

All is More: Energy Flow Networks for Jet Quenching

J. A. Gonçalves^{a,b}

^a*Laboratório de Instrumentação e Física Experimental de Partículas (LIP), Av. Professor Gama Pinto 2, 1649-003 Lisboa, Portugal*

^b*Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1609-001, Lisboa, Portugal*

E-mail: jgponcalves@lip.pt

ABSTRACT: Jet quenching, the modification of jets by the quark-gluon plasma in heavy-ion collisions, provides a sensitive probe of the properties of the medium. A jet-by-jet discrimination study between proton-proton and lead-lead jets using energy flow networks and simple baselines, explicitly retaining medium response and underlying event contamination is presented. As references, linear discriminants and neural networks have been trained on high-level observables such as N -subjettiness and energy flow polynomials, including an extended energy flow polynomial set, in order to quantify the achievable performance without constituent-level learning. Energy flow networks are then trained on jet constituents and extended to observable-enhanced energy flow networks that concatenate standardized N -subjettiness and/or energy flow polynomials to the energy flow network latent space. In the realistic scenario, including both underlying event contamination and medium response, observable-enhanced energy flow networks set state-of-the-art performance with receiver operating characteristic area under the curve of $\simeq 0.83$, improving markedly over linear and non-linear baselines and previous work with different architectures. Finally, results from moment energy flow networks, an energy flow network variant that attains comparable area under the curve with a substantially more compact and interpretable latent space are shown. These results establish energy-flow-network-based approaches (especially when enhanced with physics-motivated observables) as practical and robust tools for jet-quenching studies.

Contents

1	Introduction	1
2	Simulation setup	3
3	Results	4
3.1	Observables: distributions and correlations	4
3.2	Linear baselines	9
3.3	Non-linear baselines	9
3.4	Energy flow networks	12
3.5	Observable-enhanced EFNs	14
3.6	Moment EFNs	16
3.7	Overall comparison	17
4	Conclusion	18
A	Stability Plots	22

1 Introduction

Jets produced in ultrarelativistic heavy-ion (HI) collisions traverse and interact with the quark-gluon plasma (QGP), undergoing medium-induced modifications collectively known as jet quenching [1]. These modifications encode space-time information about the plasma and its transport properties, motivating per-jet methods that can identify, jet-by-jet, which reconstructed jets are substantially modified and which remain effectively vacuum-like [1–5]. As a practical proxy for per-jet modification, a binary classifier is trained to separate jets from proton-proton (pp) events — taken as vacuum-like — from jets in PbPb events, which contain an admixture of vacuum-like and quenched jets. This surrogate is not identical to a “quenched vs. unquenched” label, but it is a controlled and measurable step toward jet-by-jet quenching assessments.

Recent machine learning (ML) work has trained supervised classifiers and regressors to expose quenching on a per-jet basis — predicting fractional energy loss and even localizing dijet production points in the nuclear overlap region [6, 7] — while complementary studies build calculable, interpretable discriminants and rank which infrared and collinear (IRC) safe or unsafe features carry the most separating power [8, 9]. Quenched-jet taggers trained on JEWEL vs. PYTHIA quantify discrimination and its degradation under underlying event (UE) and subtraction, and show strong gains when medium response (MR) is present [10, 11]. A machine-learning-driven scan of jet-substructure observables further identifies robust, portable choices for experimental use [12]. To prevent background-driven

shortcuts, an “apples-to-apples” baseline — pp jets embedded in a realistic heavy-ion UE and subtracted identically to PbPb — has been advocated so any residual differences can be cleanly attributed to genuine quenching [13]. For a broader perspective on machine learning in particle physics, see the living review [14].

Realistic interpretation demands that two effects be handled explicitly. MR, i.e. recoil partons induced by jet-medium interactions, populates soft, often large-angle radiation and can become an unphysical shortcut for discrimination if mishandled. In **JEWEL**, recoils are therefore retained and the dedicated thermal-recoil subtraction procedure that has been validated for jet-shape systematics [15] is applied. In parallel, the heavy-ion UE degrades discrimination and can mask genuine quenching features; to avoid background-driven artifacts the “apples-to-apples” prescription [13] has been followed, embedding the pp reference in a realistic HI UE and treating subtraction analogously to the PbPb case so that any residual differences are attributable to quenching rather than UE contamination.

Our approach combines physics-motivated observables with constituent-level learning. We examine several observables, namely N -subjettiness bases with N up to 21, according to Refs. [16, 17] given by:

$$\left\{ \tau_1^{1/2}, \tau_1^1, \tau_1^2, \tau_2^{1/2}, \tau_2^1, \tau_2^2, \dots, \tau_{N-2}^{1/2}, \tau_{N-2}^1, \tau_{N-2}^2, \tau_{N-1}^1, \tau_{N-1}^2 \right\} \quad (1.1)$$

and, to span multiparticle energy-angle correlations with a complete linear IRC-safe basis, energy flow polynomials (EFPs) [18], given by:

$$\text{EFP}_G = \sum_{i_1, \dots, i_N} z_{i_1} \dots z_{i_N} \prod_{(k, \ell) \in G} \theta_{i_k i_\ell}, \quad (1.2)$$

where G is a given graph structure, z_i are transverse-momentum fractions defined as $z_i = p_{T,i}/p_T^{\text{jet}}$ with $\sum_i z_i = 1$, and $\theta_{ij} = \sqrt{(\Delta\eta_i - \Delta\eta_j)^2 + (\Delta\phi_i - \Delta\phi_j)^2}$ are distances in pseudorapidity-azimuth space. An extended variant was also considered with a set of exponents for z and θ . On the constituent-learning side, energy flow networks (EFNs) were adopted, which specialize deep sets to collider data and realize IRC safety at the representation level through transverse-momentum-fraction-weighted, permutation-invariant pooling over latent space variables [19, 20]. These networks can be mathematically described as:

$$\text{EFN} : F \left(\sum_i z_i \Phi(\hat{p}_i) \right), \quad (1.3)$$

where $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^L$ is a learned per-particle mapping into a latent space of dimension L , $F : \mathbb{R}^L \rightarrow \mathbb{R}^k$ is a second neural network acting on the aggregated representation and $\hat{p}_i = (\Delta\eta_i, \Delta\phi_i)$ are constituent cylindrical detector coordinates relative to the jet axis.

Two further EFN variants have been considered. Observable-enhanced EFNs (oEFNs), introduced in this paper, concatenate standardized high-level observables (e.g. N -subjettiness and/or EFPs) to the EFN latent space summary, injecting domain knowledge into an IRC-aware architecture. Moment EFNs (MEFNs) replace the sum in Eq. 1.3 by moment pooling defined as:

$$\text{MEFN} : F_k \left(\langle \Phi^a \rangle_P, \langle \Phi^{a_1} \Phi^{a_2} \rangle_P, \dots, \langle \Phi^{a_1} \dots \Phi^{a_k} \rangle_P \right), \quad (1.4)$$

where k is the highest order moment considered, and a are indices running through all latent space dimensions. This results in a significant increase in the input dimension of the F network, L_{eff} relative to the chosen latent space dimension. In fact:

$$L_{\text{eff}} = \binom{L+k}{k}. \quad (1.5)$$

This generalizes the expectation value of the EFN from a simple average to a vector of the moments of the latent space. These models, exposing these moments of the latent features, yield compact and more interpretable representations while maintaining competitive performance [21]. Where appropriate, weak supervision results via classification without labels (CWoLa) are reported to reflect the experimental reality that definitive “quenching labels” are currently unavailable [22].

Hadron-level jets from **JEWEL** 2.3.0 with MR retained were considered and an experimentally motivated UE, as well as correct UE handling under the apples-to-apples prescription [13] was employed. Linear and non-linear baselines have been established on N -subjettiness, EFPs, and extended EFP sets. EFNs trained on constituents were shown to significantly outperform observable-only baselines in the realistic (UE+MR) regime. Furthermore, oEFNs are shown to deliver further, robust gains. MEFNs that match EFN-level AUC with markedly smaller, interpretable latent spaces, were trained, with their observable-enhanced counterparts (oMEFNs) reaching oEFN-level performance. Throughout, K -fold averaged ROC AUC with cross-validation was used as the primary performance metric and the standard deviation across folds as the primary uncertainty metric. In the UE+MR setting, our best oEFN/oMEFN configurations set state-of-the-art per-jet discrimination consistent with, and improving upon, the apples-to-apples benchmark [13]. Sec. 2 summarizes the simulated samples, recoil subtraction, and UE handling. Sec. 3 presents observable diagnostics, linear/non-linear baselines, EFN/oEFN results, and MEFN studies. Conclusions follow in Sec. 4.

2 Simulation setup

All results use hadron-level jets from **JEWEL** 2.3.0 [23] at $\sqrt{s_{NN}} = 5.02$ TeV with MR and an experimentally motivated UE model, following the same samples and processing as Ref. [13], where the UE is subtracted using iterative constituent subtraction (ICS) [24]. Baselines are constructed in two regimes: (i) “no UE” (optimistic ceiling) and (ii) “with UE contamination” (realistic), and focus on the latter for the main results. Main results come from classifiers trained with identical K -fold splits.

Two samples were generated with identical hard-process settings: pp and PbPb in the $[0, 10]\%$ centrality bin. In both cases, jets are reconstructed at hadron level after hadronization, with **JEWEL** recoils enabled and kept for subsequent subtraction (thermal-recoil subtraction validated for **JEWEL**) [25]. Jets are clustered using the anti- k_t algorithm with jet parameter $R = 0.4$ as implemented in FASTJET [26, 27]. Both charged and neutral particles are included, and no grooming is applied.

For observable baselines, N -subjettiness and EFPs are computed on the subtracted constituents. The extended EFPs use $(\kappa, \beta) \in \{0.5, 1\} \times \{0.5, 1, 2\}$ on the same constituent

set [18]. For EFN-based models, inputs are the constituent transverse-momentum fractions $z_i = p_{T,i}/p_T^{\text{jet}}$ and angular coordinates, namely translated pseudorapidity-azimuth, $(\Delta\eta_i, \Delta\phi_i)$ relative to the jet axis; jets are centered to $(0, 0)$ in (η, ϕ) with $\phi \in (-\pi, \pi]$, and no additional per-jet rescaling beyond z_i normalization is applied.

Evaluation uses K -fold cross-validation with fixed, splits shared across all models and input sets to enable one-to-one comparisons. Unless noted otherwise, headline results correspond to the realistic UE+MR configuration; the no UE configuration is used only as a guiding baseline. Where indicated, weak supervision via CWoLa (see Ref. [22]) is performed by considering PbPb jets as a mixture of a vacuum-like distribution and a quenched-jet distribution, while pp jets with UE contamination are taken to be a pure sample of vacuum-like jets, equivalent to the vacuum-like component of the PbPb case.

All observable inputs are standardized using training-fold statistics. The non-linear model is a small one-layer neural network (NN); oEFN concatenates the standardized observable vector to the EFN latent before F . MEFN replaces sum pooling by moments up to order k ; oMEFN concatenates observables analogously. All networks are optimized using Adam [28], early stopping on validation loss, and fixed seeds. Evaluation uses identical K -fold splits shared across all models. The neural architectures and training hyperparameters used throughout are summarized in Tab. 1. Quoted uncertainties are the standard deviation across all validation folds.

Model	Layers	Activation	Patience	Dropout	L_2
NN	(2048)	ReLU	30	0.20	0.005
EFN	Φ : (100, 100, 126), F : (100, 100, 100)	ReLU	30	0.075	—
oEFN	Φ : (100, 100, 126), F : (100, 100, 100)	ReLU	30	0.20	—
MEFN	Φ : (100, 100, L), F : (100, 100, 100)	ReLU	30	0.20	—
oMEFN	Φ : (100, 100, L), F : (100, 100, 100)	ReLU	30	0.20	—

Table 1: Network architectures and training hyperparameters. “Patience” is the early-stopping patience (epochs without validation loss improvement). For MEFN, L denotes the latent size scanned in Sec. 3.6.

3 Results

3.1 Observables: distributions and correlations

Before turning to physics-informed architectures, baseline performance is established using high-level observable sets (N -subjettiness, EFPs, and an extended EFP basis). These baselines serve as reference points and sanity checks for discrimination power under controlled complexity and are evaluated both without and with UE contamination for completeness.

