Eur. Phys. J. C (2026) 86:1211
https://doi.org/10.1140/epjc/s10052-026-16489-z

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Computing, Software and Data Science

Quantum-inspired tensor network autoencoders for anomaly detection: a MERA-based approach

Emre Gurkanli[1,a], **Michael Spannowsky**[2,3,b]

[1] Department of Physics, Sinop University, Sinop, Turkey
[2] Institute for Theoretical Physics, Karlsruhe Institute of Technology (KIT), Campus Süd, 76128 Karlsruhe, Germany
[3] Institute for Quantum Materials and Technologies, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

Received: 13 April 2026 / Accepted: 7 October 2026
© The Author(s) 2026

Abstract We investigate whether a multiscale tensor-network architecture provides a useful inductive bias for reconstruction-based anomaly detection in collider jets. Jets are produced by a branching cascade, so their internal structure is naturally organised across angular and momentum scales. This motivates an autoencoder that compresses information hierarchically and can reorganise short-range correlations before coarse-graining. Guided by this picture, we formulate a Multiscale Entanglement Renormalization Ansatz (MERA)-inspired autoencoder acting directly on ordered jet constituents. To our knowledge, a MERA-inspired autoencoder has not previously been proposed, and this architecture has not been explored in collider anomaly detection. We compare this architecture to a dense autoencoder, the corresponding tree-tensor-network limit, and standard classical baselines within a background-only reconstruction framework. The paper is organised around two main questions: whether locality-aware hierarchical compression is genuinely supported by the data, and whether the disentangling layers of MERA contribute beyond a simpler tree hierarchy. To address these questions, we combine benchmark comparisons with a training-free local-compressibility diagnostic and a direct identity-disentangler ablation. In the nominal benchmark, the MERA-inspired model attains a higher mean AUC than the dense autoencoder at all three latent dimensions while using roughly one-third of its trainable parameters. The locality-preserving multiscale structure is well matched to jet data, and, at the strongest compression bottleneck, the trained disentanglers move substantially away from the identity and are accompanied by a positive mean performance shift. Overall, the study supports locality-aware hierarchical compression as a useful and parameter-efficient inductive bias for reconstruction-based jet anomaly detection in this benchmark setting.

Contents

1 Introduction . . . . . .
2 Constructing a MERA-inspired anomaly detector . . .
3 Benchmarking and data preprocessing . . . . . .
4 Model implementations and reference baselines . . . .
5 Results and discussion . . . . . .
6 Summary and conclusions . . . . . .
References . . . . . .

1 Introduction

The Standard Model (SM) gives an excellent description of collider data, but it does not explain several basic observations, including dark matter, neutrino masses, and the baryon asymmetry of the Universe. The search for physics beyond the SM, therefore, remains a central goal of the LHC programme. In practice, however, most searches are optimised for specific benchmark signatures. This is a powerful strategy when the signal is known in advance, but it can reduce sensitivity to unexpected phenomena that do not fit a standard search template. For this reason, there is increasing interest in search strategies that rely less strongly on a detailed signal hypothesis and instead look more broadly for deviations from the dominant SM background [1,2].

Anomaly detection provides one such framework. The basic idea is to learn what typical background events look like and then assign a large anomaly score to events that are poorly described by that learned background model [1]. In

[a] e-mail: egurkanli@sinop.edu.tr (corresponding author)
[b] e-mail: michael.spannowsky@kit.edu (corresponding author)

Published online: 22 October 2026

collider physics, this programme has grown rapidly in recent years, supported by community benchmarks and systematic comparisons that clarified both the promise and the limitations of model-agnostic searches [2,3]. A recurring lesson is that the choice of data representation and the built-in structure of the model are crucial: an anomaly detector should be flexible enough to respond to a wide class of signals, while remaining robust against ordinary fluctuations of the background.

Among the many possible approaches, reconstruction-based methods are especially attractive because they are conceptually simple and easy to compare across architectures. In these methods, a model is trained only on background events to compress and reconstruct its input. The reconstruction error is then used as the anomaly score, under the expectation that atypical events will be reconstructed less accurately. Autoencoders have therefore become a standard reference point in collider anomaly detection, including for jet-based searches, resonant-search studies, and more specialized adversarial, normalized, or graph-based constructions [4–11]. At the same time, recent studies have emphasized that reconstruction quality alone is not enough: performance can depend strongly on the input representation, the architecture, and the evaluation protocol [3,11]. Related work has also shown that non-neural or weakly supervised alternatives, such as tree-based strategies, can be highly competitive in collider anomaly searches [12]. This makes it especially important to ask what kind of architecture is most naturally matched to the physics of jets.

Jets are collimated sprays of hadrons produced by the fragmentation and hadronisation of energetic quarks or gluons in high-energy collisions. In proton-proton collisions, ordinary QCD jets constitute the dominant background, while jets originating from boosted heavy-particle decays, such as hadronic top-quark decays, typically exhibit richer internal substructure and more complex radiation patterns. These differences make jet substructure a natural setting for anomaly detection studies.

This substructure arises from a branching cascade: a high-energy quark or gluon radiates, those daughters radiate again, and the observable jet is built up through a nested sequence of splittings across several angular and momentum scales [13]. Physically relevant information is therefore not distributed uniformly across constituents. Hard, large-angle splittings determine the coarse prong structure of the jet, while softer and more collinear radiation refines that structure at progressively smaller scales. For anomaly detection, this is precisely the type of organisation one would like a model to capture: signal jets may differ from QCD not only through isolated features, but through how correlations persist or reorganise across scales.

This motivates looking beyond dense autoencoders. A fully connected autoencoder is flexible, but it treats the input as a largely unstructured vector and must learn any hierarchical organisation from the data alone. In other words, the network is capable of representing multiscale correlations, but it is not encouraged to do so in any particular way. This makes a concrete hypothesis testable: if the ordered jet representation is locally compressible in a way that follows the shower geometry, then a hierarchical architecture should model the background more economically than a dense map. Whether the additional disentangling structure of Multiscale Entanglement Renormalization Ansatz (MERA) is genuinely useful beyond a simpler tree hierarchy is, however, an empirical question rather than an assumption.

We study a multiscale architecture inspired by the Multiscale Entanglement Renormalisation Ansatz (MERA) [14,15]. MERA was originally developed in quantum many-body physics as a tensor-network construction that alternates coarse-graining with disentangling transformations, allowing relevant correlations to be retained across length scales. Related tensor-network ideas have also been adapted to machine learning, where they offer structured and often parameter-efficient representations of high-dimensional data [16]. In collider physics, tensor-network-inspired approaches have already been explored for event reconstruction, comparisons between classical and quantum tensor-network circuits, and generative or anomaly-oriented modelling [17–19]. For jet analysis, the appeal of MERA is, that its coarse-graining steps resemble the progressive aggregation of information from nearby constituents into larger-scale features, while the disentangling operations are designed to keep short-range correlations from being lost before compression. In a jet context, this means that local geometric structure can be reorganised before information is passed to the next scale, rather than being mixed globally from the start.

To the best of our knowledge, this is the first study to formulate collider anomaly detection with a MERA-inspired autoencoder. Earlier anomaly-detection applications in high-energy physics have considered quantum autoencoders and, more recently, tensor-network models based on matrix product states acting in an externally learned latent space, but not a MERA-inspired encoder–decoder operating directly on ordered jet constituents [20,21].

This is why we consider the MERA-type anomaly detection worth studying. If the dominant background really lies on a multiscale manifold shaped by the QCD shower, then a MERA-inspired autoencoder should be able to model that background economically: it should retain the correlations that matter, discard redundant short-distance structure efficiently, and compress the jet representation with fewer parameters than a dense network of comparable reach. The public Top Quark Tagging Reference Dataset, introduced as part of a broad comparison of modern jet-tagging methods, provides a clean and widely used benchmark for testing this idea in a controlled setting [22]. Thus, we turn that motivation

into two concrete tests: whether a locality-preserving constituent chain is measurably more compressible at the local tensor-network level, and whether trainable MERA disentanglers improve on the corresponding identity-disentangler limit. The goal of the present work is therefore to investigate whether a hierarchical MERA-inspired architecture provides a useful inductive bias for jet representations in a controlled anomaly-detection benchmark.

To this end, we perform a comparison between a dense autoencoder baseline, a MERA-inspired neural architecture, the corresponding tree tensor network (TTN) limit, and classical unsupervised methods, including principal component analysis (PCA), Gaussian modelling, and Isolation Forest. All methods are trained on background-only QCD jets and evaluated against hadronically decaying top jets using identical inputs, preprocessing, and performance metrics. This common setup allows observed differences to be attributed directly to the underlying representation rather than to differences in data handling or optimisation.

We study the models across several latent dimensions and assess their behaviour through ROC curves, AUC values, local-compressibility diagnostics, and direct architectural ablations. Particular emphasis is placed on identifying which part of the multiscale construction is actually supported by the data.

The rest of the paper is organised as follows. In Sect. 2, we introduce the canonical MERA as the physical and architectural motivation for the proposed MERA-inspired autoencoder and develop the mathematical framework underlying its construction. In Sect. 3, we describe the jet representation, the locality-preserving ordering, and the anomaly-detection setup. In Sect. 4, we present the concrete MERA-inspired, TTN, dense-autoencoder, and classical baseline models used in the study. In Sect. 5, we report the nominal benchmark, the local-compressibility and ordering analyses, and the disentangler ablation. Finally, Sect. 6 summarises the main findings and their implications.

2 Constructing a MERA-inspired anomaly detector

Tensor networks provide structured factorisations of high-order maps into products of low-order tensors. In many-body physics they are used to represent states with constrained correlation structure; in machine learning the same idea has been adapted to supervised classification, multiscale feature extraction, and generative modelling [16,23–29]. Collider applications of tensor-network-inspired methods have also begun to appear, including matrix-product-state event reconstruction and quantum-probabilistic generative or anomaly-detection frameworks [17–19]. A reconstruction-based anomaly detector built from MERA combines this tensor-network viewpoint with the standard background-only autoencoder strategy used in collider anomaly detection: one constructs a constrained multiscale encoder–decoder map, fits it only on background jets, and then uses the reconstruction error as the anomaly score. The overall workflow of the proposed MERA-inspired autoencoder is illustrated in Fig. 1.

We emphasise at the outset that the model studied here is a MERA-inspired neural architecture rather than a canonical tensor-network MERA. Canonical binary MERA uses staggered disentanglers acting across the boundaries of the blocks subsequently coarse-grained by isometries. The implemented model instead applies a local mixing map and a coarse-graining map successively to the same constituent pair, introduces nonlinear activations between hierarchical levels, and uses an untied decoder with soft rather than exact structural constraints. We use the term MERA-inspired because the architecture retains the motivating alternation of local correlation reorganisation and scale-by-scale compression. All numerical results below concern this explicitly defined neural architecture and should not be interpreted as results for canonical MERA.

2.1 From canonical MERA to a MERA-inspired encoder

In this section, we first describe the canonical MERA construction that motivates the model. The exact finite MERA-inspired neural architecture used in the numerical benchmark, including its modified connectivity and nonlinearities, is specified later in Sect. 4.

