MACHINE LEARNING
Science and Technology

PAPER • OPEN ACCESS

Optimal equivariant architectures from the symmetries of matrix-element likelihoods

To cite this article: Daniel Maître *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 015059

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
15 November 2024

REVISED
1 February 2025

ACCEPTED FOR PUBLICATION
26 February 2025

PUBLISHED
7 March 2025

PAPER

Optimal equivariant architectures from the symmetries of matrix-element likelihoods

Daniel Maître, Vishal S Ngairangbam[*] and Michael Spannowsky

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

[*] Author to whom any correspondence should be addressed.

E-mail: vishal.s.ngairangbam@durham.ac.uk, daniel.maitre@durham.ac.uk and michael.spannowsky@durham.ac.uk

Keywords: matrix-element method, optimal equivariance, symmetries

Abstract

The Matrix-Element Method (MEM) has long been a cornerstone of data analysis in high-energy physics. It leverages theoretical knowledge of parton-level processes and symmetries to evaluate the likelihood of observed events. In parallel, the advent of geometric deep learning has enabled neural network architectures that incorporate known symmetries directly into their design, leading to more efficient learning. This paper presents a novel approach that combines MEM-inspired symmetry considerations with equivariant neural network design for particle physics analysis. Even though Lorentz invariance and permutation invariance over all reconstructed objects are the largest and most natural symmetry in the input domain, we find that they are sub-optimal in most practical search scenarios. We propose a longitudinal boost-equivariant message-passing neural network architecture that preserves relevant discrete symmetries. We present numerical studies demonstrating MEM-inspired architectures achieve new state-of-the-art performance in distinguishing di-Higgs decays to four bottom quarks from the QCD background, with enhanced sample and parameter efficiencies. This synergy between MEM and equivariant deep learning opens new directions for physics-informed architecture design, promising more powerful tools for probing physics beyond the Standard Model.

1. Introduction

The search for new physics at the Large Hadron Collider (LHC) is a complex and data-intensive challenge. As particle collisions produce high-dimensional data, distinguishing between Standard Model events and potential new physics requires sophisticated analysis techniques. Traditionally, matrix-element methods (MEM) [1–17] have been used to compare observed data to theoretical predictions by evaluating the likelihood of various hypothesized processes. In parallel, the advent of deep learning has enabled the development of powerful algorithms capable of learning complex patterns in data [18–46], often outperforming conventional methods in classification tasks.

In recent years, geometric deep learning [47–54] has emerged as a promising framework for physics analysis, incorporating known symmetries of physical laws directly into the neural network architecture. This approach, which could be called equivariant neural network design, seeks to restrict the learning task to a smaller yet appropriate class of functions by embedding symmetries such as Lorentz and permutation invariances into the model structure [55–62]. The general intuition that guides such architecture design is the invariance of physical observables under group transformations.

Despite the natural synergy between MEM, which explicitly utilises theoretical knowledge of symmetries through matrix element calculations and equivariant neural networks, a systematic connection between these two approaches has not been fully established. This work aims to bridge this gap by demonstrating how MEM-inspired symmetries can guide the design of equivariant neural network architectures for event classification tasks at the LHC. When deciding on which symmetries to embed in the model, we will show that the considerations that should guide the choice are the symmetry of the *target function* rather than the physical symmetries of the network input. For example, in the case we consider here, the symmetry to use is

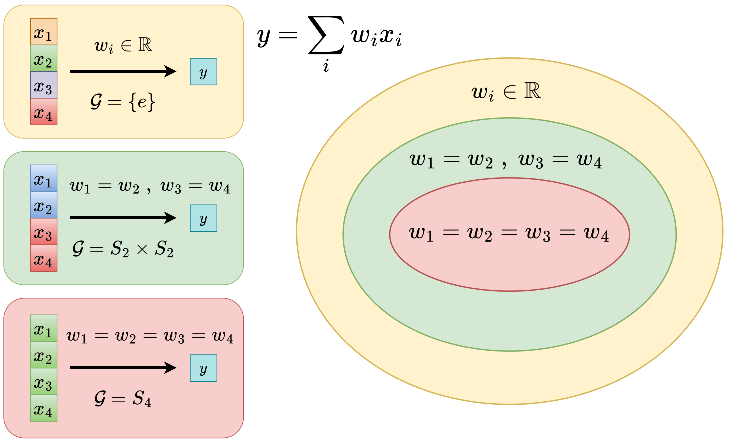

Figure 1. Representing the hierarchy in group invariant function approximation where a larger group (S_4) imposes additional constraints on the weights compared to a proper subgroup ($S_2 \times S_2$). Although the constraints of $S_2 \times S_2$ can become those of S_4, the stronger constraints of S_4 cannot become a function that is $S_2 \times S_2$ invariant but not S_4 invariant, as its weights lie strictly outside the red ellipse with the constraint $w_1 = w_2 \neq w_3 = w_4$. Therefore, even though S_4 contains the group $S_2 \times S_2$, an S_4-invariant function cannot become a purely $S_2 \times S_2$-invariant function. This holds for general group invariant functions due to the structure of fibres induced by invariance in the function's domain (see figure 3).

that of the likelihood ratio and not the full Lorentz invariance of the input momenta. We highlight the benefits of embedding optimal symmetries derived from the matrix-element calculations into the neural network to enhance classification accuracy, sample efficiency, and generalisation capabilities.

We begin by discussing the role of symmetries in function approximation, where they manifest themselves as group orbits in the equivalence classes of a target function's fibre in section 2. We then explore optimal symmetry group choices for classification problems using the Neyman–Pearson lemma and their connection to the fibres of group-equivariant functions in section 3. While the arguments based on group actions are more general, a simplified example which helps explain the hierarchy of group invariant function approximation is shown in figure 1. The general intuition can be stated as follows:

> In signal vs. background classification tasks, one can infer the (approximate) symmetries of the target function (not the data) from the specific processes' underlying likelihood based on the differential cross-sections. A universally approximating equivariant architecture on the space of functions with smaller or the same symmetries can approximate the target function but not one with a strictly larger symmetry.

Incorporating Lorentz symmetry and permutation invariance, essential in evaluating cross-sections, provides a foundation for developing equivariant architectures. Building on this theoretical groundwork, we investigate the optimality of the Lorentz invariance and S_n permutation invariance over all n reconstructed objects for the evaluated MEM-likelihoods in section 4. The former is suboptimal due to the dependence of the event likelihood on the transfer function, which is invariant only under longitudinal boosts and rotations along the z-axis. The S_n group is optimal when the final state consists of a single type of reconstructed object[1]. Therefore, we devise a longitudinal boost invariant homogeneous[2] message passing neural network, where the smaller permutation symmetries are maintained by concatenated sub-graph readouts.

To illustrate the practical implications of our approach, in section 5, we present a case study of di-Higgs production with four bottom jets in the final state, a channel of particular interest for probing the Higgs self-coupling at the LHC. We demonstrate that MEM-inspired symmetries improve network performance in classification metrics compared to state-of-the-art results [41] and maintain better performance metrics with up to three orders of magnitude fewer parameters.

Our findings suggest that by integrating the principles of the MEM with modern equivariant deep learning techniques, we can develop more efficient and physically informed architectures for LHC data

[1] Any event-level analyses on reconstructed objects with point cloud architectures which assume S_n invariance is, therefore, suboptimal in the sense of Neyman–Pearson when there is more than one type of reconstructed object. While their good performance may be due to the negligible null orbits of finite group symmetries in an uncountable domain, even in this suboptimal situation, they mostly outperform shallow machine learning on high-level features, which is a testament to the expressibility of modern deep learning algorithms.

[2] The requirement of permutation symmetry under separate classes of reconstructed objects allows for a heterogeneous graph construction. We do not consider such an approach as it has a factorial growth of learnable functions based on the number of edge and node types.

analysis. This synergy paves the way for new methodologies in the ongoing search for physics beyond the Standard Model.

2. Symmetries as strong inductive biases

The theoretical reasoning behind symmetries becomes evident from its relation to conserved quantities, i.e. a symmetry transformation on a physical system does not change observable quantities. This carries over to function approximation, as the value of physically meaningful functions should not change under a symmetry transformation in the input feature space. Even without such symmetry considerations, defining a function requires each element on its domain to be associated with only one element in its co-domain (not one-to-many). Therefore, any given function divides the domain into mutually exclusive subsets mapped to the same element on its co-domain. These subsets called the function's fibres, are a particular *partition* of the input domain unique to a family of functions.

A partition of a set is a collection of subsets that do not have any element in common and, together, contain all elements of the parent set. Each set in this collection forms an *equivalence class* in the set we refer to as *blocks*. There are infinitely many ways of constructing such partitions of a set with infinite elements. These are diagrammatically illustrated on the left in figure 2. If a subset of the parent set can be expressed as a union of blocks of a partition, this subset is said to be *saturated* in the said partition. If the subset has a non-empty intersection with a block but without containing all of its contents, it is called *unsaturated*. For example, on the top right in figure 2, the yellow rectangle is saturated in the partitions of P_1, while the blue ellipse is unsaturated.

In function approximation, the target function's fibre corresponds to a unique partition out of all possible partitions of the input feature space. Therefore, the process of function approximation can be broken down into two stages:

1. finding a partition on the domain which matches that of the target function
2. matching the target function's value on these domain partitions over the family of functions having the same fibres.

By strong inductive biases, we mean the assumption of a partition on the domain, which helps in the first stage and is related to the specifics of the data representation and its associated architecture. The second part is related to the actual function finding via an optimisation algorithm where one can include additional information as weaker forms of inductive biases without hard restrictions on the partitions. For instance, regularisation terms on the loss function will prioritise a region of the weight space without a hard boundary. While our definition can be made more general to encompass such biases, we do not consider such a generalisation since the former is a necessary condition for the latter and is the primary motif of the work.

One can now define a strong inductive bias in terms of assumed partitions on the domain:

Strong Inductive Bias. Given an approximation problem where we want to learn a continuous target function $f:\mathcal{D}\to\mathcal{H}$ between the domain $\mathcal{D}$ and the co-domain $\mathcal{H}$ via an approximator $\hat{f}:\mathcal{D}\to\mathcal{H}$ belonging to a family of functions Σ, a strong inductive bias is an assumption of a partition of the domain $\mathcal{D}$, such that $\hat{f}(\mathbf{x})$ has constant value in each block of the partition for all $\hat{f}\in\Sigma$.

Let the partitions be represented as $\hat{\mathsf{P}}=\{[\mathbf{x}]^a_\Sigma : a\in I\}$, with I being a set which indexes each block $[\mathbf{x}]^a_\Sigma$. Since the collection $\hat{\mathsf{P}}$ is a partition of the domain, $\mathcal{D}=\bigcup_{a\in I}[\mathbf{x}]^a_\Sigma$, and $[\mathbf{x}]^a_\Sigma\cap[\mathbf{x}]^b_\Sigma=\emptyset$ for $a\neq b$ and $[\mathbf{x}]^a_\Sigma=[\mathbf{x}]^b_\Sigma$ otherwise. The assumed partitions define the smallest mutually exclusive subsets of the domain, where an approximated function should be equal. Therefore, a strong inductive bias defines a function space on the domain where any function's value has to be constant within a single block while they can be different in separate blocks as a whole. Additionally, there is no restriction to the functions becoming equal in two distinct blocks. Therefore, the partitions $\hat{\mathsf{P}}$ are the *minimal fibres* over the function space Σ.