Figure 1 shows the distributions and pairwise correlations of four N -subjettiness observables, $\tau_1^{1/2}$, τ_1^1 , τ_1^2 , and τ_2^1 , for jets reconstructed in pp and PbPb samples. In the absence of UE contamination, the marginal distributions display slight differences between the two collision systems for most variables, with the notable exception of $\tau_1^{1/2}$, whose distributions

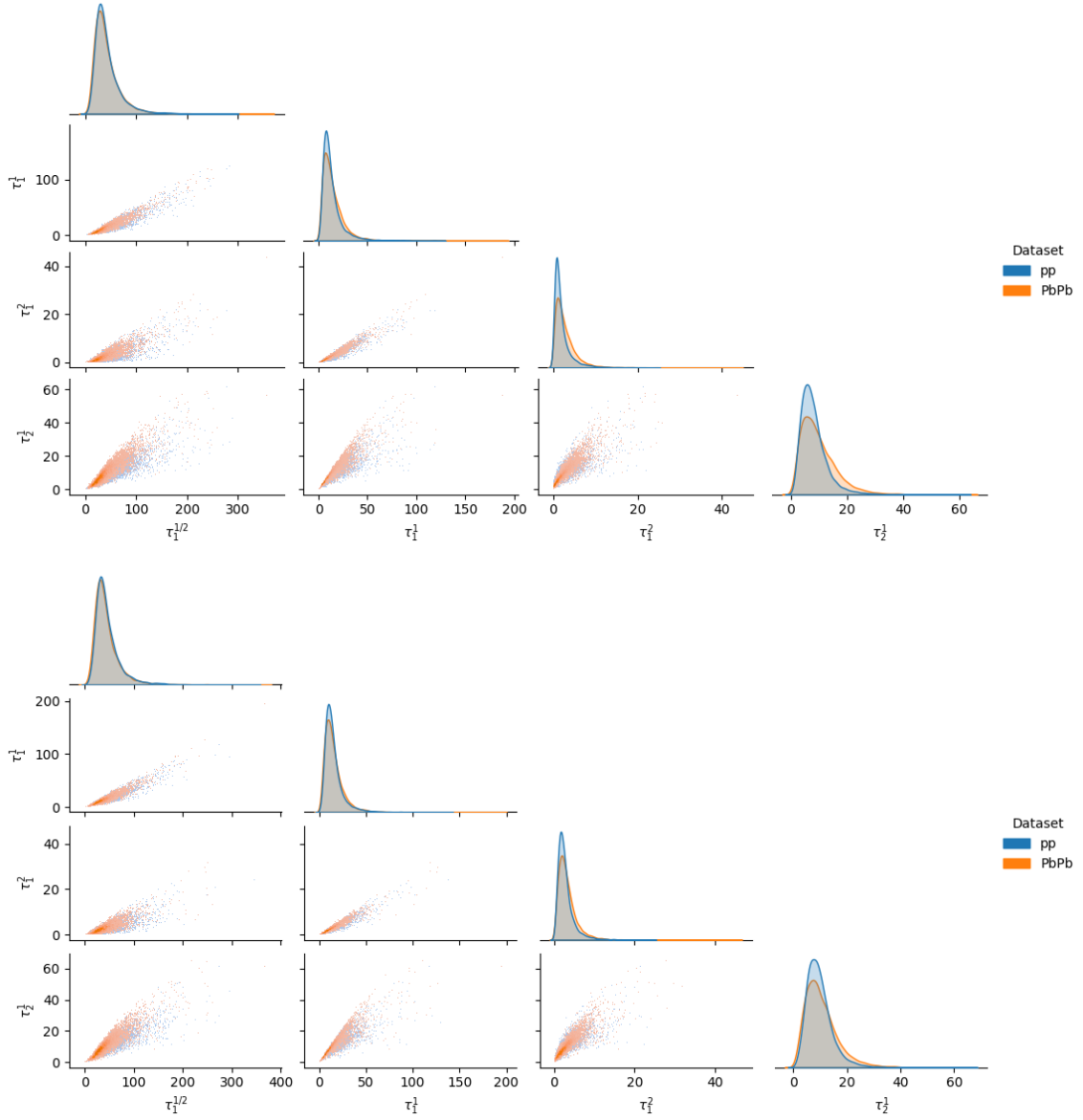


Figure 1: Distributions of 4 N -subjettiness variables and their correlations (pairwise scatter with marginal histograms); (top) without UE contamination; (bottom) with UE contamination.

remain nearly identical. Despite this similarity, correlations of $\tau_1^{1/2}$ with the remaining observables still reveal discernible separation between pp and PbPb jets, a trend that is consistently observed across all correlation panels. Once UE contamination is included, the distributions become more similar across the two systems, with very slightly broader shapes for pp and slightly narrower for PbPb with an accompanying reduced difference in the maximum of the distributions, indicating that background activity partially smears out the quenching-driven differences. This seems to indicate that discriminatory power in

N -subjettiness arises less from individual observables and slightly more from the correlated patterns among them, and that these correlations are less vulnerable to soft contamination, when compared to the absolute distributions.

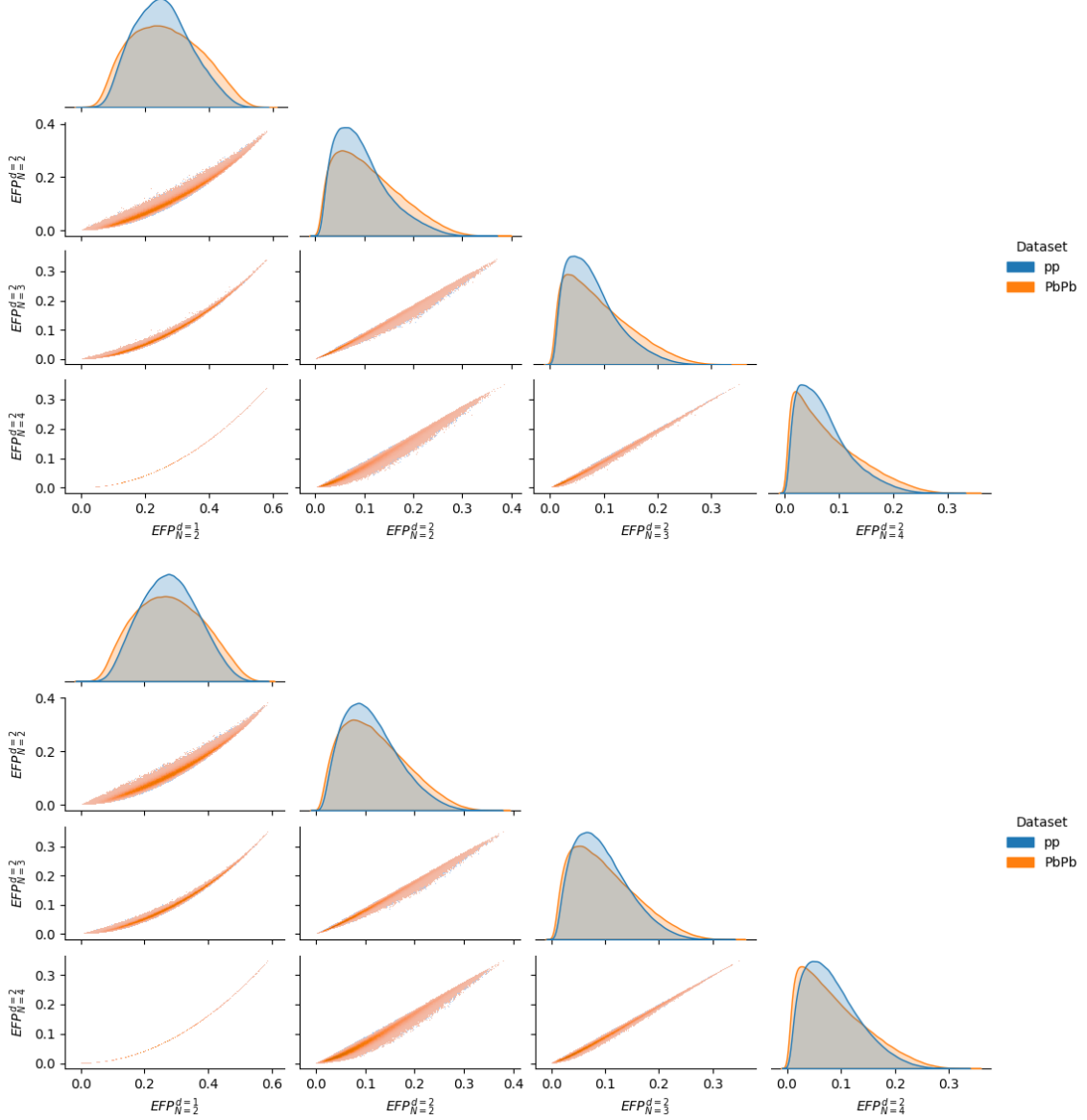
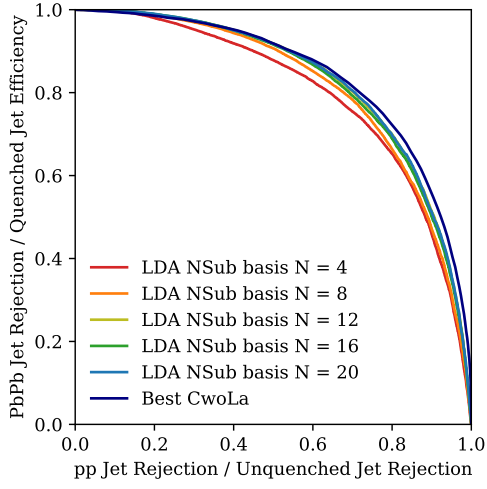
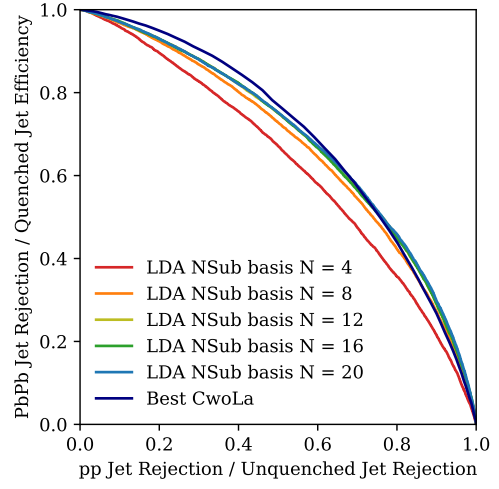


Figure 2: Distributions of 4 EFPs and their correlations (pairwise scatter with marginal histograms); (top) without UE contamination; (bottom) with UE contamination.