Let a jet be represented by an ordered list of N constituent-level feature vectors,

$$x = (x_1, \ldots, x_N), \qquad x_i \in \mathbb{R}^{d_0}. \tag{2.1}$$

In the present work, $d_0 = 3$ and $x_i = (p_T^i, \Delta\eta_i, \Delta\phi_i)$, but the construction is more general. One may either work directly with these site vectors or first apply a local feature map

$$\phi : \mathbb{R}^{d_0} \to \mathbb{R}^{\chi_0}, \qquad h_i^{(0)} = \phi(x_i), \tag{2.2}$$

where χ_0 is the local site dimension. The full input is then viewed as an element of a tensor-product space,

$$h^{(0)} \in \bigotimes_{i=1}^{N} \mathbb{R}^{\chi_0}. \tag{2.3}$$

The essential modelling decision is to interpret the ordered constituents as sites on a one-dimensional hierarchical graph. Once a site ordering has been fixed, a binary MERA encoder acts by alternating local basis changes (the disentanglers) with local projections (the isometries). At layer ℓ, with n_ℓ active sites of dimension χ_ℓ, we write schematically

$$h^{(\ell+1)} = \mathcal{W}^{(\ell)}\, \mathcal{U}^{(\ell)}\, h^{(\ell)}, \qquad \ell = 0, \ldots, L-1, \tag{2.4}$$

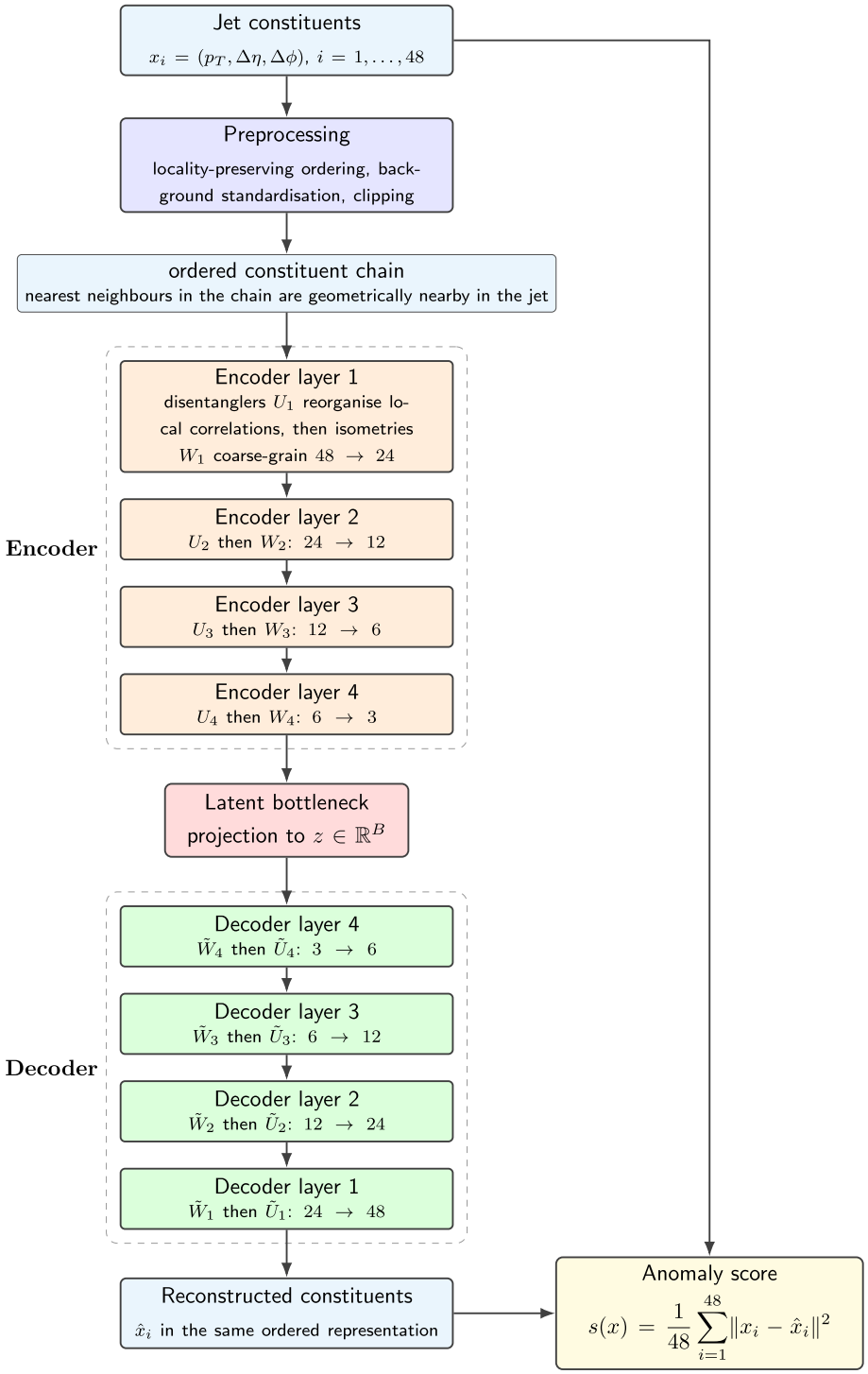


Fig. 1 Schematic overview of the MERA-inspired neural autoencoder used in this work. After locality-preserving ordering and standardisation, the encoder alternates local mixing maps U_ℓ and coarse-graining maps W_ℓ recursively across all four layers of the hierarchy $48 \rightarrow 24 \rightarrow 12 \rightarrow 6 \rightarrow 3$, followed by a projection to a latent vector $z \in \mathbb{R}^B$. In the implemented architecture, U_ℓ and W_ℓ act successively on the same local pair; this differs from the staggered cross-boundary connectivity of canonical binary MERA. The decoder mirrors the hierarchy with untied tensors $\tilde{U}_\ell$ and $\tilde{W}_\ell$ and reconstructs the ordered jet constituents. The anomaly score is computed from the reconstruction error between the original jet and its reconstruction, as indicated by the direct comparison arrow into the score box

where $\mathcal{U}^{(\ell)}$ is a product of local disentanglers and $\mathcal{W}^{(\ell)}$ is a product of local isometries. In a finite binary MERA-like layout, the local operations may be written as

$$p_j^{(\ell)} = h_{a_j}^{(\ell)} \otimes h_{b_j}^{(\ell)} \in \mathbb{R}^{\chi_\ell^2}, \tag{2.5}$$

$$\tilde{p}_j^{(\ell)} = U_j^{(\ell)}\, p_j^{(\ell)}, \qquad U_j^{(\ell)} \in \mathbb{R}^{\chi_\ell^2 \times \chi_\ell^2}, \tag{2.6}$$

$$h_j^{(\ell+1)} = W_j^{(\ell)}\, \tilde{p}_j^{(\ell)}, \qquad W_j^{(\ell)} \in \mathbb{R}^{\chi_{\ell+1} \times \chi_\ell^2}. \tag{2.7}$$

Here, the pair labels (a_j, b_j) are determined by the chosen MERA connectivity. In the standard binary construction, the disentanglers act across the boundaries of neighbouring blocks and the isometries then coarse-grain those disentangled blocks into the next layer [14,15]. The defining structural constraints are

$$(U_j^{(\ell)})^\top U_j^{(\ell)} = I_{\chi_\ell^2}, \qquad W_j^{(\ell)} (W_j^{(\ell)})^\top = I_{\chi_{\ell+1}}, \tag{2.8}$$

so each $U_j^{(\ell)}$ is orthogonal and each $W_j^{(\ell)}$ is an isometry. The action of a single MERA layer, including the disentangling and coarse-graining operations, is illustrated schematically in Fig. 2.

After L coarse-graining layers, the remaining top-level tensors are concatenated and mapped to a latent vector

$$z = P\,\mathrm{vec}\Big(h^{(L)}\Big), \qquad P \in \mathbb{R}^{B\times(n_L\chi_L)}, \tag{2.9}$$

where B denotes the latent dimension of the final representation. In the concrete implementation studied below, this abstract top-level map is specialised to a finite hierarchy whose top tensor has dimension $n_L\chi_L = 9$, followed by a final linear projection of width B. The encoder is, therefore, a structured map

$$E_\theta : \mathbb{R}^{Nd_0} \to \mathbb{R}^B, \qquad z = E_\theta(x), \tag{2.10}$$

with parameter set $\theta = \{U^{(\ell)}, W^{(\ell)}, P\}$.

2.2 Decoder and anomaly score

To use MERA for reconstruction-based anomaly detection, the encoder must be paired with a decoder

$$D_\vartheta : \mathbb{R}^B \to \mathbb{R}^{Nd_0}, \qquad \hat{x} = D_\vartheta(z). \tag{2.11}$$

If one insists on an exactly tied construction, the decoder can be built from the transposes of the encoder tensors, since for orthogonal U and isometric W the natural expansion maps are $U^\top$ and $W^\top$. In a more flexible autoencoder realization one may instead introduce an untied decoder,

$$\hat{h}^{(\ell)} = \bar{\mathcal{U}}^{(\ell)}\,\bar{\mathcal{W}}^{(\ell)}\,\hat{h}^{(\ell+1)}, \tag{2.12}$$

with independent parameters $\vartheta = \{\bar{U}^{(\ell)}, \bar{W}^{(\ell)}, \bar{P}\}$. This generic choice is useful because exact inversion is impossible once information has been discarded by coarse-graining. In Sect. 4.1 we adopt this untied option for the concrete benchmark model.

At the conceptual level, the encoder–decoder is trained on background jets and the anomaly score is the reconstruction loss,

$$S(x) = \mathcal{L}_{\mathrm{rec}}\big(x,\, D_{\vartheta^\star}(E_{\theta^\star}(x))\big)\,. \tag{2.13}$$

Large values of $S(x)$ indicate that the jet lies away from the multiscale background manifold learned by the encoder–decoder. This subsection defines only that general reconstruction principle; the exact finite decoder, per-jet loss, soft regularisation terms, and numerical hyperparameters used in the benchmark are specified together in Sect. 4.1.