Given the input feature space, the assumption of a partition reduces the learning process (the optimisation stage) to learning over single representatives from the equivalence classes. While it is most straightforward to encode the target function's fibre as partitions of the domain, their exact fibres are never known in practice. As a result, partitioning the domain to help achieve the target function's fibre demands a notion of compatibility. For a given target function $f:\mathcal{D}\to\mathcal{H}$, it essentially boils down to the comparison of two partitions in $\mathcal{D}$:

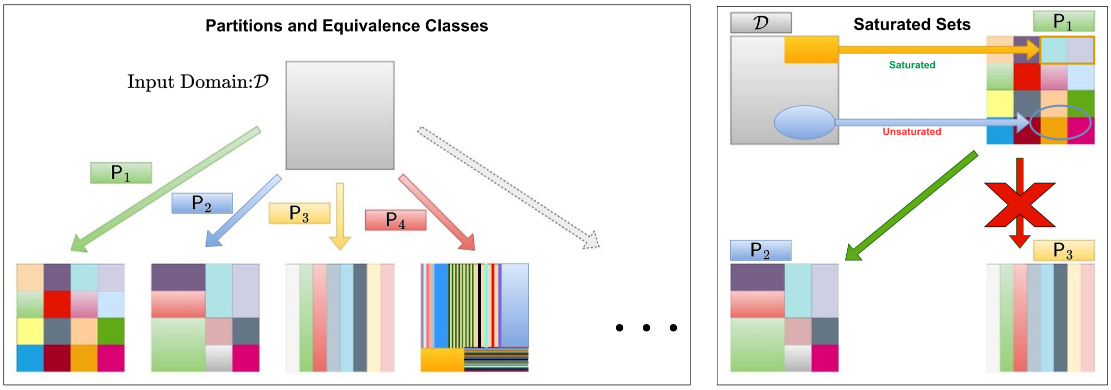

Figure 2. The left shows some possible partitions of a bounded domain $\mathcal{D}$ in 2D out of infinitely many possibilities. On the top right, the yellow rectangle is a saturated set in P_1 while the blue ellipse is not. Consequently, if one restricts the smallest possible fibres that a function approximator can have to be those in P_1, it can accommodate a target function with P_2 (bottom left of figure on the right) as its fibres since all partitions in P_2 are saturated under P_1. However, if it had P_3, no amount of function approximation on P_1 will agree over the whole domain $\mathcal{D}$ since all of its fibres are unsaturated in P_1. Note that for incompatibility, one fibre being unsaturated is sufficient for incorrectness of P_1.

- The partitions P induced by the target function's fibres, say $[\mathbf{x}]_f$, where the function is equal in each block $[\mathbf{x}]_f \in \mathsf{P}$
- The smallest possible fibres restricted via the inductive bias in all functions $\hat{f}: \mathcal{D} \to \mathcal{H}$ represented by the neural network architecture class, say $\hat{\mathsf{P}} \ni [\mathbf{x}]_\Sigma$.

In the sense of an exact representation[3], the requirement is that a strong inductive bias (or partitions of the domain) is compatible with a target function if all of its fibres are saturated sets in the assumed partitions i.e.

$$[\mathbf{x}]_f^b = \bigcup_{a \in I_b} [\mathbf{x}]_\Sigma^a$$

for every $[\mathbf{x}]_f^b$ in P, with I_b an index set for each $[\mathbf{x}]_f^b$. If this does not hold true, the target function has two distinct fibres in some partition $[\mathbf{x}]_\Sigma$ in $\hat{\mathsf{P}}$ and any $\hat{f}$ cannot simultaneously become equal to both values in $[\mathbf{x}]_\Sigma$. Going back to the bottom right of figure 2, an inductive bias of P_1 is correct if the target function has fibres that correspond to P_2, and incorrect if its fibres are P_3.

In particle physics applications, target functions are generally invariant under a group and therefore, elements belonging to each partition $[\mathbf{x}]_f$ are related by symmetry transformations. Due to the nice algebraic properties of elements in each fibre of the target function, it is comparatively straightforward to construct architectures which respect these symmetries. Therefore, symmetries play an important role in function approximation tasks. As we shall see, the main difference to the usual notion of symmetries is that the largest possible physical symmetry in the input domain is not necessarily the best choice since it enlarges the minimal fibres compared to its subgroup symmetries.

3. Optimal symmetries in group invariant classification

For a group $\mathcal{G}$ with corresponding transformations $\rho_\mathcal{D}$ and $\rho_\mathcal{H}$ on the domain and co-domain, respectively, a function $f: \mathcal{D} \to \mathcal{H}$ is equivariant with respect to these transformations if

$$f(\rho_\mathcal{D}(g)\,\mathbf{x}) = \rho_\mathcal{H}(g)\, f(\mathbf{x}) \quad . \tag{3.1}$$

If the representation $\rho(g)_\mathcal{H}$ is trivial ($\rho_\mathcal{H}(g) = \mathbf{1}\, \forall g \in \mathcal{G}$), then f is called $\mathcal{G}$-invariant. Particle fields in Quantum Field Theory are classified in terms of their transformation properties under the Lorentz group, and interacting theories are written down with Lorentz invariant Lagrangians and additional internal symmetries. This is the origin of the symmetries of the differential cross-section. For instance, take the transformation

$$\psi\left(\Lambda(g)\; p_\nu\right) = W(g)\;\psi\left(p_\nu\right) \quad ,$$

[3] The case involving an ϵ-accurate approximation with $\epsilon > 0$ is more involved and will be touched upon in a future work [63]. For the present work, it suffices to regard that the exact representation belongs to the restricted function space where the relevant architecture class with an inductive bias has (or should have) the universal approximation property.

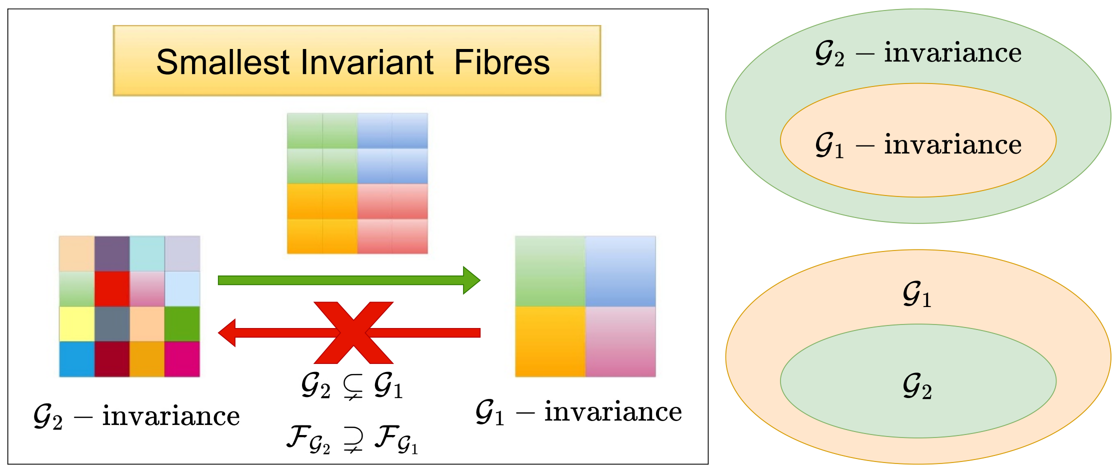
Smallest Invariant Fibres
$\mathcal{G}_2$ − invariance
$\mathcal{G}_2 \subsetneq \mathcal{G}_1$
$\mathcal{F}_{\mathcal{G}_2} \supsetneq \mathcal{F}_{\mathcal{G}_1}$
$\mathcal{G}_1$ − invariance
$\mathcal{G}_2$ − invariance
$\mathcal{G}_1$ − invariance
$\mathcal{G}_1$
$\mathcal{G}_2$

Figure 3. The difference between the smallest *fibres* (the set of points in the domain where the function's value is equal) of a function invariant under a group $\mathcal{G}_1$ and its proper subgroup $\mathcal{G}_2$. The smaller squares denote the coarsening of the domain with similar colours, signifying equality of the function's value. $\mathcal{G}_1$-invariance assumes larger fibres from the start. In contrast, $\mathcal{G}_2$-invariance assumes smaller and *compatible partitions* with $\mathcal{G}_1$-invariance, i.e. they can become enlarged so that the function becomes equal on the smallest fibres of $\mathcal{G}_1$-invariant functions.

of the Weyl spinor $\psi(p_\mu)$, $\psi:\mathbb{R}^4\to\mathbb{C}^2$, under a Lorentz group element g, where $W(g)$ and $\Lambda(g)$, respectively, are the Weyl and vector representations of the Lorentz group. While the term 'equivariance' is seldom used in QFT textbooks, classically, ψ is a Lorentz equivariant function under the defined group transformations.

The nature of perturbative differential cross-sections already contains a rich structure of symmetries without going into the specific details of processes. On the other hand, the search for new physics is essentially a hypothesis test with the null background-only hypothesis vs the alternate signal and background hypothesis. With optimality of the likelihood ratio, guaranteed by the Neyman–Pearson lemma [64], one can study the optimality of an imposed group equivariance by checking whether the space of a family of group equivariant functions contains monotonic functions of the likelihood ratio. We briefly discuss this connection by describing the structure of fibres of group equivariant functions and its relation to the Neyman–Pearson optimality of group invariant likelihood ratios. This is essentially a condensed summary of [65].

3.1. Equivariant function spaces

As mentioned in figure 1, the underlying motivation for choosing correct symmetries is the comparable constraints of a hierarchical set of group invariant functions. More precisely, there is a set-inclusion relationship within the space of invariant functions of a group and its subgroup, which goes in the opposite direction of group inclusions. Take a group $\mathcal{G}_1$ and its proper sub-group $\mathcal{G}_2$, i.e. $\mathcal{G}_2 \subsetneq \mathcal{G}_1$. On the same domain and a given group action of the group $\mathcal{G}_1$, restricting the group elements to those in $\mathcal{G}_2$ creates a $\mathcal{G}_2$-action. For group invariance (i.e. trivial action on the co-domain), this creates two invariant function spaces on the domain $\mathcal{D}$: say $\mathcal{F}_{\mathcal{G}_1}$ and $\mathcal{F}_{\mathcal{G}_2}$. Since all $\mathcal{G}_1$ invariant functions are $\mathcal{G}_2$ invariant, but not all $\mathcal{G}_2$ invariant functions are $\mathcal{G}_1$ invariant, we have : $\mathcal{F}_{\mathcal{G}_2} \supsetneq \mathcal{F}_{\mathcal{G}_1}$. This means that functions which are $\mathcal{G}_2$ invariant but not $\mathcal{G}_1$ invariant do not belong to $\mathcal{F}_{\mathcal{G}_1}$. A schematic diagram depicting this inverted hierarchy in the function space is shown in figure 3. Assuming the target function is always group invariant, a qualitative explanation of why this happens is given separately for the $\mathcal{G}$-invariant classification and $\mathcal{G}$-equivariant feature extraction.

3.1.1. $\mathcal{G}$-invariance

Suppose a given function $f:\mathcal{D}\to\mathcal{H}$ is invariant under a transformation $\rho_{\mathcal{D}}(g)$ of a group $\mathcal{G}$. This means that $f(\rho_{\mathcal{D}}(g)\,\mathbf{x}) = f(\mathbf{x}) \equiv \mathbf{y}$ for all $g\in\mathcal{G}$, i.e. the fibre of an element $\mathbf{y}$ in the image of the function $\mathrm{Im}(f)$, is at least as large as all those elements which can be traversed from $\mathbf{x}$ via the group action $\rho(g)_{\mathcal{D}}\,\mathbf{x}$. This subset of elements in $\mathcal{D}$ is the orbit of $\mathbf{x}$ under the $\mathcal{G}$-action. While a like-for-like comparison between different group invariant neural networks is highly non-trivial, the structure of the smallest fibres (see figure 3) induced by group invariance in the input domain $\mathcal{D}$ provide a mathematically consistent mechanism of checking the suitability of a particular group invariance in the input domain even without recourse to the specific detail of the architecture or the universal approximation property. The important observation which allows such an inspection is that restricting the group action on the domain $\mathcal{D}$ to group elements of a proper subgroup generally[4] results in smaller minimal fibres as they have smaller orbits. Crucially, larger group invariance forces a function to be equal in different orbits of the subgroup action and is, therefore, not a correct symmetry when the target function is invariant only under a proper subgroup but not under the parent group. On the other hand, since group invariance only fixes the smallest fibres of a function, an expressive enough invariant network of a smaller group can approximate a function invariant under a larger group. The state-of-the-art performance of transformers [66] for jet-tagging [67] which match or outperform equivariant ones [55, 57, 59, 62] is an extreme example of an architecture learning the relevant fibre structure of the target function without continuous group algebraic constraints in the domain. A diagrammatic representation of the compatibility of a subgroup invariance for a target function invariant under a larger group and incompatibility of a larger group invariance for a proper subgroup invariant target function is shown in figure 3.