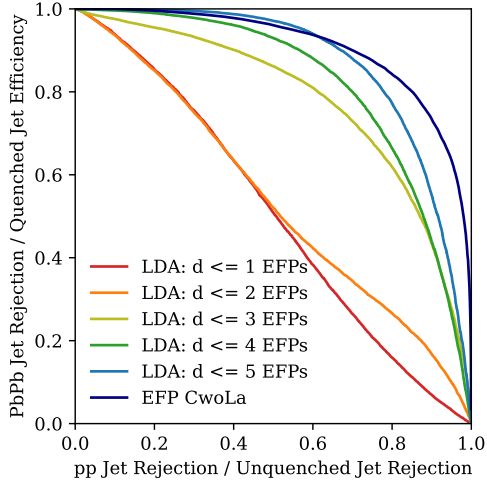
Figure 2 presents the distributions and pairwise correlations of four selected EFPs: $EFP_{N=2}^{d=1}$, $EFP_{N=2}^{d=2}$, $EFP_{N=3}^{d=2}$, and $EFP_{N=4}^{d=2}$, as defined in Eq. 1.2. In contrast to the N -subjettiness case, the marginal distributions of all four observables are shown to be more distinct between the pp and PbPb samples, both with and without UE contamination. The correlations, however, tend to overlap more strongly, suggesting that while each EFP



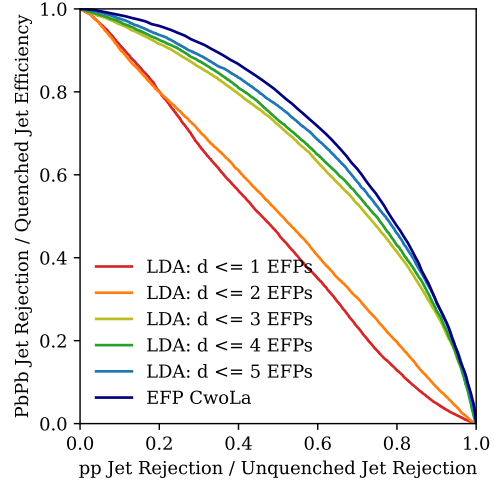
(a) no UE, N -subjettiness



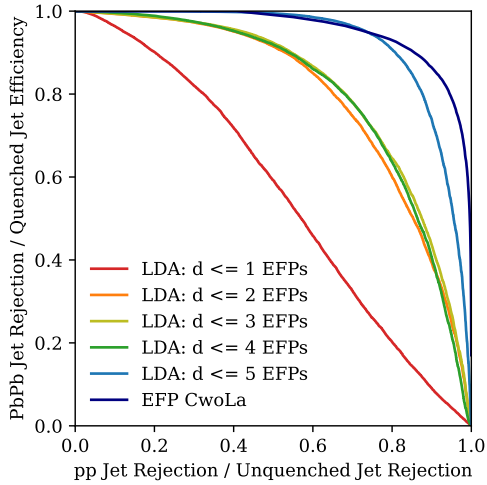
(b) with UE, N -subjettiness



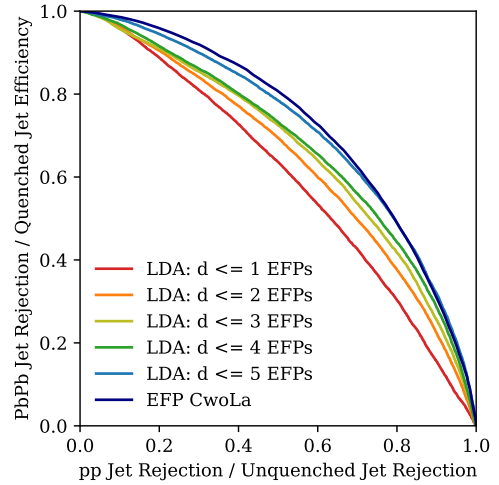
(c) no UE, EFPs



(d) with UE, EFPs

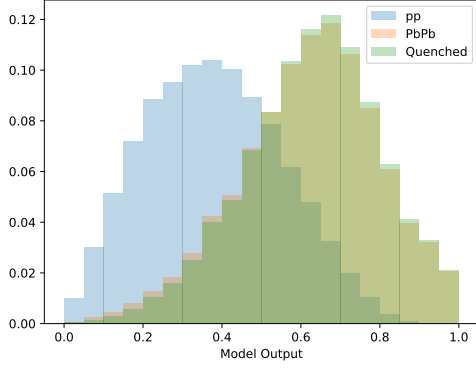


(e) no UE, extended EFPs

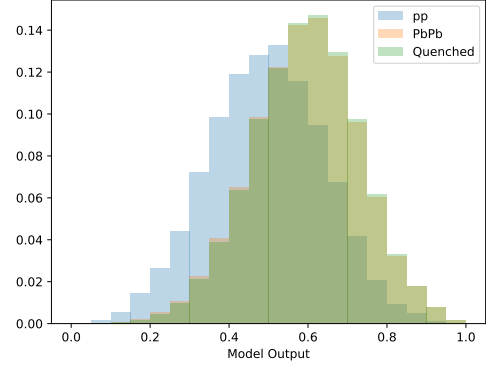


(f) with UE, extended EFPs

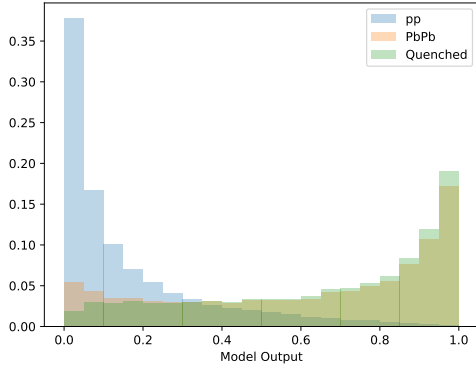
Figure 3: ROC curves for linear models across input sets and UE configurations.



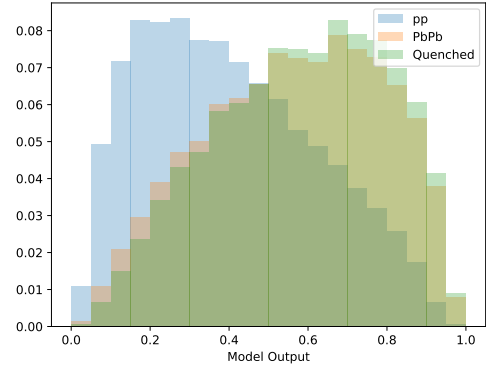
(a) no UE, N -subjettiness



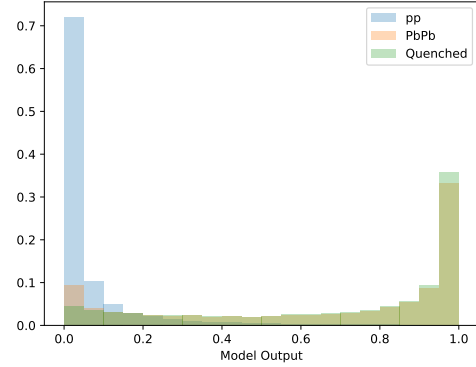
(b) with UE, N -subjettiness



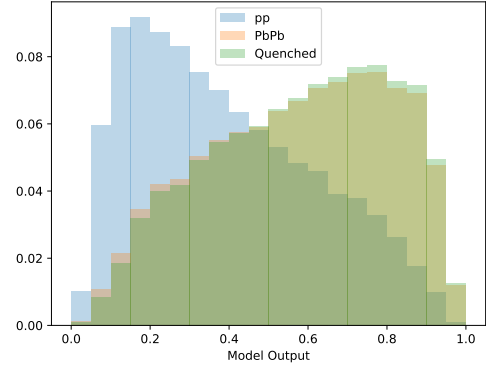
(c) no UE, EFPs



(d) with UE, EFPs



(e) no UE, extended EFPs



(f) with UE, extended EFPs

Figure 4: Classifier output score distributions for linear models across input sets and UE configurations.

encodes quenching sensitivity individually, their joint structures are less distinctive across collision systems, at least at two-dimensional correlations. An interesting feature is the apparent non-linear one-to-one correspondence between $EFP_{N=2}^{d=1}$ and $EFP_{N=4}^{d=2}$, visible

as a narrow curved band in their pairwise correlation panel. This relation reflects the polynomial hierarchy inherent in the EFP construction: for the specific graph choices used here, $EFP_{N=4}^{d=2}$ and $EFP_{N=2}^{d=1}$ are strongly correlated, with the former scaling approximately like the square of the latter.

3.2 Linear baselines

The performance of linear classifiers trained on the different observable bases is summarized in Fig. 3. Linear discriminant analysis (LDA) was considered. For the N -subjettiness basis, the ROC curves in Figs. 3a and 3b show that increasing the number of observables yields only modest gains. In the absence of UE contamination, the presence of MR leads to a reasonably good separation (Fig. 3a); once UE is included (Fig. 3b), the curve moves toward the diagonal and the gap between smaller and larger N -subjettiness bases compresses. This pattern is consistent with the class-score behavior (Figs. 4a and 4b), where UE contamination reduces the mean separation between pp and PbPb scores.

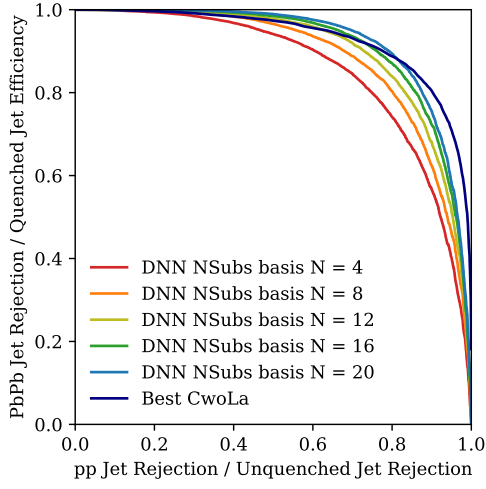
For the EFP basis, Figs. 3c and 3d show decent discrimination in the no UE configuration — slightly better than N -subjettiness at comparable operating points — and a degradation under UE that nevertheless remains above the N -subjettiness curves. In contrast to N -subjettiness, including an increasing number of EFPs produces a clear and significant improvement of the ROC, reflecting the added value of higher-order angular correlations even for a linear decision boundary. The corresponding score distributions (Figs. 4c and 4d) retain a more pronounced valley between classes than in the N -subjettiness case, aligning with the stronger background rejection at fixed signal efficiency.

The extended EFPs, shown in Figs. 3e and 3f, generalize the exponents on momentum fractions and angular distances, $(\kappa, \beta) \in \{0.5, 1\} \times \{0.5, 1, 2\}$, effectively scaling the basis size by a factor of six relative to the $(\kappa, \beta) = (1, 1)$ set. This extension yields the best linear performance across both UE settings: a slight lift over standard EFPs without UE and a persistent advantage with UE. Notably, smaller subsets of the extended EFPs already match or exceed non-extended sets most likely due to their high-dimensionality, indicating that exponent tuning captures discriminating structure that simple basis growth in the $(1, 1)$ family cannot — again reflected in tighter and less-overlapping score densities (Figs. 4e and 4f).

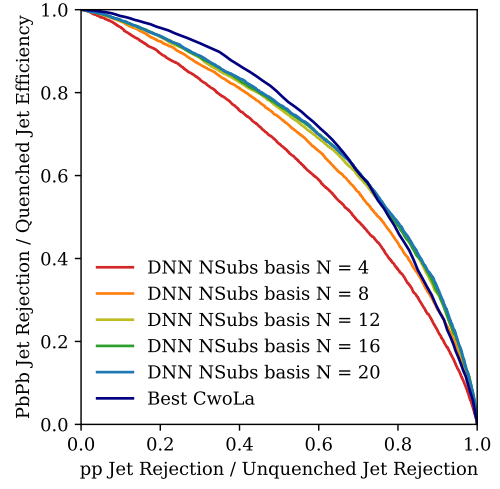
3.3 Non-linear baselines

Linear decision boundaries are replaced with a feed-forward NN, trained on the same observable sets (N -subjettiness, EFPs, extended EFPs). This captures non-linear correlations among observables without leveraging low-level particle information. Models are trained and evaluated with consistent preprocessing and validation, both without and with UE contamination.

