IRC-safety from them. In that case the values of weights corresponding to collinear inputs necessarily match due to permutation symmetry. Indeed, IRC-safety implies that if, say, p_1 and p_2 are collinear, then both c_1 and c_2 are functions of only their spatial direction and of $E_1 + E_2$. Since the direction and the total energy are permutation-invariant but c_1 and c_2 are permutation-*equivariant*, i.e. $c_1(p_1, p_2, \dots) = c_2(p_2, p_1, \dots)$, we conclude that $c_1 = c_2$. This in turn also guarantees the IRC-safety of the reconstructed 4-vector. Nonetheless, whether IRC-safe PELICAN can approximate *any* IRC-safe vector-valued observable, and whether there is a more general way of generating such observables, is an open problem deserving future investigation.

Finally, since the multiplicity N is not IR-safe, the aggregation function cannot be defined using means. However, there exist IRC-safe analogs of N such as the Soft Drop multiplicity defined in ref. [68].⁵ Nevertheless, even they can't be directly used in PELICAN due to their non-Lorentz-invariance. For PELICAN tests, we have modified the Soft Drop algorithm by executing it *in the jet frame*, making it manifestly Lorentz-invariant, and defined the Lorentz-invariant Soft Drop multiplicity n_{SD} that way. This is also equivalent to replacing energies and angles in the original definitions of the Cambridge/Aachen [69, 70] and Soft Drop algorithms with their Lorentz-invariant analogs defined in eq. (10.4) and eq. (10.5). With this, it is safe to use aggregators of the form $n_{\text{SD}}^\alpha \sum_{i=1}^N \bullet_i$.

⁵The Soft Drop multiplicity actually measures the depth of the Cambridge-Aachen branching tree along the hard core of the jet. Its definition also involves several new hyperparameters denoted z_{cut} , θ_{cut} and β , see ref. [68].

10.3 Testing IRC-safe PELICAN models

First we quantify the deviation in PELICAN’s outputs that occurs under soft and collinear splittings and observe how training affects them. We define an IR-splitting as adding a zero 4-vector to the list of input constituents. Then PELICAN’s output on IR-split data is directly compared to the original output. Defining a C-splitting is more difficult since realistic events never contain any exactly collinear constituents, and we want to avoid changing the number of particles so as to make this test independent of IR-safety. Therefore we prepare the data by inserting two copies of the vector $(1, 0, 0, 1)$ to each event. Then the C-splitting will amount to replacing these two vectors with $(1.5, 0, 0, 1.5)$ and $(0.5, 0, 0, 0.5)$ (recall that C-safety does not apply when $\lambda = 1$ in eq. (10.2) because the zero vector is “collinear” with anything). The outputs on the same event prepared in these two ways can be directly compared.

To compare two outputs $p_{\text{reco}}, p'_{\text{reco}}$ we compute the relative deviation $|(p'_{\text{reco}} - p_{\text{reco}})/p_{\text{reco}}|$, where the division is component-wise. To estimate the effect of an IR- or C-splitting on PELICAN’s predictions, we take the median value of this deviation over a batch of events. The same can also be done with PELICAN weights as the outputs. The splittings are applied to 100-event batches of events from one of our datasets and the relative deviations are averaged over 300 batches. We test 5 randomly-initialized models and 5 models trained on the full variable W mass dataset from section 8.

We find that a randomly-initialized PELICAN regression model’s output 4-vector deviates by 0.5%-14% (measured by the deviation of each of the 4 Cartesian components) under an IR-split, and the PELICAN weights deviate by up to 9%. After training on one of our datasets the range of these deviations doesn’t appear to change. Under the C-safety test defined above, the 4-vector prediction of a randomly initialized model deviates by an absolute amount of up to 15 GeV, and the median absolute deviation of the PELICAN weights can be as high as 0.03 (depending on the random seed). With IRC-safety enabled these go down to 10^{-4} GeV and 10^{-7} , respectively. The main reason these deviations aren’t even smaller under IRC-safety is the accumulation of numerical errors after repeated aggregation with the energy weights z_i . The values of these weights on realistic data are highly concentrated near zero, and overall they can span up to 8 orders of magnitude. After several layers of aggregation with weighting by $z_i z_j$ the spread of the values goes way beyond the size of the standard floating point types, which limits the precision of IRC-safety. However, for practical purposes the current precision of PELICAN is likely to be sufficient, seeing as it is still dominated by the uncertainties originating from the random initialization.

The resolutions σ_{p_T}, σ_m , and $\sigma_{\Delta R}$ of the IRC-safe DELPHES models (using aggregations weighted by Soft Drop multiplicities) trained on the variable W mass dataset targeting p_{true}^W are, respectively, 6.8%, 8.5%, and 3.1 centirad, which is only about 10%, 3%, and 8% worse (higher) than the resolutions of the original non-IRC-safe models as reported in table 12. Here is how these deviations in the values of the outputs reflect on the overall quality of the predictions. Under an IR-splitting (adding a zero vector to every event) the resolutions of trained non-IRC-safe models get worse by between 0.3% and 2% depending on the initialization. Under an “IRC-splitting” (splitting the first beam $(1, 0, 0, 1)$ into two constituents $(0.5, 0, 0, 0.5)$) the resolutions get worse by between 10% and 25%. Therefore non-IRC-safe PELICAN models can be quite sensitive to IRC splittings, especially collinear ones.

Aside from regression, we also trained IRC-safe PELICAN classifiers for top-tagging and quark-gluon tagging. These results are included alongside the non-safe models in table 1 and table 4. In both cases PELICAN provides state-of-the-art performance among IRC-safe models. In quark-gluon tagging, IRC-safe classifiers perform almost as well as the non-IRC-safe analogs, which confirms the conclusions of the recent study in ref. [71]. Notably, the IRC-safe PELICAN top-tagger used the Lorentz-invariant Soft Drop multiplicity in place of N in its aggregators, whereas the quark-gluon tagger only used simple summation aggregators (i.e. N was replaced with a fixed constant). This makes the comparison in table 4 fair, but at the same time we expect even better performance from a model that uses Soft Drop multiplicity.

11 Conclusion

We have presented a full description of PELICAN: a network that respects particle permutation and Lorentz symmetries important in particle physics. PELICAN is a general network which is performant at 4-vector regression and provides state-of-the-art performance in the tasks of top-tagging and quark-gluon tagging. The IRC-safe modification of the network is similarly leading among other such architectures. PELICAN also achieves state-of-the-art performance in the more difficult task of multi-class jet identification. Due to the equivariant architecture, all of this is made possible despite PELICAN's relatively compact model size.

To demonstrate PELICAN's regression capabilities, we chose the reconstruction of the W -boson's 4-momentum from a full top quark jet, and to our knowledge PELICAN is the first ML method applied to this problem. Even within these tasks there is room to improve PELICAN's performance by introducing additional scalar information such as particle charges, which would allow the network to account for the simulated collider's magnetic field.

PELICAN's most unique features, however, have to do with its relatively low complexity (in terms of the number of parameters) and explainability. For example, it provides highly competitive top-tagging performance with only several hundred parameters. At the same time, in regression tasks the "PELICAN weights" represent a new kind of output that not only has a direct physical interpretation, but also allows one to perform unprecedented particle-level ML-based analysis of particle jets. PELICAN's architecture, its flexibility, and generalizability may also allow for future applications to charged-particle track reconstruction, pile-up identification, and full-event reconstruction. Being a general architecture, PELICAN is not limited to top quark decays or even jets. This network inherently provides powerful tools for investigating its own behavior due to the combination of Lorentz and permutation equivariance and shows promise as a tool which can be thoroughly investigated if deployed in real world scenarios.

Acknowledgments

The authors would like to thank the Data Science Institute at the University of Chicago for its generous support of this research. TH is supported by the Department of Physics at the University of Chicago. DWM and JTO are supported by the National Science Foundation under Grant PHY-2013010. The computations in this work were, in part, run at facilities supported by the Scientific Computing Core at the Flatiron Institute, a division of the

Simons Foundation. The Center for Computational Mathematics at the Flatiron Institute is supported by the Simons Foundation. In addition, we are grateful to Andrew Larkoski for insightful comments.

A Additional results and plots

A.1 W -boson 4-momentum reconstruction

Below is the list of additional figures for the models trained on the W -boson 4-momentum regression dataset from section 7:

- Figure 21: histograms corresponding to table 8 (models trained targeting p_{true}^W);
- Figure 22: histograms corresponding to table 9 (models trained targeting p_{cont}^W);
- Figure 23: histograms for models trained targeting p_{cont}^W .
- Figure 24 distributions of PELICAN weights for models trained targeting p_{cont}^W ;
- Figure 25: target vs. reconstructed mass correlation histograms for models trained targeting p_{true}^W ;
- Figure 26: true W -boson vs. reconstructed mass correlation histograms for models trained targeting p_{cont}^W ;
- Figure 27: target vs. reconstructed mass correlation histograms for models trained targeting p_{cont}^W .

A.2 W -boson mass measurement

Below is the list of additional tables and figures for the models trained on the variable mass dataset from section 8:

- Table 12: resolutions for models trained targeting p_{true}^W ;
- Figure 28 histograms corresponding to table 12 (models trained targeting p_{true}^W);
- Figure 29: histograms corresponding to table 10 (models trained targeting p_{cont}^W and compared to p_{true}^W);
- Figure 30: histograms for models trained targeting p_{cont}^W and compared to p_{cont}^W ;
- Figure 31: target vs. reconstructed mass correlation histograms for models trained targeting p_{cont}^W .

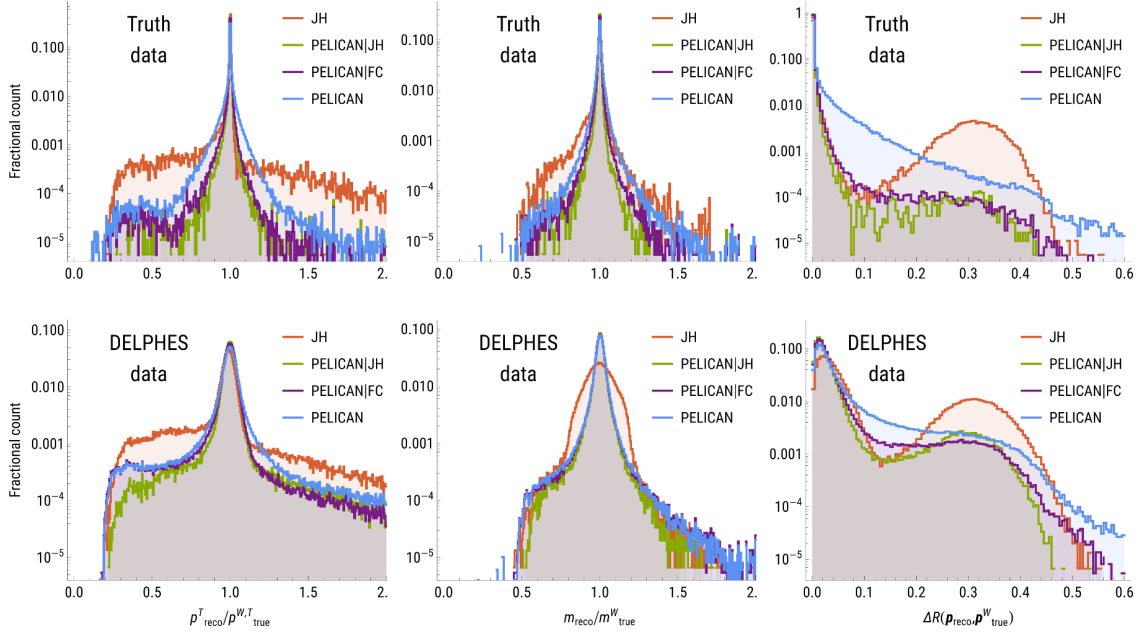


Figure 21. Full set of histograms corresponding to the entries in table 8 (models trained targeting p_{true}^W , truth-level and DELPHES versions).

