Recent ML developments in jet clustering and substructure techniques

Marco Letizia – Machine Learning Genoa Center and INFN

Intro and outline

- Meaningfull data representation.
- Physics-informed strategies.
- Robustness and efficiency.

- Optimal transport and jet physics.
- Foundation models in HEP.

Optimal transport and jets

Geometric description of event/jet shapes using optimal transport:

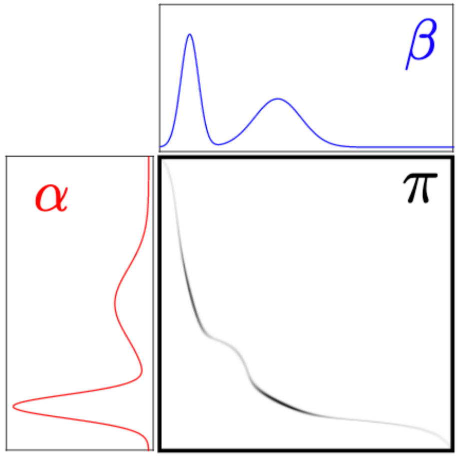
- Komiske, Metodiev, Thaler PRL 2019 1902.02346
- Komiske, Metodiev, Thaler, JHEP 2020 2004.04159
- Romao, Castro, Milhano, Pedro, Vale EPJC 2021 2004.09360
- Tianji Cai, Junyi Cheng, Katy Craig, Nathaniel Craig PRD 2020 2008.08604
- Park, Harris, Ostdiek, JHEP 2023 2208.05484
- Ba, Dogra, Gambhir, Tasissa, Thaler JHEP 2023 2302.12266
- ATLAS, JHEP 2023 2305.16930

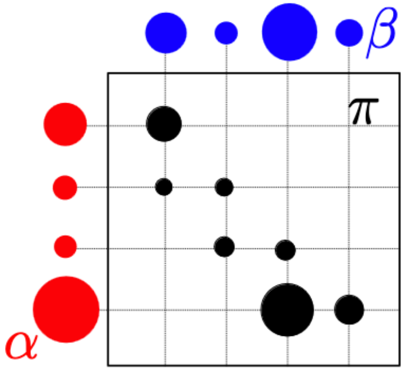
Optimal transport and jets

(Kantorovich problem) Wasserstein distance between two measures μ and ν

$$\mathcal{W}_p = \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\mathbb{E}_{(x,y)\sim\gamma}\, d(x,y)^p \right)^{1/p}$$

$d(x,y)$ is the ground metric, $p \in [1,+\infty]$ and $\Gamma(\mu,\nu)$ set of all joint measures with marginals μ and ν.





Canonical way to lift a ground metric between points to a metric between measures.

If $d(x,y)$ is a distance then $\mathcal{W}_p$ is a distance

- symmetric,
- nonnegative,
- $\mathcal{W}_p(\alpha,\beta) = 0$ if and only if $\alpha = \beta$
- it satisfies the triangle inequality

Peyré, Cuturi, *Computational Optimal Transport,* Foundations and Trends in Machine Learning 1803.00567

Optimal transport and jets

Energy flow density $\mathcal{E}(x) = \sum_{i=1}^{M} E_i \, \delta(x - x_i)$ Detector space $\mathcal{X} = [y_{\min}, y_{\max}] \times S^1$

Event/jet shapes: $\mathcal{O}(p_1, \ldots, p_M) = \min_{\theta \in \mathcal{M}} F\left(\sum_{i=1}^{M} E_i \phi_\theta(x_i)\right)$

IRC-safe weighted sum over the four-momenta of the particles in an event/jet

Shape info: $\mathcal{M}, F, \phi_\theta$.