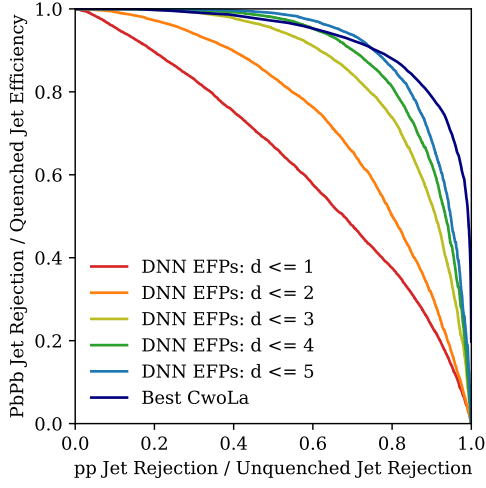
The ROC curves in Fig. 5 show the same ordering as in the linear case: the extended EFP basis provides the strongest discrimination, the standard EFP basis is next, and N -subjettiness performs the weakest of the three, although with comparable results. For N -subjettiness, Figs. 5a and 5b: in the no UE configuration the NN achieves a rather good separation, which is degraded but still above the state-of-the-art boosted decision tree



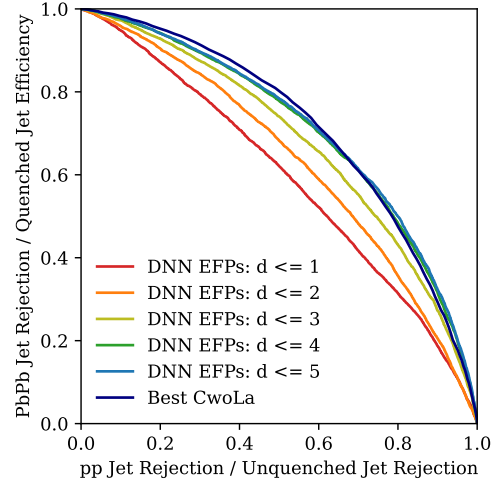
(a) no UE, N -subjettiness



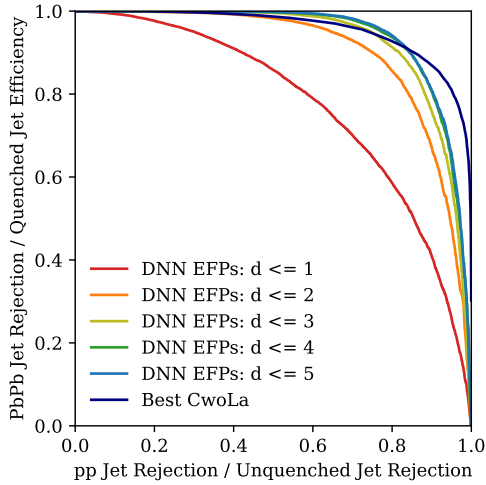
(b) with UE, N -subjettiness



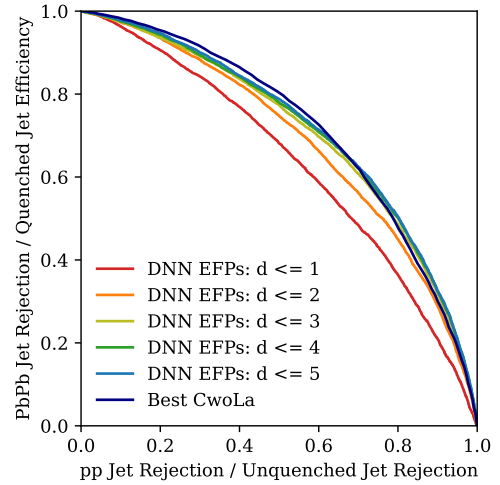
(c) no UE, EFPs



(d) with UE, EFPs

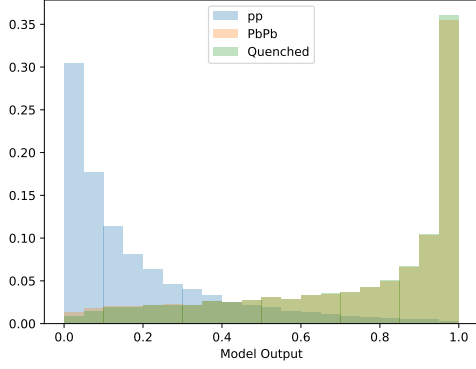


(e) no UE, extended EFPs

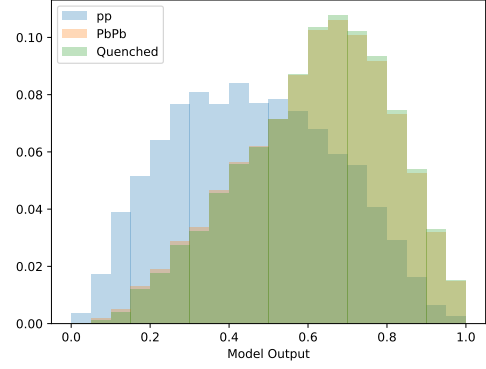


(f) with UE, extended EFPs

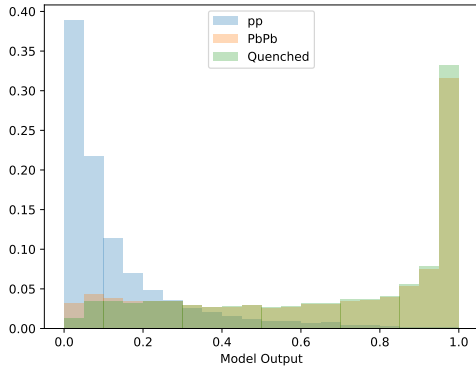
Figure 5: ROC curves for NN classifiers across input sets and UE configurations.



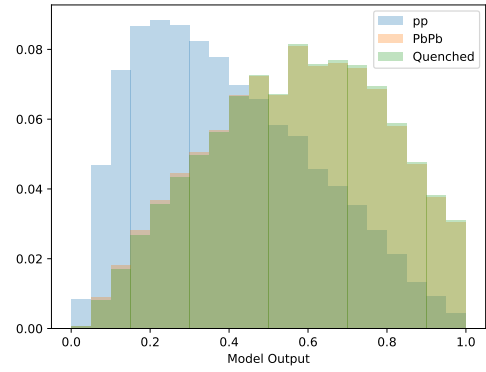
(a) no UE, N -subjettiness



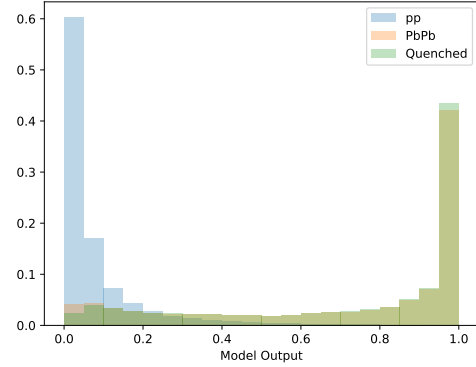
(b) with UE, N -subjettiness



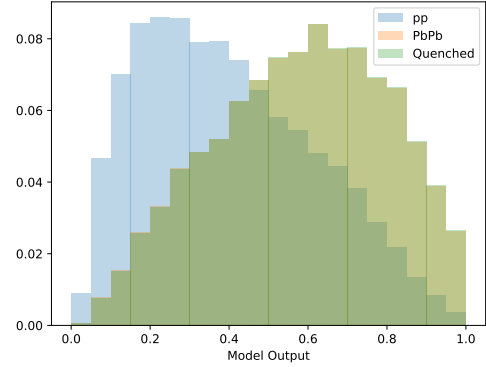
(c) no UE, EFPs



(d) with UE, EFPs



(e) no UE, extended EFPs



(f) with UE, extended EFPs

Figure 6: Classifier output score distributions for NN models across input sets and UE configurations.

results presented in Ref. [13]. Relative to the linear baselines in Fig. 3, the NN uplift is modest but systematic — AUC gains are typically at the level of about 2–9% in the no UE case and about 0.5–2% with UE — while preserving the hierarchy across bases. The

EFP ROCs (Figs. 5c, 5d) retain higher performance, when compared to the N -subjettiness curves, and the extended EFPs (Figs. 5e, 5f) yield the best overall discrimination in both settings.

The class-conditional score distributions in Fig. 6 mirror these trends. For N -subjettiness without UE (Fig. 6a), the pp and PbPb peaks are essentially disjoint with a clear valley, in marked contrast to the substantial overlap seen with LDA (cf. Figs. 4a, 4b). With UE (Fig. 6b), the separation reduces but remains visibly better than the linear case. EFPs (Figs. 6c, 6d) show narrower overlaps than N -subjettiness under both conditions, and the extended EFPs (Figs. 6e, 6f) exhibit the cleanest separation overall, fully consistent with the ROC ordering.

3.4 Energy flow networks

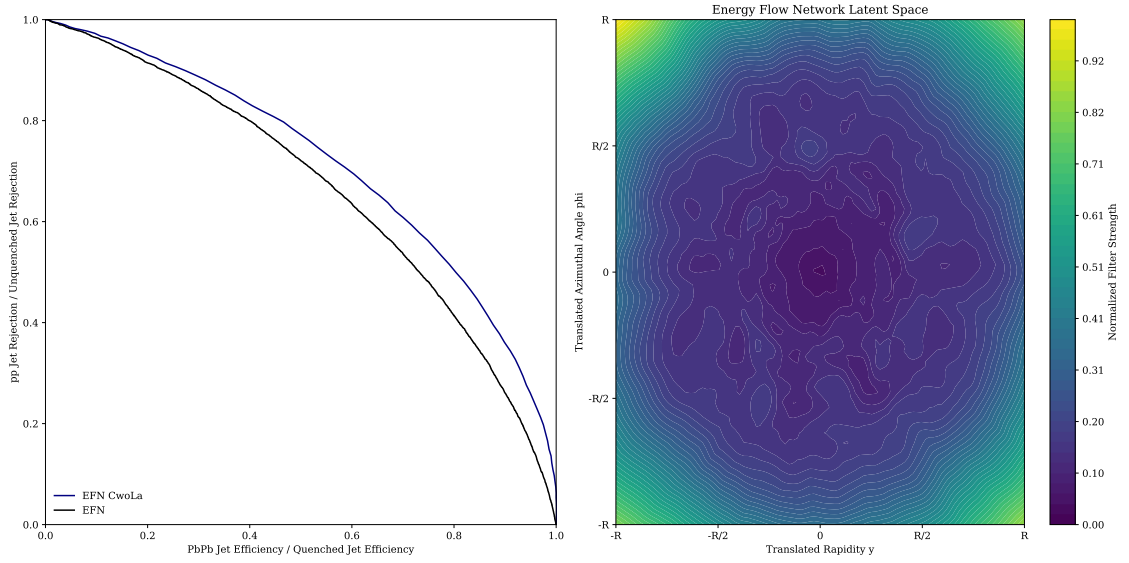


Figure 7: ROC curve for EFN (left) and latent space average visualization (right).

EFNs process sets of jet constituents with permutation invariance and transverse-momentum-weighted latent space, targeting IRC safety at the representation level. EFNs are evaluated with consistent preprocessing, cross-validation, and early stopping, and ROC AUC, latent-space structure, output distributions, learning curves, and cross-validation stability plots are reported.

All EFN results in this subsection include UE contamination. The left panel of Fig. 7 shows the supervised ROC together with the CWoLa ROC on the test set; CWoLa yields a small performance increase. The right panel of Fig. 7 displays the average latent space response in the translated pseudorapidity-azimuth plane, revealing a radially symmetric structure, with fluctuations in the mid-range.

The class-conditional score densities in Fig. 8a exhibit some separation between pp and PbPb jets, with a more pronounced valley and slightly less overlap when compared to the observable baselines. Consistently, the EFN achieves an AUC improvement of about

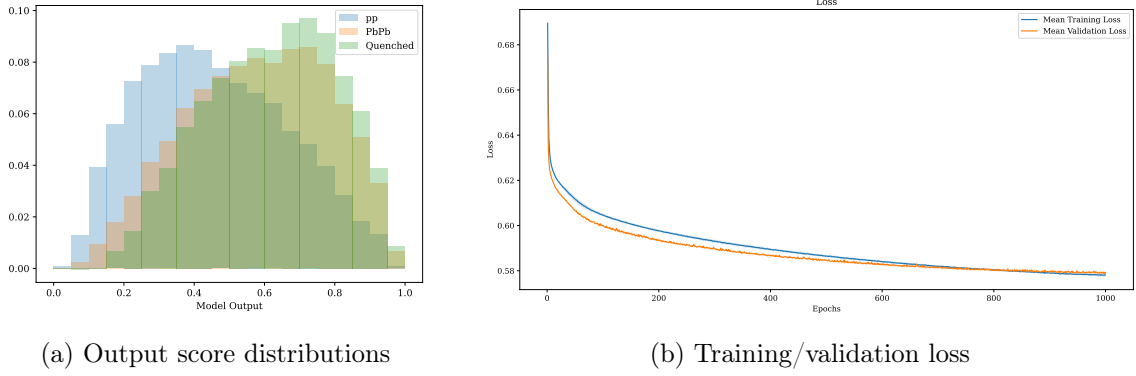


Figure 8: EFN output distributions and learning curve.