3.1.2. $\mathcal{G}$-equivariance

Now consider that the function $f: \mathcal{D} \to \mathcal{H}$ is equivariant with respect to is a corresponding non-trivial transformation $\rho_{\mathcal{H}}(g)$ of the group acting on the co-domain, i.e. $f(\rho_{\mathcal{D}}(g)\,\mathbf{x}) = \rho_{\mathcal{H}}(g) f(\mathbf{x})$. In such a case, the function is equal for at least those elements $\mathbf{x}' = \rho_{\mathcal{D}}(g)\,\mathbf{x}$ in the domain $\mathcal{D}$, transformed by group elements g which fixes $\mathbf{y} = f(\mathbf{x})$, i.e. $\mathbf{y} = \rho_{\mathcal{H}}(g)\,\mathbf{y}$. This subgroup of $\mathcal{G}$, dependent on the representation $\rho_{\mathcal{H}}(g)$ and the particular element $\mathbf{y}$, is the little group [68] of the group transformation for $\mathbf{y}$. For group invariant binary classification of signal and background events, one can consider that $\mathcal{H}$ is a hidden representation where we extract the relevant features as the target function. Within this, there are two extremes depending on the nature of the representation $\rho_{\mathcal{H}}(g)$ in the co-domain $\mathcal{H}$. If the action is *free*, i.e. the little group of every element in $\mathcal{H}$ is the trivial group consisting only of the identity, equivariant feature extraction does not assume any larger fibres than the one assumed by an invertible function between the input domains. Therefore, for any noticeable gain in inductive biases, the group action on $\mathcal{H}$ should not be free. At the other extreme, if the little group of all elements in $\mathcal{H}$ is the group itself, then f is $\mathcal{G}$-invariant. Therefore, in the case of group equivariant feature extraction for an invariant target function, the little group of all the elements in the co-domain should be no larger than the largest subgroup under which the target function is invariant. One should remember that our discussions relate to the equivariant approximation of an invariant target function. For equivariant target functions, the purpose of equivariance beyond the assumption of a fibre structure is an efficient generalisation to unseen input data related via group transformations. Here, a correct free group action on the co-domain will offer advantages compared to non-equivariant ones in generalisation capabilities. Moreover, given a free group action on the co-domain, one can build subgroup invariants out of the equivariant quantities, manually inducing appropriate little groups. Such an approach would be suitable, for instance, in multi-class classification tasks where the different likelihood ratios are invariant under different subgroups of a parent group.

3.2. Neyman–Pearson optimality and group equivariance

Consider the binary classification problem of a signal hypothesis $\mathcal{P}_{\mathrm{S}}$ with the corresponding set of processes $\mathcal{P}_{\mathrm{B}}$ from the known sector of the Standard Model forming the background hypothesis. For an observed event E, each hypothesis $H \in \{S, B\}$ has a normalised probability densities $p_H(\mathbf{E}) = \frac{1}{\sigma_H}\frac{\mathrm{d}\sigma_H}{\mathrm{d}\mathbf{E}}$, with $\mathrm{d}\sigma_H$ and σ_H, the differential and integrated cross section for the set of processes $\mathcal{P}_H$. From the Neyman–Pearson lemma, an optimal classifier between the two hypotheses is a monotonic function of the likelihood ratio[5] $\lambda(\mathbf{E}) = p_{\mathrm{S}}(\mathbf{E})/p_{\mathrm{B}}(\mathbf{E})$. Thus, for a group-equivariant neural network to approximate a monotonic function of the likelihood ratio, the smallest fibres assumed via group equivariance should be comparable to that of the likelihood ratio. Recalling the nature of group equivariant fibres as discussed above, one can construct the following guidelines for an optimal choice of the group $\hat{\mathcal{G}}$ given a $\mathcal{G}$-invariant likelihood ratio:

- **$\hat{\mathcal{G}}$-invariance**: $\hat{\mathcal{G}}$ can be a subgroup of $\mathcal{G}$ but not larger
- **$\hat{\mathcal{G}}$-equivariance**: The little group of the $\hat{\mathcal{G}}$-action on the co-domain should not be larger than $\mathcal{G}$.

For the $\hat{\mathcal{G}}$-equivariant case, a free action on the codomain will be compatible with any target function. Still, it will not provide any noticeable gain in generalisation ability compared to non-equivariant architectures. These guidelines also hold for any general $\mathcal{G}$-invariant target function.

[4] Mathematically, the group action should be *effective* in that any non-identity group element has at least one non-trivial action on an element of the domain. This property is generally satisfied by group actions utilised in particle physics.

[5] Generally, the alternate hypothesis for a signal search at the LHC is the presence of both signal and background processes, in which case the probability distribution is $p_1(\mathbf{E}) = \frac{1}{\sigma_{\mathrm{S}}+\sigma_{\mathrm{B}}}(\frac{\mathrm{d}\sigma_{\mathrm{S}}}{\mathrm{d}\mathbf{E}} + \frac{\mathrm{d}\sigma_{\mathrm{B}}}{\mathrm{d}\mathbf{E}})$. Therefore, the likelihood ratio is $\lambda(\mathbf{E}) = \frac{\sigma_{\mathrm{B}}}{\sigma_{\mathrm{S}}+\sigma_{\mathrm{B}}}(1 + \frac{\mathrm{d}\sigma_{\mathrm{S}}/\mathrm{d}\mathbf{E}}{\mathrm{d}\sigma_{\mathrm{B}}/\mathrm{d}\mathbf{E}})$. Our case considers the behaviour of the non-constant second term as the symmetry properties depend on this term alone.

The guidelines provide little utility in binary classification tasks when one knows the group $\mathcal{G}$. The real utility of these guiding principles arises in $\hat{\mathcal{G}}$-equivariant feature extraction for multi-class classification where each class c, has a possibly different $\mathcal{G}_c$-invariant probability distribution. One can then use knowledge of the invariant probabilities to identify the group invariant likelihood for one-vs-one and one-vs-many classification scenarios and construct a $\hat{\mathcal{G}}$-equivariant function, which contains all these possibilities as its subgroup and has a compatible action with the final subgroup invariances in the intermediate feature extraction layers. One can enforce the appropriate little group invariances of the different possibilities at the final feature extraction layer to feed into the classifier head.

In signal searches, different processes in $\mathcal{P}_\mathrm{H}$ may have a different set of permutation symmetry. For instance, the event weight in the resonant decay of a Z boson to a pair of leptons will be invariant under their exchange, while it will not be if they originate from a pair of $W^\pm$ bosons. Such physical arguments open up an avenue for the design of equivariant architectures tailored to particular search scenarios, which on top of theoretically[6] being able to approximate a monotonic function of the likelihood ratio will have better parameter and sample efficiency. They can also be used to modify the architecture of foundation models before fine-tuning for particular search scenarios.

4. Equivariant architectures from the MEM

In the point cloud representation, one regards the input as a set and learns a permutation-invariant function for all possible permutations of the elements. They generally utilise sum-decomposition in a latent space to account for variable cardinalities of the samples, which is known to have universally approximating properties as set [69] and multi-set [70] functions. However, to study the optimality of the permutation group action on the squared matrix elements, we will consider a point cloud sample as an ordered n-tuple where we define functions to be invariant under possibly different permutation groups $S_{n'}$, acting on $n' \leqslant n$ elements. In this section, we first discuss the Lorentz and permutation symmetries of fixed-order differential cross-sections. We then discuss optimal symmetries that are present in the matrix-element likelihoods and present a longitudinal boost equivariant architecture which respects these symmetries.

4.1. Symmetries in fixed-order differential cross sections

4.1.1. Lorentz symmetry

Let $\mathbf{X} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n)$ be the four-vectors of a measured event at LHC. In addition to these four-vectors, we have a corresponding vector $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots., \mathbf{h}_n)$ containing additional information such as the type of the reconstructed object, flavour, charge etc. These properties determine the information available on the partonic process at reconstruction and the permutation symmetry of the differential cross-sections in addition to the quantum mechanical indistinguishability of identical particles. Representing the combined observed information of $\mathbf{X}$ and $\mathbf{H}$ as $\mathbf{E} = (\mathbf{p}_1 \oplus \mathbf{h}_1, \mathbf{p}_2 \oplus \mathbf{h}_2, \ldots, \mathbf{p}_n \oplus \mathbf{h}_n)$, consider that there are r incoherent but observationally identical (i.e. at reconstruction) processes $\mathcal{P} = \{a_1\, b_1 \to F_1, a_2\, b_2 \to F_2, \ldots, a_r\, b_r \to F_r\}$ that can lead to the production of this event. Here, a_i and b_i are the incoming partons, and F_i represents the partonic final state. The leading-order differential cross section dependent on theory parameters θ can be written as

$$\begin{aligned}\mathrm{d}\sigma_{\mathcal{P}}\left(\mathbf{q}_1, \mathbf{q}_2, \mathbf{E}, \theta\right) = \sum_{a_i b_i \to F_i \in \mathcal{P}} \int \mathrm{d}x_1\, \mathrm{d}x_2\, \frac{f_{a_i}\left(x_1\right)\ f_{b_i}\left(x_2\right)}{2E_{cm}\, x_1 x_2}\, \delta^{(4)}\left(x_1 \mathbf{q}_1 + x_2 \mathbf{q}_2 - \sum_{j=1}^{n} \mathbf{p}_j\right) \\ \times \left|\mathcal{M}_i\left(x_1 \mathbf{q}_1, x_2 \mathbf{q}_2, \mathbf{E}, \theta\right)\right|^2 \mathrm{d}\Pi_n \quad ,\end{aligned} \tag{4.1}$$

where[7] $\mathbf{q}_1 = (0, 0, E_{cm}/2, E_{cm}/2)$ and $\mathbf{q}_2 = (0, 0, -E_{cm}/2, E_{cm}/2)$ are the incoming proton momenta with centre-of-mass energy E_{cm}, $|\mathcal{M}_i|^2$ is the Lorentz invariant squared matrix-element for the parton-level process $a_i b_i \to F_i$, f_{a_i} and f_{b_i} are the proton parton distribution functions of the parton a_i and b_i, respectively, and $\mathrm{d}\Pi_n$ is the Lorentz invariant phase space (LIPS) of the n-body final state $\mathrm{d}\Pi_n = \prod_{j=1}^{n} \frac{d^3 p_j}{(2\pi)^3 2E_j}$. Given a Lorentz group element g, the corresponding transformation of the final state $\mathbf{E}$ is

$$\Lambda_{\mathbf{E}}\left(g\right)\mathbf{E} = \left(\Lambda\left(g\right)\mathbf{p}_1 \oplus \mathbf{h}_1, \Lambda\left(g\right)\mathbf{p}_2 \oplus \mathbf{h}_2, \ldots, \Lambda\left(g\right)\mathbf{p}_n \oplus \mathbf{h}_n\right) \quad , \tag{4.2}$$

where the matrix representation $\Lambda_{\mathbf{E}}(g)$ can be built from the vector representation $\Lambda(g)$ acting on four vectors $\mathbf{p}_i$, and the trivial identity matrix representation acting on scalars $\mathbf{h}_i$. Events correspond to different

[6] depending on the universal approximation property of the equivariant architecture class.

[7] We use the convention $\mathbf{p} = (p_x, p_y, p_z, E)$ for easier discussion of the transverse and longitudinal components in later sections.

points in the phase space whose relative weight is determined by the Lorentz invariant matrix-element squared $|\mathcal{M}_i|^2$, i.e. the probability distribution of a given final state signature under a hypothesised process $a_i b_i \to F_i$ is Lorentz invariant. At this point, the sum over all processes is also Lorentz invariant. However, experimental considerations render event likelihoods that do not respect the full Lorentz invariance. This will be discussed further in section 4.1.1.

4.1.2. Permutation symmetries

Let the observed event be $\mathbf{E} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ such that $\mathbf{r}_i = \mathbf{p}_i \oplus \mathbf{h}_i$. The action of the n-object permutation group S_n on $\mathbf{E}$, permutes each $\mathbf{r}_i$ as a whole

$$\rho(\sigma)(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) = \left(\mathbf{r}_{\sigma(1)}, \mathbf{r}_{\sigma(2)}, \ldots, \mathbf{r}_{\sigma(n)}\right) \quad ,$$

where $\rho : S_n \to \mathrm{GL}(n \times (4+m), \mathbb{R})$ is a matrix representation of S_n built as $\rho(\sigma) = \rho_n(\sigma) \otimes \mathbf{1}_{4+m}$, out of the canonical representation $\rho_n(\sigma)$ of S_n in $\mathrm{GL}(n, \mathbb{R})$, with m being the dimensions of $\mathbf{h}_i$. **$\mathrm{GL}(n, \mathbb{R})$ is the general linear group of real invertible $n \times n$ matrices.** Similarly for some $n' < n$, one can also define the permutation action on n' elements via a representation $\rho_{n'} : S_{n'} \to \mathrm{GL}(n, \mathbb{R})$ of $S_{n'}$ in $\mathrm{GL}(n, \mathbb{R})$. Clearly, there are $\binom{n}{n'}$ ways of choosing subsets of cardinality n' from $\mathbf{E}$, each having a particular form of the matrix $\rho_{n'}(\sigma') \in \mathrm{GL}(n, \mathbb{R})$, $\sigma' \in S_{n'}$ reflecting the chosen subset. A function $f : \mathcal{E} \to \mathcal{H}$, where $\mathcal{E}$ is the space of measured events is permutation invariant if

$$f(\rho(\sigma)(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)) = f(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) \quad , \tag{4.3}$$

for all $\sigma \in S_n$. The differential cross-section is not symmetric concerning the exchange of distinct particles, which results in the non-invariance of the likelihood under the exchange of reconstructed objects belonging to different classes. This will be discussed further in section 4.1.1.