Event Shape	Description	Expression
Thrust [1, 7, 8]	How Pencil-Like?	$t(\mathcal{E}) = 2\min_{\hat{n}} \left(\sum_i E_i (1 - \lvert\hat{n}_i \cdot \hat{n}\rvert)\right)$
Spherocity [55]	How Tranverse-Planar?	$s(\mathcal{E}) = \min_{\hat{n}} \left(\sum_i E_i \lvert\hat{n}_i \times \hat{n}\rvert\right)^2$
Broadening [56]	How 2-Pronged?	$b(\mathcal{E}) = \min_{\hat{n}_1, \hat{n}_2} \left(\sum_i E_i \min(d_{i1}, d_{i2})\right)$
N-jettiness [54, 57]	How N-particle like?	$\mathcal{T}_N^{(\beta)}(\mathcal{E}) = \min_{\hat{n}_1, \ldots, \hat{n}_N} \left(\sum_i E_i \min(R^\beta, d_{i1}^\beta, \ldots, d_{iN}^\beta)\right)$
Isotropy [40]	How Uniform?	$\mathcal{I}^{(\beta)}(\mathcal{E}) = \min_{\mathcal{U} \in \mathcal{M}} \left(\mathrm{EMD}^{(\beta,R)}(\mathcal{E}, \mathcal{U})\right)$
XCone [54]	Which N-particles?	$\hat{n}_i(\mathcal{E}) = \mathrm{argmin}_{\hat{n}_1, \ldots, \hat{n}_N} \left(\sum_i E_i \min(R^\beta, d_{i1}^\beta, \ldots, d_{iN}^\beta)\right)$
S. Recomb. [60–64]	Clustering History?	$d_{ij}^N(\mathcal{E}) = \min(E_i^{2p}, E_j^{2p}) \frac{d_{ij}^2}{R^2};\ d_{iR}^N(\mathcal{E}) = E_i^{2p}$

Jet Shape	Description	Expression
Angularities [9, 10]	Angular Moments?	$\lambda_\beta(\mathcal{J}) = \sum_i E_i d_{iJ}^\beta$
	... Recoil Free?	$\lambda_\beta(\mathcal{J}) = \min_{\hat{n}} \left(\sum_i E_i d_{in}^\beta\right)$
N-subjettiness [68]	How N-Particle Like?	$\mathcal{T}_N^{(\beta)}(\mathcal{J}) = \min_{\hat{n}_1, \ldots, \hat{n}_N} \left(\sum_i E_i \min(d_{i1}^\beta, \ldots, d_{iN}^\beta)\right)$
Int. Shape [65, 66]	Radial Energy CDF?	$\psi_\mathcal{J}(r/R) = \left(\sum_i E_i \Theta(r - d_{iJ})\right) / \left(\sum_i E_i \Theta(R - d_{iJ})\right)$

Find $\mathcal{L}$: $\mathcal{O}_\mathcal{M}(\mathcal{E}) = \min_{\mathcal{E}_\theta \in \mathcal{M}} [\mathcal{L}(\mathcal{E}; \mathcal{E}_\theta)]$ How close, in event space, is my event to looking like an optimal $\mathcal{E}^*$?

Ba, Dogra, Gambhir, Tasissa, Thaler JHEP 2023 [2302.12266](2302.12266);

Optimal transport and jets

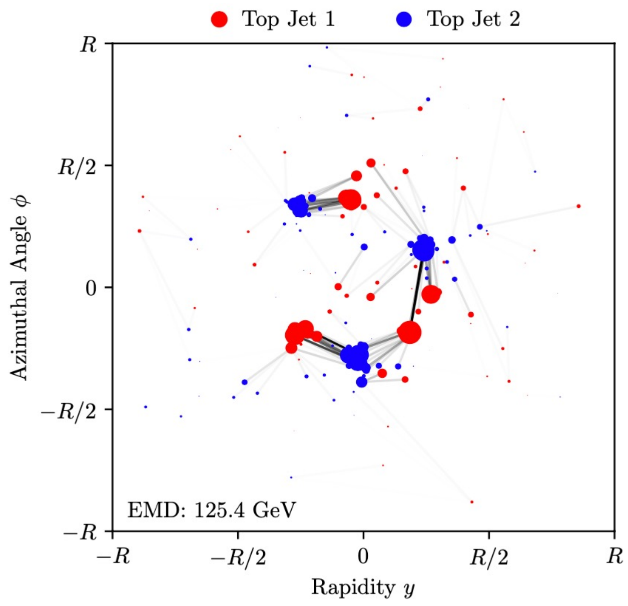
- **Infrared safety:** For any atomic event $\mathcal{E}$, adding or removing an ϵ-soft emission to $\mathcal{E}$ eaves $\mathcal{O}$ unchanged as $\epsilon \to 0$.
- **Collinear safety:** For any atomic event $\mathcal{E}$, splitting any particle into two particles at the same location with the same total energy leaves $\mathcal{O}$ unchanged. Moreover, translating either particle by an ϵ -small displacement leaves O unchanged as $\epsilon \to 0$.