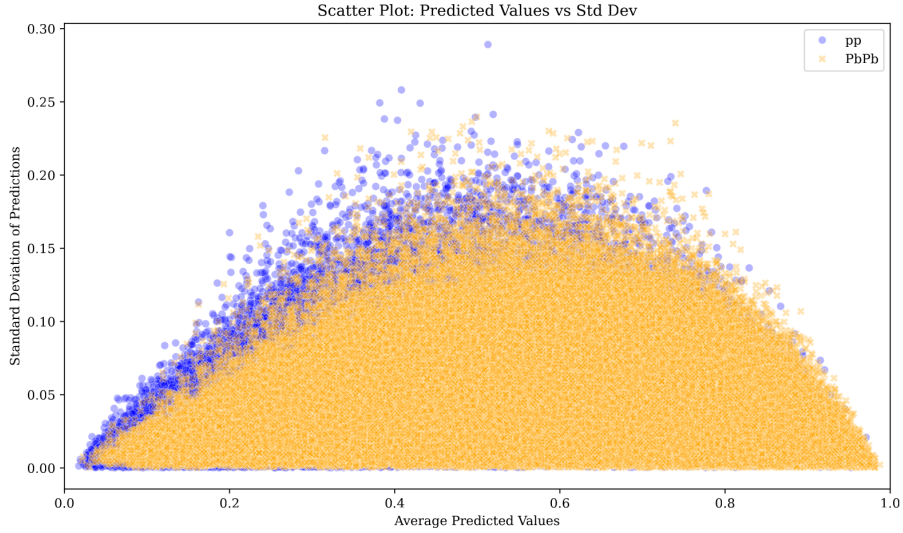


Figure 9: Cross-validation stability: AUC per fold (with mean and variance) or output mean vs. standard deviation.

five percentage points relative to the best linear/non-linear baseline that also includes UE, confirming the benefit of learning directly from constituents within an IRC-aware architecture. The training and validation losses in Fig. 8b decrease smoothly and plateau with a negligible gap; there are no signs of underfitting or overfitting, and early-stopping behavior is consistent with stable convergence.

To assess training stability across folds, Fig. 9 presents the scatter of the fold-averaged classifier output versus its standard deviation across folds: the taco plot. The dispersion is smallest near scores close to 0 or 1 and remains below about 0.25 through the central region, indicating limited fold-to-fold variability. This level of stability and performance sets a new baseline for the observable-augmented variants considered next.

3.5 Observable-enhanced EFNs

While EFNs encode the full set of jet constituents in an IRC-safe way, they do not directly incorporate high-level observables known to capture important aspects of quenching. Observable-enhanced EFNs (oEFNs) address this by concatenating global observables such as N -subjettiness or EFPs to the latent representation learned by the EFN. This hybrid approach allows the model to benefit from explicit, physics-motivated features, while keeping permutation invariance if the features also have this property. This subsection gauges the gains from such enhancements.

The aggregated ROC comparison in Fig. 10 shows a monotonic improvement as observables are concatenated to the EFN latent space. Relative to the vanilla EFN (AUC $\simeq 0.765$), extending the latent space with either N -subjettiness or EFPs lifts the AUC to about 0.805, using extended EFPs yields about 0.810, combining N -subjettiness with standard EFPs reaches 0.821, and combining N -subjettiness with extended EFPs attains about 0.827. It is worth underlining that this is under operation in the UE-contaminated setting with MR, and hence these results quantify real gains in a realistic regime.

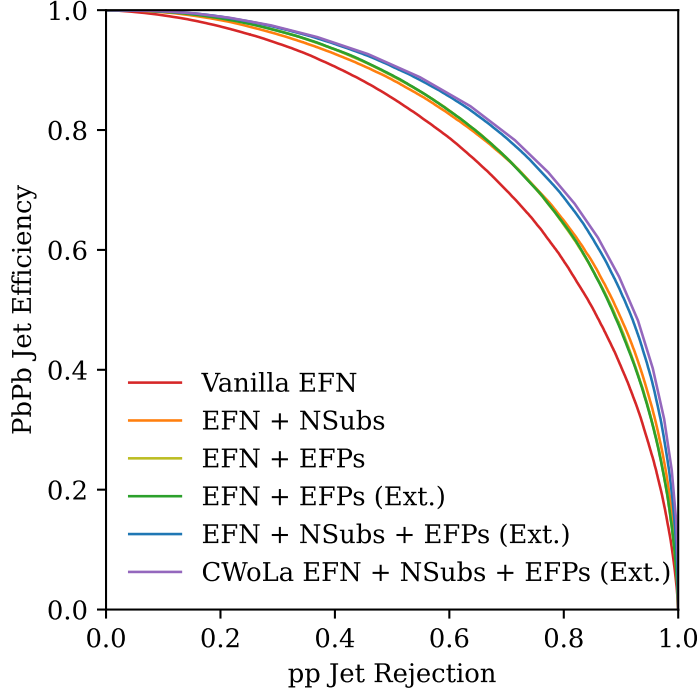
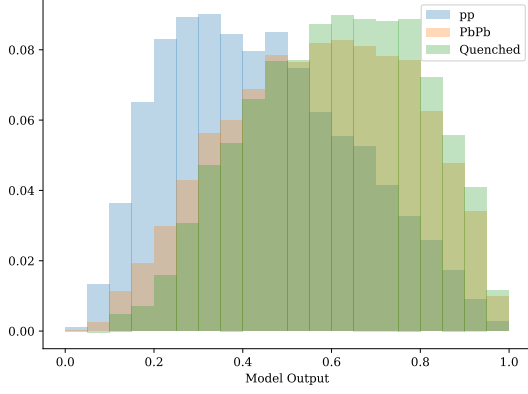


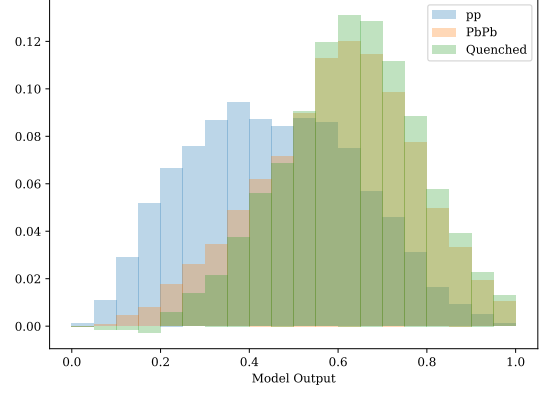
Figure 10: ROC curves for oEFN classifiers under different observable augmentations, with and without UE contamination.

The class-conditional score densities in Fig. 11 reflect the same ordering. Comparing panels 11a, 11b, and 11c, the pp and PbPb peaks separate progressively as the AUC rises, with a slightly wider valley. The extended EFP augmentation (Fig. 11d) adds a further, smaller improvement. The strongest separation appears when combining observables (Figs. 11e and 11f); in the best case the PbPb distribution is substantially shifted to the

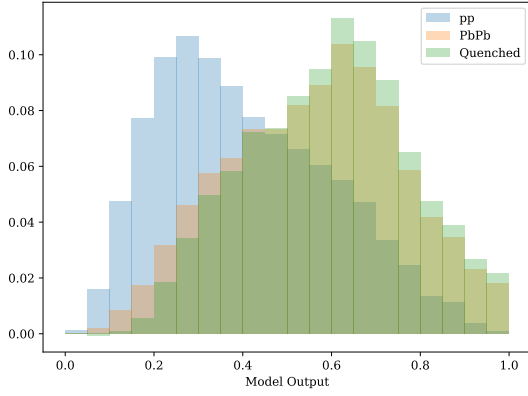
right relative to the vanilla configuration, consistent with the ROC uplift and improved background rejection.



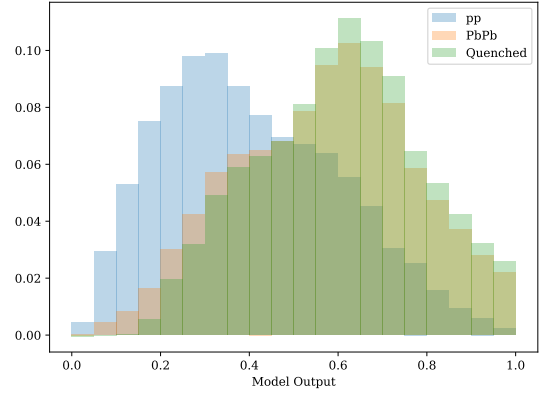
(a) Vanilla



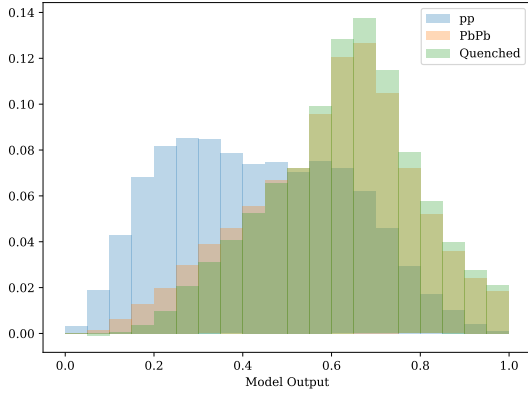
(b) N -subjettiness



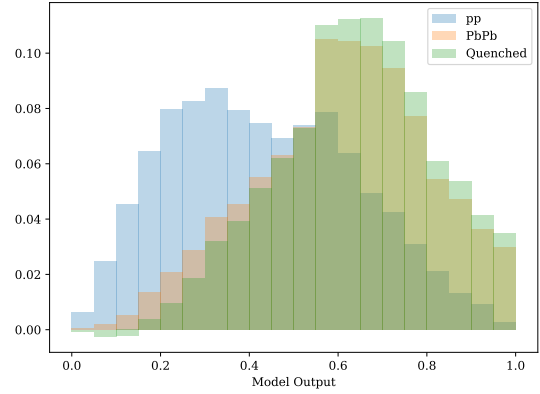
(c) EFPs



(d) Extended EFPs



(e) N -subjettiness + EFPs



(f) N -subjettiness + Extended EFPs

Figure 11: Classifier output score distributions for oEFNs under different input augmentations.

3.6 Moment EFNs

Moment energy flow networks (MEFNs) extend EFNs by passing higher-order moments of the learned latent coordinates to the classifier (an operation known as moment pooling), enabling EFN-level discrimination with far smaller latent space size. A scan across order k and latent dimension L under UE contamination is presented and summarizes the landscape in a two-dimensional AUC map (Fig. 12). The grid exhibits clear structure across (k, L) rather than random variation, delineating regions where compact models already approach the best observed performance.

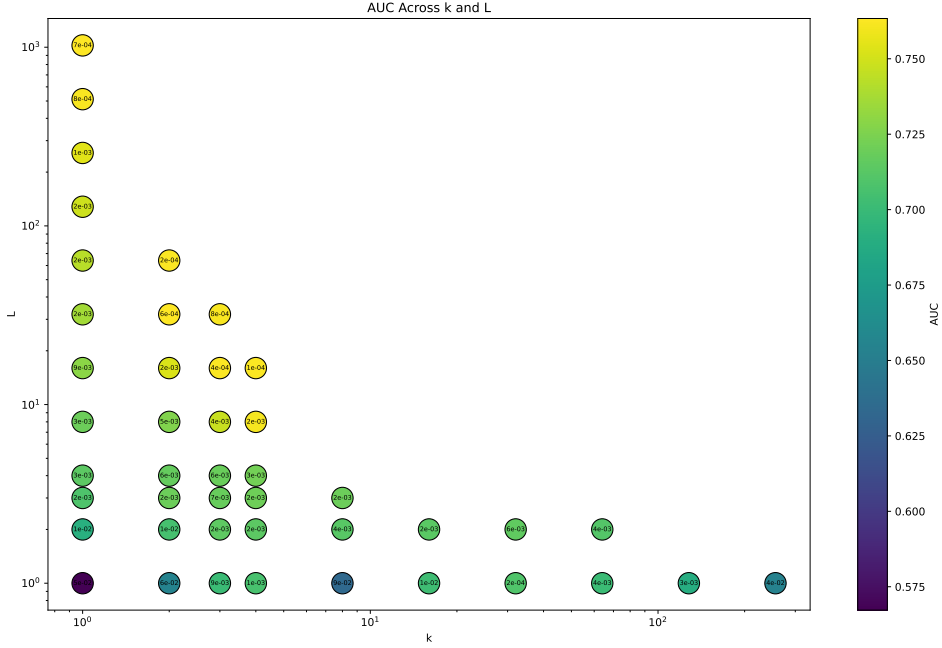


Figure 12: Two-dimensional map of AUC values across latent space dimension L and MEFN order k .