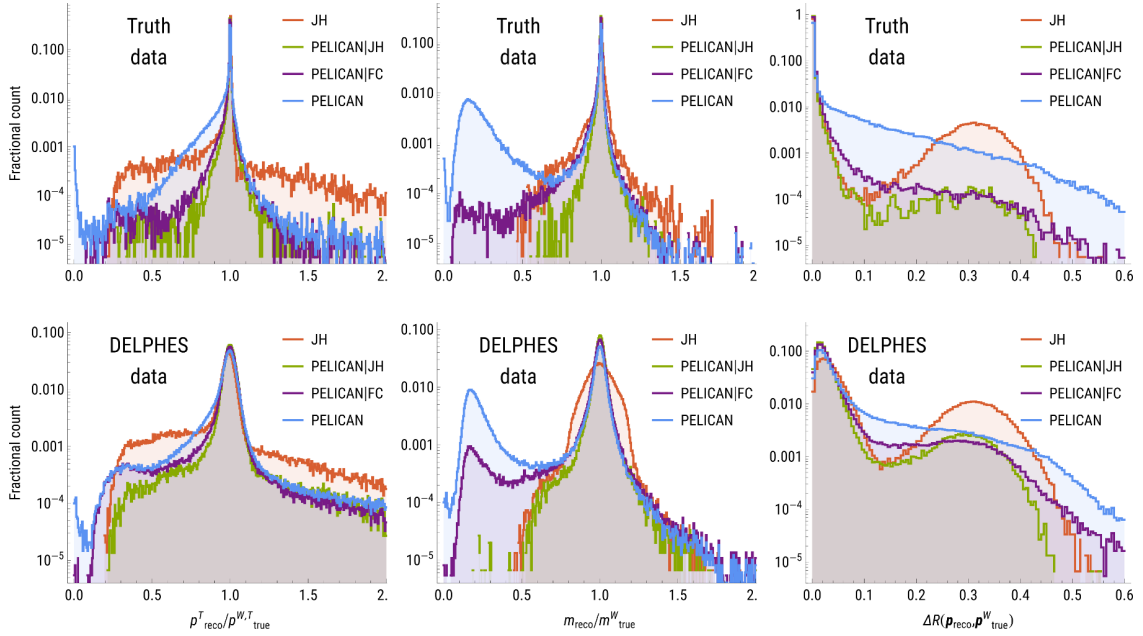


Figure 22. Full set of histograms corresponding to the entries in table 9 (models trained targeting p_{cont}^W , truth-level and DELPHES versions, and compared to p_{true}^W).

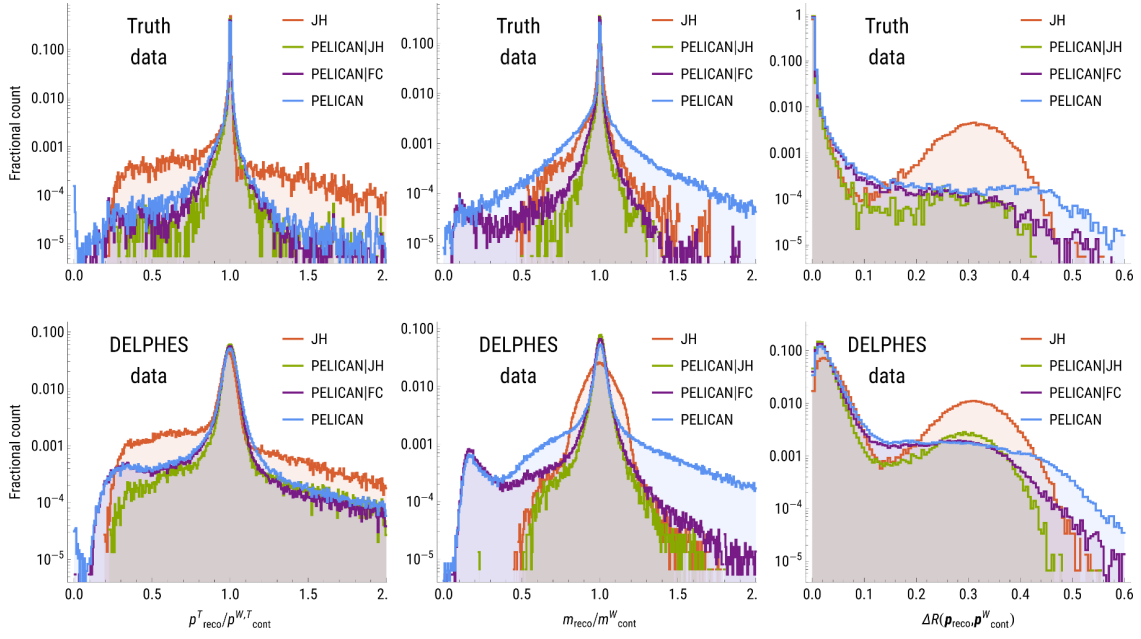


Figure 23. Full set of histograms for models trained targeting p_{cont}^W and compared to p_{cont}^W .

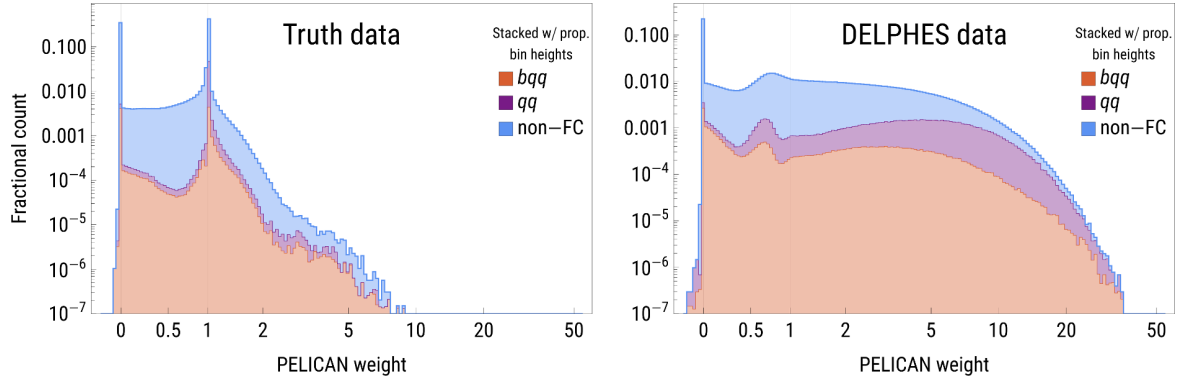


Figure 24. Distributions of PELICAN weights analogous to figure 13 but for models trained targeting p_{cont}^W .

	Method	σ_{p^T} (%)	σ_m (%)	$\sigma_{\Delta R}$ (centirad)
Without DELPHES	JH	7.98%	4.75%	22.180
	PELICAN JH	0.26%	0.58%	0.111
	PELICAN FC	0.31%	0.76%	0.142
	PELICAN	0.88%	1.67%	0.548
With DELPHES	JH	16.0 %	12.0 %	25.4
	PELICAN JH	4.1 %	6.3 %	3.2
	PELICAN FC	4.7 %	7.3 %	3.5
	PELICAN	6.2 %	8.2 %	5.6

Table 12. PELICAN resolutions for models trained to reconstruct p_{true}^W with variable W mass.

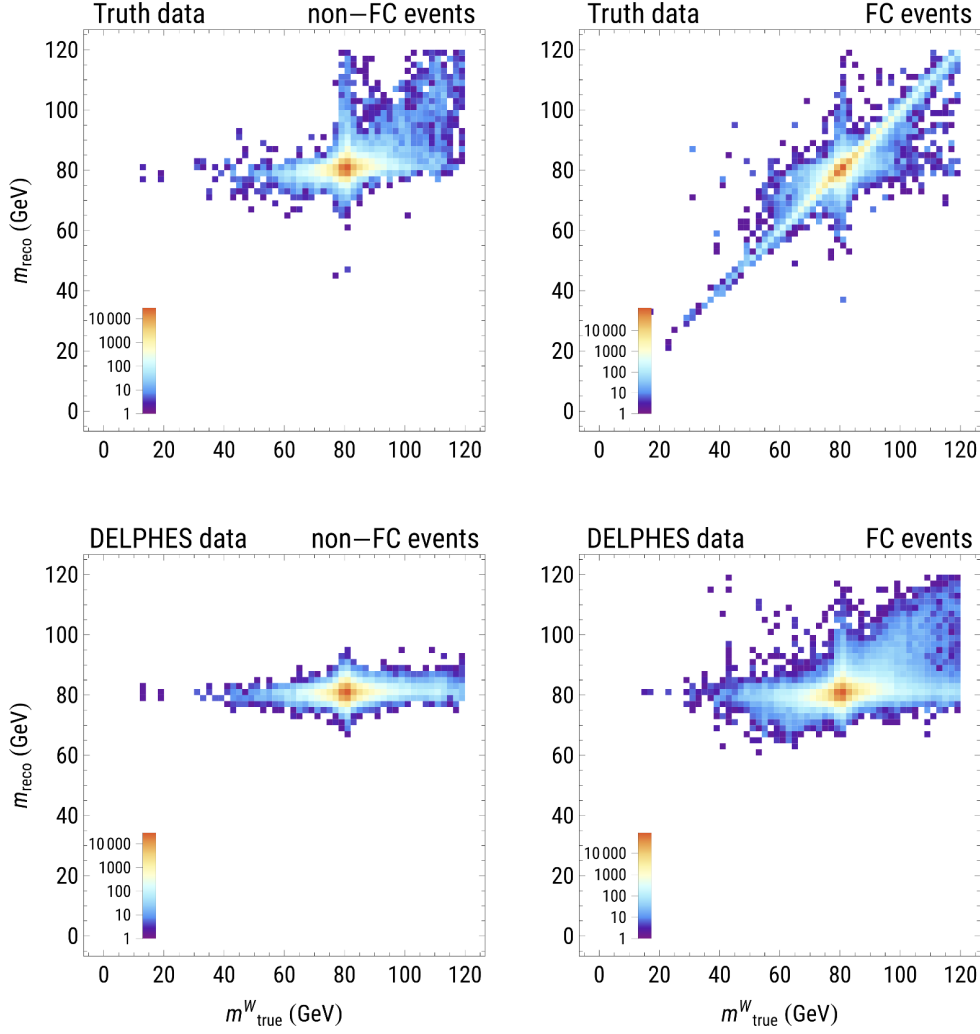


Figure 25. 2D histograms of target vs. reconstructed masses for models (top: truth data; bottom: DELPHES data) trained targeting p_{true}^W , broken up into two populations based on containment (left: non-FC events; right: FC events).