Definition. An observable $\mathcal{O}$ is IRC safe if it is continuous with respect to the weak* topology on energy flows.

Must depend on $d(x, y)$ and requiring a faithful lift:

$$\mathrm{EMD}^{(\beta,R)}(\mathcal{E},\mathcal{E}') = \min_{\pi\in\mathcal{M}(\mathcal{X}\times\mathcal{X})} \left[\frac{1}{\beta R^\beta}\langle \pi, d(x,y)^\beta\rangle\right] + |\Delta E_{\mathrm{tot}}|,$$

$$\pi(\mathcal{X}, Y) \leq \mathcal{E}'(Y), \pi(X,\mathcal{X}) \leq \mathcal{E}(X), \pi(\mathcal{X},\mathcal{X}) = \min(E_{\mathrm{tot}}, E'_{\mathrm{tot}})$$



Optimal transport and jets

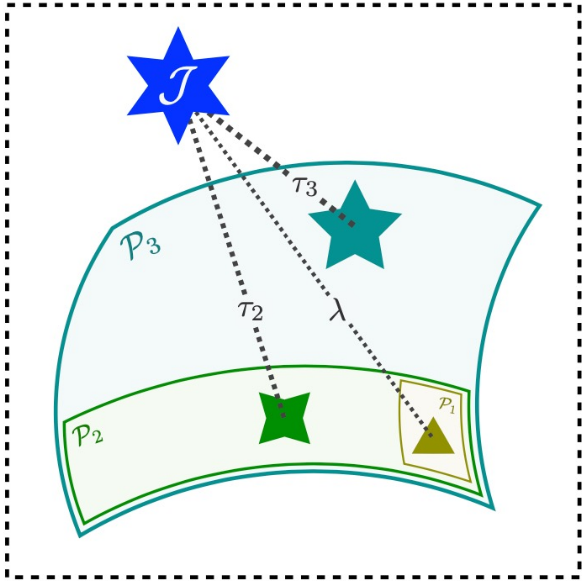
N-subjettiness: how N-particle like

$$\tau_N(\mathcal{J}) = \min_{N\text{ axes}} \sum_i E_i \min\{\theta_{1i}, \theta_{2i}, \dots, \theta_{Ni}\}$$

distance between the jet and the manifold of all N-particle jets

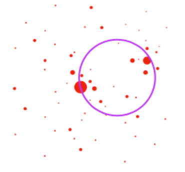
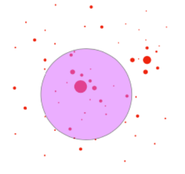
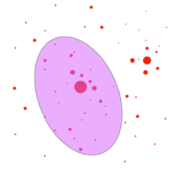
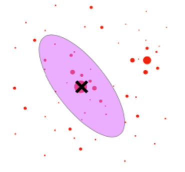
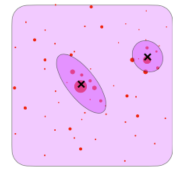
$$\rightarrow \tau_N(\mathcal{J}) = \min_{\mathcal{J}' \in \mathcal{P}_N} \mathrm{EMD}\,(\mathcal{J}, \mathcal{J}')$$

(jet angularity: distance from $\mathcal{P}_1$)