Scaling trends are explicit in the one-dimensional sweeps. At fixed order, increasing the latent size improves AUC up to a visible plateau; conversely, at fixed latent size, raising the order yields gains that taper as k grows (Figs. 13). These curves make the diminishing-returns trade-off transparent and guide the choice of a compact working point.

Re-expressing performance against the effective latent dimension L_{eff} reduces the spread across configurations relative to plotting against L alone (Fig. 14). The corresponding maps of L_{eff} over (k, L) (Fig. 15) further show how higher orders with large latent space dimension L quickly scale L_{eff} .

In practice, this yields efficient choices that retain EFN-level AUC with markedly fewer channels (as reflected in the scan table): representative points such as $k = 2, L = 32$ and $k = 3, L = 16$ lie near the efficient frontier and match baseline EFN performance within

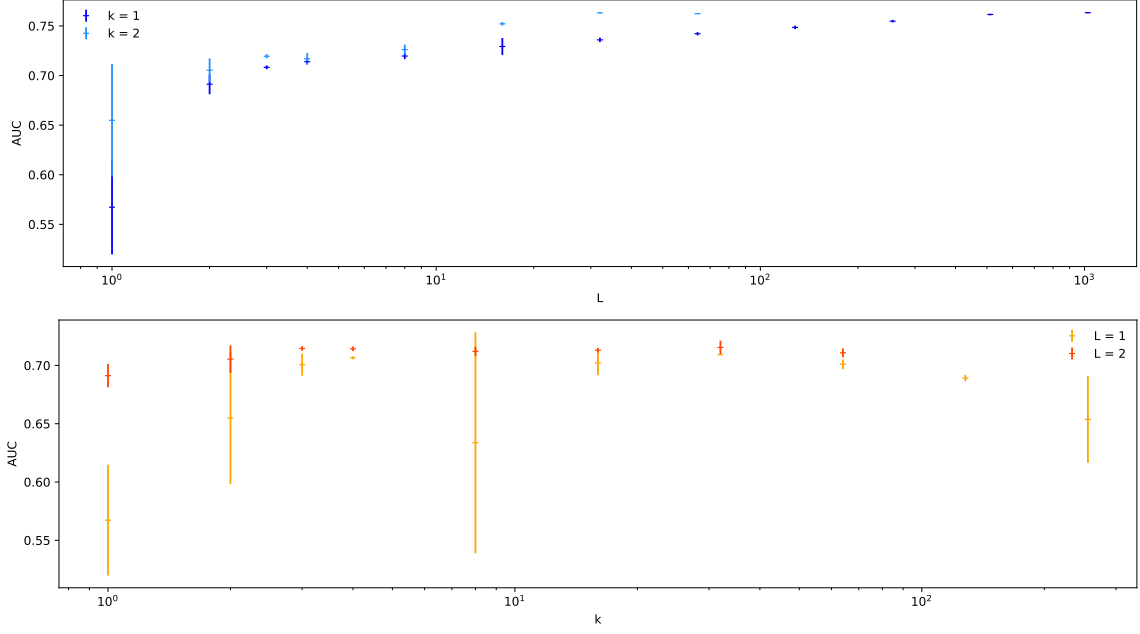


Figure 13: Evolution of $1 - \text{AUC}$ with latent space dimension (top) and MEFN order k (bottom).

uncertainties while reducing model size and enhancing interpretability. The AUC for the complete set of runs is consolidated in Tab. 4a alongside a similar table for the observable-enhanced version, Tab. 4b.

3.7 Overall comparison

Tables 2–3 consolidate the performance of all tested models in terms of ROC AUC, except for the moment variants. Results without UE contamination are presented in Tab. 2, while results with UE contamination are presented in Tab. 3. In the realistic setting, EFNs improve substantially over observable-only baselines, and oEFNs provide the best overall AUCs, reaching $\simeq 0.83$. A detailed MEFN/oMEFN scan is presented in Tabs. 4a–4b and shows many compact configurations achieving oEFN-level performance, clarifying the trade-off between interpretability/compactness and raw capacity.

Model	AUC
LDA EFP	0.8675
LDA EFP Ext.	0.9234
LDA N -subjettiness	0.8314
NN EFP	0.9067
NN EFP Ext.	0.9420
NN N -subjettiness	0.9232

Table 2: Model comparison (supervised) without UE contamination.

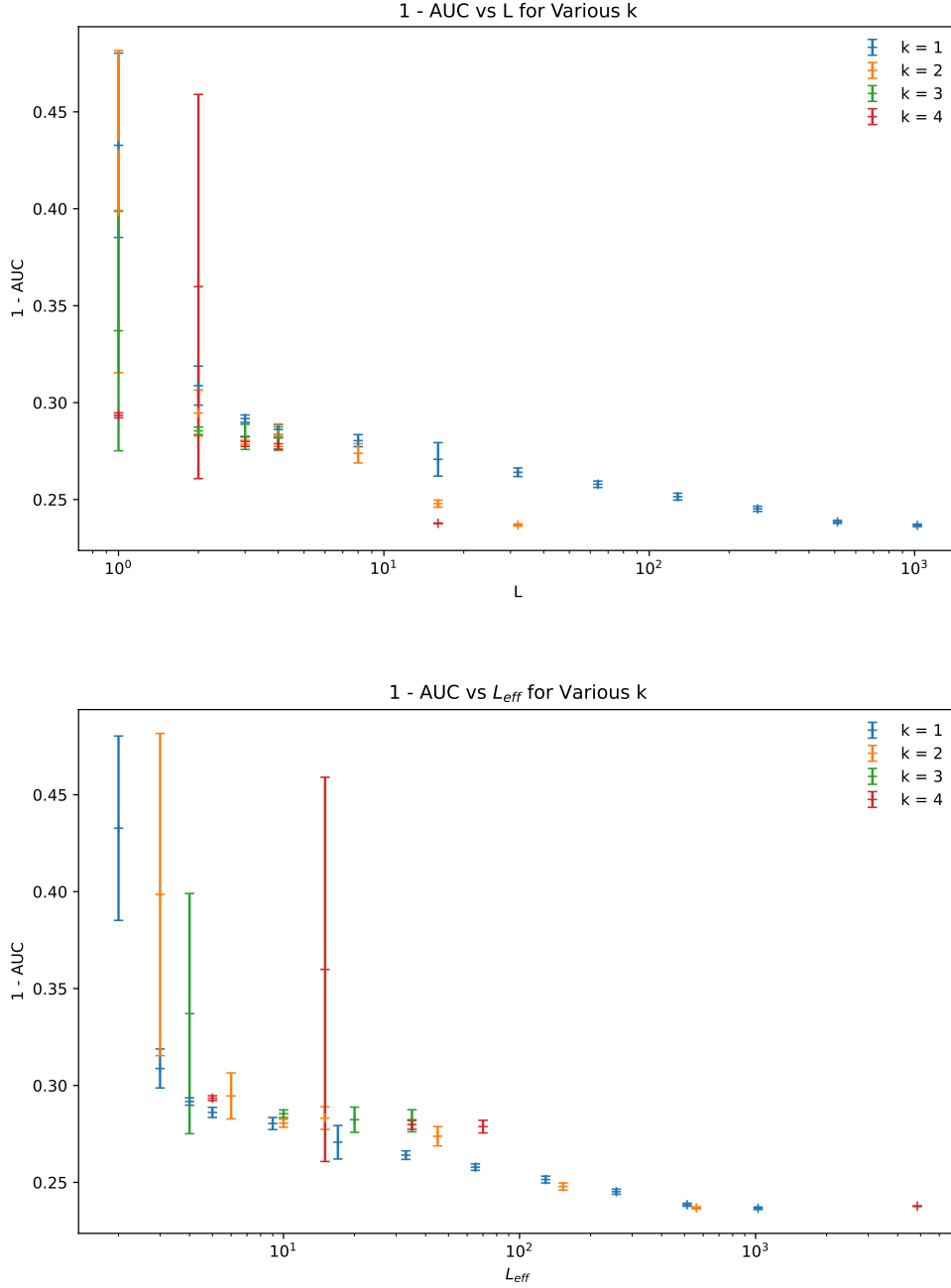


Figure 14: $1 - \text{AUC}$ vs. latent space dimension (top) and effective latent dimension (bottom).

4 Conclusion

A per-jet discrimination study between pp and PbPb jets in a realistic (UE+MR) setting, combining physics-motivated observables with constituent-level, IRC-aware representations

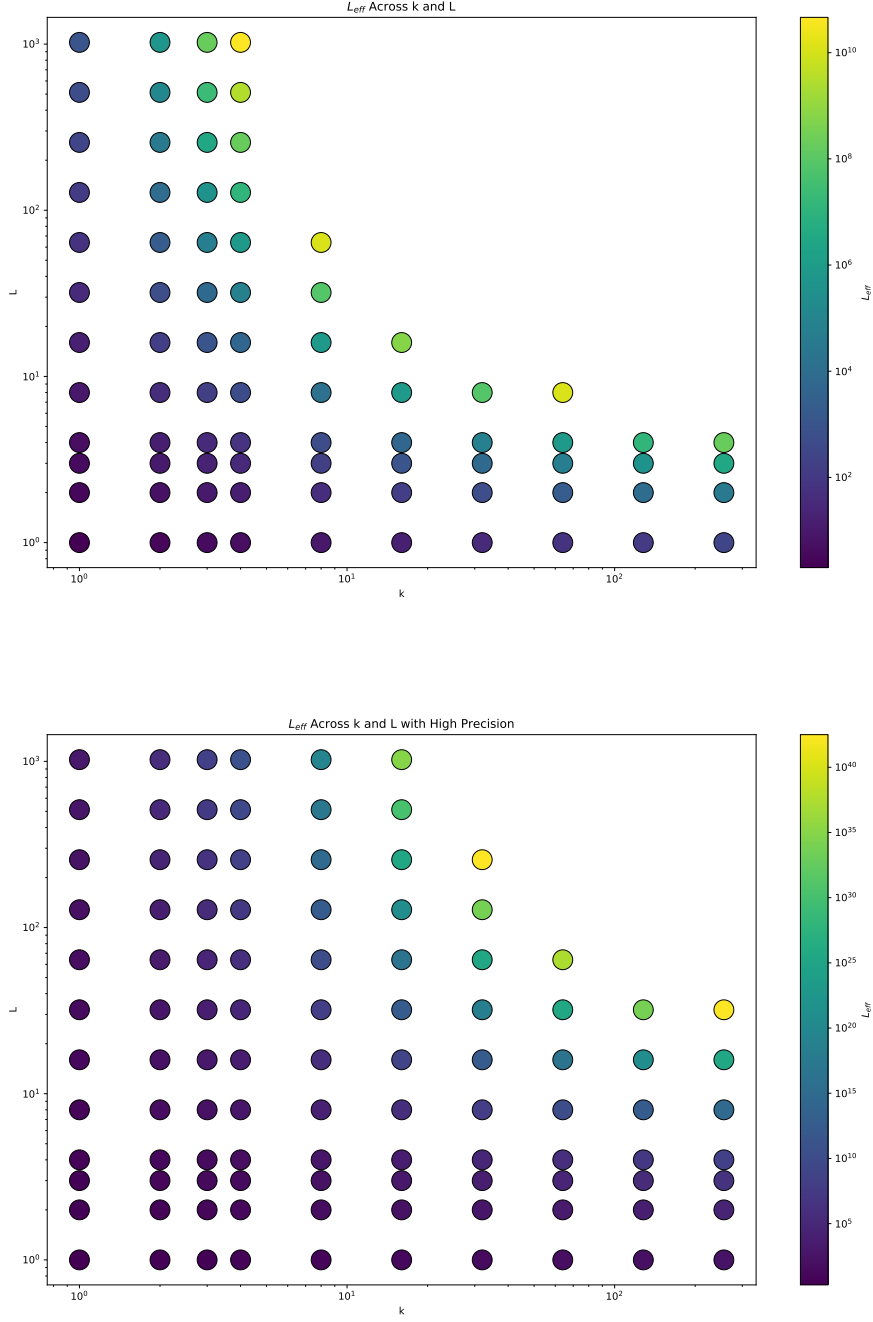


Figure 15: Effective latent space dimension across MEFN order k (top) and latent space dimension L (bottom).

is presented. Linear and shallow non-linear baselines on N -subjettiness and (extended) EFPs establish a clear hierarchy, with extended EFPs performing best among high-level

Model	AUC
LDA EFP	0.6964
LDA EFP Ext.	0.7132
LDA N -subjettiness	0.6900
NN EFP	0.7142
NN EFP Ext.	0.7176
NN N -subjettiness	0.7075
EFN	0.7651 ± 0.0004
EFN + EFP	0.8050 ± 0.0002
EFN + EFP Ext.	0.8104 ± 0.0004
EFN + N -subjettiness	0.8053 ± 0.0003
EFN + EFP + N -subjettiness	0.8206 ± 0.0001
EFN + EFP Ext. + N -subjettiness	0.8265 ± 0.0011

Table 3: Model comparison (supervised) with UE contamination. Uncertainties denote the cross-validation standard deviation across all folds.

observable-only models. EFNs trained on constituents deliver a marked uplift over these baselines. Augmenting EFNs with standardized high-level observables (oEFNs) yields further, robust gains, achieving ROC AUC $\simeq 0.83$ with cross-validation stability.