A.3 Additional event displays

Below is the list of additional event displays:

- Figure 32 an event display showing how PELICAN processes a qq event that JH fails to tag;
- Figure 33: an event display showing how PELICAN processes a bqq event that JH fails to tag.

B IRC-safety and Lorentz symmetry

Let us now try to characterize the IRC-safe Lorentz invariants of a set of jet constituents. When working with Lorentz-invariant observables, it is convenient to rewrite everything in

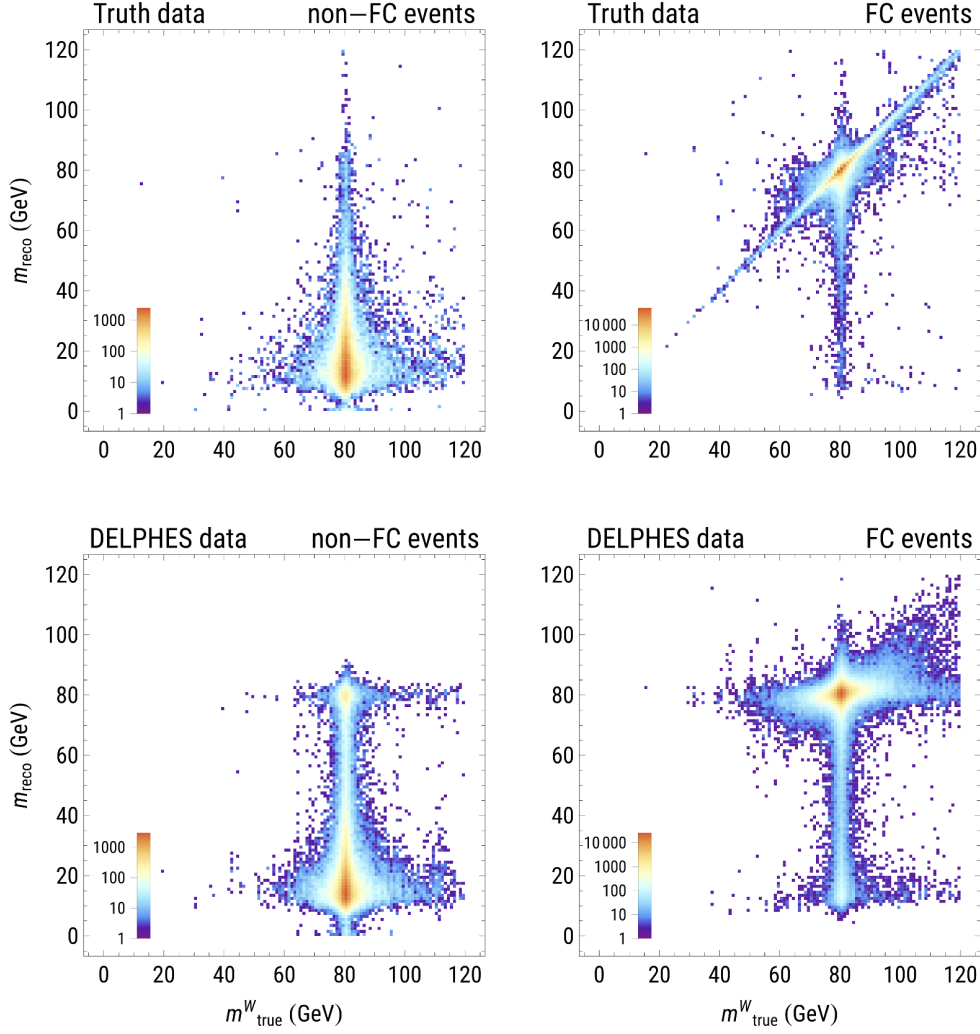


Figure 26. 2D histograms of true W mass vs. reconstructed contained mass for models (top: truth data; bottom: DELPHES data) trained targeting p_{cont}^W , broken up into two populations based on containment (left: non-FC events; right: FC events).

terms of Lorentz invariant coordinates — the dot products $d_{ij} = p_i \cdot p_j$ (note that these coordinates are in general not independent since any 5×5 minor of the matrix $\{d_{ij}\}$ must vanish due to the fact that any five 4-vectors are linearly dependent, but they will suffice for our purposes). For two massless particles, collinearity is equivalent to the vanishing of the dot product. Conversely, the dot product between two time- or light-like vectors is zero only if both are light-like and collinear. Thus the C-safety condition amounts to the equality of the gradients of f with respect to any two of the rows (or columns) of $\{d_{ij}\}$ when evaluated on the corresponding coordinate hyperplane:

$$\left. \frac{\partial f}{\partial d_{ki}} \right|_{d_{ij}=0} = \left. \frac{\partial f}{\partial d_{kj}} \right|_{d_{ij}=0}. \quad (\text{B.1})$$

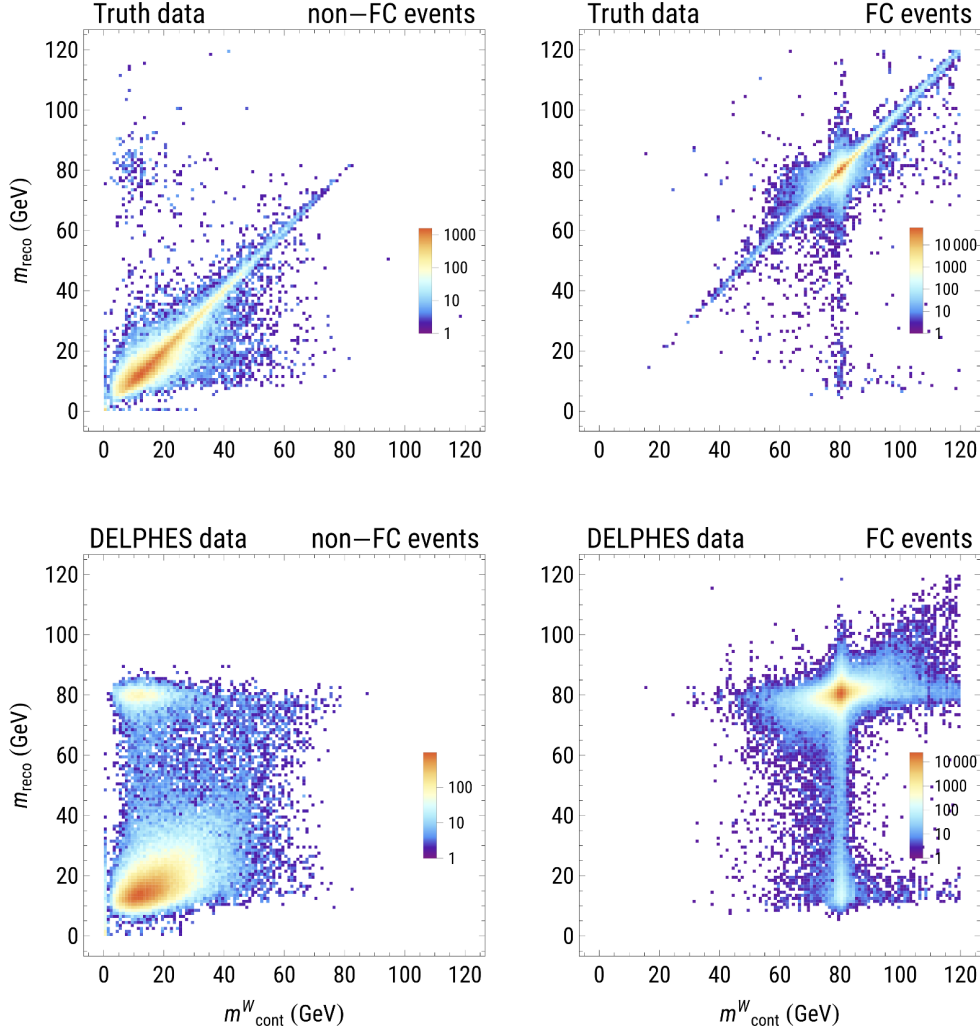


Figure 27. 2D histograms of target vs. reconstructed contained masses for models (top: truth data; bottom: DELPHES data) trained targeting p_{cont}^W , broken up into two populations based on containment (left: non-FC events; right: FC events).

In general this condition is difficult to solve, however it turns out to be extremely powerful in one special case. If we assume that f is *analytic* in the dot products d_{ij} (i.e. it can be expanded into a convergent multivariate Taylor series in some vicinity of the origin), then by the Identity theorem from complex analysis eq. (B.1) must hold in an entire neighborhood of the origin, even when $d_{ij} \neq 0$. However, if this formula holds for all values of k in an entire domain, then f is necessarily only a function of $p_i + p_j$. Moreover, if all inputs are assumed to be massless, then this applies to all pairs of indices i, j , and f becomes a function of only the jet mass m_J^2 . Thus we arrive at the following rather disappointing result.

Theorem B.1. *If an IRC-safe Lorentz-invariant observable with a mixture of massless and massive inputs is real analytic in the pairwise dot products d_{ij} near the origin, then it can depend on the massless inputs only through their sum. If all inputs are massless, this observable can be reduced to an analytic function of a single scalar — the jet mass $m_J^2 = \sum_{i,j} d_{ij}$.*

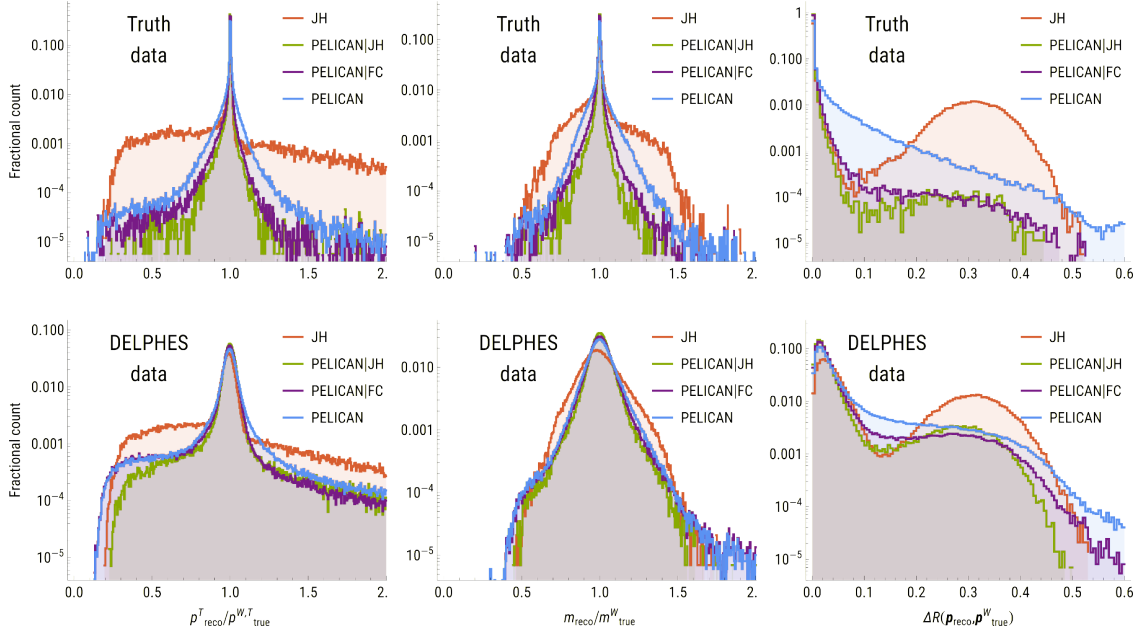


Figure 28. Full set of histograms corresponding to the entries in table 12 (models trained targeting p_{true}^W , truth-level and DELPHES versions).