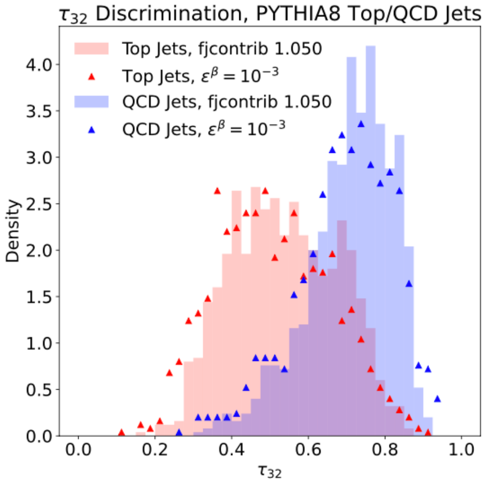
Optimal transport and jets

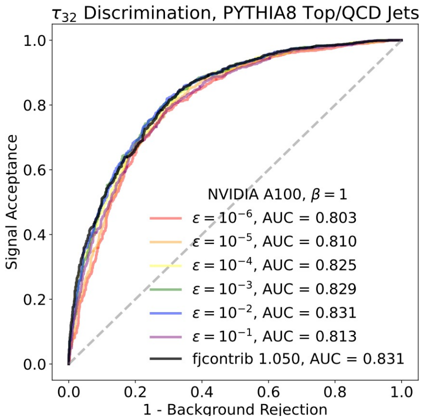
Novel shapes

Sec.	Shape	Specification	Illustration
3.3.1	Ringiness $\mathcal{O}_R$	**Manifold of Rings** $\mathcal{E}_{x_0,R_0}(x) = \frac{1}{2\pi R_0}$ for $\lvert x - x_0 \rvert = R_0$; x_0 = Center, R_0 = Radius	
3.3.2	Diskiness $\mathcal{O}_D$	**Manifold of Disks** $\mathcal{E}_{x_0,R_0}(x) = \frac{1}{\pi R_0^2}$ for $\lvert x - x_0 \rvert \leq R_0$; x_0 = Center, R_0 = Radius	
3.3.3	Ellipsiness $\mathcal{O}_E$	**Manifold of Ellipses** $\mathcal{E}_{x_0,a,b,\varphi}(x) = \frac{1}{\pi ab}$ for $x \in \text{Ellipse}_{x_0,a,b,\varphi}$; x_0 = Center, a, b = Semi-axes, φ = Tilt	
3.3.4	(Ellipse +Point)iness	**Composite Shape** $\mathcal{O}_E \oplus \tau_1$; Fixed to same center x_0	
3.3.5	N-(Ellipse +Point)iness +Pileup	**Composite Shape** $N \times (\mathcal{O}_E \oplus \tau_1) \oplus \mathcal{I}$	

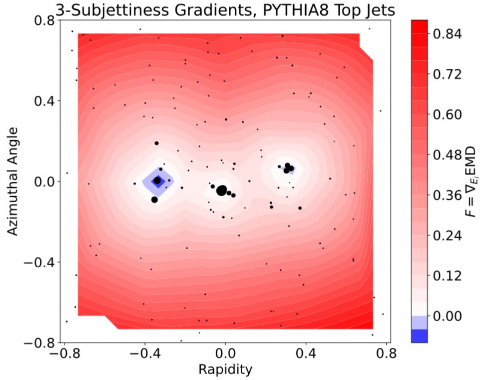
SHAPER

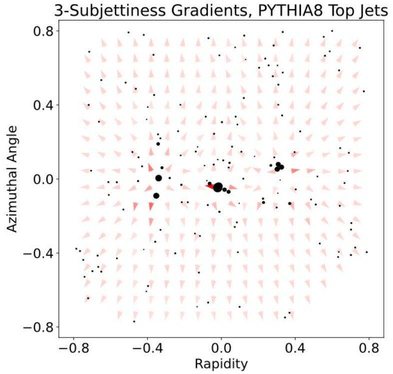
N-subjettiness (FastJet)