Moment EFNs (MEFNs), which make use of moment pooling (defined in Eq. 1.4), have been trained and shown to match EFN-level AUC in the jet quenching problem. Their observable-enhanced counterparts (oMEFNs) reach oEFN-level performance across a broad (k, L) band, providing a practical route to smaller and more interpretable latent spaces without sacrificing accuracy.

Several aspects are natural targets for future work. On the physics side: centrality, jet- p_T and jet- R dependence would further complete the study for the usability of the approach across these observables. More importantly, studies across different physics models would allow a Pareto Frontier study of the jet quenching problem, such as in Ref. [29]. On the learning side: comparisons to alternative set/graph architectures (e.g. ParticleNet/transformers), calibration and uncertainty estimation, and systematic robustness to UE modeling and subtraction choices. Finally, applying (o)EFN/(o)MEFN weak supervision directly to data, in the apples-to-apples setup, would close the loop to experimental deployment [13].

Taken together, these results establish EFN-based approaches — especially oEFNs and oMEFNs — as practical, robust tools for jet-quenching classification in realistic UE+MR environments. The main message of this paper is that combining constituent-level information with high-level observables is the best approach to the jet quenching problem: “all is more”. Crucially, by operating on constituents they capture non-linear, multi-particle structure that escapes simple correlations among a handful of high-level observables, and they do so in a compact and relatively more interpretable form [13].

k	L	AUC $\pm$ CV std.
1	1024	0.7634 ± 0.0007
2	32	0.7632 ± 0.0006
3	16	0.7624 ± 0.0004
2	64	0.7623 ± 0.0002
4	16	0.7623 ± 0.0001
1	512	0.7615 ± 0.0008
4	8	0.7608 ± 0.0016
1	256	0.7548 ± 0.0013
2	16	0.7521 ± 0.0018
1	128	0.7485 ± 0.0017
1	64	0.7421 ± 0.0017
1	32	0.7359 ± 0.0022
1	16	0.7293 ± 0.0086
2	8	0.7262 ± 0.0050
4	4	0.7212 ± 0.0033
4	3	0.7201 ± 0.0025
1	8	0.7196 ± 0.0031
8	3	0.7193 ± 0.0020
3	3	0.7177 ± 0.0065
3	2	0.7146 ± 0.0020
4	2	0.7142 ± 0.0020
16	2	0.7131 ± 0.0020
8	2	0.7121 ± 0.0039
64	2	0.7108 ± 0.0038
32	1	0.7092 ± 0.0002
1	3	0.7083 ± 0.0019
4	1	0.7066 ± 0.0013
2	2	0.7054 ± 0.0119
16	1	0.7021 ± 0.0105
64	1	0.7009 ± 0.0042
3	1	0.7006 ± 0.0094
1	2	0.6912 ± 0.0101
128	1	0.6891 ± 0.0028
2	1	0.6549 ± 0.0567
256	1	0.6537 ± 0.0373
8	1	0.6337 ± 0.0948
1	1	0.5673 ± 0.0476

(a) MEFN results

k	L	AUC $\pm$ CV std.
64	1	0.8277 ± 0.0005
64	2	0.8277 ± 0.0004
16	1	0.8277 ± 0.0005
32	2	0.8276 ± 0.0005
3	1	0.8276 ± 0.0006
128	1	0.8275 ± 0.0007
1	2	0.8275 ± 0.0006
1	3	0.8275 ± 0.0005
1	4	0.8275 ± 0.0008
256	1	0.8275 ± 0.0006
2	2	0.8275 ± 0.0007
3	2	0.8275 ± 0.0004
32	1	0.8275 ± 0.0005
4	2	0.8275 ± 0.0008
2	1	0.8274 ± 0.0008
8	1	0.8274 ± 0.0008
16	2	0.8273 ± 0.0005
1	1	0.8273 ± 0.0006
1	32	0.8273 ± 0.0007
1	8	0.8273 ± 0.0008
2	3	0.8273 ± 0.0007
2	8	0.8273 ± 0.0007
3	3	0.8273 ± 0.0009
1	128	0.8272 ± 0.0005
2	16	0.8272 ± 0.0007
2	4	0.8272 ± 0.0007
8	2	0.8272 ± 0.0011
4	3	0.8272 ± 0.0010
1	16	0.8271 ± 0.0012
2	32	0.8271 ± 0.0006
3	4	0.8271 ± 0.0010
1	256	0.8270 ± 0.0007
1	64	0.8270 ± 0.0010
4	4	0.8270 ± 0.0010
4	1	0.8269 ± 0.0009
2	64	0.8268 ± 0.0008
1	512	0.8267 ± 0.0009
1	1024	0.8265 ± 0.0013

(b) oMEFN results

Table 4: AUC across (k, L) settings for MEFN and oMEFN. Uncertainties denote the cross-validation standard deviation across folds.

Acknowledgments

The author thanks José Guilherme Milhano and André Cordeiro for helpful discussions and for support by European Research Council under project ERC-2018-ADG-835105 YoctoLHC, as well as by FCT under contract PRT/BD/151554/2021.

A Stability Plots

The training and validation losses for the non-linear baselines are shown in Fig. 16 and indicate stable convergence across all six configurations. For N -subjettiness (Figs. 16a, 16b), EFPs (Figs. 16c, 16d), and extended EFPs (Figs. 16e, 16f), the curves decrease smoothly and plateau with no visible train-validation gaps, showing no signs of underfitting or overfitting. The very slightly higher validation floor with UE is consistent with the harder classification task under background contamination.

Training and validation losses for the oEFN models are shown in Fig. 17 and indicate stable optimization across all six configurations. The vanilla EFN curves (Fig. 17a) and the augmented variants (Figs. 17b–17f) decrease smoothly and plateau with negligible train-validation gaps. All losses show a non-existent or small gap between validation and training loss, showing no significant evidence of underfitting or overfitting.

Cross-validation stability, for the oEFN models, summarized by the taco plots in Fig. 18, shows small fold-to-fold variability near classifier outputs close to 0 or 1 for all cases. Around intermediate scores, augmenting with N -subjettiness and/or standard EFPs reduces the standard deviation relative to the vanilla EFN (cf. 18a vs. 18b, 18c, 18e), indicating more stable decision boundaries. In contrast, using extended EFPs increases the variability in the central region for some data points (Figs. 18d and 18f), in some ranges exceeding the vanilla level; this is consistent with expectation and suggests mild sensitivity to fold composition, even though overall AUC improves.

References

- [1] S. Cao, A. Majumder, R. Modarresi-Yazdi, I. Soudi, and Y. Tachibana, *Jet quenching: From theory to simulation*, *Int. J. Mod. Phys. E* **33** (2024), no. 08 2430002, [[arXiv:2401.10026](#)].
- [2] W. Busza, K. Rajagopal, and W. van der Schee, *Heavy Ion Collisions: The Big Picture, and the Big Questions*, *Ann. Rev. Nucl. Part. Sci.* **68** (2018) 339–376, [[arXiv:1802.04801](#)].
- [3] L. Apolinário, Y.-J. Lee, and M. Winn, *Heavy quarks and jets as probes of the QGP*, *Prog. Part. Nucl. Phys.* **127** (2022) 103990, [[arXiv:2203.16352](#)].
- [4] L. Cunqueiro and A. M. Sickles, *Studying the QGP with Jets at the LHC and RHIC*, *Prog. Part. Nucl. Phys.* **124** (2022) 103940, [[arXiv:2110.14490](#)].
- [5] **ATLAS, CMS, ALICE** Collaboration, R. Vertesi, *Jet substructure measurements in heavy-ion collisions*, *Int. J. Mod. Phys. A* **40** (2025), no. 09 2444002, [[arXiv:2405.16955](#)].
- [6] Y.-L. Du, D. Pablos, and K. Tywoniuk, *Deep learning jet modifications in heavy-ion collisions*, *JHEP* **21** (2020) 206, [[arXiv:2012.07797](#)].

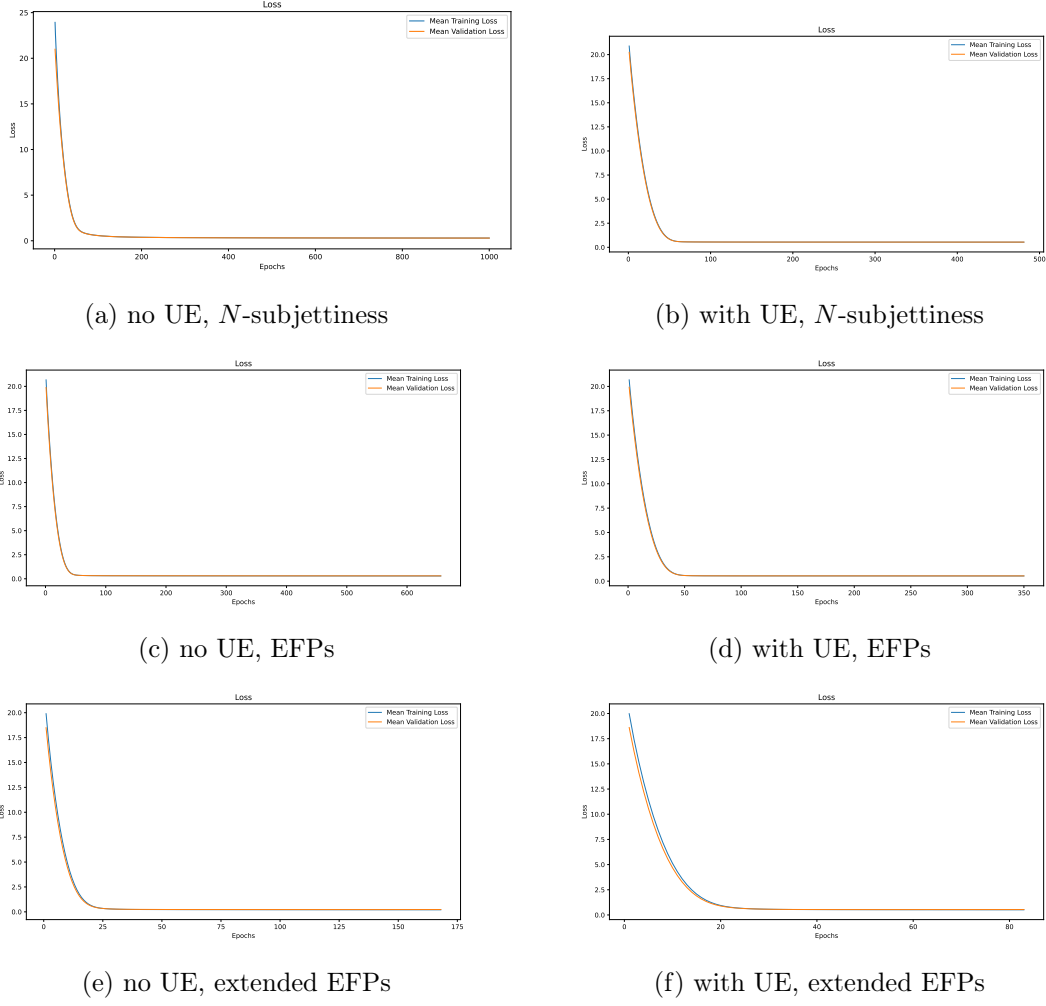


Figure 16: Training/validation loss evolution for NNs across input sets and UE configurations.