4.2. Optimal symmetries from the MEM

The MEM is a theoretically motivated multivariate data analysis approach which evaluates the likelihood of an event arising from a set of parton-level processes $\mathcal{P}$. With a slight modification of equation (4.1) to account for detector effects and implicitly considering momentum conservation, the likelihood of an event $\mathbf{E}$ arising due the ith parton level process in $\mathcal{P}$ say $ab \to F$, is

$$p_i(\mathbf{E}|\theta) = \frac{1}{\sigma_i} \int \mathrm{d}\Pi_n(\mathbf{P})\ \mathrm{d}x_1\,\mathrm{d}x_2\, \frac{f_a(x_1) f_b(x_2)}{2E_{cm}\, x_1 x_2} |\mathcal{M}_i(x_1 \mathbf{q}_1, x_2 \mathbf{q}_2, \mathbf{P}, \theta)|^2\, T(\mathbf{E}, \mathbf{P}) \quad . \tag{4.4}$$

Here, $T(\mathbf{E}, \mathbf{P})$ is the transfer function modelling the probability of the event $\mathbf{E}$ arising from the final state four-vectors $\mathbf{P}$ of the partonic configuration F. In conjunction with the integration over the parton-level LIPS $\mathrm{d}\Pi_n(\mathbf{P})$, the transfer function accounts for detector effects which decide up to what extent the exact symmetries of $|\mathcal{M}_i|^2$ are carried over to the likelihood $p_i(\mathbf{E}|\theta)$ or add new discrete symmetries by making quantum mechanically non-identical partons indistinguishable due to experimental considerations. The likelihood for the hypothesis set $\mathcal{P}_\mathrm{H}$, with $\sigma_\mathrm{H} = \sum_i \sigma_i$ is

$$p_\mathrm{H}(\mathbf{E}|\theta) = \frac{1}{\sigma_H} \sum_{i \in \mathcal{P}_H} \sigma_i p_i(\mathbf{E}|\theta) \quad .$$

Therefore, in such a set-up, one can construct the likelihood and likelihood ratios of any set of non-interfering parton level processes. Moreover, for equivariant feature extraction, one can infer the (approximate) optimal group symmetries from each $p_i(\mathbf{E}|\theta)$.

In this section, we highlight the general structure of symmetries inherent in the likelihoods while consistently taking resonant and non-resonant production of di-Higgs decaying to four bottom jets as an example to concretely illustrate the synergy between group equivariant architecture design and the probabilities $p_\mathrm{H}(\mathbf{E})$. This is one of the most promising channels for looking into the quartic Higgs self-coupling at LHC, as it has the highest branching ratios but is plagued by a very high QCD multi-jet background, and we will consider it for the numerical analysis in the next section.

4.2.1. Continuous symmetry

In most searches, we are interested in a fixed number of primary partons: the four bottom quarks in the di-Higgs case. Due to the inevitability of additional QCD radiation at the very high energies of LHC, a rigid cut on the number of jets is sub-optimal as it throws away many possible signal events. To consider many events, one includes hard processes with additional QCD radiation beyond the four bottom quarks in the signal and the background sets of processes. Possibly coherent processes in the unresolved regime within

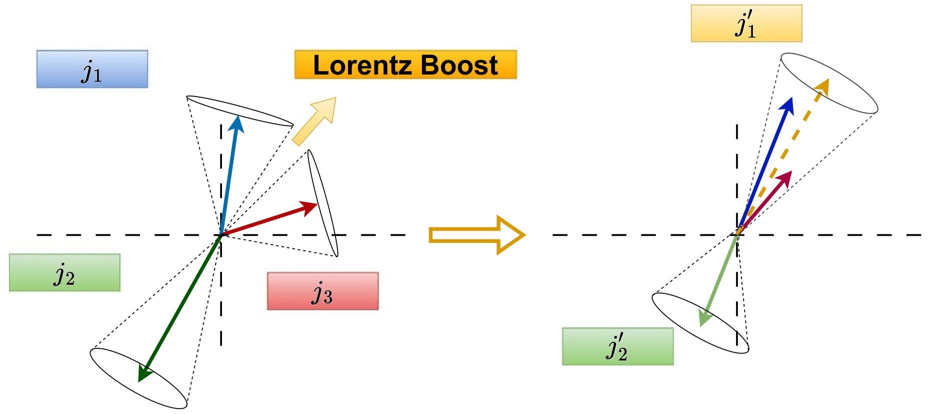

Figure 4. A Lorentz boost in the direction opposite to the leading jet in a three-jet event (on the left) will transform it into a two-jet event (on the right). For a baseline selection criteria which allows for more than two jets in the final state, the MEM-likelihood evaluated as a sum of three jet final state processes and two jet final state processes will not be the same for either event, making the event likelihood violate Lorentz invariance.

these processes must be matched and merged with the parton-shower-generated additional radiation to avoid over-counting in the overlapping phase space regions. Additionally, these processes involve a variable number of final state particles that do not live in the same phase space. Special care needs to be taken to evaluate such weights [4–7, 10–12]. One mechanism is to introduce kinematic corrections on an event-by-event basis for manageable number of additional hard radiations, [4, 6, 7] so that the weights are evaluated in the phase space involving fixed number of primary partons. Such kinematic corrections are essentially a preprocessing stage in machine learning terminology.

To bring in MEM-inspired symmetries into equivariant architecture design, we do not consider a kinematic preprocessing stage and consider the group invariance of MEM-weights of the sum over processes with a variable number of final state particles. In such a case, since the transfer function $T(\mathbf{E},\mathbf{P})$ involves the reconstruction algorithm and baseline selection criterion, the likelihood $p_i(\mathbf{E}|\theta)$ is not necessarily Lorentz invariant. For example, commonly used jet algorithms depending on p_T and ΔR are longitudinal boost invariant but not fully Lorentz invariant. In figure 4, we show a fully visible final state with three jets on the left, becoming a two-jet event on the right with an appropriate Lorentz boost. In the three jet event, the green leading jet has a large transverse momentum compared to the two sub-leading jets, a boost along the direction opposite to the leading jet will result in its momentum becoming lower with the two sub-leading jets coming closer. Once the sub-leading jets' angular separation is reduced to within the jet radius, the event will become a two-jet event, as shown on the right. The situation becomes more severe for signatures with invisible particles in the final state where there is no upper limit on the missing transverse momentum as there are many possible boost directions, which will result in two separated objects becoming unresolvable in the sample space of selected events since the momentum mismatch in the lab-frame will be regarded as belonging to the invisible particles and therefore belong to the sample space of selected events.

From this example, one can see that the event likelihood is not Lorentz invariant because of the non-invariance of the jet algorithm where the radius is kept fixed and the measures ΔR_{ij} transform non-trivially under general Lorentz boosts or rotations. The isolation criteria on other types of reconstructed objects and the jet algorithm are generally invariant under rotations, and Lorentz boosts along the z-axis, with the likelihood maintaining invariance under such a sub-group. As a group which mixes the orbits under longitudinal boosts and rotations along the z-axis, the Lorentz group is strictly larger and, hence, an incorrect group.

4.2.2. Discrete symmetry

An event consists of sets of different reconstructed objects like leptons, light jets, bottom jets, photons, etc which may be grouped into a single class or separated depending on the signal and background hypotheses. Denoting each object type as a vector $\mathbf{E}_\alpha$ with each $\alpha \in \{1,2,..,k\}$ specifying the object type of k classes of reconstructed objects with cardinality n_α, an event is represented as a vector[8] $\mathbf{E} = \bigoplus_\alpha \mathbf{E}_\alpha$. Since a $\hat{\mathcal{G}}$-invariant function approximator cannot efficiently approximate any $\mathcal{G}$-invariant functions when $\mathcal{G}$ is a proper subgroup, we need to determine the largest possible permutation symmetry of the likelihood and the

[8] Strictly speaking, each $\mathbf{E}_\alpha$ as well as the full representation $\mathbf{E}$ is also a direct sum over $\mathbf{r}_i$. However, when considering the object properties, we will write all capital boldfaced vectors as a tuple of elements $\mathbf{r}_i$ to avoid confusion between the two situations.

likelihood ratios. Again, this is entirely determined by $T(\mathbf{E},\mathbf{P})$: for each reconstructed object $\mathbf{r}_i$ in $\mathbf{E}$, $T(\mathbf{E},\mathbf{P})$ assigns it all possible parton flavours within a sum. This renders $T(\mathbf{E},\mathbf{P})$ and hence $p_i(\mathbf{E}|\theta)$ invariant under the exchange of elements within the same reconstructed object class that have no charge information (i.e. jets, bottom-tagged jets, and photons but not electrons, muons, and tau jets). Therefore, even if two particles are (considered) indistinguishable at reconstruction, they may be separate particles in the partonic final states like gluons and quarks. On the other hand, distinguishable particles at reconstruction are always non-identical at the parton level, and first-principle arguments do not guarantee permutation invariance of the matrix-element squared under their exchange in the final state. Therefore, *if an observed event* $\mathbf{E}$ *with* n *objects contains more than one reconstructed object type, or if it contains a single object type with at least two objects having different observed charges, the process likelihood* $p_i(\mathbf{E}|\theta)$ *is not* S_n*-invariant.*

As a concrete example, let us consider a signature with two photons and three jets represented as $\mathbf{E}_\gamma = (\mathbf{r}_1^\gamma, \mathbf{r}_2^\gamma)$ and $\mathbf{E}_J = (\mathbf{r}_1^J, \mathbf{r}_2^J, \mathbf{r}_3^J)$. The largest symmetry in the underlying matrix elements is when all three jets originate from a gluon at the parton level. For this process, the matrix-element squared $|\mathcal{M}(\mathbf{r}_1^\gamma, \mathbf{r}_2^\gamma, \mathbf{r}_1^J, \mathbf{r}_2^J, \mathbf{r}_3^J)|^2$ is permutation invariant under the exchange of the two photons or within the exchange of gluons amongst themselves but not in the interchange of a photon and a gluon. Therefore, the MEM-likelihood is not S_5 permutation invariant. Almost all point cloud-based architectures studied for event-level analyses implicitly consider a full permutation invariant representation over the reconstructed objects regardless of the final state's composition. Even though this contains the smaller permutation symmetries of the MEM-likelihood ratio, S_n permutation symmetry is a larger symmetry unless all reconstructed objects belong to the same type and are, therefore, not a correct symmetry for any given final state.

A straightforward solution which fixes the non-invariance of the target function under the exchange of elements belonging to different blocks in any point cloud approach, including graph neural networks, is to operate a sub-graph readout over the different classes $\hat{\mathbf{E}}_\alpha$ which segregates the reconstructed objects based on distinguishability and then concatenate these sub-graph representations. For instance, in a mean readout operation, the event representation

$$\hat{\mathbf{E}} = \bigoplus_{\alpha=1}^{k} \left(\frac{1}{n_\alpha} \sum_{i=1}^{n_\alpha} \hat{\mathbf{r}}_i^\alpha \right) \quad , \tag{4.5}$$

fixes a particular ordering of the reconstructed object classes and is invariant only under permutations that act separately on each block vector $\hat{\mathbf{E}}_\alpha$'s constituents. So far, we have considered object reconstruction to have perfect accuracy. One should relax such rigid division of the reconstructed objects to account for experimental realities, including the possible absence of some classes in an event depending on the baseline selection criterion. This can be done by assigning relative weights $w_{\alpha_1 \leftarrow \alpha_2} \in (0,1]$ not necessarily symmetric, which controls the relative contribution of class α_2 to the readout operation of α_1. These weights could be learnt as an attention mechanism modified with the concatenation operation over the α_1 axis. However, as proof of principle, we do not consider such modifications and set the weights beforehand in the architecture design for the numerical experiments. Even though the modified structure may not affect the performance of highly expressive networks, we speculate it will affect the theoretical uncertainties when merging additional radiations at higher perturbative accuracies. Since understanding such theoretical uncertainties is crucial in deploying deep learning algorithms for phenomenological studies, we leave an in-depth analysis of such an impact for independent future work.