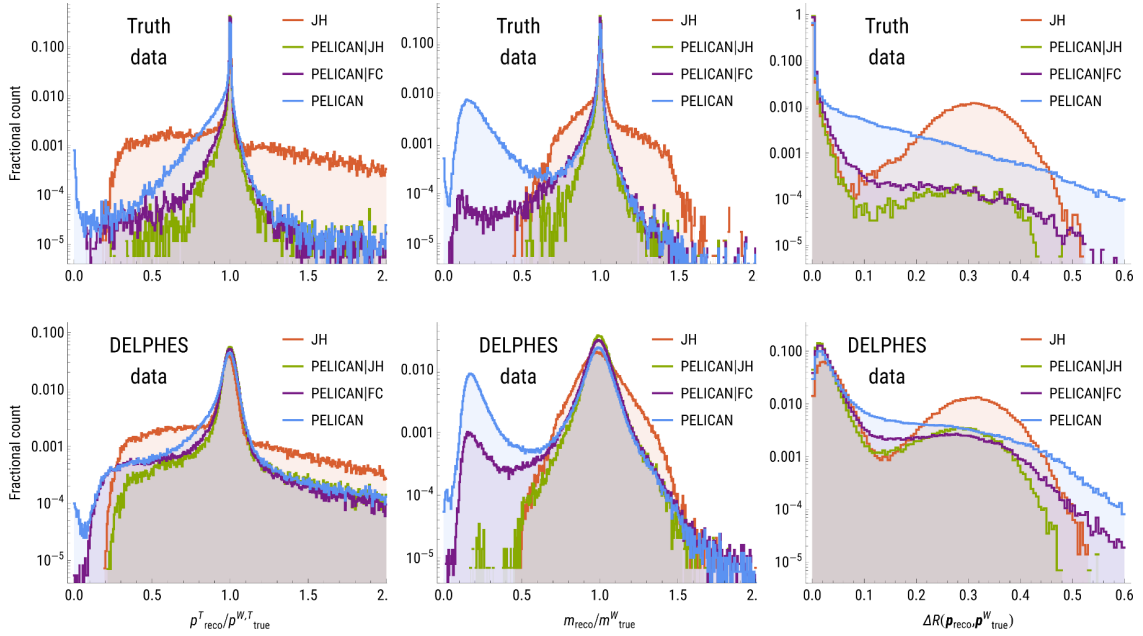


Figure 29. Full set of histograms corresponding to the entries in table 10 (models trained targeting p_{cont}^W , truth-level and DELPHES versions, and compared to p_{true}^W).

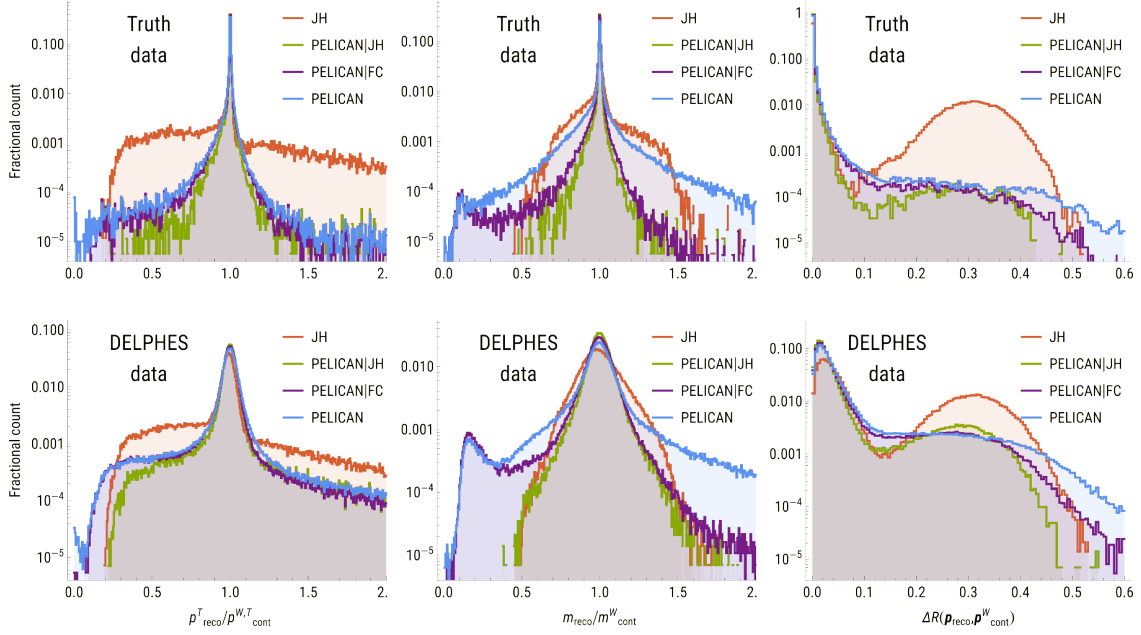


Figure 30. Full set of histograms for models trained on the variable mass dataset targeting p_{cont}^W and compared to p_{cont}^W .

In this way, IRC-safety is a significantly more powerful restriction than one could naïvely expect based on the definitions of IR and C-safety alone. As an example, consider the quantity $\prod_{i<j} d_{ij} = d_{12}d_{23}d_{31}$ on $N = 3$ massless inputs. It is in fact C-safe because it simply vanishes whenever any two inputs are massless and collinear. However it is not IR-safe because sending $p_3 \rightarrow 0$ reduces the expression to zero instead of the expected d_{12} . Alternatively, as we will see in the following section, the natural IR-safe extension of $d_{12}d_{23}d_{31}$ to arbitrary N will not be C-safe.

Clearly, by restricting ourselves to analytic functions, i.e. functions that can be approximated by polynomials in the variables d_{ij} , we have arrived at an extremely narrow set of IRC-safe observables which will be of little practical use. This is where our construction of IRC-safe PELICAN comes in handy, seeing as it clearly produces many more IRC-safe observables. In terms of the jet-frame coordinates defined in eq. (10.4) and eq. (10.5), the C-safety condition reads

$$\left. \frac{\partial f}{\partial \mathcal{E}_i} \right|_{\hat{d}_{ij}=0} = \left. \frac{\partial f}{\partial \mathcal{E}_j} \right|_{\hat{d}_{ij}=0}. \quad (\text{B.2})$$

The crucial observation here is that $\mathcal{E}_i$ and $\hat{d}_{ij}$ are *not* analytic functions of the original coordinates d_{ij} , as is evident from eq. (10.4). Therefore the set of analytic symmetric functions in these coordinates is very different, and in fact much larger, as we already know from our construction of IRC-safe PELICAN.

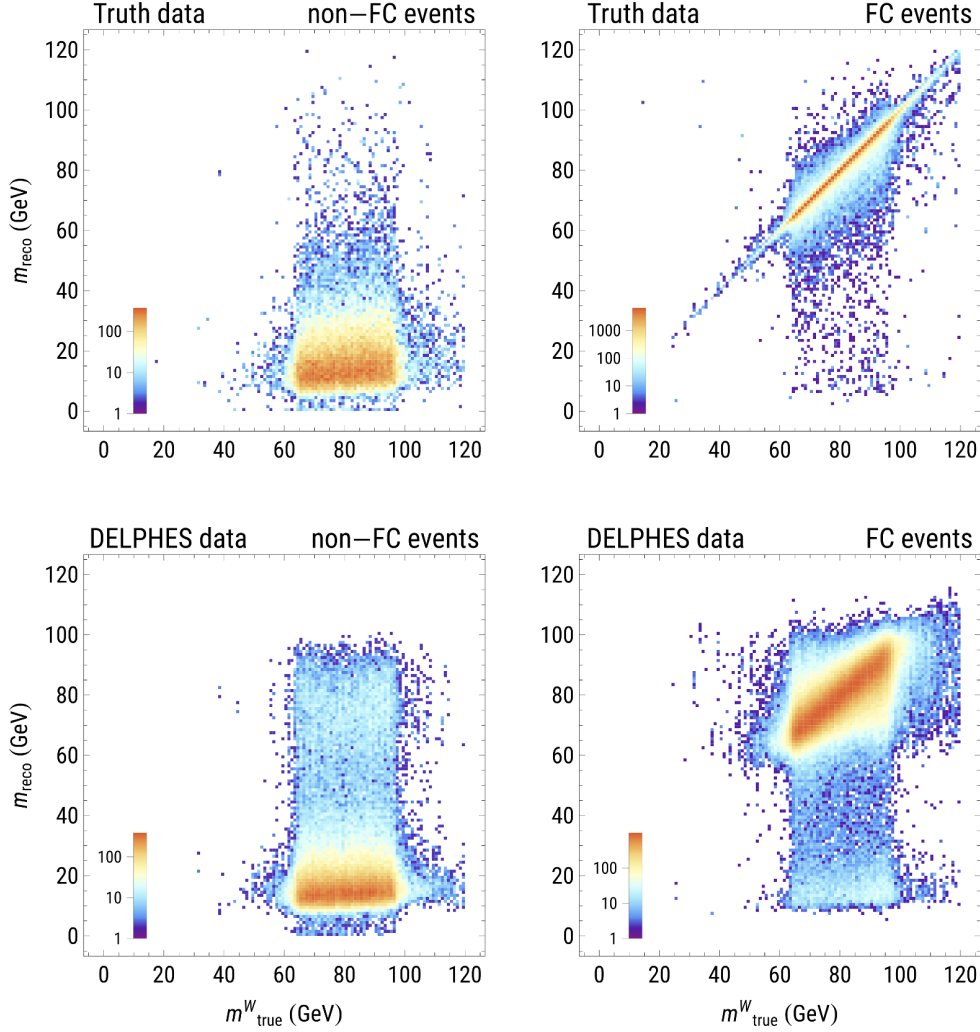


Figure 31. 2D histograms of target vs. reconstructed masses for models (top: truth data; bottom: DELPHES data) trained on the variable mass dataset targeting p_{cont}^W , broken up into two populations based on containment (left: non-FC events; right: FC events).