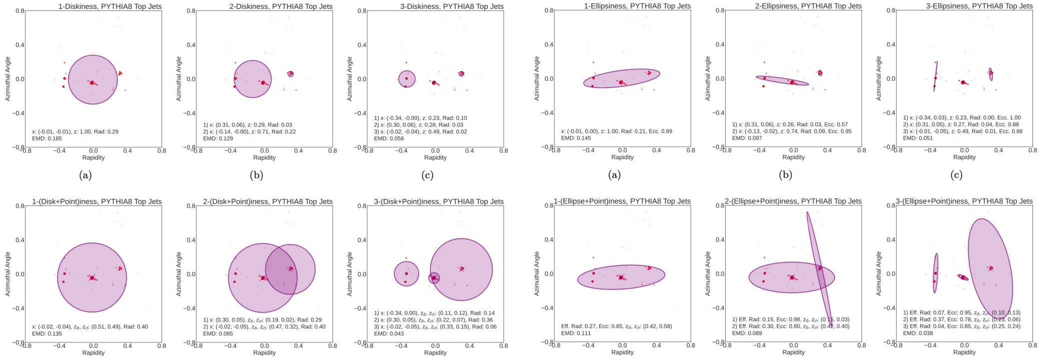


3-subjettiness: grandients





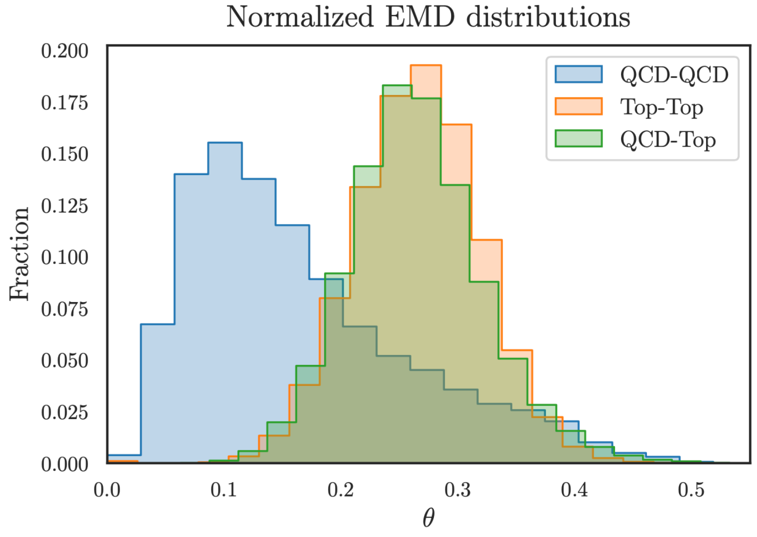
New shapes



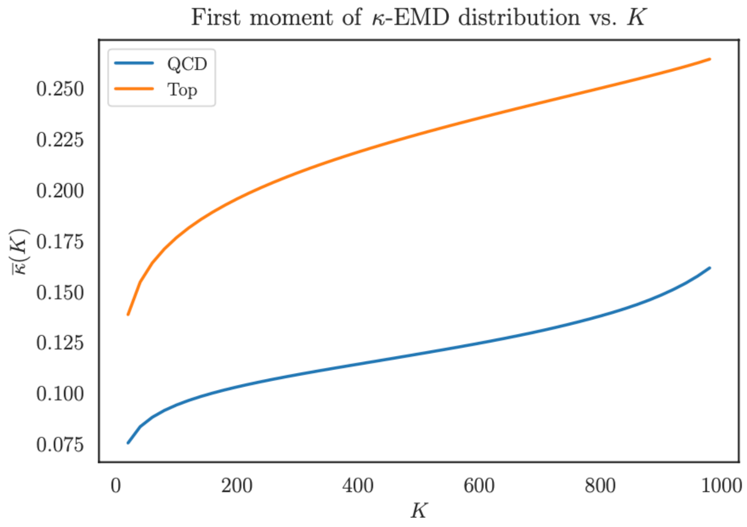
Optimal transport and jets

Unsupervised jet tagging Gaertner, Reiten [2312.06948](https://arxiv.org/abs/2312.06948)

$$\mathcal{J}(\eta,\phi) = \sum_{k\in\text{jet}} \frac{p_{T_k}}{p_T^{\text{jet}}}\delta(\eta-\eta_k)\delta(\phi-\phi_k)$$

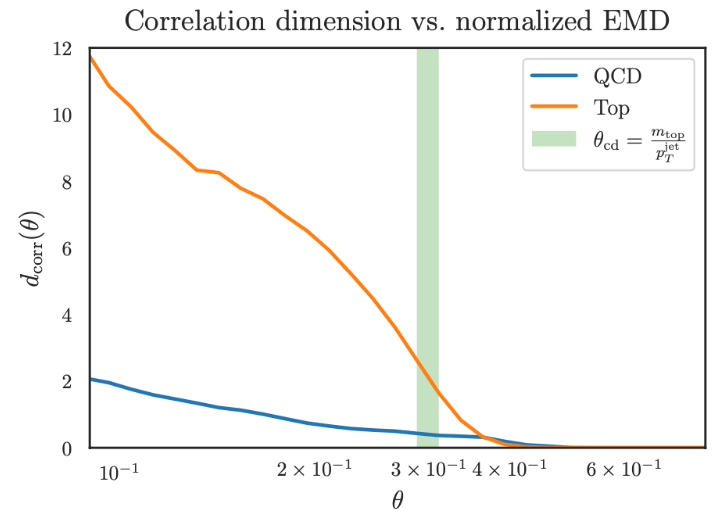


$$\mathrm{EMD}_{ij} \equiv \mathrm{EMD}(\mathcal{J}_i,\mathcal{J}_j)$$



$$\bar{\kappa}(K) = \int d\kappa\, k\, p_f(\kappa;K)$$

$$\kappa_i \equiv \mathrm{EMD}_{i(K)}$$



$$d_{corr}(\theta) = \frac{\partial}{\partial\log\theta}\log\sum_{1\le i<j\le N}\Theta\big(\mathrm{EMD}_{ij}<\theta\big)$$

Optimal transport and jets

DBSCAN

Density-based clustering.