4.3. Approximate symmetries under the narrow width approximation (NWA)

As we have seen above, the largest permutation symmetry in an event is the product group $\otimes_{\alpha=1}^{k} S_{n_\alpha}$ permuting elements within the same class of reconstructed objects. However, these symmetries may change within the NWA where the decay dynamics of a narrow resonance is factorised from its production. For QCD background processes producing at least four bottom jets, the event weight is S_4 permutation invariant. In contrast, for the SM di-Higgs production within the NWA, out of the three possible partitions into two pairs of bottom quarks, the phase space volume where more than one of them lies near the mass peak is very small and hence, for most events, two out of the three distinct parton level pairings will have a negligible contribution to the overall event weight, giving us a reduced $S_2 \times S_2$ approximate symmetry. On the other hand, if instead of the SM di-Higgs production, there is a resonant heavy Higgs with a very small width, the complete S_4 symmetry is approximately restored as the dominant contribution will come from the larger resonant mass peak of the heavier Higgs boson. The situation becomes increasingly complex when, in a given set of processes for a hypothesis, some have intermediate resonances while others do not. Nevertheless, such permutation symmetric arguments could effectively guide architecture design for cascade decays.

One must, however, be cautious against the limitations of the narrow-width approximation [71]. The important takeaway message is that smaller group invariant approximations are not as overly constrained as larger ones: the smallest fibres of smaller group symmetries can become enlarged to those demanded by the larger one during training, but those of larger group invariant functions can not become smaller. Therefore, for the case of observationally indistinguishable particles, the restriction to a smaller permutation symmetry does not induce any additional restrictions beyond the ones dictated by measurements. Enlarging the symmetry in the case of distinguishable particles at reconstruction, like oppositely charged leptons, imposes the restrictions of NWA on the feature extraction even when the input data may contain effects beyond the NWA. An invariant graph readout over oppositely charged leptons, therefore, restricts the network to effects within the NWA in the case of resonant decays. To combine different processes into hypotheses, we would choose a permutation symmetry shared by all constituent processes.

4.4. Longitudinal boost equivariant message passing neural network

Let us now construct an equivariant architecture looking into longitudinal boost equivariant quantities for a given final state **E**. While the same can be done within the formalism of [55, 56] or that of [62], we choose the invariant theoretic formalism of [52, 53, 57, 58], where one builds invariants and equivariant functions out of the basis of $\binom{n}{2}$ combinatorial dot products. Before going into detail, let us clarify the nature of the Lorentz group and its appropriate little groups concerning the fibre structures discussed above to guide the mathematical form of the architecture.

Since we are eventually interested in invariant quantities, the graph readout should only propagate the invariant information. Within such an architecture, the feature extraction module by design has the smallest fibres of an invariant function, and one may erroneously conclude that intermediate equivariant updates are unimportant. However, the utility of function compositions (i.e. depth) in a neural network is to precisely induce successive topological changes in the data as evidenced in various studies [72, 73]. Therefore, one cannot *a priori* conclude that an invariant message passing update which induces larger minimal fibres of invariance from the beginning will behave the same as an equivariant update even though there is an invariant stage as one goes deeper in either network. Now, the equivariant updates of the longitudinal components (p_z, E), already take care of the $\mathrm{O}(2)$ symmetry along the z-axis since it is the little group of the longitudinal boost action of the full Lorentz action, i.e. the longitudinally equivariant update of (p_z, E) alone, make the fibres consists of rotations along the z-axis from the start. If one has a covariant expression of the complete four-vector update

$$\mathbf{p}'_{\mu,i} = \mathbf{p}_{\mu,i} + \sum_j \mathbf{p}_{\mu,j} \Phi\left(\mathbf{p}_1, \mathbf{p}_2, \ldots.\right) \quad ,$$

Φ being a longitudinal boost invariant function, the transverse components will respect the vector action of the $\mathrm{O}(2)$ rotations around the z-axis, and hence be able to capture the equivariant information of the rotation. This is because in the 4×4 matrix representation, longitudinal boosts and rotations along z-axis commute, i.e. we can break down the four-vector space as a direct sum of transverse and longitudinal components $p_\mu = (p_x, p_y) \oplus (p_z, E)$. However, in our final experiments, we only updated the longitudinal components and kept the $\mathrm{O}(2)$ invariant fibres from the beginning, as we did not find any additional performance gain[9]. As we shall see in section 5.4, a scalar-only update performs just as well as the scalar-vector update for both the resonant and non-resonant di-Higgs searches.

At the $(l+1)$th stage of message passing, $l \geqslant 0$, let $\tilde{\mathbf{h}}_i^{(l)}$, $\tilde{\mathbf{e}}_{ij}^{(l)}$ be Lorentz scalar node-representation and edge representations, respectively. Similarly, let $\mathbf{h}_i^{(l)}$ and $\mathbf{e}_{ij}^{(l)}$ be longitudinal boost invariant representations. With $\tilde{\mathbf{x}}_i^{(0)} = (p_x, p_y)$ and $\mathbf{x}_i^{(l)} = (p_z^{(l)}, E^{(l)})_i$, the transverse and longitudinal components of the covariant four-vector $\mathbf{p}_i^{(l)} = (p_x, p_y, p_z^{(l)}, E^{(l)})_i$ we have

$$\mathbf{p}_i^{(l)} = \tilde{\mathbf{x}}_i^{(0)} \oplus \mathbf{x}_i^{(l)} \quad . \tag{4.6}$$

Since all invariants of the Lorentz group are longitudinal boost invariant, let $\bar{\mathbf{h}}_i^{(l)}$ and $\bar{\mathbf{e}}_{ij}^{(l)}$ be longitudinal boost invariant quantities that are not fully Lorentz invariant, so that we have $\mathbf{h}_i^{(l)} = \bar{\mathbf{h}}_i \oplus \tilde{\mathbf{h}}_i$ and $\mathbf{e}_{ij}^{(l)} = \bar{\mathbf{e}}_{ij}^{(l)} \oplus \tilde{\mathbf{e}}_{ij}^{(l)}$. The transverse component $\tilde{\mathbf{x}}_i^{(l)}$ being longitudinal boost invariant can be included in $\bar{\mathbf{h}}_i^{(l)}$, if

[9] This is done using $\mathbf{p}_i$ instead of $\mathbf{x}_i$ in the vector update expression in eq 4.7 since the scalars utilised are invariant under z-axis rotations. We did not find any difference in performance with such an update compared to the longitudinal-only update reported in section 5.4. However, no hyperparameter scan was conducted for either choices.

one chooses only to update the longitudinal components, but must be left out if we want an O(2) equivariant update of the transverse components.

With the notations clarified and abbreviating $\mathbf{p}_i^{(l)} + \mathbf{p}_j^{(l)} = \mathbf{p}_{ij}^{(l)}$, we can construct a longitudinal equivariant message passing operation which updates $\mathbf{h}_i^{(l)}$ and $\mathbf{x}_i^{(l)}$ as

$$\begin{aligned}
\mathbf{m}_{ij}^{(l+1)} &= \Phi_e^{(l+1)}\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, \mathbf{e}_{ij}^{(l)}, |\mathbf{p}_{ij}^{(l)}|^2_{(1,2)}, \langle \mathbf{p}_i^{(l)}, \mathbf{p}_j^{(l)} \rangle_{(1,2)}, |\mathbf{p}_{ij}^{(l)}|^2, \langle \mathbf{p}_i^{(l)}, \mathbf{p}_j^{(l)} \rangle\right) \quad , \\
\mathbf{x}_i^{(l+1)} &= \mathbf{x}_i^{(l)} + \sum_{j\in\mathcal{N}(i)} \mathbf{x}_j^{(l)}\, \Phi_x^{(l+1)}\left(\mathbf{m}_{ij}^{(l+1)}\right) \quad , \\
\mathbf{m}_i^{(l+1)} &= \frac{1}{|\mathcal{N}(i)|} \sum_{j\in\mathcal{N}(i)} \mathbf{m}_{ij}^{(l+1)} \quad , \\
\mathbf{h}_i^{(l+1)} &= \Phi_h^{(l+1)}(\mathbf{h}_i^{(l)}, \mathbf{m}_i^{(l+1)}) \quad .
\end{aligned} \tag{4.7}$$

Here, $\mathcal{N}(i)$ denotes the neighbourhood of node i, while $(1,2)$ in the subscripts denotes taking the relevant operation over the x and y axis only. The functions $\Phi_e^{(l+1)}$, $\Phi_x^{(l+1)}$, and $\Phi_h^{(l+1)}$ are multi-layer perceptrons (MLP), with $\Phi_x^{(l+1)}$ giving a one-dimensional weight after a sigmoid activation on the final layer. While we have included a node-update function $\Phi_h^{(l+1)}$, we have used $\mathbf{m}_i^{(l+1)} = \mathbf{h}_i^{(l+1)}$ in our experiments as there was no relative difference in the performance.

5. Illustrative example: Di-Higgs to four bottom jets

We employ the challenging but important di-Higgs search in the four bottom decay channels to test the methodology developed in the previous section. A recent work [41] utilising Symmetry Preserving Attention Networks (SPA-NET) [38–40] achieved state-of-the-art performance in the resonant and non-resonant production channel of the two Higgs boson, where in the former, there is an additional BSM heavy scalar boson which then resonantly decays to the two SM Higgs. As discussed above, while the final signatures are the same for both signals, they have inherently different approximate permutation symmetries. Moreover, we use the same data made public [74] by the authors, with the only essential difference coming from the network analysis.

5.1. Dataset description

We highlight the important elements of the utilised dataset. Parton level events were generated using `MadGraph5_aMC@NLO (v3.3.1)` [75] at $E_{cm} = 13$ TeV, which were showered and hadronised with `Pythia8.306` [76]. All stable hadrons went through a detector simulation in `Delphes (v3.5.0)` [77]. In the object reconstruction, `FastJet (v3.3.4)` [78] was used to cluster anti-k_t [79] jets with radius $R = 0.4$ and transverse momentum $p_\mathrm{T} \geqslant 20$ GeV.For the resonant analysis, the b-tagging efficiencies were modified at the 70% working point of the ATLAS MV2c10 b-tagger [80, 81]. In contrast, the non-resonant case was modified to the 77% working point of ATLAS DL1r tagger [82]. Selected events contain at least four b-tagged jets with $p_\mathrm{T} > 40$ GeV and $|\eta| < 2.5$. We refer interested readers to [41] for more data generation and baseline selection details.

5.2. Preprocessing and data representation

In each event, we use the four hardest b-tagged jets to form the two Higgs candidates using the $\Delta R + \min D_{hh}$ cut-based pairing motivated by the ATLAS analysis [83] also utilised in the cut-based pairing in the dense neural network input in [41] with a minor difference. For the ΔR requirement, defining the candidate with leading p_T as h_1 and the other as h_2, one considers the cut

$$\begin{aligned}
\frac{360\ \mathrm{GeV}}{m_{4j}} - 0.5 < \Delta R_{bb}^{h_1} < \frac{653\ \mathrm{GeV}}{m_{4j}} + 0.475 \\
\frac{235\ \mathrm{GeV}}{m_{4j}} < \Delta R_{bb}^{h_2} < \frac{875\ \mathrm{GeV}}{m_{4j}} + 0.35
\end{aligned} \tag{5.1}$$

if $m_{4b} < 1250$ GeV over the possible bottom jet pairings and

$$\begin{aligned}
0 < \Delta R_{bb}^{h_1} < 1 \\
0 < \Delta R_{bb}^{h_2} < 1
\end{aligned} \tag{5.2}$$

if $m_{4b} > 1250$ GeV. For those events having more than one instance of the partitions passing the above requirements, the one with the minimum D_{hh} defined as

$$D_{hh} = \left| m_{h_1} - \frac{120}{110} m_{h_2} \right| \left(1 + \frac{120^2}{110^2} \right)^{-1/2} \quad , \tag{5.3}$$

is chosen to be the Higgs candidate. In contrast to the above-mentioned analyses, we do not drop the event if no partitions pass the ΔR criterion and use the minimum D_{hh} pair over all possible pairs in such events to specify the possible Higgs candidates. These possible Higgs candidates segregate the four bottom jets into two classes of reconstructed objects: H_1 and H_2. Any other jet in the reconstructed event, including additional b-jets, is classified under a single jet class J.

