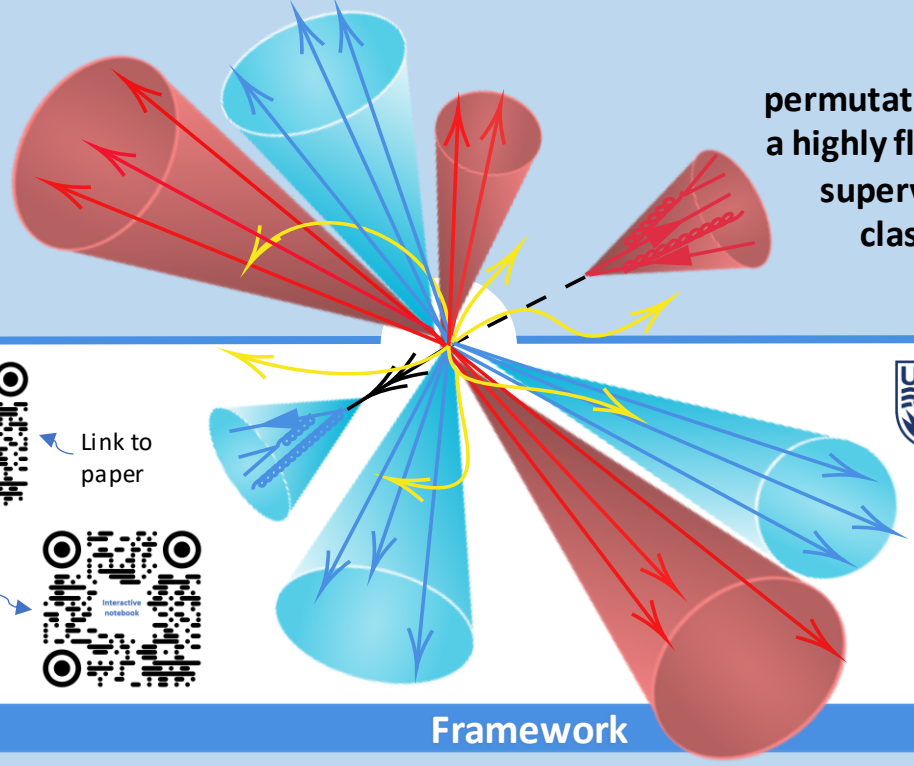
Boost Invariant Polynomials

for Efficient Jet Tagging



We construct highly efficient features invariant under permutations, rotations, and boosts in the jet direction generating a highly flexible and interpretable scheme with SOTA accuracies in supervised and unsupervised approaches using out-of-the-box classifiers with hundreds of parameters. And second, micro-second training and inference times on CPU.

Jose M Munoz

Ilyes Batatia

UNIVERSITY OF CAMBRIDGE

Christoph Ortner

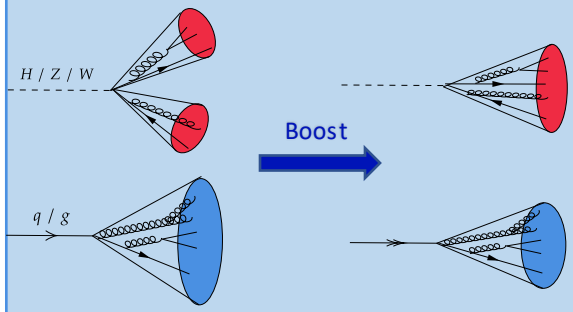
Link to paper

Play with the model

The problem

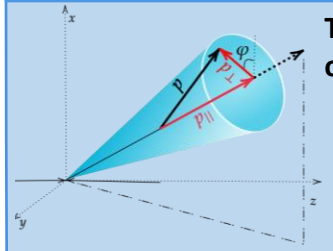
DL approaches are the go-to methods for data analysis related to jet-tagging. Specially to tackle the large amount of data generated in particle colliders, from which only some events are physically interesting.

Thus, it is necessary to distinguish events using their substructure. Nevertheless, since the events are highly boosted, most of the times the structure is not accessible from the laboratory frame:



Highest accuracies are currently achieved by models incorporating boost invariance, resulting in computationally costly and complex parametrizations and thus lack interpretability.

Framework



The BIP basis composes a complete basis for the jet representation via orthogonal polynomials via:

$$A_{nkl\mu} = \sum_{i=1}^{N} e^{il\varphi_i} e^{-k y_i} B_n(\tilde{p}_{\perp,i}) \psi_\mu(\theta_i)$$

IRC Safe + invariant polynomial

Detected rapidity

Log-normalized transverse momentum

Additional features

Bessel polynomials

$$\sum_\xi l_\xi = \sum_\xi k_\xi = 0$$

Invariance condition

BIP embeddings

$$A_{lkn\mu} = \prod_{\xi=1}^{\nu} A_{l_\xi k_\xi n_\xi \mu_\xi}$$

$\nu \leq \bar{\nu}$: correlation order

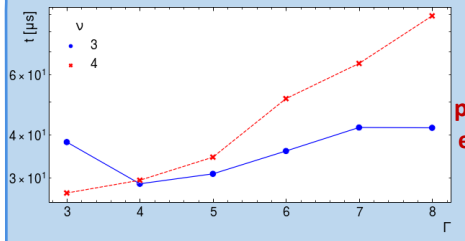
Outlooks

Extend geometric deep-learning framework based on our construction.

Generalization of the BIP embedding generation.

Application of the framework to further problems in HEP.

Results

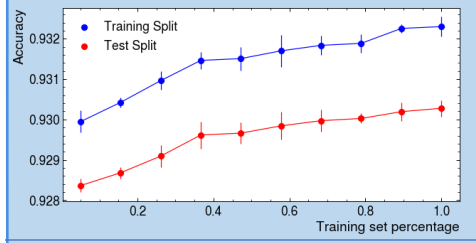


High performance embeddings

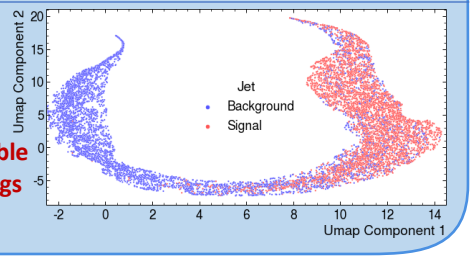
Model - independent SOTA-range accuracies

Architecture	#Param	Accuracy	AUC
Supervised			
BIP(4, 8, XGBoost)	2k	0.931	0.981
BIP(3, 6, LogReg)	120	0.928	0.977
BIP(2, 3, ARD)	71	0.926	0.976
Unsupervised			
BIP(3, 4, UMAP+GMM)	5	0.884	0.864

Top Tagging Dataset



Data Efficiency



Interpretable Embeddings