B.1 Lorentz meets EFPs

B.1.1 Review of EFPs

Particle data is often analyzed via IRC-safe observables such as N -subjettiness [16]. A general polynomial basis for all IRC-safe observables was obtained in ref. [18]. Here we first retrace some of the steps from the original derivation, and introduce a Lorentz-invariant analog at the end. Massless vectors can be written as $p = (E, E\hat{p})$ with a unit 3-vector $\hat{p}$. Then any (not necessarily Lorentz-invariant) smooth,⁶ symmetric observable of a set of

⁶It is important to note that smoothness is generally an excessively powerful restriction, and many useful IRC-safe observables cannot be expanded in Taylor series this way. For instance, staying within the realm of Lorentz-invariant observables, the quantity $\sum_{i,j} \mathcal{E}_i \mathcal{E}_j (\hat{d}_{ij})^\beta$ is IRC-safe for any β , but when written out in terms of d_{ij} it is clearly not differentiable at the origin. We thank Andrew Larkoski for this example, prompting much of our discussion of IRC-safety as presented in this work.

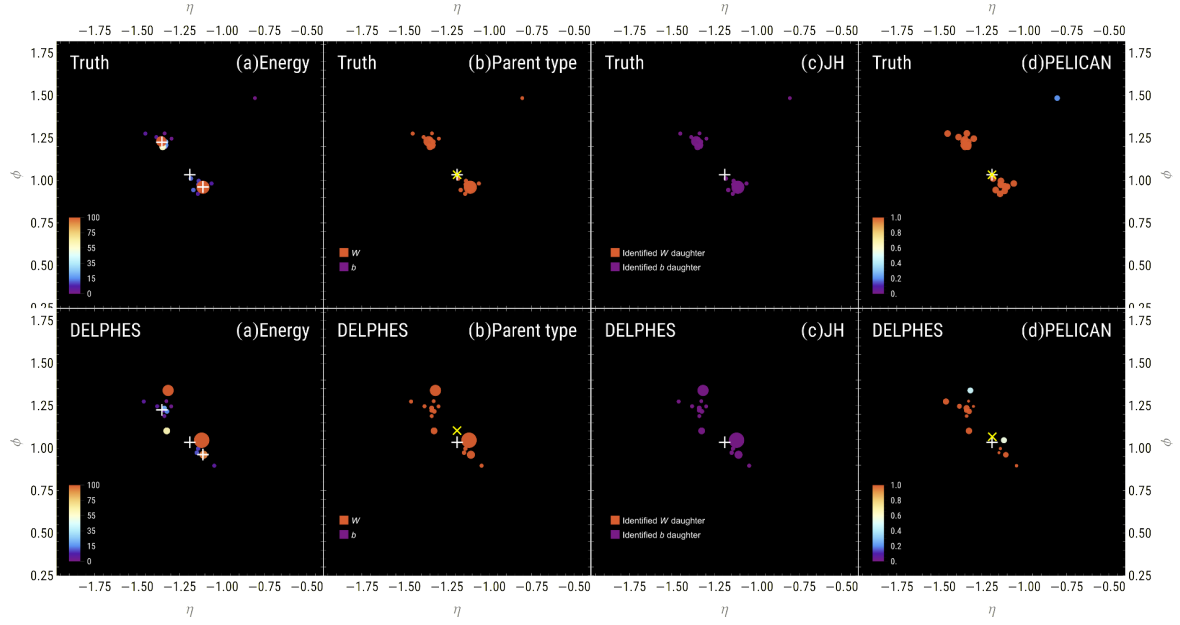


Figure 32. Event display combining the kind of information shown in figures 20–19. This is a qq event, i.e. one where the b quark from the $t \rightarrow bW$ decay fell outside of the jet radius. The JH tagger failed to tag this event as a top quark decay, whereas PELICAN correctly assigned weights close to 1 to most of the constituents and obtained an accurate reconstruction of the W -boson. The legend in pane (c) loses its direct meaning here, but it can be understood as JH effectively setting all clustering weights to zero.

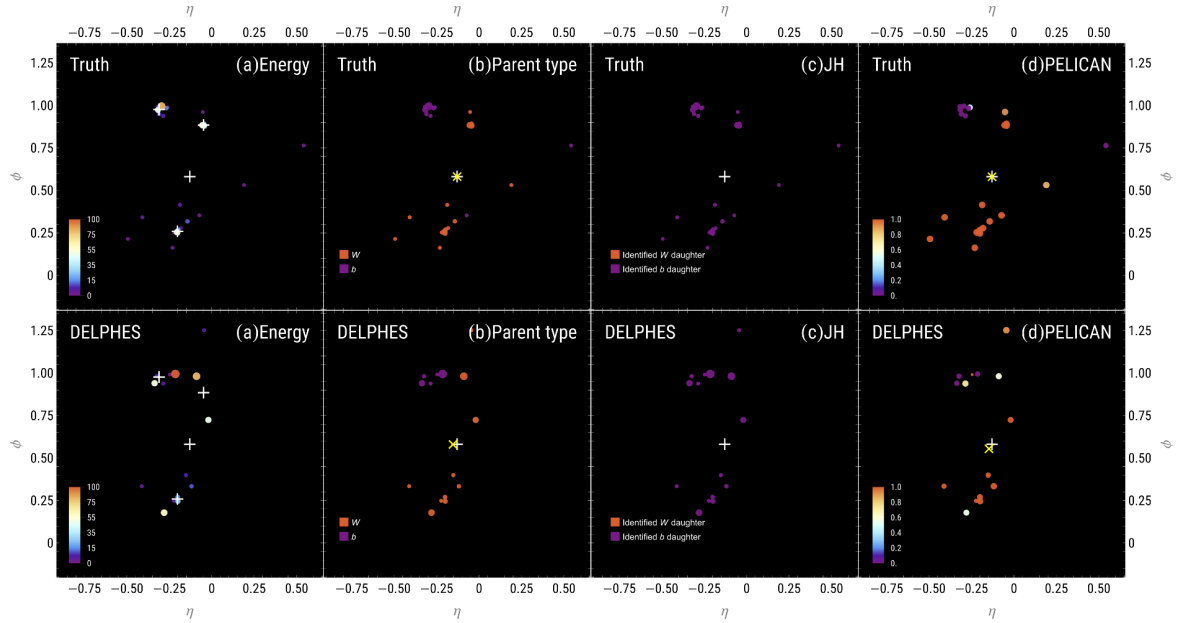


Figure 33. Yet another event display, this time a bq event that was still mistagged by the JH tagger. PELICAN correctly recognized the b quark subjet and obtained a fairly accurate reconstruction of the W -boson.

N massless vectors can be expanded at low energies as a combination of terms (omitting the constant term) of the form

$$\sum_{i_1, \dots, i_M=1}^N \sum_{j_1, \dots, j_L=1}^N E_{i_1} \dots E_{i_M} \cdot f_{i_1, i_2, \dots, i_M}^{(N)}(\hat{p}_1, \dots, \hat{p}_N) \quad (\text{B.3})$$

with some “angular functions” $f_{i_1, i_2, \dots, i_M}^{(N)}$. As shown in ref. [18], IR-safety for such a series amounts to requiring that the angular function $f_{i_1, i_2, \dots, i_M}^{(N)}$ in fact depends only on the particles whose indices are listed in the label of the function and is also independent of the total number N , i.e. it can be replaced by $f_{i_1, i_2, \dots, i_M}(\hat{p}_{i_1}, \dots, \hat{p}_{i_M})$. Furthermore, permutation symmetry of the entire observable descends to the total permutation symmetry of each angular function. Finally, C-safety requires that equating any two unit vectors $\hat{p}_i \rightarrow \hat{p}_j$ should lead to a function of only $E_1 + E_2$. At the level of angular functions, this implies that whenever i appears as an index in the subscript of the angular function, it can be replaced with j , and this can be done one index at a time. Ultimately this reduces the entire set of angular functions to just one $f_M = f_{1,2,\dots,M}$, and therefore the order M term in the low-energy expansion of any IRC-safe observable becomes

$$\sum_{i_1, \dots, i_M=1}^N E_{i_1} \dots E_{i_M} \cdot f_M(\hat{p}_{i_1}, \dots, \hat{p}_{i_M}). \quad (\text{B.4})$$

From here, EFP’s are derived by expanding the angular functions into a Taylor series around small angles $\cos \theta_{ij} = \hat{p}_i \cdot \hat{p}_j$. The resulting polynomial expansion can be broken up into a linear basis enumerated by the set of isomorphism classes of multigraphs G with no loops (edges connecting a vertex to itself):

$$\text{EFP}_G = \sum_{i_1, \dots, i_M=1}^N \prod_{j \in V(G)} E_{i_j} \cdot \prod_{(k,l) \in E(G)} \theta_{i_k i_l}, \quad (\text{B.5})$$

where $V(G) = \{1, 2, \dots, M\}$ is the set of vertices of G , and $E(G)$ is the set of edges. The EFP is a homogenous polynomial in the energies of degree $|V(G)| = M$ and in the angles of degree $D = |E(G)|$. Notice that the indices i_j can coincide but the corresponding terms vanish since $\theta_{ii} = 0$, which is why multigraphs with loops are excluded from the basis. As a trivial but important example, the total energy $E = \sum_{i=1}^N E_i$ corresponds to $G = \bullet$. If G consists of multiple connected components, the resulting EFP is the product of the EFP’s corresponding to the components, so the entire basis is algebraically generated by just the connected multigraphs.

It is instructive to see how IRC-safety works in EFP’s. IR-safety is guaranteed by the mere presence of the energy pre-factors: sending $E_N \rightarrow 0$ will simply recover the same EFP for the remaining $N - 1$ particles. Meanwhile, C-safety is observed because if two particles, say 1 and 2, become collinear, then the angular factor is completely invariant under their permutations. In other words, all terms where the list $(i_1, i_2, \dots, i_M)$ contains a fixed number, say L , of 1’s or 2’s, can be grouped together, and after summation over those indices the energy pre-factor manifestly depends on E_1 and E_2 only through an overall factor of $(E_1 + E_2)^L$. This is exactly the statement of C-safety.

B.1.2 MFPs: Lorentz-invariant analytic EFPs

Now let us task ourselves with identifying the subset of Lorentz-invariant IRC-safe observables. Obviously none of the EFP's are Lorentz-invariant due to the direct dependence on spatial angles, but smooth Lorentz-invariant observables can still be expanded in the EFP basis at small energies and angles. The only Lorentz-invariant of two massless 4-vectors p_i, p_j is $p_i \cdot p_j = E_i E_j (1 - \cos \theta_{ij})$. At small angles the approximately boost-invariant combination is then $E_i E_j \theta_{ij}^2$. The expansion of any permutation-symmetric function of such dot products into a series as above will differ from general EFP's as follows: every angle θ_{ij} must appear in the product an even number of times, and the pre-factor must consist of nothing but one factor of $E_i E_j$ for every factor of θ_{ij}^2 (or $1 - \cos \theta_{ij}$). The Lorentz-invariant version of the polynomial can then be obtained by replacing each θ_{ij}^2 with $2\hat{d}_{ij}$. In each EFP only a subset of terms will generally satisfy these conditions, so a better representation of the Lorentz-invariant basis is warranted.