2.3 Potential advantages of a MERA approach for anomaly detection and interpretability

The advantage of MERA over a plain tree tensor network or an unstructured dense map can be made explicit at the level of local compression. Consider a local pair or block vector $p \in \mathbb{R}^{\chi_\ell^2}$ with mean $\mu = \mathbb{E}[p]$ and covariance matrix

$$\Sigma = \mathbb{E}\Big[(p-\mu)(p-\mu)^\top\Big]. \tag{2.14}$$

Writing $q = p - \mu$ for the centred local fluctuation, a tree tensor network without disentanglers compresses q by projection onto a low-dimensional subspace,

$$\hat{q}_{\mathrm{TTN}} = W^\top W\, q. \tag{2.15}$$

By contrast, MERA first rotates the local basis and only then truncates,

$$\hat{q}_{\mathrm{MERA}} = U^\top W^\top W U\, q. \tag{2.16}$$

The corresponding expected local truncation error is

$$\begin{aligned}\varepsilon(U,W) &= \mathbb{E}\Big[\|q - U^\top W^\top W U q\|_2^2\Big] \\ &= \mathrm{tr}(\Sigma) - \mathrm{tr}\Big(WU\Sigma U^\top W^\top\Big).\end{aligned} \tag{2.17}$$

For fixed U, the optimal W is obtained by choosing the $\chi_{\ell+1}$ leading eigenvectors of $U\Sigma U^\top$; this is just PCA in the rotated basis. The role of the disentangler is therefore clear: it seeks a basis in which the degrees of freedom that must be retained by the subsequent isometry are as informative as possible. In jet language, this is exactly the mechanism one would like if short-distance angular correlations have to be reorganised before a coarse description is formed. This is the key mathematical reason to expect MERA to outperform a pure tree, and potentially to reach a given reconstruction quality with fewer parameters than a dense autoencoder.

The same argument also explains why the choice of geometry matters. A dense autoencoder compresses the entire vector $x \in \mathbb{R}^{Nd_0}$ using global matrix multiplications. It can represent multiscale structure, but it does not encode where that structure is expected to live. A MERA, by construction, allocates different tensors to different scales and only allows information to propagate upward through a prescribed hierarchy. If that hierarchy is well aligned with the QCD shower, the network spends its capacity on the correlations one expects to be physically relevant, rather than learning them indirectly from scratch. The capacity available at each scale is in turn controlled by the bond dimension χ, which denotes the dimension of the internal tensor indices connecting neighbouring tensors. Physically, larger values of χ allow more correlations and information to be retained between different parts of the system, at the cost of increased computational complexity.

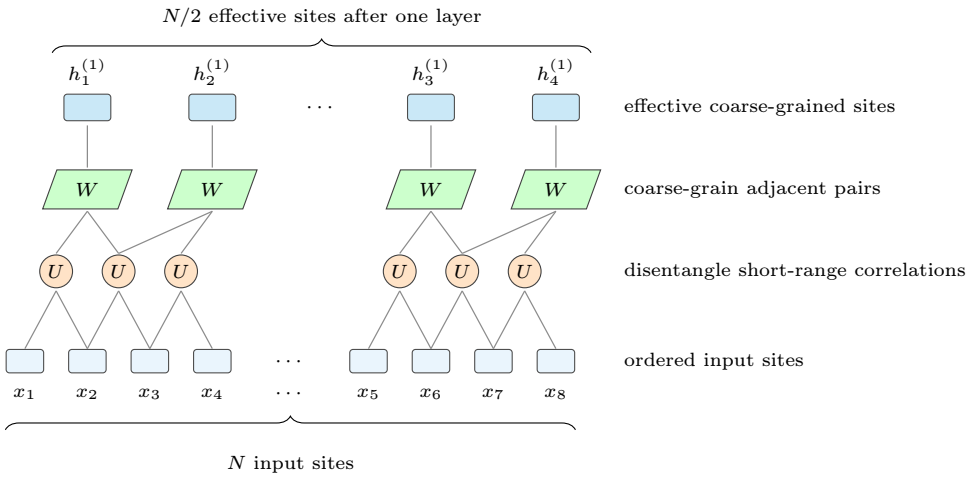


Fig. 2 Schematic action of one layer in a canonical binary MERA, shown here as the motivation for the proposed architecture. Canonical disentanglers U reorganise short-range correlations across block boundaries before isometries W coarse-grain adjacent pairs into a reduced set of effective sites. The implemented MERA-inspired neural architecture differs in that each $U_j^{(\ell)}$ and $W_j^{(\ell)}$ acts successively on the same local pair, as specified in Sect. 4.1

For moderate bond dimensions, this leads to favourable scaling. If $\chi_\ell \sim \chi$ across layers, then each disentangler carries $\mathcal{O}(\chi^4)$ parameters and each isometry carries $\mathcal{O}(\chi^3)$ parameters, while the total number of local tensors is $\mathcal{O}(N)$ in a binary hierarchy. The resulting parameter count is therefore

$$N_{\text{param}}^{\text{MERA}} = \mathcal{O}(N\chi^4), \tag{2.18}$$

up to the small top tensor. The point is not that this is universally smaller than every dense autoencoder, but that the scaling is linear in the number of sites for fixed bond dimension and that the retained parameters have a clear multiscale interpretation.

2.4 Mathematical challenges

Although the construction above is conceptually natural, designing a MERA for anomaly detection is mathematically non-trivial.

Geometry and site ordering. MERA assumes a graph on which locality is defined, whereas a jet is fundamentally an unordered set of constituents. One must therefore specify a map

$$\pi : \{1, \dots, N\} \to \{1, \dots, N\}, \qquad x \mapsto (x_{\pi(1)}, \dots, x_{\pi(N)}), \tag{2.19}$$

that turns the set into an ordered sequence. If this ordering fails to place strongly correlated constituents near each other, then the local disentanglers and isometries will act on the wrong degrees of freedom and the MERA inductive bias is weakened. This is one of the central modelling choices in collider applications, and it is why constituent ordering is not a cosmetic preprocessing step but part of the architecture itself.

Irregular system size. The cleanest binary MERA is defined for a dyadic number of sites, $N = 2^L$. Jet representations rarely satisfy this exactly, and even a fixed choice such as $N = 48$ leads to an irregular top of the hierarchy:

$$48 \to 24 \to 12 \to 6 \to 3. \tag{2.20}$$

One must then decide whether to pad to a dyadic size, use mixed-arity tensors near the top, or terminate the hierarchy with an additional dense projection. Each choice changes the balance between exact MERA structure and practical flexibility.

Constrained optimisation. The exact parameter spaces are not Euclidean:

$$U_j^{(\ell)} \in \mathrm{O}(\chi_\ell^2), \qquad W_j^{(\ell)} \in \mathrm{St}(\chi_{\ell+1}, \chi_\ell^2), \tag{2.21}$$

where $\mathrm{O}(m)$ is the orthogonal group and $\mathrm{St}(r, m)$ is the Stiefel manifold[1] of r orthonormal rows in $\mathbb{R}^m$ [30]. Training directly in these constrained spaces is possible, but more involved than ordinary gradient descent. Alternatives include soft penalties of the form used in Sect. 4.1, explicit retractions onto the constraint manifold [31], or layerwise constructions based on truncated higher-order singular value decompositions and orthogonal Procrustes updates [32–34]. The practical challenge is that one wants enough freedom to fit the background distribution, while preserving enough orthogonality that the MERA interpretation is not lost.

Non-convex coupling across layers. Even the local objective in Eq. (2.17) becomes difficult once all layers are trained jointly, because the covariance seen at layer ℓ depends on every lower-layer tensor. The optimisation problem is therefore highly non-convex and strongly coupled across scales. Moreover, internal tensor-network gauge freedoms imply

[1] The Stiefel manifold $\mathrm{St}(r, m)$ is the space of all ordered sets of r mutually orthonormal vectors in $\mathbb{R}^m$, or equivalently all $r \times m$ matrices with orthonormal rows. It is the natural parameter space for the MERA isometries, because each W must choose and preserve an orthonormal r-dimensional subspace of the local m-dimensional block that is retained under coarse-graining.

that many parameter configurations represent the same global map, producing gauge-induced flat directions in the loss landscape [23]. This complicates both optimisation and interpretation of the learned tensors.

No exact inverse after truncation. Once $\chi_{\ell+1} < \chi_\ell^2$, the projector $W^\top W$ is not the identity:

$$W^\top W \neq I_{\chi_\ell^2}. \tag{2.22}$$

Therefore, a MERA autoencoder is never an exactly invertible transform after coarse-graining; information has been discarded by construction. For anomaly detection, this is not a bug but the central mechanism, since the model is supposed to keep the dominant background modes and discard the rest. Nevertheless, it means that the decoder must learn a stable approximate inverse of a sequence of projections and rotations, which can be numerically delicate when the retained subspaces vary strongly across layers.

These issues explain why constructing a useful MERA anomaly detector is not equivalent to simply replacing dense layers with tensor contractions. The success of the method depends on simultaneously choosing a meaningful geometry for the data, enforcing a tractable multiscale compression scheme, and training under non-trivial algebraic constraints. The concrete architecture used in our numerical study is one specific realisation of this general construction.

3 Benchmarking and data preprocessing

3.1 Benchmark dataset and task definition

Our study uses the publicly available Top Quark Tagging Reference Dataset, introduced as a common benchmark for comparing jet-substructure and machine-learning methods [22,35]. This benchmark is well-suited to the present analysis for two reasons. First, it is already widely used, which makes the results easy to compare with the existing literature. Second, it provides an environment in which differences between methods can be traced primarily to model design rather than to variations in event generation, detector simulation, or train–test splitting.

The events correspond to proton–proton collisions at a centre-of-mass energy of $\sqrt{s} = 14$ TeV. They are generated with PYTHIA8 [36] and passed through a fast detector simulation based on DELPHES [37]. Jets are reconstructed with the anti-k_T clustering algorithm [38] using radius parameter $R = 0.8$. Throughout the analysis, we restrict attention to jets with transverse momentum

$$550\ \text{GeV} \leq p_T^{\text{jet}} \leq 650\ \text{GeV}$$

and pseudorapidity

$$|\eta^{\text{jet}}| < 2.$$

A jet is a collimated spray of hadrons produced when an energetic quark or gluon fragments in the detector. In this benchmark, the background sample consists of ordinary QCD jets from dijet production, while the signal sample consists of jets initiated by hadronically decaying top quarks. The top jets are therefore not used here as a literal model of new physics. Rather, they serve as a well-understood example of jets with richer internal structure than the dominant QCD background, making them a useful proxy anomaly when testing reconstruction-based methods. A broader study involving multiple signal classes would be valuable for assessing anomaly-detection generalisation, but lies beyond the scope of the present work.

The dataset is provided with fixed splits of 1.2M training jets, 400k validation jets, and 400k test jets. We keep these predefined splits unchanged for every method considered in this paper. In the anomaly-detection setup, all models are trained or fitted only on background jets. Signal jets enter only at validation and test time. This is conceptually important: the task is not supervised classification of top versus QCD jets, but the identification of jets that are poorly described by a model of typical background structure.

3.2 Jet representation

Each jet is stored as a list of up to 200 reconstructed constituents. For each constituent, the dataset provides its four-momentum,

$$p_i^\mu = (E_i, p_{x,i}, p_{y,i}, p_{z,i}), \qquad i = 1, \ldots, N_{\text{CONST}}.$$

Jets with fewer than 200 constituents are zero-padded, while jets with more constituents are truncated by construction of the benchmark. In the released dataset, the constituents are ordered by decreasing transverse momentum.

For the present study, we retain only the first

$$N_{\text{CONST}} = 48$$

constituents. This choice keeps the input dimension manageable while preserving the dominant substructure information in the kinematic regime under consideration. Just as importantly, it provides every method with the same fixed-size input, which is necessary for a controlled comparison.