$B_\epsilon(\mathcal{J}_i) = \{\mathcal{J}_i \in D | \mathrm{EMD}_{ij} \leq \epsilon\}$

$|B_\epsilon(\mathcal{J}_i)| \geq \mu$

Ricci flow on graphs

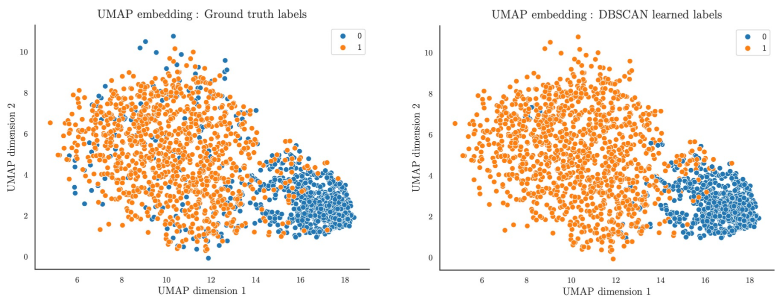
Unsuperv. clustering with Ricci curvature.

$G = (V, E, \omega)$ fully connected

→ reduce the edge set $G_K = (V, E_k, \omega)$

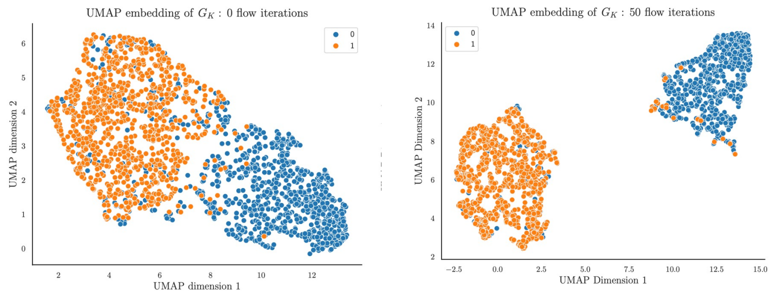
$$R^{(t)}(v_i v_j) = 1 - \frac{W_1\left(P_i^{(t)}, P_j^{(t)}\right)}{d_G^{(t)}(v_i, v_j)},$$

$$\omega^{(t+1)}(v_i v_j) = \left(1 - R^{(t)}(v_i v_j)\right) \times d_G^{(t)}(v_i, v_j)$$



Architecture	Accuracy	Parameters	Learning
ResNeXt [49]	0.9360	1.46e6	Supervised
ParticleNET [50]	0.9380	4.98e5	Supervised
PFN [51]	0.9320	8.20e4	Supervised
LGN [52]	0.9290	4.50e4	Supervised
nPELICAN$_{\mathrm{hidden}=1}$ [53]	0.8951	11	Supervised
DBSCAN$_{\mathrm{EMD}}$	0.9003	2	Unsupervised
Ricci-Flow$_{\mathrm{Curvature}}$	0.9113	2	Unsupervised
Ricci-Flow$_{\mathrm{UMAP}}$	0.9104	2	Unsupervised

TABLE I: Comparison to a limited selection of top-taggers from the literature.