Let us start again with the most general Lorentz-invariant and permutation-invariant observable of N 4-momenta that is *analytic* when expressed in terms of the dot products $d_{ij} = p_i \cdot p_j$. Such a function f can be expanded at small arguments into a linear combination $f \sim \sum_G w_G f_G^{(N)}$ of terms of the form

$$f_G^{(N)} = \sum'_{i_1, \dots, i_M=1}^N \prod_{(k,l) \in E(G)} d_{i_k i_l}, \quad (\text{B.6})$$

where the prime indicates that the terms where any two of the indices coincide are excluded. Including them would also produce a valid basis, but we leave them out to reduce the complexity of each polynomial. Such terms simply correspond to other multigraphs obtained from G by vertex identification, and those already appear in the expansion with independent coefficients. This is still an over-complete basis since the matrix of dot products d_{ij} is highly degenerate for $N \geq 5$. Unlike in EFP's, we also allow multigraphs with loops since we are not restricting ourselves to purely massless inputs. However, if all inputs are massless, all graphs with loops will produce vanishing polynomials.

Now we can easily see how to enforce IRC-safety in these polynomials. IR-safety requires that sending any of the 4-momenta to zero, say $p_N \rightarrow 0$, must be equivalent to simply restricting the same expression to only the other $N - 1$ particles. We observe that the only case when this is not already satisfied in the expression above is when the multigraph G contains isolated vertices (vertices of degree zero). Isolated vertices lead to summations over dummy indices corresponding to those vertices, which results in an overall factor of some power of N . The multiplicity N is not an IR-safe variable, and it is sufficient to ban graphs with isolated vertices to enforce IR-safety. We can thus drop the notation $f_G^{(N)}$ and simply write f_G with the implicit understanding that eq. (B.6) defines the full infinite family of observables for all N .

Now we address C-safety. Much like in EFP's, C-safety in f_G 's means that whenever, say, p_1 and p_2 are massless, each monomial that contains a certain number of indices equal to 1 or 2 must match its coefficient with any other monomial that differs only by flipping any number of the indices from 2 to 1 or vice versa. In the language of multigraphs, this means that each monomial that assigns a vertex of degree n_1 to p_1 and another vertex of

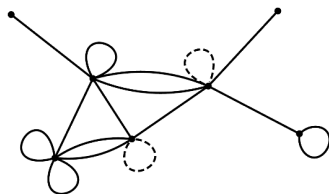


Figure 34. The addition of the dashed loops makes this multigraph *loop-saturated*: all non-leaf vertices come with at least one loop. Loop-saturated multigraphs enumerate all IRC-safe polynomials of d_{ij} .

degree n_2 to p_2 must in fact result as just one among all possible monomials obtained via vertex identification from a graph G' where those two vertices have been “blown up” into $n_1 + n_2$ vertices of degree one. We also notice that this condition applies only to vertices that do not have any loops attached to them because otherwise the corresponding terms would vanish as p_1 and p_2 become massless, thereby making any additional restrictions unnecessary. To summarize, any IRC-safe polynomial contains with equal coefficients all f_G ’s with G ’s obtained by any combination of vertex identifications from some multigraph G' such that all of its loopless vertices are *leaves* (i.e. of degree one). We shall call such multigraphs *loop-saturated*. In this process it is sufficient to allow identifications of only the leaf vertices because all other vertex identifications will result in a polynomial independently defined by another loop-saturated multigraph.

We thus define an IRC-safe Lorentz-invariant polynomial basis indexed by non-isomorphic loop-saturated multigraphs G with no isolated vertices. Assuming the vertices of G are identified with the set $\{1, 2, \dots, M\}$, we write

$$\text{MFP}_G = \sum_{i_1, \dots, i_M=1}^N \prod_{(k,l) \in E(G)} d_{i_k i_l} = \sum_{i_1, \dots, i_M=1}^N \prod_{j \in V(G)} \mathcal{E}_{i_j}^{\delta_j} \prod_{(k,l) \in E(G)} \hat{d}_{i_k i_l}, \quad (\text{B.7})$$

where the sum is taken over all M -tuples and $\delta_1, \dots, \delta_M$ are the degrees of the vertices of G (each loop increases the degree by 2). An alternative basis is

$$\text{MFP}'_G = \sum'_{i_1, \dots, i_M=1}^N \prod_{(k,l) \in E(G)} d_{i_k i_l}, \quad (\text{B.8})$$

where the coincidence of two or more indices $i_k = i_l$ is allowed only if their corresponding vertices (k and l) are leaves of G . MFP_G is simply MFP'_G plus a linear combination of certain MFP'_H ’s with fewer vertices, namely such that H is a graph obtained from G by one or more vertex identifications each of which involves at least one non-leaf vertex of G . Indeed, such identifications preserve the property of being loop-saturated, and all other identifications are already included in the definition of MFP'_G .

Note that just like for EFP’s, disconnected multigraphs correspond to products of their connected components (this holds for MFP_G but not for MFP'_G):

$$\text{MFP}_{G \sqcup H} = \text{MFP}_G \cdot \text{MFP}_H. \quad (\text{B.9})$$

As an example, if all inputs are known to be massless, then any connected multigraph with loops will produce a vanishing polynomial. And the only loopless loop-saturated connected multigraph is the graph consisting of just one edge, which evaluates to the jet mass: $\text{MFP}_{\bullet-\bullet} = m_J^2 = \sum_{i,j} d_{ij} = (\sum_i \mathcal{E}_i)^2$. Therefore in the massless case MFP's generate all analytic functions of m_J^2 , which is a disappointingly small subset of all IRC-safe Lorentz-invariant observables. As a final example, we can consider the polynomial $d_{12}d_{23}d_{31}$, which nominally seems C-safe. It can only be interpreted as an IR-safe polynomial if it corresponds to the triangle graph. However, such a graph is not loop-saturated, and therefore this polynomial is not IRC-safe despite being C-safe for $N = 3$ (which can be confirmed by observing that for $N = 4$ the triangle graph generates a non-C-safe polynomial), and the “correct” IRC-safe completion is simply m_J^6 .

What we found is congruent with eq. (B.1) in that MFP's represent a polynomial expansion basis for all IRC-safe observables that are analytic in the dot products d_{ij} , and depend on massless inputs only through their sum. Since we already know how restrictive this result is, we can now move on to a more useful generalization of EFPs.

B.1.3 JFPs: Lorentz-invariant jet-frame EFPs

As already discussed above, by transforming our inputs from d_{ij} to the jet-frame coordinates $\mathcal{E}_i$ and $\hat{d}_{ij}$ we get a much larger space of IRC-safe analytic observables. Ideologically, the transformation to the jet-frame coordinates is extremely simple: $\mathcal{E}_i$ is nothing but the regular energy E_i as measured in the rest frame of the jet, and the same is true for the angular parts which take the form $\hat{d}_{ij} = 1 - \cos \Theta_{ij}$ for a pair of massless particles. This coordinate transformation itself is IRC-safe in the sense that IR and collinear splittings that don't involve constituent p_i will not affect its jet-frame energy and angles with the other constituents. Therefore the notion of IRC-safe *continuous* observables is the same in both coordinate systems. On the other hand, as we have seen above, the restriction to analyticity makes the two sets of IRC-safe observables very different.

All of this implies that the set of IRC-safe observables which are invariant under spatial rotations and restricted to the manifold of events whose jet momentum J is purely time-like is in one-to-one correspondence with Lorentz-invariant IRC-safe observables. Indeed, given a rotationally invariant observable $f(p_1, \dots, p_N)$, consider

$$f_L(p_1, \dots, p_N) = f(\Lambda_J p_1, \dots, \Lambda_J p_N), \quad (\text{B.10})$$

where Λ_J is a Lorentz transformation that maps the jet momentum $J = \sum_i p_i$ to $(m_J, 0, 0, 0)$. Note that Λ_J is not unique, namely $R\Lambda_J$ is also a solution for any spatial rotation R . However, since we have assumed f to be rotationally invariant, the resulting observable f_L is independent of the specific choice of Λ_J and is *fully Lorentz-invariant*. And since Λ_J is invertible, the transformation $f \mapsto f_L$ is also invertible, which proves the one-to-one correspondence.⁷

We conclude that a polynomial expansion basis for analytic IRC-safe Lorentz-invariant observables in the jet-frame coordinates can be trivially obtained by taking the original EFPs, replacing θ_{ij}^2 with their rotationally invariant analogs $2(1 - \cos \theta_{ij})$, and then replacing

⁷This principle, of course, applies not only to observables but to entire algorithms, such as jet clustering. E.g. we defined our Lorentz-invariant analog of Soft Drop multiplicity using exactly this method.

$E_i \mapsto \mathcal{E}_i$ and $(1 - \cos \theta_{ij}) \mapsto \hat{d}_{ij}$. We do have to abandon EFPs that include odd powers of any angle due to their non-analyticity, but that doesn't affect the completeness of the basis since all the angular coordinates are non-negative (however it means that every edge in our multigraphs corresponds to two edges in the analogous EFP). The only subtlety here is that the original EFPs were built only for all-massless inputs, so they don't depend on masses whereas $\hat{d}_{ij}$ do. But all the original arguments in the derivation of the EFPs still apply even when some of the inputs are massive (as long as those masses are also perturbatively small), so we still get a complete basis of IRC-safe observables that are analytic in the jet-frame coordinates. We call these polynomials the *Jet Flow Polynomials* (JFPs):

$$\text{JFP}_G = \sum_{i_1, \dots, i_M=1}^N \prod_{j \in V(G)} \mathcal{E}_{i_j} \cdot \prod_{(k,l) \in E(G)} \hat{d}_{i_k i_l}, \quad (\text{B.11})$$

where the multigraphs G have no restrictions on them and can contain loops. Comparing to eq. (B.7), we see that the powers of the energies are no longer tethered to the structure of the multigraph, which is why C-safety doesn't place any constraints on the multigraph unlike in the case of MFPs. Namely, the symmetry that C-safety forces on the angular parts $\prod_{(k,l)} \hat{d}_{i_k i_l}$ (that one must be able to switch the value of an index i_k from one collinear particle to another and obtain another monomial in the same polynomial) is no longer coupled to the power counting of the energy factors, which makes any multigraph G permissible as long as we sum over all possible values of the indices i_k .

B.2 Universality of PELICAN

While the Deep Sets theorem [33] shows that permutation-invariant networks are universal for set learning problems, it is known that message passing networks are strictly non-universal for *graph* learning problems, i.e. in the presence of rank 2 data such as an adjacency matrix. Permutation-equivariant networks that take graph data as input (PELICAN's inputs can be interpreted as such) and use layers like $\text{Eq}_{2 \rightarrow 2}$ are known to have the expressivity of the 2-WL (Weisfeiler-Lehman) graph isomorphism test, matching the expressivity of message passing networks, see ref. [72]. By adding matrix multiplication to the list of equivariant aggregators (thus turning the linear equivariant layer into a quadratic or polynomial one), one can raise the expressivity to the “folklore” version of the WL test, i.e. 2-FWL for PELICAN, which is equivalent to 3-WL [43]. However, in our tests we have not been able to see any performance improvements from the addition of an aggregator of this kind.