To quantify the information retained by this choice, we evaluated the cumulative transverse-momentum coverage of the leading constituents in the benchmark dataset. The leading 48 constituents retain on average 98.9% of the total constituent transverse momentum, compared with 95.8% for 32

constituents and 99.8% for 64 constituents. We therefore adopt $N_{\mathrm{CONST}} = 48$ as a practical compromise between information retention and computational cost. The fact that $48 = 3 \times 2^4$ is compatible with the four-layer binary MERA hierarchy used in this work, but this was not the primary motivation for the choice.

Rather than using raw Cartesian momentum components directly, we express each jet in a jet-centred coordinate system. The four-momentum reconstructed from the retained constituents is

$$p^{\mu}_{\mathrm{jet}} = \sum_{i=1}^{N_{\mathrm{CONST}}} p^{\mu}_{i},$$

from which the jet axis $(p_{T,\mathrm{jet}}, \eta_{\mathrm{jet}}, \phi_{\mathrm{jet}})$ is obtained. For each constituent, we compute the transverse momentum

$$p^{i}_{T} = \sqrt{p^2_{x,i} + p^2_{y,i}},$$

the azimuthal angle

$$\phi_i = \mathrm{atan2}(p_{y,i}, p_{x,i}),$$

and the pseudorapidity

$$\eta_i = \frac{1}{2} \ln \left(\frac{|\mathbf{p}_i| + p_{z,i}}{|\mathbf{p}_i| - p_{z,i}} \right), \qquad |\mathbf{p}_i| = \sqrt{p^2_{x,i} + p^2_{y,i} + p^2_{z,i}}.$$

The constituent coordinates used as inputs are then

$$\left(p^{i}_{T}, \Delta\eta_i, \Delta\phi_i \right),$$

with

$$\Delta\eta_i = \eta_i - \eta_{\mathrm{jet}}, \qquad \Delta\phi_i = \mathrm{wrap}\left(\phi_i - \phi_{\mathrm{jet}}\right),$$

where wrap maps the azimuthal difference to the interval $(-\pi, \pi]$.

This representation removes the trivial dependence on the global jet direction and focuses the model on the internal energy flow within the jet. In other words, the models see how momentum is distributed *inside* the jet rather than where the jet happens to point in the detector. The final input is obtained by concatenating the constituent triplets into a vector of dimension

$$3 \times N_{\mathrm{CONST}} = 144.$$

The same representation is used for all neural and classical baselines.

3.3 Locality-preserving ordering and standardisation

The original benchmark orders constituents by decreasing p_T. This is a sensible default, but it is not ideal for a multi-scale architecture whose local tensor operations act on neighbouring sites. For a MERA-type model, the notion of "neighbouring" constituents matters: one would like adjacent sites to correspond, as far as possible, to constituents that are also close in the jet.

For this reason, we reorder the active constituents using a simple geometry-based procedure in the $(\Delta\eta, \Delta\phi)$ plane. Constituents with

$$p_T \leq 10^{-6}\,\mathrm{GeV}$$

are treated as inactive padding entries. This is not a physical transverse-momentum cut, but only a numerical zero threshold used to identify padded constituents in the native dataset units. Among the active constituents, we begin with the highest-p_T constituent and then append the nearest unvisited neighbour according to the angular distance

$$\Delta R^2_{ij} = (\Delta\eta_i - \Delta\eta_j)^2 + (\Delta\phi_i - \Delta\phi_j)^2. \tag{3.1}$$

Inactive padded entries are placed at the end of the sequence. This does not change the jet itself, but it produces a fixed ordering that better preserves local geometric correlations. The same reordered inputs are used for every model in the benchmark, so any performance differences cannot be attributed to giving MERA privileged information.

After ordering, each input feature is standardised using only the background portion of the training set. If x_j denotes a given input component, we define

$$\tilde{x}_j = \frac{x_j - \mu_j}{\sigma_j + 10^{-6}}, \tag{3.2}$$

where μ_j and σ_j are the mean and standard deviation computed from background training jets alone. The additive term 10^{-6} is a small numerical stabiliser: some components can have vanishing or extremely small variance, especially for padded entries, and the offset prevents division by zero or by an anomalously small denominator. The same transformation is then applied to validation and test data. This prevents information leakage from the signal sample and ensures that every method sees identically normalised inputs. To improve numerical stability, we additionally clip the standardised features to the range

$$\tilde{x}_j \leftarrow \mathrm{clip}(\tilde{x}_j, -z_{\max}, z_{\max}), \qquad z_{\max} = 5. \tag{3.3}$$

We checked explicitly how often this clipping threshold is reached before clipping is applied. Only approximately 0.12% of feature entries in QCD jets and 0.086% in top jets lie outside the interval $[-5, 5]$. Clipping therefore affects a negligible fraction of entries in both samples and does not

preferentially remove extreme values from the top-jet sample.

We also decomposed the per-jet mean-squared reconstruction loss into its p_T, $\Delta\eta$, and $\Delta\phi$ contributions after standardisation. Averaged over three independent seeds at $B = 8$, the p_T component accounts for approximately 60% of the dense-autoencoder loss on QCD jets and 72% on top jets. For the MERA-inspired model, the corresponding fractions are approximately 52% and 61%. Thus, p_T provides the largest single contribution for both architectures and both samples, while the two angular variables together account for the remainder. These fractions refer to the reconstruction loss and do not alter the equal component-wise standardisation applied to the inputs.

3.4 Benchmarking protocol

All methods are compared in the same unsupervised setting. Neural models are trained on background jets only, and the classical baselines are likewise fitted only to the background training sample. Signal jets are withheld until validation and testing. This benchmarking protocol is designed to answer a specific question: how well can each method learn the structure of ordinary QCD jets, and how strongly does a top jet deviate from that learned background description?

For the neural models studied here, anomaly scores are defined by the per-jet reconstruction loss. A jet that is well represented by the learned background manifold receives a small score, while a jet that cannot be reconstructed accurately receives a larger score and is interpreted as more anomalous. For the classical baselines, we use the natural score associated with each method: the reconstruction residual for PCA, the Mahalanobis distance for the Gaussian model, and the isolation-depth-based score for Isolation Forest.

Because the dataset, preprocessing, train–validation–test splits, and scoring protocol are fixed across all models, the benchmark isolates the effect of representational structure as cleanly as possible. This is the main reason for organising the study in this way.

4 Model implementations and reference baselines

This section specifies the concrete models used in the numerical study. All methods act on the common representation defined in Sect. 3, namely a standardised vector

$$x \in \mathbb{R}^{144}, \tag{4.1}$$

obtained from 48 ordered jet constituents with three features each. For the capacity-controlled comparison, MERA, the dense autoencoder, and PCA are studied at effective latent dimension $B \in \{8, 16, 32\}$. The Gaussian and Isolation Forest baselines do not contain an explicit bottleneck, but they are trained on the same background-only sample and evaluated on the same validation and test sets. Details of optimisation and model selection are summarised at the beginning of Sect. 5; here, the focus is on the structure of the models themselves.

4.1 MERA-inspired neural implementation

The canonical MERA construction and its motivation for anomaly detection were discussed in Sect. 2. We now define the modified, finite MERA-inspired neural architecture used in the benchmark. In particular, the implementation uses within-pair rather than staggered cross-boundary mixing, nonlinear activations between hierarchical levels, an untied decoder, and soft structural constraints. The input vector is reshaped into a sequence of site tensors,

$$x \mapsto X^{(0)} = \{s_i^{(0)}\}_{i=1}^{48}, \qquad s_i^{(0)} \in \mathbb{R}^3, \tag{4.2}$$

so that each site corresponds to one ordered constituent described by $(p_T, \Delta\eta, \Delta\phi)$. We then apply four binary coarse-graining stages, giving

$$n_0 = 48, n_1 = 24, n_2 = 12, \qquad n_3 = 6, \qquad n_4 = 3, \tag{4.3}$$

where n_ℓ denotes the number of active sites at level ℓ. In the implementation used here, the site dimension is kept fixed at $\chi = 3$ at every scale. This keeps the local tensors small and makes the trainable complexity grow mainly with the number of sites rather than with an increasing bond dimension.

At each layer $\ell = 0, \ldots, 3$, neighbouring sites are paired according to the geometry-preserving ordering introduced in Sect. 3. For the j-th pair, we form the local vector

$$p_j^{(\ell)} = \begin{pmatrix} s_{2j-1}^{(\ell)} \\ s_{2j}^{(\ell)} \end{pmatrix} \in \mathbb{R}^6. \tag{4.4}$$

A learned local mixing map, playing the role of a disentangler in the MERA-inspired construction,

$$U_j^{(\ell)} \in \mathbb{R}^{6\times 6} \tag{4.5}$$

first mixes the two neighbouring sites, and a learned coarse-graining map, playing the role of an isometry,

$$W_j^{(\ell)} \in \mathbb{R}^{3\times 6} \tag{4.6}$$

then projects the result to the next scale:

$$\tilde{p}_j^{(\ell)} = U_j^{(\ell)} p_j^{(\ell)}, \qquad s_j^{(\ell+1)} = \sigma\left(W_j^{(\ell)} \tilde{p}_j^{(\ell)}\right), \tag{4.7}$$

where σ denotes a leaky-ReLU activation with slope parameter $\alpha = 0.1$ at the intermediate hierarchical levels and the identity at the final encoder level. Thus, the nonlinearity is applied after each intermediate coarse-graining operation $W_j^{(\ell)}$. Unlike canonical MERA, the implemented layer

is nonlinear and its $U_j^{(\ell)}$ and $W_j^{(\ell)}$ maps act successively on the same pair $(s_{2j-1}^{(\ell)}, s_{2j}^{(\ell)})$. The tensors are not shared across positions or layers. This is a deliberate choice: the ordered jet is finite, not translationally invariant, and the upper part of the hierarchy is irregular because $48 \neq 2^L$.

After the final coarse-graining stage, the remaining three sites are concatenated into a top tensor

$$h \in \mathbb{R}^9. \tag{4.8}$$

A linear map then produces the latent representation,

$$z = P_{\text{enc}} h, \qquad P_{\text{enc}} \in \mathbb{R}^{B\times 9}, \tag{4.9}$$

with $B \in \{8, 16, 32\}$. For $B = 16$ and 32, this final map expands the 9-dimensional top tensor, so the main compression is already carried by the hierarchical coarse-graining rather than by this last linear layer alone. The decoder begins with an untied linear expansion

$$\hat{h} = P_{\text{dec}} z, \qquad P_{\text{dec}} \in \mathbb{R}^{9\times B}, \tag{4.10}$$

and then reconstructs finer scales through learned expansion maps

$$\bar{W}_j^{(\ell)} \in \mathbb{R}^{6\times 3} \tag{4.11}$$

and local mixing maps

$$\bar{U}_j^{(\ell)} \in \mathbb{R}^{6\times 6}, \tag{4.12}$$

according to

$$\tilde{q}_j^{(\ell)} = \bar{W}_j^{(\ell)} s_j^{(\ell+1)}, \qquad \hat{q}_j^{(\ell)} = \bar{U}_j^{(\ell)} \tilde{q}_j^{(\ell)}, \tag{4.13}$$

followed by reshaping $\hat{q}_j^{(\ell)}$ into two daughter sites at the next finer level. The final output layer is linear, so the model reconstructs

$$\hat{x} \in \mathbb{R}^{144} \tag{4.14}$$

in the same feature space in which it was trained.