- [7] Y.-L. Du, D. Pablos, and K. Tywoniuk, *Jet Tomography in Heavy-Ion Collisions with Deep Learning*, *Phys. Rev. Lett.* **128** (2022), no. 1 012301, [[arXiv:2106.11271](#)].
- [8] L. Apolinário, N. F. Castro, M. Crispim Romão, J. G. Milhano, R. Pedro, and F. C. R. Peres, *Deep Learning for the classification of quenched jets*, *JHEP* **11** (2021) 219, [[arXiv:2106.08869](#)].
- [9] Y. S. Lai, J. Mulligan, M. Płoskoń, and F. Ringer, *The information content of jet quenching and machine learning assisted observable design*, *JHEP* **10** (2022) 011, [[arXiv:2111.14589](#)].
- [10] L. Liu, J. Velkovska, Y. Wu, and M. Verweij, *Identifying quenched jets in heavy ion collisions with machine learning*, *JHEP* **04** (2023) 140, [[arXiv:2206.01628](#)].
- [11] U. S. Qureshi and R. Kunnawalkam Elayavalli, *Model-Agnostic Tagging of Quenched Jets in Heavy-Ion Collisions*, [arXiv:2411.19389](#).
- [12] M. Crispim Romão, J. G. Milhano, and M. van Leeuwen, *Jet substructure observables for jet*

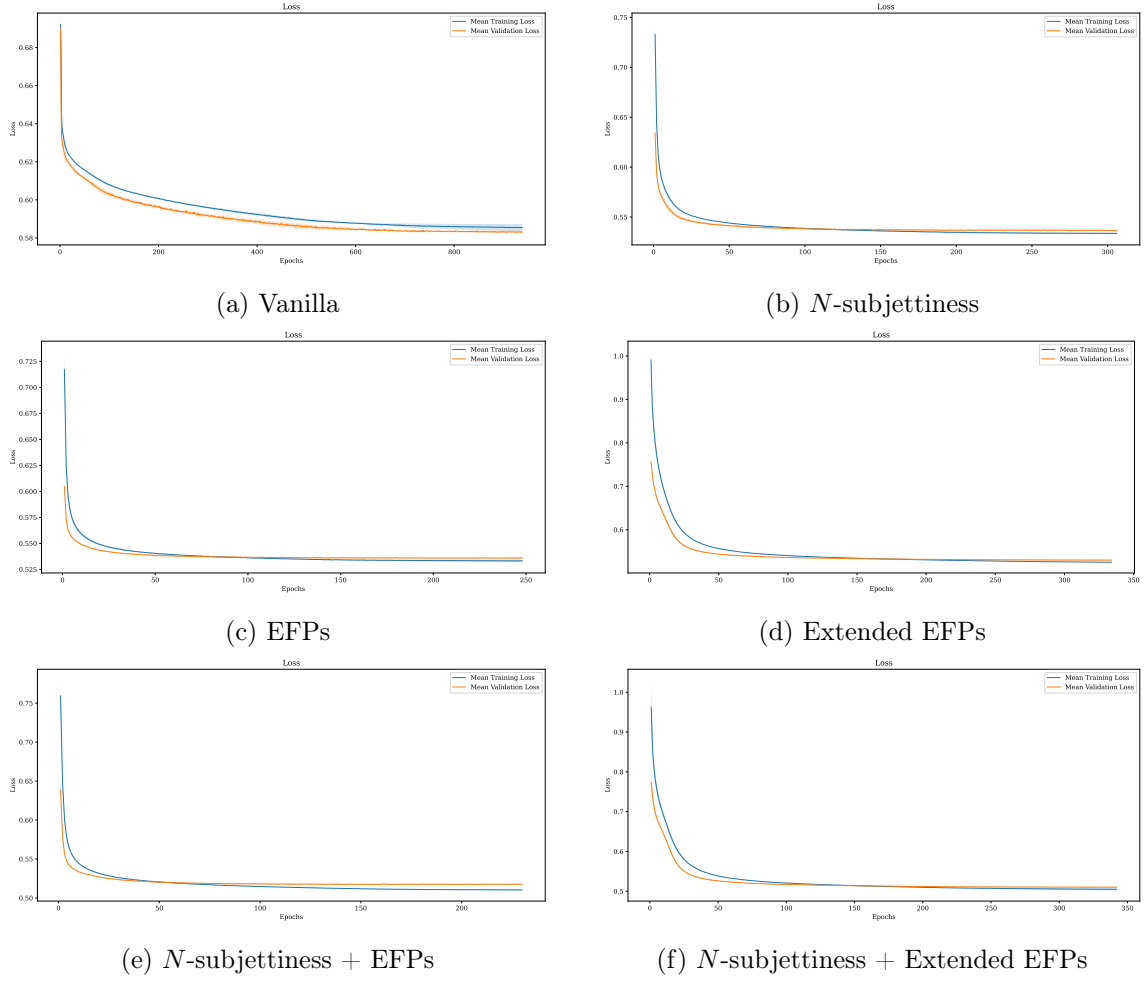


Figure 17: Training and validation losses for oEFNs under different augmentations.

quenching in quark gluon plasma: A machine learning driven analysis, *SciPost Phys.* **16** (2024), no. 1 015, [[arXiv:2304.07196](#)].

- [13] J. Arruda Gonçalves and J. G. Milhano, *Apples to apples in jet quenching: robustness of Machine Learning classification of quenched jets to underlying event contamination*, *JHEP* **09** (2025) 028, [[arXiv:2501.14015](#)].
- [14] HEP ML Community, “A Living Review of Machine Learning for Particle Physics.”
- [15] R. Kunnawalkam Elayavalli and K. C. Zapp, *Medium response in JEWEL and its impact on jet shape observables in heavy ion collisions*, *JHEP* **07** (2017) 141, [[arXiv:1707.01539](#)].
- [16] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N -subjettiness*, *JHEP* **03** (2011) 015, [[arXiv:1011.2268](#)].
- [17] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N -subjettiness*, *JHEP* **02** (2012) 093, [[arXiv:1108.2701](#)].
- [18] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, *JHEP* **04** (2018) 013, [[arXiv:1712.07124](#)].

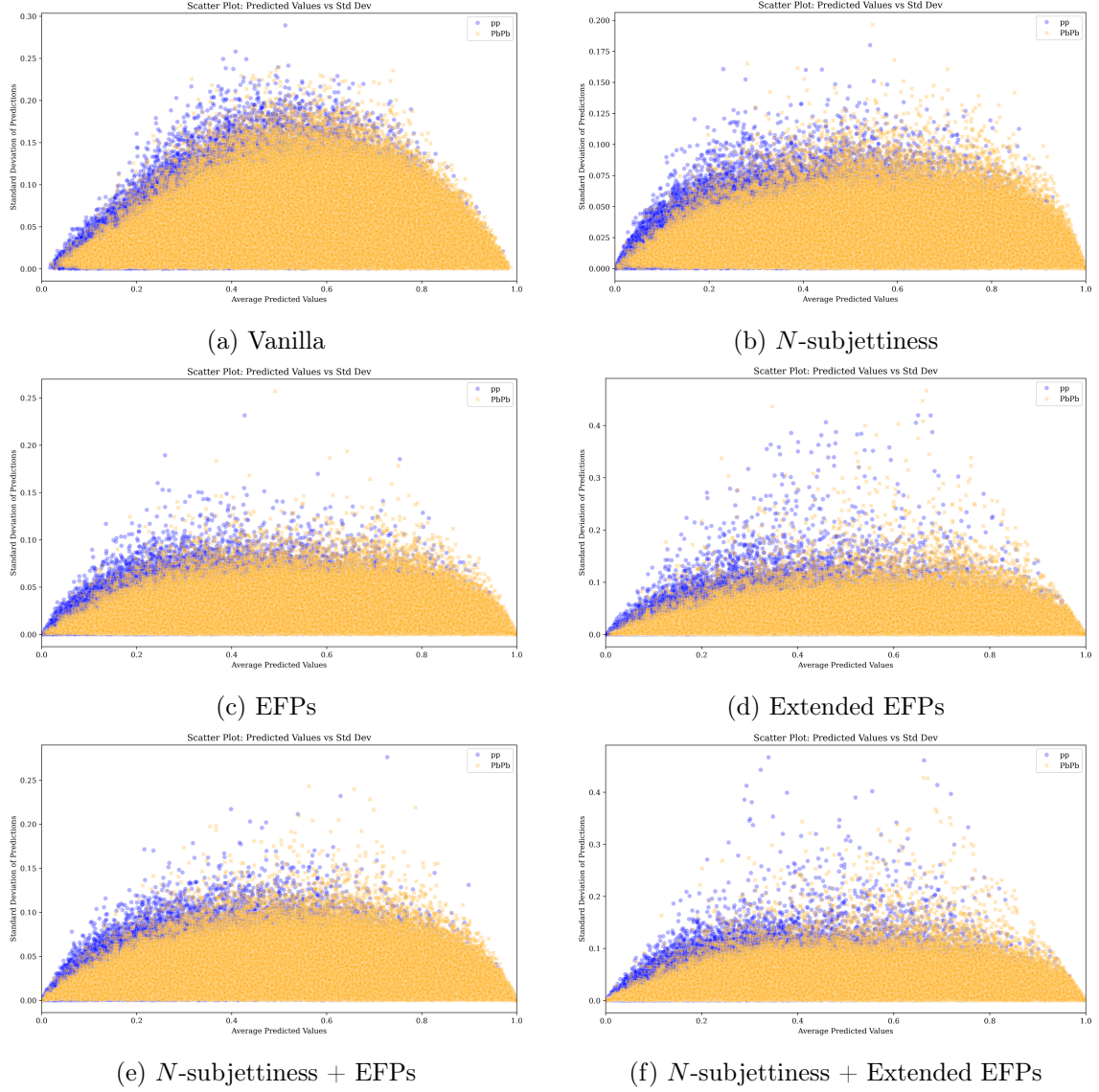


Figure 18: Cross-validation stability of oEFNs: mean vs. standard deviation across folds.

- [19] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Energy Flow Networks: Deep Sets for Particle Jets*, *JHEP* **01** (2019) 121, [[arXiv:1810.05165](#)].
- [20] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola, *Deep Sets*, [arXiv:1703.06114](#).
- [21] R. Gambhir, A. Osathapan, and J. Thaler, *Streamlining latent spaces in machine learning using moment pooling*, *Phys. Rev. D* **110** (2024), no. 7 074020, [[arXiv:2403.08854](#)].
- [22] E. M. Metodiev, B. Nachman, and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174, [[arXiv:1708.02949](#)].
- [23] K. C. Zapp, *JEWEL 2.0.0: directions for use*, *Eur. Phys. J. C* **74** (2014), no. 2 2762, [[arXiv:1311.0048](#)].

- [24] P. Berta, M. Spousta, D. W. Miller, and R. Leitner, *Particle-level pileup subtraction for jets and jet shapes*, *JHEP* **06** (2014) 092, [[arXiv:1403.3108](#)].
- [25] J. G. Milhano and K. Zapp, *Improved background subtraction and a fresh look at jet sub-structure in JEWEL*, *Eur. Phys. J. C* **82** (2022), no. 11 1010, [[arXiv:2207.14814](#)].
- [26] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, *Eur. Phys. J. C* **72** (2012) 1896, [[arXiv:1111.6097](#)].
- [27] M. Cacciari, G. P. Salam, and G. Soyez, *The anti- k_t jet clustering algorithm*, *JHEP* **04** (2008) 063, [[arXiv:0802.1189](#)].
- [28] D. P. Kingma, *Adam: A method for stochastic optimization*, *arXiv preprint arXiv:1412.6980* (2014).
- [29] R. Gambhir, M. LeBlanc, and Y. Zhou, *The Pareto Frontier of Resilient Jet Tagging*, in *39th Annual Conference on Neural Information Processing Systems: Includes Machine Learning and the Physical Sciences (ML4PS)*, 9, 2025. [arXiv:2509.19431](#).