We can see the limited expressivity of PELICAN in terms of the graphs G in eq. (B.6) that can be generated by PELICAN when expanded at small values of the inputs (assuming a smooth activation function). Indeed, the messaging layer combined with single-index aggregation $\sum_i \bullet_{ij}$ can attach new edges (d_{ij}), loops (d_{jj}), and multi-edges (d_{ij}^n) to a given vertex j . By applying more of these layers we can only keep joining such “starfish” graphs at their central vertex j , whereas even something as simple as the triangle graph can never be generated due to the fact that it requires three free indices at an intermediate step. The only graphs with two free indices generated by PELICAN are the multi-edges d_{ij}^n , but joining them together will never produce a more complex graph. At best, PELICAN can only generate multigraphs that can be broken up into subgraphs with at most two severed edges on each

of them (since each severed edge requires a free index on the corresponding vertices for us to be able to “glue” the vertex back together via aggregation).

Finally, it is curious to compare PELICAN-like architectures to architectures that work with irreducible representations of the continuous symmetry group, like LGN [25]. As was shown in ref. [73], the latter kind of network is universal, provided that the set of the finite-dimensional irreducible representations that are stored is not limited. In particular, the proof involves showing that a universal network based on the tensor product nonlinearity will be able to generate all symmetric polynomial tensors of the form $P_\alpha = \sum_{i_1, \dots, i_r} p_{i_1}^{\alpha_1} \otimes \dots \otimes p_{i_r}^{\alpha_r}$ where $\alpha \in \mathbb{Z}^r$. It is easy to see that the Lorentz-invariant part of such a tensor produces all possible multigraphs G with vertices of degrees $\alpha_1, \dots, \alpha_r$. Therefore, indeed, the ability of a network to generate all such tensors implies its universality on the space of Lorentz-invariant observables.

However, in practice the dimensionality of these tensor representations has to be quite limited. For example, LGN kept only representations up to spin 2, that is, tensors of the form $p_i \otimes p_j$, but not higher than that (of course, these can also be multiplied by invariant scalars without increasing the dimension). This effectively limits the number of free particle indices in all latent values of the network to 2 (right before aggregation reduces it back to 1 again), just like in PELICAN. Therefore there is ample reason to believe that the expressivity of LGN-like networks that go up to spin k is not universal, but is equivalent to the expressivity of PELICAN-like networks that go up to the permutation-equivariant rank of k (using $\text{Eq}_{k \rightarrow k}$ blocks). Moreover, this expressivity can be stated in terms of the class of multigraphs G that any analytic network of this kind can generate. We leave this problem for future work.

B.3 Quantifying IRC-safety

This section is an attempt to define some Lorentz-invariant quantities that can measure *whether* and *how* IR/C-safe any given observable is. It is mostly of theoretical interest since it does not yet have a software implementation.

Let $f^{(N)}$ be a Lorentz-invariant permutation-symmetric observable defined for varying number N of inputs, for instance the output of a PELICAN instance. We define the following difference operator that measures the obstruction to f being IR-safe:

$$\Delta^{\text{IR}} f^{(N)} = f^{(N+1)}(p_1, \dots, p_N, 0) - f^{(N)}(p_1, \dots, p_N). \quad (\text{B.12})$$

IR-safety is equivalent to the vanishing of this operator:

$$\text{IR} - \text{safety} : \quad \Delta^{\text{IR}} f = 0. \quad (\text{B.13})$$

Furthermore, we can define *IR-robustness* as the partial derivative corresponding to the injection of an infinitesimally soft but non-zero 4-momentum:

$$\text{IR} - \text{robustness} : \quad R_p^{\text{IR}} f^{(N)} = \left. \frac{d}{d\epsilon} \right|_{\epsilon=0} f^{(N+1)}(p_1, \dots, p_N, \epsilon p). \quad (\text{B.14})$$

All together, we have the following series expansion for a general observable near $p_{N+1} = 0$:

$$f^{(N+1)} = f^{(N)} + \Delta^{\text{IR}} f^{(N)} + \left(R_{p_{N+1}}^{\text{IR}} f^{(N)} \right) \cdot p_{N+1} + \mathcal{O}(p_{N+1}^2). \quad (\text{B.15})$$

An IR-safe observable must always have $\Delta^{\text{IR}} f = 0$, and lower values of IR-robustness indicate lower sensitivity to soft particles. In practice, we can evaluate the ratio $(\Delta^{\text{IR}} f^{(N)})/f^{(N)}$ and average it over a testing dataset to quantify the IR-safety of a model.

For C-safety, we pick a massless 4-vector p and define the differential operator

$$D_p^{\text{C}} f = \left. \frac{d}{d\lambda} \right|_{\lambda=0} f(p_1 + \lambda p, p_2 - \lambda p, p_3, \dots). \quad (\text{B.16})$$

C-safety then is equivalent to the requirement that $D_p^{\text{C}} f = 0$ for any massless p whenever p_1 and p_2 are collinear with p :

$$\text{C-safety : } D_p^{\text{C}} f \Big|_{p_1 \parallel p_2 \parallel p} = 0, \quad p^2 = 0. \quad (\text{B.17})$$

C-robustness can be easily defined as the second-order analog of D_p^{C} . Namely, we pick a second vector $\hat{p}$ such that $p \cdot \hat{p} = 0$ (which means that $\hat{p}$ is tangent to the light cone at p , as required by the massless constraint) and measure how quickly $D_p^{\text{C}} f$ deviates from zero as we deform p_1 and p_2 away from collinearity:

$$\text{C-robustness : } R_{\hat{p},p}^{\text{C}} f = D_{\hat{p}}^{\text{C}} D_p^{\text{C}} f, \quad \hat{p} \cdot p = p^2 = 0, \quad p_1 \parallel p_2 \parallel p. \quad (\text{B.18})$$

It is instructive to express these quantities in terms of the dot products d_{ij} . We have

$$D_p^{\text{C}} f = \sum_{j=1}^N (\partial_{1j} f - \partial_{2j} f) (p_j \cdot p), \quad (\text{B.19})$$

where ∂_{kl} indicates the partial derivative with respect to d_{kl} . Note that individual partial derivatives aren't really well-defined due to the fact that $\{d_{ij}\}$ is not a set of independent coordinates on the manifold of N 4-vectors for $N \geq 5$. However, combinations such as above are valid since they are nothing but a different way of expressing the original well-defined operator. We also immediately notice that if $N = 2$ and $p_1 \parallel p_2 \parallel p$, then $D_p^{\text{C}} f = 0$ automatically. Indeed, every Lorentz-invariant observable for two particles is C-safe due to the fact that $d_{12} = \frac{1}{2}(p_1 + p_2)^2$ for massless inputs.

C-robustness is similarly expressed by

$$R_{\hat{p},p}^{\text{C}} f = \sum_{i,j=1}^N \left(\partial_{1i,1j}^2 + \partial_{2i,2j}^2 - \partial_{1i,2j}^2 - \partial_{1j,2i}^2 \right) f \cdot (p_i \cdot \hat{p})(p_j \cdot p). \quad (\text{B.20})$$

Open Access. This article is distributed under the terms of the Creative Commons Attribution License ([CC-BY4.0](https://creativecommons.org/licenses/by/4.0/)), which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited.

References

- [1] J. Gallicchio and M.D. Schwartz, *Quark and Gluon Jet Substructure*, [*JHEP* **04** \(2013\) 090](#) [[arXiv:1211.7038](#)] [[INSPIRE](#)].
- [2] A.J. Larkoski, J. Thaler and W.J. Waalewijn, *Gaining (Mutual) Information about Quark/Gluon Discrimination*, [*JHEP* **11** \(2014\) 129](#) [[arXiv:1408.3122](#)] [[INSPIRE](#)].

- [3] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, *JHEP* **01** (2017) 110 [[arXiv:1612.01551](#)] [[INSPIRE](#)].
- [4] E. Alvarez, M. Spannowsky and M. Szewc, *Unsupervised Quark/Gluon Jet Tagging With Poissonian Mixture Models*, *Front. Artif. Intell.* **5** (2022) 852970 [[arXiv:2112.11352](#)] [[INSPIRE](#)].
- [5] G. Kasieczka, N. Kiefer, T. Plehn and J.M. Thompson, *Quark-Gluon Tagging: Machine Learning vs Detector*, *SciPost Phys.* **6** (2019) 069 [[arXiv:1812.09223](#)] [[INSPIRE](#)].
- [6] J.M. Butterworth, A.R. Davison, M. Rubin and G.P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys. Rev. Lett.* **100** (2008) 242001 [[arXiv:0802.2470](#)] [[INSPIRE](#)].
- [7] D.E. Kaplan, K. Rehermann, M.D. Schwartz and B. Tweedie, *Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks*, *Phys. Rev. Lett.* **101** (2008) 142001 [[arXiv:0806.0848](#)] [[INSPIRE](#)].
- [8] A. Butter et al., *The Machine Learning landscape of top taggers*, *SciPost Phys.* **7** (2019) 014 [[arXiv:1902.09914](#)] [[INSPIRE](#)].
- [9] J. Thaler and L.-T. Wang, *Strategies to Identify Boosted Tops*, *JHEP* **07** (2008) 092 [[arXiv:0806.0023](#)] [[INSPIRE](#)].
- [10] D.E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, *Phys. Rev. D* **87** (2013) 054012 [[arXiv:1211.3140](#)] [[INSPIRE](#)].
- [11] I. Feige, M.D. Schwartz, I.W. Stewart and J. Thaler, *Precision Jet Substructure from Boosted Event Shapes*, *Phys. Rev. Lett.* **109** (2012) 092001 [[arXiv:1204.3898](#)] [[INSPIRE](#)].
- [12] Y.-T. Chien and I. Vitev, *Jet Shape Resummation Using Soft-Collinear Effective Theory*, *JHEP* **12** (2014) 061 [[arXiv:1405.4293](#)] [[INSPIRE](#)].
- [13] S. Marzani, L. Schunk and G. Soyez, *The jet mass distribution after Soft Drop*, *Eur. Phys. J. C* **78** (2018) 96 [[arXiv:1712.05105](#)] [[INSPIRE](#)].
- [14] F.A. Dreyer, G.P. Salam and G. Soyez, *The Lund Jet Plane*, *JHEP* **12** (2018) 064 [[arXiv:1807.04758](#)] [[INSPIRE](#)].
- [15] R. Kogler, *Advances in Jet Substructure at the LHC: Algorithms, Measurements and Searches for New Physical Phenomena*, *Springer Tracts Mod. Phys. (STMP)* **284** (2021). [[INSPIRE](#)].
- [16] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N -subjettiness*, *JHEP* **03** (2011) 015 [[arXiv:1011.2268](#)] [[INSPIRE](#)].
- [17] CMS collaboration, *First Measurement of Hadronic Event Shapes in pp Collisions at $\sqrt{s} = 7$ TeV*, *Phys. Lett. B* **699** (2011) 48 [[arXiv:1102.0068](#)] [[INSPIRE](#)].
- [18] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, *JHEP* **04** (2018) 013 [[arXiv:1712.07124](#)] [[INSPIRE](#)].
- [19] R. Kogler et al., *Jet Substructure at the Large Hadron Collider: Experimental Review*, *Rev. Mod. Phys.* **91** (2019) 045003 [[arXiv:1803.06991](#)] [[INSPIRE](#)].
- [20] S. Marzani, G. Soyez and M. Spannowsky, *Looking inside jets: an introduction to jet substructure and boosted-object phenomenology*, Springer (2019) [[DOI:10.1007/978-3-030-15709-8](#)] [[INSPIRE](#)].
- [21] M. Feickert and B. Nachman, *A Living Review of Machine Learning for Particle Physics*, [arXiv:2102.02770](#) [[INSPIRE](#)].
- [22] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned Top Tagging with a Lorentz Layer*, *SciPost Phys.* **5** (2018) 028 [[arXiv:1707.08966](#)] [[INSPIRE](#)].