The network is trained by minimising the per-jet mean-squared reconstruction loss

$$\mathcal{L}_{\text{rec}}(x) = \frac{1}{144}\|x - \hat{x}\|_2^2, \tag{4.15}$$

augmented by soft orthogonality penalties on the tensors that play the role of disentanglers and isometries,

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_U \sum_{\ell,j} \left\| U_j^{(\ell)\top} U_j^{(\ell)} - I_6 \right\|_F^2 + \lambda_W \sum_{\ell,j} \left\| W_j^{(\ell)} W_j^{(\ell)\top} - I_3 \right\|_F^2. \tag{4.16}$$

In the nominal benchmark, we use $\lambda_U = \lambda_W = 10^{-3}$. Rather than enforcing exact constraints on orthogonal or Stiefel manifolds, we use these penalties to keep the learned maps close to the intended MERA structure while retaining the flexibility of end-to-end optimisation [14,15,23,25,26,34]. The decoder is kept untied for a related reason: after coarse-graining, exact inversion is impossible, so a learned approximate inverse is more suitable than a rigid transpose construction. Once trained on background jets only, the anomaly score is taken to be the reconstruction error,

$$S_{\text{MERA}}(x) = \mathcal{L}_{\text{rec}}(x). \tag{4.17}$$

The relationship between the concrete penalised training objective and the inference-time anomaly score is illustrated in Fig. 3. During training, the reconstruction and regularisation terms act on the same encoder–decoder parameters, whereas at inference time only the reconstruction error is used as the anomaly score.

4.2 Dense autoencoder baseline

Fully connected autoencoders provide the standard nonlinear reconstruction baseline for unsupervised anomaly detection and have been widely used in collider applications [4–8,10,11,39]. Their role in the present study is clear: unlike MERA, a dense autoencoder can mix all input features globally, but it carries no explicit notion of scale or locality. Any hierarchical organisation in the jet therefore has to be discovered implicitly from the training data rather than being built into the architecture.

Our reference autoencoder takes the same 144-dimensional input and uses a symmetric fully connected encoder–decoder of the form

$$144 \to 48 \to 24 \to B \to 24 \to 48 \to 144, \tag{4.18}$$

with $B \in \{8, 16, 32\}$. Writing the encoder and decoder as f_{enc} and f_{dec}, the model computes

$$z = f_{\text{enc}}(x) \in \mathbb{R}^B, \qquad \hat{x} = f_{\text{dec}}(z) \in \mathbb{R}^{144}. \tag{4.19}$$

Leaky-ReLU activations with slope $\alpha = 0.1$ are used in the hidden layers, while the bottleneck and output layers are kept linear. The architecture is intentionally modest: it is expressive enough to model global nonlinear correlations, but still simple enough that differences with respect to MERA can be interpreted as differences in inductive bias rather than brute-force network size.

The autoencoder is trained on background jets only with the same reconstruction objective used for MERA,

$$\mathcal{L}_{\text{rec}}(x) = \frac{1}{144}\|x - \hat{x}\|_2^2, \tag{4.20}$$

and the anomaly score is defined by

$$S_{\text{AE}}(x) = \mathcal{L}_{\text{rec}}(x). \tag{4.21}$$

Using the same input representation, the same loss, and the same latent sizes makes this model a direct test of whether explicit multiscale structure improves background compression over a generic dense map.

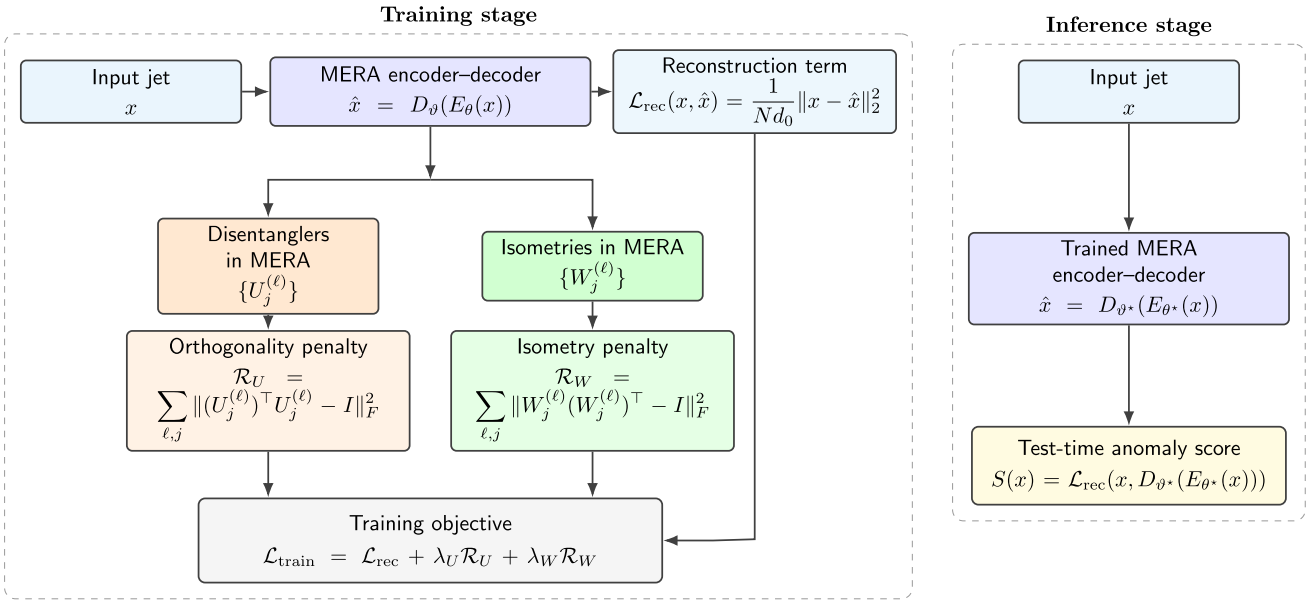


Fig. 3 Schematic relationship between the training objective and inference-time anomaly score for the implemented MERA-inspired network. During training, the reconstruction loss is supplemented by soft orthogonality and isometry penalties on the local mixing and coarse-graining maps. At inference time, only the per-jet reconstruction error is used as the anomaly score

4.3 Classical reference baselines

To complement the neural comparison, we also include three standard unsupervised baselines that embody different notions of normality: a linear low-rank model (PCA), a single-component density model (Gaussian/Mahalanobis), and a tree-based isolation method (Isolation Forest). All are fitted on the same standardised background sample used for the neural architectures.

Principal Component Analysis. PCA is the canonical linear compression baseline [40,41]. Let $\hat{\mu}$ denote the empirical background mean and let

$$V_B = \big(v_1, \ldots, v_B\big) \in \mathbb{R}^{144\times B} \tag{4.22}$$

contain the B leading eigenvectors of the empirical covariance matrix. The rank-B reconstruction is

$$\hat{x}_{\rm PCA} = \hat{\mu} + V_B V_B^\top (x - \hat{\mu}), \tag{4.23}$$

and the anomaly score is the squared residual

$$S_{\rm PCA}(x) = \|x - \hat{x}_{\rm PCA}\|_2^2 = \big\|(I - V_B V_B^\top)(x-\hat{\mu})\big\|_2^2. \tag{4.24}$$

PCA is especially informative in the present context because each MERA isometry performs a local linear truncation. It therefore provides the cleanest global linear reference against which one can assess the value of multiscale nonlinear structure.

Gaussian background model. As a minimal density-estimation baseline, we fit a multivariate Gaussian distribution to the background sample using the empirical mean $\hat{\mu}$ and a ridge-regularised covariance matrix $\hat{\Sigma} + \lambda_{\rm G} I$ with $\lambda_{\rm G} = 10^{-3}$. Under this model, the negative log-likelihood differs from the squared Mahalanobis distance only by additive terms that are common across events, so we use the corresponding regularised Mahalanobis score [42]

$$S_{\rm G}(x) = (x-\hat{\mu})^\top \big(\hat{\Sigma} + \lambda_{\rm G} I\big)^{-1}(x-\hat{\mu}). \tag{4.25}$$

This baseline asks whether the standardised background can already be described adequately by a single ellipsoidal distribution in feature space. If the background manifold is curved, multimodal, or strongly non-Gaussian, this model should underperform, which helps to interpret the gains from more structured methods.

Isolation forest. Isolation Forest provides a conceptually different unsupervised baseline based on recursive random partitioning rather than reconstruction or density estimation [43]. We use an ensemble of $T = 100$ isolation trees trained on background events. For an event x, each tree returns a path length $h_t(x)$, and the average path length

$$\bar{h}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{4.26}$$

is converted into the standard isolation score

$$S_{\mathrm{IF}}(x) = 2^{-\bar{h}(x)/c(n)}, \qquad c(n) = 2H_{n-1} - \frac{2(n-1)}{n}, \tag{4.27}$$

where n is the subsample size used to construct a tree and H_k denotes the k-th harmonic number. Jets that are isolated after fewer random splits receive larger anomaly scores. Any monotonic equivalent of this standard score would lead to the same ROC ranking, but the expression above makes the statistical interpretation transparent.

5 Results and discussion

For both MERA and the dense autoencoder, hyperparameters were selected separately for each bottleneck dimension B using the validation sample. The selected configurations were then fixed and evaluated over independent random seeds disjoint from those used for model selection. All methods are trained on the QCD background sample only, with signal events used exclusively for validation and final testing. For the neural architectures, optimisation is carried out with Adam [44] using the reconstruction losses defined in Sect. 4, while the classical reference models are fitted with their standard unsupervised objectives. For stochastic models, the final performance is reported as a mean over repeated runs with independent random seeds.

The main question of this section is no longer just which curve sits highest in one benchmark table. We instead separate two issues. First, we establish the nominal benchmark comparison against the dense autoencoder and the classical baselines. Second, we test the locality argument directly through a training-free local-compressibility diagnostic and then ask whether the disentanglers improve on the corresponding identity-disentangler limit of the same hierarchy.

5.1 Nominal benchmark

Figure 4 and Table 1 summarise the nominal benchmark for the locality-preserving representation introduced in Sect. 3. Within this specific scan, MERA attains the largest test AUC at all three latent dimensions, reaching 0.7959, 0.7885, and 0.7963 at $B = 8, 16, 32$, respectively. The corresponding dense-autoencoder values are 0.7616, 0.7487, and 0.7550, while PCA becomes competitive only once the retained dimension is large enough to capture a sizeable fraction of the global background variance.