After segregating the reconstructed jets into the three classes, we construct a complete graph with edges connecting all distinct objects, i.e. without self-loops. The input node representations consist of the Lorentz four-vector $\mathbf{p}_i^{(0)}$, and the longitudinal scalar node representation[10]

$$\mathbf{h}_i^{(0)} = (\phi_i, \log p_i^t, \log m_i^t, b_i, \log m_i) \quad ,$$

consists of the jets' azimuthal angle ϕ_i , transverse momentum p_i^t , transverse mass $m_i^t = \sqrt{E_i^2 - p_z^2}$, b-tagging information $b_i \in \{-1, 1\}$, and mass m_i. We set $b_i = 1$ for a b-tagged jet. Each edge has a longitudinal scalar edge-representation

$$\mathbf{e}_{ij}^{(0)} = \left(\log p_{ij}^t, \log\left(p_i^t p_j^t \right), \Delta\eta_{ij}, \Delta\phi_{ij}, \Delta R_{ij} \right) \quad ,$$

where p_{ij}^t is the transverse momentum of $\mathbf{p}_i + \mathbf{p}_j$, $\Delta\eta_{ij}$ is the difference in pseudorapidity, $\Delta\phi_{ij}$ the azimuthal separation and $\Delta R_{ij} = \sqrt{\Delta\eta_{ij}^2 + \Delta\phi_{ij}^2}$. For the O(1,3) network, we consider only the fully Lorentz invariant node features[11], $\tilde{\mathbf{h}}_i^{(0)} = (b_i, \log m_i)$ and do not provide any additional edge feature since the message passing operation automatically evaluates the relevant edge invariants. While one could argue that the Lorentz invariant model has less information supplied, this is a mandatory requirement: larger group invariances assume that information contained within the separate orbits of its proper sub-groups are the same and therefore not relevant. Thus, $\mathbf{e}_{ij}^{(0)}$ being O(1,1) invariant but not O (1,3) invariant cannot be used as scalar edge features in an O(1,3) invariant model[12]**.** The classes H_1 and H_2 undergo a mean global mean readout either separately (for $S_2 \times S_2$ group) or together (for S_4 group), along with any additional jets which are uniformly given a weight of $w_{\alpha \leftarrow J} = 0.001$ for $\alpha \in \{H_1, H_2, H_1 \cup H_2\}$.

5.3. Network analysis

Looking into graph-based architectures, a segregation of the reconstructed objects allows for a heterogeneous graph message-passing operation, which preserves all symmetries of the likelihood ratio. On the other hand, we want to learn the kinematic correlations between the different classes efficiently. This can be achieved in the heterogeneous set-up with multiple copies of learnable functions for the node and edge type combinatorics. Since this scales factorially, if we consider edge directions, we choose the simpler homogeneous message-passing operation with the learnable functions shared between all nodes and edges. All network analyses uses PYTORCH-GEOMETRIC (V2.5.0) [84] and PYTORCH (V2.0.0) [85]. The training was done using two NVIDIA A100 GPUs using the inbuilt `DistributedDataParallel` module with equally divided batches. We consider three base architectures with different message-passing heads:

1. O(1,1)-S : a scalar-only longitudinal boost invariant message passing head. This is essentially a `EdgeConv` [86] network that takes $\mathbf{h}_i^{(0)}$ and $\mathbf{e}_{ij}^{(0)}$ as inputs.
2. O(1,1)-SV : a scalar-and-vector update longitudinal boost equivariant message passing head
3. O(1,3) : a Lorentz Group Equivariant Block [57] modified so that Φ_e takes $|\mathbf{p}_i^{(l)} + \mathbf{p}_j^{(l)}|^2$ instead of their choice of momentum difference squared inputs and no[13] Φ_h.

[10] A statistically negligible amount of events in the dataset had jets with zero mass and were excluded from all numerical analyses.

[11] Strictly speaking, the b-tagging information b_i being dependent on reconstruction is not Lorentz invariant. However, as is usually done in most applications, we assume that it reflects the true flavour of the underlying primaeval parton.

[12] In general, a proper subgroup has more invariant quantities as a result of the inverted set-inclusion relationship (see figure 3).

[13] We did not find any noticeable performance difference with the addition of Φ_h.

Table 1. The best AUC score out of all experiments conducted for each base architecture on the full dataset of each signal scenario. The mean and standard deviation is taken over ten training instances from random weight initialisation. For comparison, we show the relevant figures for Spa-Net. The data consists of 1 M train and 100k test samples for the resonant case, while for the non-resonant case it has 180k train, and 18k test samples.

Arch.	Signal	Disc. Sym.	Num. Param.	AUC
O(1,1)-S	Resonant	S_4	293k	0.9652 ± 0.0002
	Non-resonant	S_4	22k	0.9165 ± 0.0005
O(1,1)-SV	Resonant	S_4	458k	0.9653 ± 0.0001
	Non-resonant	S_4	33k	0.9169 ± 0.0009
O(1,3)	Resonant	$S_2 \times S_2$	743k	0.9550 ± 0.0016
	Non-resonant	$S_2 \times S_2$	743k	0.9000 ± 0.0009
Spa-Net (Reference [41])	Resonant	S_n	37.9 M	0.961 ± 0.001
	Non-resonant	S_n	541k	0.911 ± 0.001

Similar to [57], all inner products and norms go through the function $R(x) = \mathrm{sign}(x)\log(|x|+1)$, so that the gradient descent is stable for the non-compact metric signature. Each model has a wide variant of 256, 128, and 64 updated scalar-node dimensions and a narrow variant of 64, 32, and 16 updated scalar-node representations. All MLPs have two hidden layers with the same dimensions as their respective scalar update dimensions with `ReLU` activation in the hidden layers. The output layers have `Linear` activations except for $\Phi_x^{(l)}$, which has a `Sigmoid` activation function. The message functions $\Phi_e^{(l)}$ in O(1,1)-SV and O(1,3) models take additional edge scalar edge features evaluated at each stage l. The O(1,1)-S model consists of only the Φ_e function in each stage, which takes the scalar representation $\mathbf{h}_i^{(l)}$ and $\mathbf{h}_j^{(l)}$ without any additional edge features beyond the initial input operation. Additionally, $\Phi_e^{(l)}$ in O(1,1)-S and O(1,1)-SV evaluates the `EdgeConv` input $\mathbf{h}^{(l)} \oplus \mathbf{h}_j^{(l)} - \mathbf{h}_i^{(l)}$ from the scalar node representations in each stage l of the message passing head. All three base architectures have a mean scalar node readout. For all networks, the updated scalar node-representations $\mathbf{h}_i^{(l)}$, for $l > 0$ undergoes a global mean readout which is either $S_2 \times S_2$ invariant or S_4 invariant depending on the discrete symmetry of the network. Consequently, the final message passing operation for O(1,1)-SV and O(1,3) does not have a vector update operation. The respective node representations and the permutation symmetry determine the inputs to the classifier head. The classifier MLP has two hidden layers of 64 nodes and `ReLU` activation for the wide message passing head, while the ones with narrow message passing heads have 32 nodes instead. With a single logit output, the networks are trained with `torch.nn.functional.binary_cross_entropy_with_logits` loss function.

Counting the wide and narrow variants of the message-passing heads and the global readout symmetry, we have four network architectures for each base architecture. These four instances are trained on two training sizes for the resonant and non-resonant cases: the full dataset and a reduced set containing 100k samples for the resonant case and 10k for the non-resonant case. We use the test dataset as the validation set during training and utilise the complete training dataset for the first case[14]. For the resonant analysis the full dataset consists of 1 M training and 100k validation/testing samples, while for the non-resonant case it consists of 180k training and 18k validation/testing samples. Networks in each experiment are trained ten times after random weight initialisation with the `Adam` [87] optimiser with an initial learning rate of 0.001 and a batch size of 128 samples-per-batch. A decay-on-plateau condition decays the learning rate if the validation loss has not improved for five epochs by a factor of 0.1 until it reaches 10^{-8}. The training runs for a maximum of one hundred epochs and is stopped if the validation loss has not decreased for twenty epochs.

5.4. Performance

For each training experiment, we evaluate the area under the curve (AUC) under the receiver operator characteristics curve over each training instance from which we form various summary statistics of the performance metrics. Here, we report the main findings while all results are tabulated in appendix. The best AUC score for each base architecture over the two datasets, along with the details of the specific architecture, is shown in table 1. The figures of Spa-Net from [41] are also shown for comparison. Lorentz invariant classification fares poorly in either scenario compared to O(1,1)-S and O(1,1)-SV and can not match the

[14] The difference of 50k and 9k training samples from [41] for the resonant and non-resonant cases, respectively, is not a major difference for the quoted results as network performance generally scales logarithmically with training size. Concretely, the smallest O(1,1)-S network with an S_4 invariant global readout with 22k trainable parameters reached an AUC of 0.9632 on the test dataset with 600k training samples on the resonant signal dataset.

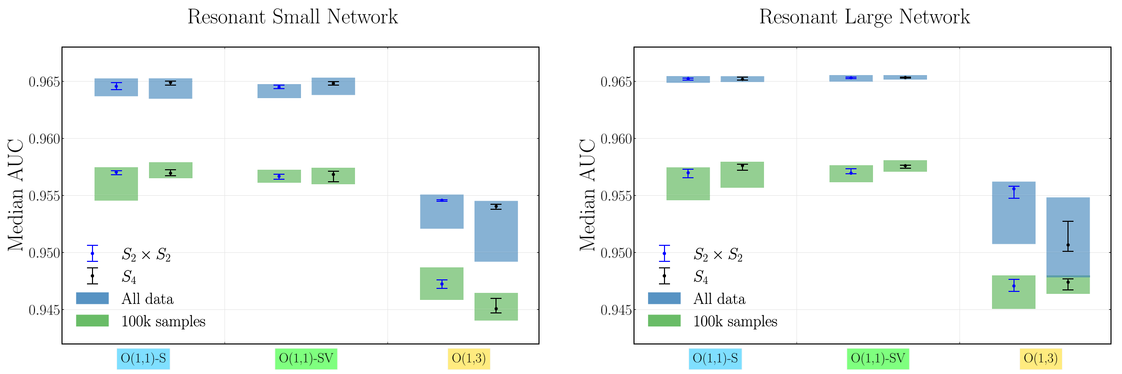

Figure 5. Median AUCs for the resonant signal for all architectures and training sizes. The shaded region shows the range, and the error bars denote the lower and upper quartiles.

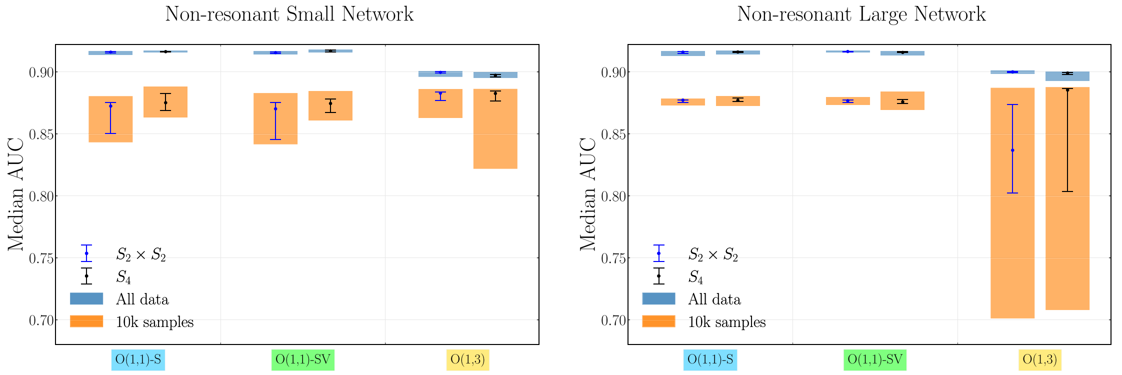

Figure 6. Median AUCs for the non-resonant signal for all architectures and training sizes. The shaded region shows the range, and the error bars denote the lower and upper quartiles.