- [23] M. Erdmann, E. Geiser, Y. Rath and M. Rieger, *Lorentz Boost Networks: Autonomous Physics-Inspired Feature Engineering*, *2019 JINST* **14** P06006 [[arXiv:1812.09722](#)] [[INSPIRE](#)].
- [24] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy Flow Networks: Deep Sets for Particle Jets*, *JHEP* **01** (2019) 121 [[arXiv:1810.05165](#)] [[INSPIRE](#)].
- [25] A. Bogatskiy et al., *Lorentz Group Equivariant Neural Network for Particle Physics*, [arXiv:2006.04780](#) [[INSPIRE](#)].
- [26] S. Gong et al., *An efficient Lorentz equivariant graph neural network for jet tagging*, *JHEP* **07** (2022) 030 [[arXiv:2201.08187](#)] [[INSPIRE](#)].
- [27] A. Bogatskiy et al., *Symmetry Group Equivariant Architectures for Physics*, in the proceedings of the *Snowmass 2021*, Seattle, U.S.A., July 17–26 (2022) [[arXiv:2203.06153](#)] [[INSPIRE](#)].
- [28] J.T. Offermann, T. Hoffman and A. Bogatskiy, *Top Jet W-Momentum Reconstruction Dataset*, *Zenodo* **1.0.0** (2023).
- [29] E. Witkowski and D. Whiteson, *Learning Broken Symmetries with Resimulation and Encouraged Invariance*, [arXiv:2311.05952](#) [[INSPIRE](#)].
- [30] S. Qiu et al., *Holistic approach to predicting top quark kinematic properties with the covariant particle transformer*, *Phys. Rev. D* **107** (2023) 114029 [[arXiv:2203.05687](#)] [[INSPIRE](#)].
- [31] H. Weyl, *The Classical Groups. Their Invariants and Representations*, 2 ed., Princeton University Press, Princeton, N.J., U.S.A. (1946).
- [32] B. Gripaios, W. Haddadin and C.G. Lester, *Lorentz and permutation invariants of particles*, *J. Phys. A* **54** (2021) 155201 [[arXiv:2003.05487](#)] [[INSPIRE](#)].
- [33] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R.R. Salakhutdinov and A.J. Smola, *Deep Sets*, in proceedings of *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, U.S.A. (2017) pp. 3391–3401 <http://papers.nips.cc/paper/6931-deep-sets.pdf>.
- [34] ATLAS collaboration, *Deep Sets based Neural Networks for Impact Parameter Flavour Tagging in ATLAS*, ATL-PHYS-PUB-2020-014 (2020) [[INSPIRE](#)].
- [35] E. Wagstaff et al., *Universal Approximation of Functions on Sets*, [arXiv:2107.01959](#).
- [36] H. Qu, C. Li and S. Qian, *Particle Transformer for Jet Tagging*, [arXiv:2202.03772](#) [[INSPIRE](#)].
- [37] H. Maron, H. Ben-Hamu, N. Shamir and Y. Lipman, *Invariant and Equivariant Graph Networks*, [arXiv:1812.09902](#).
- [38] H. Pan and R. Kondor, *Permutation equivariant layers for higher order interactions*, in proceedings of *The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, *Proc. Machine Learning Research* **151** (2022) pp. 5987–6001, <https://proceedings.mlr.press/v151/pan22a.html>.
- [39] G. Corso et al., *Principal Neighbourhood Aggregation for Graph Nets*, [arXiv:2004.05718](#).
- [40] N. Keriven and G. Peyré, *Universal Invariant and Equivariant Graph Neural Networks*, *Adv. Neural Inf. Process.* **32** (2019) 7090 [[arXiv:1905.04943](#)].
- [41] A. Sannai, Y. Takai and M. Cordonnier, *Universal approximations of permutation invariant/equivariant functions by deep neural networks*, [arXiv:1903.01939](#).
- [42] H. Maron, E. Fetaya, N. Segol and Y. Lipman, *On the Universality of Invariant Networks*, in proceedings of the *36th International Conference on Machine Learning*, *Proc. Machine Learning Research* **97** (2019) 4363, <http://proceedings.mlr.press/v97/maron19a.html> [[arXiv:1901.09342](#)].

- [43] H. Maron, H. Ben-Hamu, H. Serviansky and Y. Lipman, *Provably Powerful Graph Networks*, in H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox and R. Garnett eds., *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, <https://proceedings.neurips.cc/paper/2019/hash/bb04af0f7ecaee4aae62035497da1387-Abstract.html> [[arXiv:1905.11136](#)].
- [44] T. Sun, A. Hands and R. Kondor, *P-tensors: a General Formalism for Constructing Higher Order Message Passing Networks*, [arXiv:2306.10767](#).
- [45] S. Villar et al., *Scalars are universal: Equivariant machine learning, structured like classical physics*, *Adv. Neural Inf. Process. Syst.* **34** (2021) 28848 [[arXiv:2106.06610](#)] [[INSPIRE](#)].
- [46] G. Kasieczka, T. Plehn, J. Thompson and M. Russel, *Top Quark Tagging Reference Dataset, Zenodo Dataset v0* (2019).
- [47] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, *SciPost Phys. Codeb.* **2022** (2022) 8 [[arXiv:2203.11601](#)] [[INSPIRE](#)].
- [48] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [[arXiv:1307.6346](#)] [[INSPIRE](#)].
- [49] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, [arXiv:1711.05101](#) [[INSPIRE](#)].
- [50] J. Pearkes, W. Fedorko, A. Lister and C. Gay, *Jet Constituents for Deep Neural Network Based Top Quark Tagging*, [arXiv:1704.02124](#) [[INSPIRE](#)].
- [51] J.M. Munoz, I. Batatia and C. Ortner, *Boost invariant polynomials for efficient jet tagging*, *Mach. Learn. Sci. Tech.* **3** (2022) 04LT05 [[arXiv:2207.08272](#)] [[INSPIRE](#)].
- [52] R. Das, G. Kasieczka and D. Shih, *Feature Selection with Distance Correlation*, [arXiv:2212.00046](#) [[INSPIRE](#)].
- [53] H. Qu and L. Gouskos, *ParticleNet: Jet Tagging via Particle Clouds*, *Phys. Rev. D* **101** (2020) 056019 [[arXiv:1902.08570](#)] [[INSPIRE](#)].
- [54] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, [arXiv:1603.02754](#) [[DOI:10.1145/2939672.2939785](#)] [[INSPIRE](#)].
- [55] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J. C* **72** (2012) 1896 [[arXiv:1111.6097](#)] [[INSPIRE](#)].
- [56] M. Cacciari, G.P. Salam and G. Soyez, *The anti- k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [[arXiv:0802.1189](#)] [[INSPIRE](#)].
- [57] P. Komiske, E. Metodiev and J. Thaler, *Pythia8 quark and gluon jets for energy flow*, *Zenodo dataset v1* (2019).
- [58] V. Mikuni and F. Canelli, *ABCNet: An attention-based method for particle tagging*, *Eur. Phys. J. Plus* **135** (2020) 463 [[arXiv:2001.05311](#)] [[INSPIRE](#)].
- [59] P. Konar, V.S. Ngairangbam and M. Spannowsky, *Energy-weighted message passing: an infra-red and collinear safe graph neural network algorithm*, *JHEP* **02** (2022) 060 [[arXiv:2109.14636](#)] [[INSPIRE](#)].
- [60] E.A. Moreno et al., *JEDI-net: a jet identification algorithm based on interaction networks*, *Eur. Phys. J. C* **80** (2020) 58 [[arXiv:1908.05318](#)] [[INSPIRE](#)].
- [61] V. Mikuni and F. Canelli, *Point cloud transformers applied to collider physics*, *Mach. Learn. Sci. Tech.* **2** (2021) 035027 [[arXiv:2102.05073](#)] [[INSPIRE](#)].

- [62] J. Duarte et al., *Fast inference of deep neural networks in FPGAs for particle physics*, [2018 JINST 13 P07027](#) [[arXiv:1804.06913](#)] [[INSPIRE](#)].
- [63] M. Pierini, J.M. Duarte, N. Tran and M. Freytsis, *HLS4ML LHC Jet Dataset (100 particles)*, [Zenodo Dataset v1](#) (2020).
- [64] M. Pierini, J.M. Duarte, N. Tran, M. Freytsis and A. Bogatskiy, *Converted HLS4ML LHC Jet Dataset (100 particles)*, [OSF Dataset v1](#) (2023).
- [65] J.T. Offermann, X. Liu and T. Hoffman, *HEPData4ML*, *GitHub* (2023), <https://github.com/janTOffermann/HEPData4ML>.
- [66] S. Qiu et al., *Parton labeling without matching: unveiling emergent labelling capabilities in regression models*, *Eur. Phys. J. C* **83** (2023) 622 [[arXiv:2304.09208](#)] [[INSPIRE](#)].
- [67] PARTICLE DATA GROUP collaboration, *Review of Particle Physics*, *PTEP* **2022** (2022) 083C01 [[INSPIRE](#)].
- [68] C. Frye, A.J. Larkoski, J. Thaler and K. Zhou, *Casimir Meets Poisson: Improved Quark/Gluon Discrimination with Counting Observables*, *JHEP* **09** (2017) 083 [[arXiv:1704.06266](#)] [[INSPIRE](#)].
- [69] Y.L. Dokshitzer, G.D. Leder, S. Moretti and B.R. Webber, *Better jet clustering algorithms*, *JHEP* **08** (1997) 001 [[hep-ph/9707323](#)] [[INSPIRE](#)].
- [70] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross-sections in deep inelastic scattering*, in the proceedings of the *Workshop on Monte Carlo Generators for HERA Physics (Plenary Starting Meeting)*, Hamburg, Germany, April 27–30 (1998) [[hep-ph/9907280](#)] [[INSPIRE](#)].
- [71] D. Athanasakos et al., *Is infrared-collinear safe information all you need for jet classification?*, [arXiv:2305.08979](#) [[INSPIRE](#)].
- [72] W. Azizian and M. Lelarge, *Expressive Power of Invariant and Equivariant Graph Neural Networks*, [arXiv:2006.15646](#).
- [73] N. Dym and H. Maron, *On the Universality of Rotation Equivariant Point Cloud Networks*, [arXiv:2010.02449](#).