The comparison to the dense autoencoder is the most important one in this subsection, because the dense AE is the standard nonlinear reconstruction baseline and carries substantially more trainable freedom than the implemented MERA-inspired model. The MERA-inspired model attains a higher mean AUC at every latent dimension in this benchmark: the gain is about 0.034 at $B = 8$, 0.040 at $B = 16$, and 0.041 at $B = 32$. At the same time, the implemented models use only about one third of the trainable parameters of the corresponding dense autoencoders: roughly 5.0–5.5 $\times 10^3$ parameters, compared with 1.68–1.80 $\times 10^4$ for AE. In other words, these numerical results show that, for the locality-preserving representation used here, the implemented MERA-inspired multiscale construction can outperform the standard dense autoencoder while doing so with a far smaller parameter budget. This empirical statement concerns the neural architecture defined in Sect. 4.1, not canonical MERA in general.

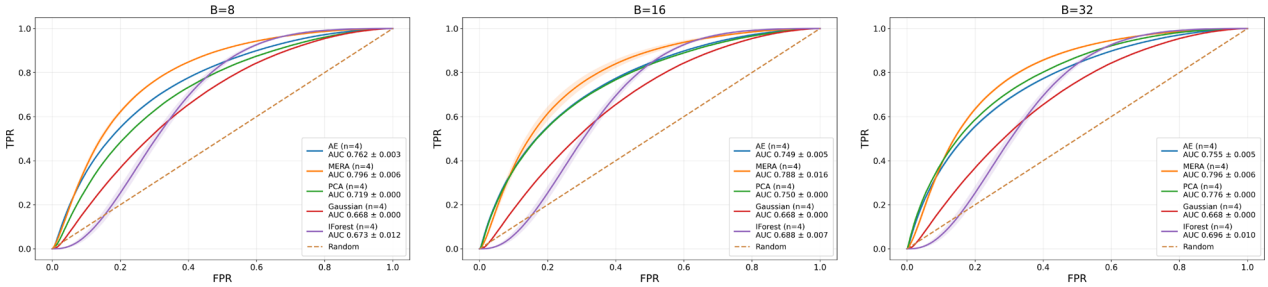
That point is important for the logic of the paper. If the MERA model had only matched the AE after introducing extra structure, the result would mainly be architectural curiosity. Instead, Table 1 shows a simultaneously stronger and more economical model in the nominal scan. This is exactly the kind of outcome that motivates asking which part of the construction is responsible: whether the gain is tied to the locality-aligned representation, to hierarchical coarse-graining more generally, or to the disentangling layers specific to MERA.

Table 1 shows the same comparison numerically. The key observation is therefore quite concrete: in the benchmark setting used throughout this paper, the implemented MERA-inspired architecture attains a higher mean AUC than the dense autoencoder with only about one-third of the dense model's trainable parameters. This makes the nominal benchmark a positive result for the specific neural construction studied here and provides the starting point for the more targeted tests below.

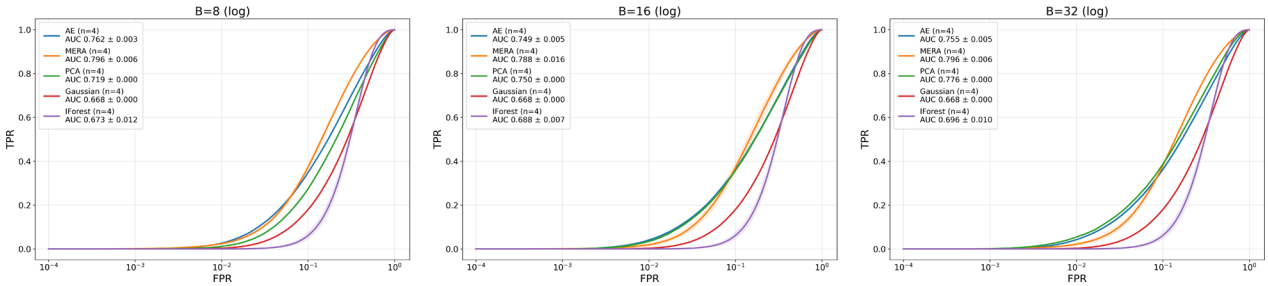
5.2 Locality and local compressibility

The most direct way to test the locality argument of Sect. 2 is to ask whether the ordered input is actually more compressible at the scale on which the local tensors of the implemented MERA-inspired architecture act. We therefore performed a training-free diagnostic on background jets for the three constituent orderings considered earlier: the locality-preserving chain used in the main benchmark, a simple p_T ordering, and random permutations. For each ordering, we measured the angular separation between adjacent sites and the retained variance fraction of the optimal three-dimensional approximation to neighbouring two-site blocks. The latter quantity is the local linear analogue of the compression step performed by the first layers of the implemented model.

Table 2 summarises the two quantities that matter most for this question. The mean adjacent-site separation $\langle \Delta R \rangle$ measures how geometrically local neighbouring entries in the one-dimensional MERA chain actually are. The retained-variance fractions then quantify how compressible those neighbouring two-site blocks are under the optimal three-



(a) ROC curves at latent dimensions $B = 8, 16, 32$ on a linear false-positive-rate scale.



(b) The same ROC curves in the low-false-positive-rate regime, shown on a logarithmic false-positive-rate scale.

Fig. 4 ROC curves for the nominal benchmark at latent dimensions $B = 8, 16, 32$. Solid lines denote the mean over random seeds where applicable, and shaded bands indicate the corresponding $\pm 1\sigma$ variation

Table 1 Test-set ROC-AUC values for all models at latent dimensions $B = 8, 16, 32$, where AUC denotes the area under the receiver operating characteristic (ROC) curve and CI denotes the confidence interval quoted for repeated runs. For the neural architectures, the corresponding number of trainable parameters is also listed. For PCA and the Gaussian baseline, the reported AUC standard deviations are exactly zero because these methods are deterministic once the dataset and preprocessing are fixed

B	Model	Architecture	Params	Test AUC	AUC(std)	CI low	CI high
8	MERA	(L = 4, latent = 8, $lr = 3 \times 10^{-4}$)	5021	0.7959	0.0065	0.7855	0.8063
8	AE	(h1 = 48, h2 = 24, bn = 8, $lr = 3 \times 10^{-4}$)	16808	0.7616	0.0038	0.7554	0.7677
8	PCA	(ncomp = 8)	–	0.7192	0.0000	0.7192	0.7192
8	Gaussian	Mahalanobis (ridge-reg. cov)	–	0.6679	0.0000	0.6679	0.6679
8	IForest	(ntrees = 100)	–	0.6726	0.0142	0.6500	0.6952
16	MERA	(L = 4, latent = 16, $lr = 2 \times 10^{-4}$)	5173	0.7885	0.0190	0.7582	0.8189
16	AE	(h1 = 48, h2 = 24, bn = 16, $lr = 3 \times 10^{-4}$)	17200	0.7487	0.0058	0.7394	0.7579
16	PCA	(ncomp = 16)	–	0.7502	0.0000	0.7502	0.7502
16	Gaussian	Mahalanobis (ridge-reg. cov)	–	0.6679	0.0000	0.6679	0.6679
16	IForest	(ntrees = 100)	–	0.6880	0.0082	0.6749	0.7011
32	MERA	(L = 4, latent = 32, $lr = 1.5 \times 10^{-4}$)	5477	0.7963	0.0074	0.7845	0.8081
32	AE	(h1 = 48, h2 = 24, bn = 32, $lr = 2 \times 10^{-4}$)	17984	0.7550	0.0053	0.7466	0.7635
32	PCA	(ncomp = 32)	–	0.7764	0.0000	0.7764	0.7764
32	Gaussian	Mahalanobis (ridge-reg. cov)	–	0.6679	0.0000	0.6679	0.6679
32	IForest	(ntrees = 100)	–	0.6957	0.0119	0.6767	0.7147

dimensional approximation, providing a linear proxy for the first local compression step of the hierarchy.

As shown in Table 2, the locality-preserving ordering reduces the mean adjacent-site separation to $\langle \Delta R \rangle = 0.131$, compared with 0.361 for p_T ordering and 0.397 ± 0.004 for

random permutations. More importantly, the same ordering yields the highest retained variance in the first two compression layers: 0.869 and 0.920, compared with 0.857 and 0.890 for p_T ordering, and only 0.721 and 0.830 for the random control. The effect is strongest precisely where the MERA architecture is most local, namely in the first layers that combine neighbouring constituents before any coarse-graining has taken place.

This is also the physics reason to expect the implemented MERA-inspired architecture to benefit from a locality-preserving input representation. Jets are produced by a branching cascade, so short-range angular correlations are not random noise: they encode the local organisation of prongs, collinear radiation, and softer emissions around them. The implemented model is designed to process such a structure hierarchically. Its local maps first rearrange correlations inside neighbouring blocks and then coarse-grain those blocks scale by scale. If neighbouring slots in the input chain correspond to constituents that are genuinely close in angle, then the relevant short-distance correlations are presented where the local maps can exploit them most efficiently. By contrast, under p_T ordering or random permutations, the first layers are forced to mix constituents that are often far apart inside the jet, so the local compression step is less well aligned with the physical organisation of the jet.

The same pattern is reflected in the performance of the implemented MERA-inspired model. Figure 5 summarises the completed ordering study for this model, based on three independently tuned runs for each ordering. The locality-preserving representation gives the highest mean test ROC-AUC at all three latent dimensions, with values of 0.780, 0.777, and 0.774 at $B = 8, 16, 32$, respectively. The corresponding p_T-ordered means are 0.745, 0.745, and 0.756, while random permutations give 0.702, 0.698, and 0.702. Thus, the ordering that makes adjacent input blocks most geometrically local and most compressible is also the ordering that gives the strongest performance for the implemented model on this benchmark. The separation is especially clear at $B = 8$ and $B = 16$, where the locality-ordered chain outperforms both alternatives by several percentage points in AUC.

To test whether this gain is a generic consequence of preprocessing rather than an architecture-specific effect, we repeated the same ordering study for the dense autoencoder and PCA. The dense autoencoder was retuned independently for each ordering and bottleneck using three selection replicates, with four evaluation seeds per replicate. Its mean AUC values for locality ordering are 0.747, 0.745, and 0.754 at $B = 8, 16, 32$, compared with 0.819, 0.832, and 0.831 for p_T ordering and 0.821, 0.835, and 0.834 for random ordering. Thus, the dense model does not share the MERA improvement under locality ordering; it performs similarly for p_T and random orderings and substantially worse for the locality-ordered representation. PCA likewise does not show a uniform locality benefit: its AUC changes from 0.834, 0.836, and 0.619 for p_T ordering to 0.719, 0.750, and 0.776 for locality ordering. As expected for a global linear reconstruction model, PCA is invariant, up to numerical precision, under the fixed random permutations used here and therefore gives the same results for the p_T and random representations. The contrast with MERA shows that the locality gain is not a shared improvement inherited by all models from the preprocessing. It arises specifically when the event-dependent local ordering is paired with an architecture whose first operations act on neighbouring sites.

Taken together, these results support the geometric premise of the construction. The locality-preserving chain is not just a convenient preprocessing choice: it aligns the ordered jet representation with the local compression steps of the MERA hierarchy and thereby places that model in the regime where its multiscale structure is most effective, while providing no corresponding general advantage to the dense or global-linear baselines.