Spa-Net results, which do not assume any continuous group equivariance. The correct continuous group symmetric design of O(1,1)-S and O(1,1)-SV outperforms Spa-Net with an order of magnitude reduction in trainable parameters. This is all the more impressive considering that the numerical experiments for the Spa-Net based analysis conducted a hyperparameter scan. Additionally, the low parameter-size networks perform nominally better for the non-resonant scenario than the highly parametrised ones (see table 4 in appendix). This could be due to the lower statistics of the training data in the non-resonant dataset, where a larger model size performs better with more training statistics. On the contrary, the incorrect invariance in O(1,3) has the wider network performing better than the smaller network, even with the limited training statistics of the non-resonant training dataset. This may be due to the assumption of an incorrect exact invariance in the domain and the presence of noise in the data, which requires more model flexibility to circumvent the exact symmetric design of the architecture. This intuition could also help explain the better performance of the smaller $S_2 \times S_2$ permutation symmetry for O(1,3) for either signal scenario, where the larger S_4 symmetry comparatively over-constrains the fibres of the target function.

The median of the AUC and its lower and upper quartiles as error bars for each base architecture and training data size are plotted in figure 5 for the resonant scenario and figure 6 for the non-resonant one. For both signal scenarios, the choice of the discrete symmetry group has nominal differences in the median values for the O(1,1)-S and O(1,1)-SV architectures that have the correct continuous invariance. The smaller networks have a larger range for both signal scenarios, suggesting a trade-off between training stability and parameter complexity. Similarly, the low training statistics cases have larger ranges for the correct continuous equivariance than the full dataset training. The situation is mostly reversed in the case of O(1,3) networks, where the continuous symmetry is incorrect. As seen above, the smaller group $S_2 \times S_2$ has better overall median AUCs than the larger S_4 symmetric readouts, barring the non-resonant large-network experiment in the low training sample scenario. However, in this situation, both networks have very erratic behaviour over the ten training instances, as can be seen by the large range and extreme position of the median values. Interestingly, all networks with the correct continuous symmetry, regardless of the network size and discrete symmetry group, outperform Spa-Net on the full dataset.

Table 2. The mean AUC, R_{30}, and R_{50} with only O(1,3) scalars and four vectors supplied to the O(1,1)-SV model taking the $S_2 \times S_2$ permutation group invariant graph readout. Without the additional O(1,1)-scalar information, the values indicate the suitability of O(1,1) invariance over O(1,3) invariance for signal-background classification tasks.

Signal	Num. Param.	AUC	R_{30}	R_{50}
Resonant	485k	0.9651 ± 0.0002	2047 ± 128	368 ± 12
	36k	0.9640 ± 0.0004	2150 ± 177	359 ± 11
Non-resonant	485k	0.9167 ± 0.0007	281 ± 37	32 ± 4
	36k	0.9162 ± 0.0013	265 ± 24	53 ± 6

To verify that the increase in performance is not solely due the additional O(1,1) invariant information provided as inputs but due to the architecture itself, we consider the O(1,1)-SV architecture with same inputs as supplied to the O(1,3) model and $S_2 \times S_2$ discrete group. Keeping the same hyper-parameters and training environment, both variants of the network are trained on the complete resonant and non-resonant dataset ten times from random initialisation. The AUC and inverse of the background acceptance at 30% and 50% signal acceptance (R_{30} and R_{50} respectively) are shown in table 2**.** The values indicate that the better performance is not due to the extra inputs supplied but due to suitability of O(1,1) invariance for the particular task and the O(1,1) invariant information that the network explicitly constructs.

6. Conclusions

In this work, we have established a novel connection between the MEM and equivariant neural network architecture design, demonstrating how MEM-inspired symmetries can guide the development of deep learning models for high-energy physics analysis. By incorporating a suitable subgroup of the physical Lorentz and permutation invariances directly into the architecture, we have shown that neural networks can achieve improved performance in classification tasks while maintaining lower parameter complexity.

Our approach uses the inherent symmetry properties embedded in fixed-order differential cross-sections and exploits the optimality of group-equivariant functions for binary classification. We demonstrated that designing neural networks with MEM-inspired equivariant updates results in architectures that better capture the kinematic correlations of events, especially for complex final states, such as di-Higgs production decaying to four bottom jets. The longitudinal boost-equivariant message-passing network proposed in this work provides a concrete example of how these principles can be applied to practical physics problems, yielding state-of-the-art performance on benchmark datasets. Moreover, the analysis reveals that smaller group invariance approximations can effectively generalise to larger symmetries during training, while larger group invariance constraints might overlook subtle details in the data.

Our findings open several avenues for future research. First, extending these principles to higher-dimensional final states and more complex processes, such as multi-jet events or processes with additional intermediate resonances, could further elucidate the benefits of MEM-inspired equivariant architectures. Additionally, integrating such symmetric architecture designs with other advanced deep learning techniques, such as transformers or attention mechanisms, could offer even more powerful tools for particle physics analysis. Furthermore, applying this framework to multi-class classification problems in physics searches, where different classes exhibit distinct symmetry properties, could improve LHC's sensitivity to new physics.

Thus, this study demonstrates that integrating MEM with equivariant deep learning techniques can significantly enhance neural networks' capabilities in high-energy physics. By grounding the architecture design in physical principles, we can improve model interpretability, reduce computational requirements, and potentially uncover new physics beyond the Standard Model.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://zenodo.org/records/10952296.

Acknowledgment

This work is carried out under the support of STFC under grant ST/X003167/1. We thank Matteo Marcoli for the helpful discussions during the project. We also thank the authors of [41] for making their dataset publicly available.

Appendix. Additional results of network analysis

This appendix shows the results of all numerical experiments conducted on the resonant and non-resonant datasets. Including the AUC, we show the R_{30} and R_{50} metrics defined as the inverse of the background acceptance (false positive rate) at 30 and 50 per cent signal acceptances (true positive rate), respectively. These are shown for the resonant and non-resonant signals in tables 3 and 4, respectively. One can confirm that the correct continuous group symmetries, regardless of the network size and permutation symmetries, outperform SPA-NET on the full dataset.

Table 3. The mean AUC, R_{30}, and R_{50} for all experiments on the resonant dataset.

Arch.	Train. Size	Disc. Sym.	Num. Param.	AUC	R_{30}	R_{50}
O(1,1)-S	All	S_4	293k	0.9652 ± 0.0002	2135 ± 303	375 ± 14
			22k	0.9647 ± 0.0005	2000 ± 139	362 ± 14
		$S_2 \times S_2$	322k	0.9652 ± 0.0002	2037 ± 269	363 ± 18
			26k	0.9646 ± 0.0005	2000 ± 227	361 ± 7
	100k	S_4	293k	0.9572 ± 0.0008	1274 ± 129	257 ± 22
			22k	0.9570 ± 0.0004	1357 ± 115	266 ± 12
		$S_2 \times S_2$	322k	0.9567 ± 0.0009	1154 ± 141	259 ± 17
			26k	0.9567 ± 0.0009	1287 ± 160	262 ± 11
O(1,1)-SV	All	S_4	458k	0.9653 ± 0.0001	2137 ± 260	370 ± 12
			33k	0.9648 ± 0.0004	2116 ± 321	356 ± 15
		$S_2 \times S_2$	487k	0.9653 ± 0.0002	2064 ± 291	357 ± 10
			36k	0.9644 ± 0.0003	2060 ± 171	367 ± 9
	100k	S_4	458k	0.9575 ± 0.0003	1281 ± 161	259 ± 13
			33k	0.9567 ± 0.0005	1394 ± 123	263 ± 10
		$S_2 \times S_2$	487k	0.9570 ± 0.0004	1149 ± 99	262 ± 10
			36k	0.9567 ± 0.0004	1343 ± 103	259 ± 13
O(1,3)	All	S_4	715k	0.9512 ± 0.0024	784 ± 100	156 ± 14
			48k	0.9536 ± 0.0016	886 ± 82	169 ± 8
		$S_2 \times S_2$	743k	0.9550 ± 0.0016	865 ± 67	180 ± 7
			52k	0.9542 ± 0.0011	864 ± 66	171 ± 7
	100k	S_4	715k	0.9472 ± 0.0006	643 ± 30	141 ± 5
			48k	0.9453 ± 0.0008	599 ± 33	135 ± 3
		$S_2 \times S_2$	743k	0.9470 ± 0.0009	618 ± 50	137 ± 4
			52k	0.9473±0.0008	630 ± 27	141 ± 4

Table 4. The mean AUC, R_{30}, and R_{50} for all experiments on the non-resonant dataset.

Arch.	Train. Size	Disc. Sym.	Num. Param.	AUC	R_{30}	R_{50}
O(1,1)-S	All	S_4	293k	0.9160±0.0009	236±20	49±2
			22k	0.9165±0.0005	256±25	50±2
		$S_2 \times S_2$	322k	0.9155±0.0013	231±21	48±2
			26k	0.9158±0.001	248±14	49±3
	10k	S_4	293k	0.8772±0.0023	82±6	22±1
			22k	0.8754±0.0087	83±19	22±4
		$S_2 \times S_2$	322k	0.8764±0.0018	78±6	22±1
			26k	0.8645±0.0148	67±16	18±4
O(1,1)-SV	All	S_4	458k	0.9157±0.001	227±18	50±2
			33k	0.9169±0.0009	272±29	50±1
		$S_2 \times S_2$	487k	0.9164±0.0004	238±28	50±2
			36k	0.9155±0.0009	243±21	49±2
	10k	S_4	458k	0.8765±0.004	81±9	22±1
			33k	0.8729±0.0076	81±20	20±3
		$S_2 \times S_2$	487k	0.8766±0.0018	81±6	22±1
			36k	0.863±0.0167	65±18	18±5
O(1,3)	All	S_4	715k	0.8981±0.0022	123±8	31±1
			48k	0.8972±0.0014	125±4	30±1
		$S_2 \times S_2$	743k	0.9000±0.0009	130±7	30±1
			52k	0.8995±0.0014	129±10	31±1
	10k	S_4	715k	0.8437±0.0662	82±47	19±9
			48k	0.8737±0.0209	92±26	22±5
		$S_2 \times S_2$	743k	0.8219±0.0663	50±32	14±8
			52k	0.8793±0.0075	93±15	23±3

ORCID iDs

Daniel Maître https://orcid.org/0000-0003-0414-9497
Vishal S Ngairangbam https://orcid.org/0000-0002-7143-715X
Michael Spannowsky https://orcid.org/0000-0002-8362-0576

References

[1] Kondo K 1988 Dynamical likelihood method for reconstruction of events with missing momentum. 1: method and toy models *J. Phys. Soc. Japan* **57** 4126

[2] Kondo K 1991 Dynamical likelihood method for reconstruction of events with missing momentum. 2: Mass spectra for 2 —> 2 processes *J. Phys. Soc. Japan* **60** 836

[3] D0 collaboration 2015 Precision measurement of the top-quark mass in lepton+jets final states *Phys. Rev.* D **91** 112003

[4] Alwall J, Freitas A and Mattelaer O 2011 The matrix element method and QCD radiation *Phys. Rev.* D **83** 074010

[5] Soper D E and Spannowsky M 2011 Finding physics signals with shower deconstruction *Phys. Rev.* D **84** 074002

[6] Andersen J R, Englert C and Spannowsky M 2013 Extracting precise Higgs couplings by using the matrix element method *Phys. Rev.* D **87** 015019

[7] Campbell J M, Giele W T and Williams C 2012 The matrix element method at next-to-leading order *J. High Energy Phys.* JHEP11(2012)043

[8] Debnath D, Gainer J S and Matchev K T 2015 Discoveries far from the lamppost with matrix elements and ranking *Phys. Lett.* B **743** 1

[9] Soper D E and Spannowsky M 2014 Finding physics signals with event deconstruction *Phys. Rev.* D **89** 094005

[10] Martini T and Uwer P 2015 Extending the Matrix Element Method beyond the Born approximation: Calculating event weights at next-to-leading order accuracy *J. High Energy Phys.* JHEP09(2015)083

[11] Ferreira de Lima D, Petrov P, Soper D and Spannowsky M 2017 Quark-Gluon tagging with shower deconstruction: unearthing dark matter and Higgs couplings *Phys. Rev.* D **95** 034001

[12] Prestel S and Spannowsky M 2019 HYTREES: combining matrix elements and parton shower for hypothesis testing *Eur. Phys. J.* C **79** 546

[13] Martini T, Kraus M, Peitzsch S and Uwer P 2020 The matrix element method as a tool for precision and accuracy *PoS EPS-HEP2019* P 673

[14] Bury F and Delaere C 2021 Matrix element regression with deep neural networks-Breaking the CPU barrier *J. High Energy Phys.* JHEP04(2021)020