5.3 Effect of the disentanglers

The direct architectural question is whether the trainable disentanglers in MERA improve anomaly detection beyond what is already achieved by the same hierarchy with purely local coarse-graining. To isolate that point as directly as possible, we compare the full MERA at the strongest-compression setting, $B = 8$, with a constrained variant in which every disentangler is fixed to the identity map, denoted MERA($U = I$). This keeps the encoder–decoder hierarchy, latent map, and local isometries unchanged, but removes any learned local basis change before compression. Operationally, this choice makes the MERA mimic the corresponding TTN limit simply by replacing each disentangler with the identity while leaving the rest of the algorithm untouched. In the present implementation, this identity-disentangler limit is mathematically equivalent to the corresponding no-disentangler tree hierarchy, and it also ensures that the ablation introduces as few algorithmic changes as possible: the data representation, connectivity, training objective, optimisation procedure, and evaluation protocol are all kept fixed, so the comparison isolates the effect of the disentangling maps themselves.

The reason to focus on $B = 8$ is also the reason to expect the clearest disentangler effect there. As discussed in Sect. 2, a TTN without disentanglers applies only the local projection $\hat{q}_{\mathrm{TTN}} = W^\top W q$, whereas MERA first rotates the local basis and only then truncates, $\hat{q}_{\mathrm{MERA}} = U^\top W^\top W U q$. The local truncation loss in (2.17) makes the role of U explicit: the disentanglers are meant to reorganise short-range correlations so that the degrees of freedom retained by the isometries capture more of the informative variance before the next coarse-

Table 2 Geometric locality and two-site compressibility of the ordered constituent representation. The retained variance fractions are computed from the optimal three-dimensional approximation to neighbouring two-site blocks on the background sample

Ordering	Mean adjacent ΔR	Retained variance (layer 0)	Retained variance (layer 1)
Locality	0.131	0.869	0.920
p_T	0.361	0.857	0.890
Random	0.397 ± 0.004	0.721	0.830

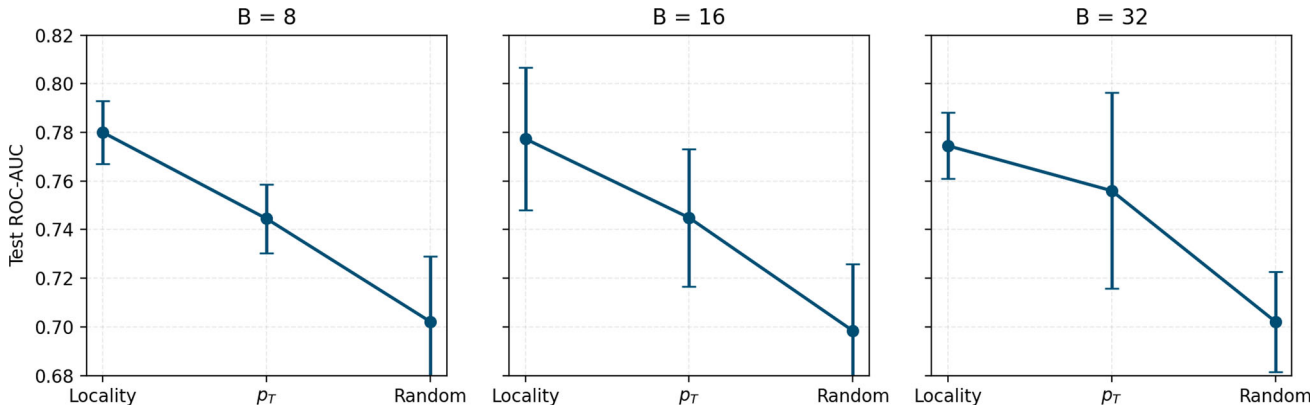


Fig. 5 MERA test ROC-AUC as a function of input ordering for $B = 8, 16, 32$, aggregated over three independently tuned runs for each ordering. Error bars indicate the corresponding confidence intervals. The locality-preserving ordering gives the strongest MERA performance at every latent dimension, consistent with the geometric and compressibility diagnostics reported in Table 2

Table 3 Focused multi-seed disentangler ablation at the strongest-compression setting $B = 8$. MERA($U = I$) denotes the identity-disentangler limit of the same hierarchy, in which the local disentangling maps are fixed rather than learned. Both models are evaluated on the same eight seeds after separate retuning at $B = 8$. The last row gives the mean matched-seed gain of the full MERA over MERA($U = I$)

Model	Params	Test AUC	TPR@10^{-3}
MERA	5021	0.8002 ± 0.0052	1.37×10^{-3}
MERA($U = I$)	1781	0.7922 ± 0.0138	1.10×10^{-3}
Gain of full MERA	–	+0.0081	$+2.71 \times 10^{-4}$

graining step. This mechanism should matter most precisely when the compression bottleneck is tight. At larger latent dimensions, the network can retain more information even without an optimal local rotation, whereas at $B = 8$ the hierarchy is forced to be economical at every stage. This interpretation is consistent with the local-compressibility results of Table 2, which already showed that the locality-preserving representation concentrates relevant structure into the neighbouring blocks on which the MERA tensors act.

The corresponding multi-seed comparison is shown in Table 3. At $B = 8$, the full MERA reaches a mean test AUC of 0.8002 ± 0.0052, compared with 0.7922 ± 0.0138 for MERA($U = I$), and the low-false-positive-rate operating point changes from 1.10×10^{-3} to 1.37×10^{-3} in mean TPR@10^{-3}. Using matched evaluation seeds, the mean differences are $\Delta\text{AUC} = 0.0081$ and $\Delta\text{TPR@}10^{-3} = 2.71 \times 10^{-4}$. The uncertainties on the AUC values overlap, however, so this finite-seed comparison should be interpreted as a positive mean shift rather than a statistically resolved performance improvement.

To determine whether the learned disentanglers nevertheless perform a nontrivial transformation, we also measured their distance from the identity at the end of training. Averaged first over the 45 encoder tensors in each model and then over the same eight evaluation seeds, we find

$$\langle \|U - I_6\|_F \rangle = 3.898 \pm 0.026,$$
$$\left\langle \frac{\|U - I_6\|_F}{\sqrt{6}} \right\rangle = 1.591 \pm 0.011, \tag{5.1}$$

where the quoted spread is the standard deviation across seeds. Thus, the trained local mixing maps do not remain close to the identity: they learn substantial transformations even though the present AUC comparison alone does not establish a statistically significant performance gain. As a

diagnostic of the soft structural constraints, the corresponding residuals are

$$\left\langle \|U^\top U - I_6\|_F \right\rangle = 1.623 \pm 0.110,$$
$$\left\langle \|WW^\top - I_3\|_F \right\rangle = 0.228 \pm 0.034. \quad (5.2)$$

The normalized residuals, obtained by dividing by $\sqrt{6}$ and $\sqrt{3}$, are 0.662 ± 0.045 and 0.131 ± 0.020, respectively. These nonzero values quantify the fact that the penalties are soft: the trained model retains the intended MERA-inspired structure but is not an exactly orthogonal/isometric canonical MERA.

This is the regime in which the MERA-inspired architectural motivation is strongest, and it is precisely there that the local-mixing ablation tests it. The result does not mean that every additional tensor must help in every possible setting. It says something more specific: when the implemented anomaly detector operates in the strongest-compression regime and the locality-preserving input representation is used, the trainable local mixing maps move substantially away from the identity and are accompanied by a positive mean performance shift. The present seed sample does not resolve that shift statistically, but the structural diagnostic shows directly that this component of the implemented MERA-inspired neural architecture is not merely an identity-like or inactive addition.

6 Summary and conclusions

We set out to test two related questions. First, can a MERA-inspired autoencoder serve as an effective background-only anomaly detector for jets when applied directly to an ordered constituent representation? Second, are the architectural ingredients that motivate MERA in the first place, namely locality-aligned hierarchical coarse-graining and local correlation-reorganisation operations, supported by the data? To answer those questions, we construct the MERA-inspired neural encoder–decoder defined in Sect. 4.1 and benchmark it on the Top Quark Tagging Reference Dataset against a dense autoencoder, the corresponding identity-mixing TTN limit, and classical unsupervised baselines. The numerical conclusions of this study apply to that implemented architecture, including its within-pair connectivity, nonlinear activations, untied decoder, and soft structural constraints; they are not claims about canonical MERA in general.

The nominal benchmark gives a clear positive result for the implemented MERA-inspired construction. We find that, for the locality-preserving representation used throughout the paper, this model achieves the largest mean test AUC at all three latent dimensions considered, with values of 0.7959, 0.7885, and 0.7963 at $B = 8, 16, 32$, respectively. The corresponding dense-autoencoder values are 0.7616, 0.7487, and 0.7550, so the mean AUC gains are about 0.034, 0.040, and 0.041. At the same time, the MERA-inspired models use only about one third of the trainable parameters of the dense autoencoder, with roughly $5.0\text{–}5.5 \times 10^3$ parameters versus $1.68\text{–}1.80 \times 10^4$ for AE. The nominal benchmark therefore shows that this particular multiscale neural architecture is both stronger in mean AUC and more economical than the standard nonlinear reconstruction baseline in the setting studied here.

The targeted follow-up studies then clarify why this happens. We showed that the locality-preserving ordering reduces the mean adjacent constituent separation to $\langle \Delta R \rangle = 0.131$, compared with 0.361 for p_T ordering and 0.397 ± 0.004 for random permutations, and that it also gives the largest retained-variance fractions in the first two local compression layers, namely 0.869 and 0.920. This is precisely the regime in which the local MERA tensors act. Consistently, the completed ordering study shows that the same locality-preserving representation gives the strongest MERA performance at every latent dimension, with mean AUC values of 0.780, 0.777, and 0.774, compared with 0.745, 0.745, and 0.756 for p_T ordering and 0.702, 0.698, and 0.702 for random orderings. Through this analysis, we showed that the ordering is not a minor preprocessing detail: it is part of the mechanism that allows the hierarchical compression to align with the physical organisation of the jet.

The disentangler study provides a more specific but equally important result. In the strongest-compression regime, $B = 8$, the full MERA exhibits a positive mean shift relative to the identity-disentangler limit MERA($U = I$), which operationally mimics the corresponding TTN by setting each disentangler to the identity while leaving the rest of the algorithm unchanged. The full MERA reaches 0.8002 ± 0.0052 in mean test AUC, compared with 0.7922 ± 0.0138 for MERA($U = I$), while the mean TPR@10^{-3} changes favourably from 1.10×10^{-3} to 1.37×10^{-3}. This is the regime in which the mathematical argument of Sect. 2 is most relevant: if the compression bottleneck is tight, then the ability of the disentanglers to rotate short-range correlations into the subspace retained by the isometries should matter most. The trained maps move substantially away from the identity and are accompanied by favourable mean shifts in both reported performance measures. The ablation therefore establishes that the disentangling step is an active, nontrivial component of the architecture, although the present eight-seed sample does not statistically resolve its performance advantage.

Thus, we established that the implemented MERA-inspired neural autoencoder is a viable and parameter-efficient architecture for collider anomaly detection on ordered jet constituents; that locality-aware preprocessing is a substantive part of why this multiscale model works; and

that, in the regime where the hierarchy is forced to compress most aggressively, its trainable local mixing maps are demonstrably active and accompany a positive mean performance shift relative to the corresponding identity-map limit. Taken together, these results provide a concrete empirical case that MERA-inspired, multiscale, locality-aware compression can be a useful inductive bias for jet anomaly detection.

Funding E.G. was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the 2219 International Postdoctoral Research Fellowship Program (Grant No. 1059B192400291).

Data Availability Statement This data will be made available on reasonable request. [Authors' comment: The publicly available Top Quark Tagging Reference Dataset analysed in this study is cited in [35] and is available at https://doi.org/10.5281/zenodo.2603256.]

Code Availability Statement This code/software cannot be made available for reasons disclosed in the code availability statement. [Authors' comment: The code used in this study is not publicly available because the research workflows require further consolidation and documentation for public distribution.]

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.
Funded by SCOAP[3].

References

1. V. Belis, P. Odagiu, T.K. Aarrestad, Machine learning for anomaly detection in particle physics. Rev. Phys. **12**, 100091 (2024). https://doi.org/10.1016/j.revip.2024.100091. arXiv:2312.14190 [physics.data-an]
2. G. Kasieczka et al., The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics. Rep. Prog. Phys. **84**(12), 124201 (2021). https://doi.org/10.1088/1361-6633/ac36b9. arXiv:2101.08320 [hep-ph]
3. K. Fraser, S. Homiller, R.K. Mishra, B. Ostdiek, M.D. Schwartz, Challenges for unsupervised anomaly detection in particle physics. JHEP **03**, 066 (2022). https://doi.org/10.1007/JHEP03(2022)066. arXiv:2110.06948 [hep-ph]
4. T. Heimel, G. Kasieczka, T. Plehn, J.M. Thompson, QCD or what? SciPost Phys. **6**(3), 030 (2019). https://doi.org/10.21468/SciPostPhys.6.3.030. arXiv:1808.08979 [hep-ph]
5. M. Farina, Y. Nakai, D. Shih, Searching for new physics with deep autoencoders. Phys. Rev. D **101**(7), 075021 (2020). https://doi.org/10.1103/PhysRevD.101.075021. arXiv:1808.08992 [hep-ph]
6. J.H. Collins, P. Martin-Ramiro, B. Nachman, D. Shih, Comparing weak- and unsupervised methods for resonant anomaly detection. Eur. Phys. J. C **81**, 617 (2021). https://doi.org/10.1140/epjc/s10052-021-09389-x
7. J. Barron, D. Curtin, G. Kasieczka, T. Plehn, A. Spourdalakis, Unsupervised hadronic SUEP at the LHC. JHEP **12**, 129 (2021). https://doi.org/10.1007/JHEP12(2021)129
8. A. Blance, M. Spannowsky, P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches. JHEP **10**, 047 (2019). https://doi.org/10.1007/JHEP10(2019)047
9. B.M. Dillon, L. Favaro, T. Plehn, P. Sorrenson, M. Krämer, A normalized autoencoder for LHC triggers. SciPost Phys. Core **6**, 074 (2023). https://doi.org/10.21468/SciPostPhysCore.6.4.074
10. O. Atkinson, A. Bhardwaj, C. Englert, P. Konar, V.S. Ngairangbam, M. Spannowsky, IRC-safe graph autoencoder for unsupervised anomaly detection. Front. Artif. Intell. **5**, 943135 (2022). https://doi.org/10.3389/frai.2022.943135
11. T. Finke, M. Krämer, A. Morandini, A. Mück, I. Oleksiyuk, Autoencoders for unsupervised anomaly detection in high energy physics. JHEP **06**, 161 (2021). https://doi.org/10.1007/JHEP06(2021)161. arXiv:2104.09051 [hep-ph]
12. T. Finke, M. Hein, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, M. Sommerhalder, Tree-based algorithms for weakly supervised anomaly detection. Phys. Rev. D **109**, 034033 (2024). https://doi.org/10.1103/PhysRevD.109.034033
13. A.J. Larkoski, I. Moult, B. Nachman, Jet substructure at the Large Hadron Collider: a review of recent advances in theory and machine learning. Phys. Rep. **841**, 1–63 (2020). https://doi.org/10.1016/j.physrep.2019.11.001. arXiv:1709.04464 [hep-ph]
14. G. Vidal, Entanglement renormalization. Phys. Rev. Lett. **99**(22), 220405 (2007). https://doi.org/10.1103/PhysRevLett.99.220405
15. G. Vidal, Class of quantum many-body states that can be efficiently simulated. Phys. Rev. Lett. **101**(11), 110501 (2008). https://doi.org/10.1103/PhysRevLett.101.110501
16. E.M. Stoudenmire, D.J. Schwab, Supervised learning with tensor networks. Adv. Neural. Inf. Process. Syst. **29**, 4799–4807 (2016). arXiv:1605.05775 [cs.LG]
17. J.Y. Araz, M. Spannowsky, Quantum-inspired event reconstruction with tensor networks: matrix product states. JHEP **08**, 112 (2021). https://doi.org/10.1007/JHEP08(2021)112. arXiv:2106.08334 [hep-ph]
18. J.Y. Araz, M. Spannowsky, Classical versus quantum: comparing tensor-network-based quantum circuits on Large Hadron Collider data. Phys. Rev. A **106**(6), 062423 (2022). https://doi.org/10.1103/PhysRevA.106.062423. arXiv:2202.10471 [quant-ph]
19. J.Y. Araz, M. Spannowsky, Quantum-probabilistic Hamiltonian learning for generative modeling and anomaly detection. Phys. Rev. A **108**(6), 062422 (2023). https://doi.org/10.1103/PhysRevA.108.062422. arXiv:2211.03803 [quant-ph]
20. V.S. Ngairangbam, M. Spannowsky, M. Takeuchi, Anomaly detection in high-energy physics using a quantum autoencoder. Phys. Rev. D **105**(9), 095004 (2022). https://doi.org/10.1103/PhysRevD.105.095004
21. E. Puljak, M. Pierini, A. Garcia-Saez, Tensor network for anomaly detection in the latent space of proton collision events at the LHC. Mach. Learn. Sci. Technol. **6**(4), 045001 (2025). https://doi.org/10.1088/2632-2153/ae0243. arXiv:2506.00102 [stat.ML]
22. G. Kasieczka et al., The machine learning landscape of top taggers. SciPost Phys. **7**(1), 014 (2019). https://doi.org/10.21468/SciPostPhys.7.1.014. arXiv:1902.09914 [hep-ph]
23. R. Orús, A practical introduction to tensor networks: matrix product states and projected entangled pair states. Ann. Phys. **349**, 117–158 (2014). https://doi.org/10.1016/j.aop.2014.06.013. arXiv:1306.2164 [cond-mat.str-el]
24. A. Cichocki, A.-H. Phan, Q. Zhao, N. Lee, I. Oseledets, M. Sugiyama, D.P. Mandic, Tensor networks for dimensionality reduction and large-scale optimization: part 2 applications and future perspectives. Found. Trends Mach. Learn. **9**(4–5), 431–673 (2017). https://doi.org/10.1561/2200000067

25. E.M. Stoudenmire, Learning relevant features of data with multi-scale tensor networks. Quantum Sci. Technol. **3**(3), 034003 (2018). https://doi.org/10.1088/2058-9565/aaba1a
26. J.A. Reyes, E.M. Stoudenmire, Multi-scale tensor network architecture for machine learning. Mach. Learn. Sci. Technol. **2**(3), 035036 (2021). https://doi.org/10.1088/2632-2153/abffe8
27. D. Liu, S.-J. Ran, P. Wittek, C. Peng, R. Blázquez García, G. Su, M. Lewenstein, Machine learning by unitary tensor network of hierarchical tree structure. New J. Phys. **21**(7), 073059 (2019). https://doi.org/10.1088/1367-2630/ab31ef. arXiv:1710.04833 [stat.ML]
28. Z.-Y. Han, J. Wang, H. Fan, L. Wang, P. Zhang, Unsupervised generative modeling using matrix product states. Phys. Rev. X **8**(3), 031012 (2018). https://doi.org/10.1103/PhysRevX.8.031012
29. S. Cheng, L. Wang, T. Xiang, P. Zhang, Tree tensor networks for generative modeling. Phys. Rev. B **99**(15), 155131 (2019). https://doi.org/10.1103/PhysRevB.99.155131. arXiv:1901.02217 [cond-mat.str-el]
30. A. Edelman, T.A. Arias, S.T. Smith, The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1998). https://doi.org/10.1137/S0895479895290954
31. P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds* (Princeton University Press, Princeton, 2008). https://doi.org/10.1515/9781400830244
32. L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000). https://doi.org/10.1137/S0895479896305696
33. P.H. Schönemann, A generalized solution of the orthogonal procrustes problem. Psychometrika **31**(1), 1–10 (1966). https://doi.org/10.1007/BF02289451
34. K. Batselier, A. Cichocki, N. Wong, MERACLE: constructive layer-wise conversion of a tensor train into a MERA. Commun. Appl. Math. Comput. **3**(2), 257–279 (2021). https://doi.org/10.1007/s42967-020-00090-6
35. G. Kasieczka, T. Plehn, J. Thompson, M. Russel, Top quark tagging reference dataset (2019). https://doi.org/10.5281/zenodo.2603256. Version v0 (2018_03_27)
36. T. Sjöstrand, S. Ask, J.R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C.O. Rasmussen, P.Z. Skands, An introduction to PYTHIA 8.2. Comput. Phys. Commun. **191**, 159–177 (2015). https://doi.org/10.1016/j.cpc.2015.01.024. arXiv:1410.3012 [hep-ph]
37. DELPHES 3 Collaboration, J. de Favereau et al., DELPHES 3, a modular framework for fast simulation of a generic collider experiment. JHEP **02**, 057 (2014). https://doi.org/10.1007/JHEP02(2014)057. arXiv:1307.6346 [hep-ex]
38. M. Cacciari, G.P. Salam, G. Soyez, The anti-k_t jet clustering algorithm. JHEP **04**, 063 (2008). https://doi.org/10.1088/1126-6708/2008/04/063. arXiv:0802.1189 [hep-ph]
39. G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006). https://doi.org/10.1126/science.1127647
40. K. Pearson, LIII. On lines and planes of closest fit to systems of points in space. Philos. Mag. **2**(11), 559–572 (1901). https://doi.org/10.1080/14786440109462720
41. H. Hotelling, Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**(6), 417–441 (1933). https://doi.org/10.1037/h0071325
42. P.C. Mahalanobis, On the generalised distance in statistics. Proc. Natl. Inst. Sci. India **2**, 49–55 (1936)
43. F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in 2008 Eighth IEEE International Conference on Data Mining (IEEE, 2008), pp. 413–422. https://doi.org/10.1109/ICDM.2008.17
44. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in *International Conference on Learning Representations (2015)*. arXiv:1412.6980 [cs.LG]