[15] Butter A, Heimel T, Martini T, Peitzsch S and Plehn T 2023 Two invertible networks for the matrix element method *SciPost Phys.* **15** 094

[16] Grossi M, Incudini M, Pellen M and Pelliccioli G 2023 Amplitude-assisted tagging of longitudinally polarised bosons using wide neural networks *Eur. Phys. J.* C **83** 759

[17] Heimel T, Huetsch N, Winterhalder R, Plehn T and Butter A 2023 Precision-machine learning for the matrix element method (arXiv:2310.07752)

[18] de Oliveira L, Kagan M, Mackey L, Nachman B and Schwartzman A 2016 Jet-images-deep learning edn *J. High Energy Phys.* JHEP07(2016)069

[19] Cranmer K, Pavez J and Louppe G 2015 Approximating likelihood ratios with calibrated discriminative classifiers (arXiv:1506.02169)

[20] Dery L M, Nachman B, Rubbo F and Schwartzman A 2017 Weakly supervised classification in high energy physics *J. High Energy Phys.* JHEP05(2017)145

[21] Metodiev E M, Nachman B and Thaler J 2017 Classification without labels: learning from mixed samples in high energy physics *J. High Energy Phys.* JHEP10(2017)174

[22] Larkoski A J, Moult I and Nachman B 2020 Jet substructure at the large hadron collider: a review of recent advances in theory and machine learning *Phys. Rept.* **841** 1

[23] Komiske P T, Metodiev E M and Thaler J 2019 Energy flow networks: deep sets for particle jets *J. High Energy Phys.* JHEP01(2019)121

[24] Brehmer J, Cranmer K, Louppe G and Pavez J 2018 A guide to constraining effective field theories with machine learning *Phys. Rev.* D **98** 052004

[25] Guest D, Cranmer K and Whiteson D 2018 Deep learning and its application to LHC physics *Ann. Rev. Nucl. Part. Sci.* **68** 161

[26] Qu H and Gouskos L 2020 ParticleNet: jet tagging via particle clouds *Phys. Rev.* D **101** 056019

[27] Brehmer J, Kling F, Espejo I and Cranmer K 2020 MadMiner: machine learning-based inference for particle physics *Comput. Softw. Big Sci.* **4** 3

[28] Butter A *et al* 2019 The machine learning landscape of top taggers *SciPost Phys.* **7** 014

[29] Karagiorgi G, Kasieczka G, Kravitz S, Nachman B and Shih D 2021 Machine learning in the search for new fundamental physics (arXiv:2112.03769)

[30] Plehn T, Butter A, Dillon B, Heimel T, Krause C and Winterhalder R 2022 Modern machine learning for LHC physicists (arXiv:2211.01421)

[31] Onyisi P, Shen D and Thaler J 2023 Comparing point cloud strategies for collider event classification *Phys. Rev.* D **108** 012001

[32] Calafiura P, Rousseau D and Terao K 2022 *Artificial Intelligence for High Energy Physics* (World Scientific)

[33] Brehmer J 2021 Simulation-based inference in particle physics *Nat. Rev. Phys.* **3** 305

[34] Maître D and Truong H 2021 A factorisation-aware Matrix element emulator *J. High Energy Phys.* JHEP11(2021)066

[35] DeZoort G, Battaglia P W, Biscarat C and Vlimant J-R 2023 Graph neural networks at the Large Hadron Collider *Nat. Rev. Phys.* **5** 281

[36] Ngairangbam V S and Spannowsky M 2024 Interpretable deep learning models for the inference and classification of LHC data *J. High Energy Phys.* JHEP05(2024)004

[37] Bhardwaj A, Englert C, Naskar W, Ngairangbam V S and Spannowsky M 2024 Equivariant, safe and sensitive-graph networks for new physics *J. High Energy Phys.* JHEP07(2024)245

[38] Fenton M J, Shmakov A, Ho T-W, Hsu S-C, Whiteson D and Baldi P 2022 Permutationless many-jet event reconstruction with symmetry preserving attention networks *Phys. Rev.* D **105** 112008

[39] Shmakov A, Fenton M J, Ho T-W, Hsu S-C, Whiteson D and Baldi P 2022 SPANet: generalized permutationless set assignment for particle physics using symmetry preserving attention *SciPost Phys.* **12** 178

[40] Fenton M J *et al* 2024 Reconstruction of unstable heavy particles using deep symmetry-preserving attention networks *Commun. Phys.* **7** 139

[41] Chiang C-W, Hsieh F-Y, Hsu S-C and Low I 2024 Deep learning to improve the sensitivity of Di-Higgs searches in the 4b channel *J. High Energy Phys.* JHEP09(2024)139

[42] Maître D and Santos-Mateos R 2023 Multi-variable integration with a neural network *J. High Energy Phys.* JHEP03(2023)221

[43] Rizvi S, Pettee M and Nachman B 2024 Learning likelihood ratios with neural network classifiers *J. High Energy Phys.* JHEP02(2024)136

[44] Janßen T, Maître D, Schumann S, Siegert F and Truong H 2023 Unweighting multijet event generation using factorisation-aware neural networks *SciPost Phys.* **15** 107

[45] Bahl H, Bresó V, De Crescenzo G and Plehn T 2024 Advancing tools for simulation-based inference (arXiv:2410.07315)

[46] Bhardwaj A, Konar P and Ngairangbam V 2024 Foundations of automatic feature extraction at lhc–point clouds and graphs *Eur. Phys. J. Spec. Top.* **233** 2619–40

[47] Bronstein M M, Bruna J, LeCun Y, Szlam A and Vandergheynst P 2017 Geometric deep learning: going beyond Euclidean data *IEEE Signal Process. Mag.* **34** 18

[48] Cohen T and Welling M 2016 Group equivariant convolutional networks *Proc. 33rd Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) vol 48, ed M F Balcan and K Q Weinberger (PMLR) pp 2990–9

[49] Kondor R and Trivedi S 2018 On the generalization of equivariance and convolution in neural networks to the action of compact groups *Proc. 35th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) vol 80, ed J Dy and A Krause (PMLR) pp 2747–55

[50] Cohen T, Weiler M, Kicanaoglu B and Welling M 2019 Gauge equivariant convolutional networks and the icosahedral CNN *Proc. 36th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) vol 97, ed K Chaudhuri and R Salakhutdinov (PMLR) pp 1321–30

[51] Bronstein M M, Bruna J, Cohen T and Veličković P 2021 Geometric deep learning: grids, groups, graphs, geodesics, and gauges (arXiv:2104.13478)

[52] Satorras V G, Hoogeboom E and Welling M 2021 E(n) equivariant graph neural networks *Proc. 38th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) vol 139, ed M Meila and T Zhang (PMLR) pp 9323–32

[53] Villar S, Hogg D W, Storey-Fisher K, Yao W and Blum-Smith B 2021 Scalars are universal: equivariant machine learning, structured like classical physics *Advances in Neural Information Processing Systems* ed A Beygelzimer, Y Dauphin, P Liang and J W Vaughan p 2021

[54] Brehmer J, de Haan P, Behrends S and Cohen T S 2023 Geometric algebra transformer *Advances in Neural Information Processing Systems* vol 36, ed A Oh, T Naumann, A Globerson, K Saenko, M Hardt and S Levine (Curran Associates, Inc.) pp 35472–96

[55] Bogatskiy A, Anderson B, Offermann J, Roussi M, Miller D and Kondor R 2020 Lorentz group equivariant neural network for particle physics *Proc. 37th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) vol 119, ed A Singh (PMLR) pp 992–1002

[56] Bogatskiy A *et al* 2022 Symmetry group equivariant architectures for physics *Snowmass 2021*

[57] Gong S *et al* 2022 An efficient Lorentz equivariant graph neural network for jet tagging *J. High Energy Phys.* JHEP07(2022)030

[58] Li C *et al* 2024 Does Lorentz-symmetric design boost network performance in jet physics? *Phys. Rev.* D **109** 056003

[59] Bogatskiy A, Hoffman T, Miller D W, Offermann J T and Liu X 2024 Explainable equivariant neural networks for particle physics: PELICAN *J. High Energy Phys.* JHEP03(2024)113

[60] Hao Z, Kansal R, Duarte J and Chernyavskaya N 2023 Lorentz group equivariant autoencoders *Eur. Phys. J.* C **83** 485

[61] Thais S and Murnane D 2023 Equivariance is not all you need: characterizing the utility of equivariant graph neural networks for particle physics tasks (arXiv:2311.03094)

[62] Spinner J, Bresó V, de Haan P, Plehn T, Thaler J and Brehmer J 2024 Lorentz-equivariant geometric algebra transformers for high-energy physics (arXiv:2405.14806)

[63] Maître D, Ngairangbam V S and Spannowsky M Towards an understanding of inductive biases for signal searches at LHC

[64] Neyman J and Pearson E S 1933 On the problem of the most efficient tests of statistical hypotheses *Phil. Trans. R. Soc.* A **231** 289

[65] Ngairangbam V S and Spannowsky M 204 Optimal symmetries in binary classification (arXiv:2408.08823)

[66] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones J, Gomez AN,Gomez AN, Kaiser K Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* vol 30, ed I Guyon *et al*

[67] Qu H, Li C and Qian S 2022 Particle transformer for jet tagging *Proc. 39th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) ed K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu and S Sabato vol 162 (PMLR) pp 18281–92

[68] Wigner E 1939 On unitary representations of the inhomogeneous Lorentz group *Ann. Math.* **40** 149

[69] Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov R R and Smola A J *et al* 2017 Deep sets *Advances in Neural Information Processing Systems* vol 30, ed I Guyon (Curran Associates, Inc.)

[70] Wagstaff E, Fuchs F, Engelcke M, Posner I and Osborne M A 2019 On the limitations of representing functions on sets *Proc. 36th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) vol 97, ed K Chaudhuri and R Salakhutdinov (PMLR) pp 6487–94

[71] Berdine D, Kauer N and Rainwater D 2007 Breakdown of the narrow width approximation for new physics *Phys. Rev. Lett.* **99** 111601

[72] Bianchini M and Scarselli F 2014 On the complexity of neural network classifiers: a comparison between shallow and deep architectures *IEEE Trans. Neural Netw. Learn. Syst.* **25** 1553–65

[73] Naitzat G, Zhitnikov A and Lim L-H 2020 Topology of deep neural networks *J. Mach. Learn. Res.* **21** 1

[74] Hsieh F Y 2024 Dataset for 'Deep Learning to Improve the Sensitivity of Di-Higgs Searches in the 4b Channel (https://doi.org/10.5281/zenodo.10952296)

[75] Alwall J *et al* 2014 The automated computation of tree-level and next-to-leading order differential cross sections and their matching to parton shower simulations *J. High Energy Phys.* JHEP07(2014)079

[76] Bierlich C *et al* 2022 A comprehensive guide to the physics and usage of PYTHIA 8.3 *SciPost Phys. Codeb.* **2022** 8

[77] DELPHES 3 collaboration and DELPHES 3 2014 A modular framework for fast simulation of a generic collider experiment *J. High Energy Phys.* JHEP02(2014)057

[78] Cacciari M, Salam G P and Soyez G 2012 FastJet user manual *Eur. Phys. J.* C **72** 1896

[79] Cacciari M, Salam G P and Soyez G 2008 The anti-k_t jet clustering algorithm *J. High Energy Phys.* JHEP04(2008)063

[80] ATLAS collaboration 2016 Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run

[81] ATLAS collaboration 2016 Performance of b-jet identification in the ATLAS experiment *JINST* **11** 04008

[82] ATLAS collaboration 2023 ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset *Eur. Phys. J.* C **83** 681

[83] ATLAS collaboration 2019 Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector *J. High Energy Phys.* JHEP01(2019)030

[84] Fey M and Lenssen J E 2019 Fast graph representation learning with PyTorch Geometric *ICLR Workshop on Representation Learning on Graphs and Manifolds*

[85] Paszke A *et al* 2019 *Pytorch: an Imperative Style, High-Performance Deep Learning Library* (Curran Associates Inc.)

[86] Wang Y, Sun Y, Liu Z, Sarma S E, Bronstein M M and Solomon J M 2019 Dynamic graph cnn for learning on point clouds *ACM Trans. Graph.* **38** 1–12

[87] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *3rd Int. Conf. on Learning Representations, ICLR 2015, (Conf. Track Proc.) (San Diego, CA, USA, 7 May–9 May, 2015)*, ed Y Bengio and Y LeCun