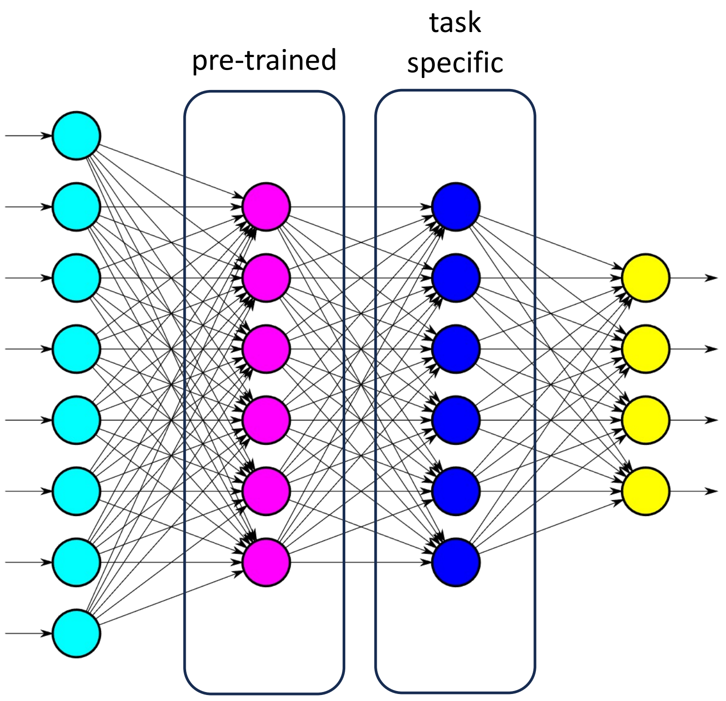


UniGe | MaLGa

Foundation models in HEP

The use of pre-trained models is ubiquitus in NLP and CV.

Learn good features: useful for a large family of downstream tasks.



- Vision Transformer Dosovitskiy et al, ICLR 2021 [2010.11929]
- GPT-4 OpenAI [2303.08774]
- BERT Devlin, Chang, Lee, Toutanova EMNLP 2018 [1810.04805]

○ Large computational cost is shared among tasks

○ Scarce data

▪ Supervised training on large datasets with several classes (Imagenet: 20M images in 20k classes)

▪ No labels → self-supervision: training on pretext objectives with self-generated labels.

Foundation models in HEP

Examples in science:

Rives et al, *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,* PNAS 2021

Irwin et al, *Chemformer: a pre-trained transformer for computational chemistry, MLST (2022)*

Lanusse et al, *Astroclip: Cross-modal pre-training for astronomical foundation models,* 2310.03024

HEP:

L. Heinrich et al*, Masked Particle Modeling on Sets,* 2401.13537

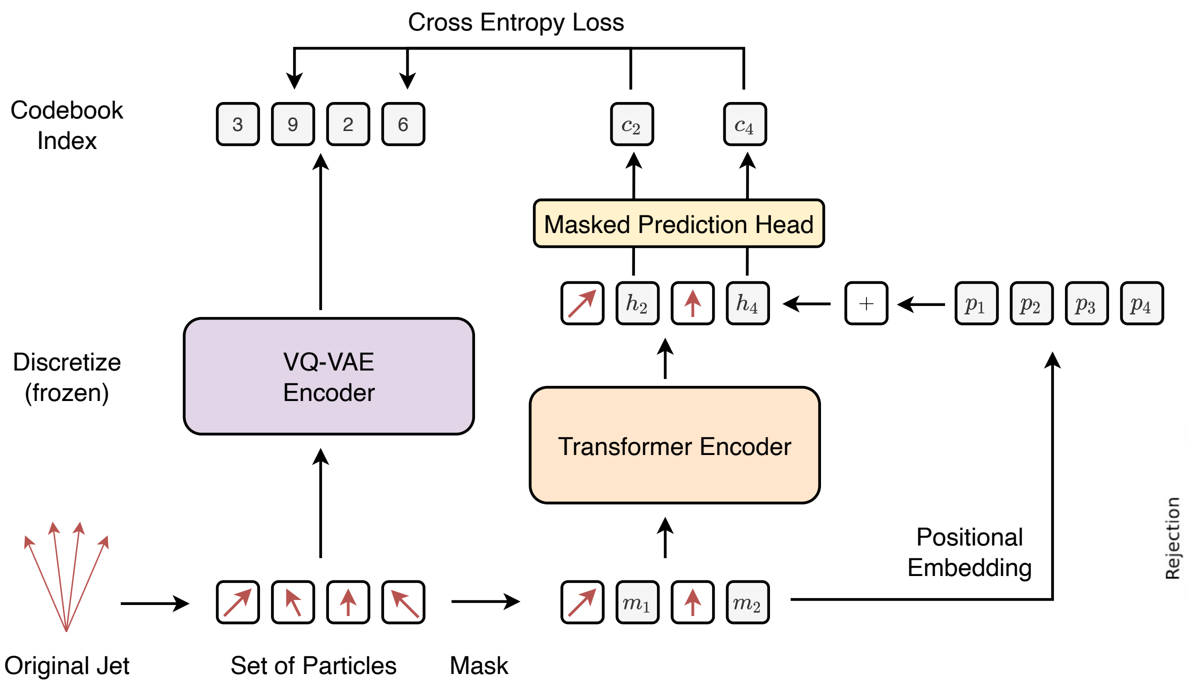
M. Vigl et al, *Finetuning Foundation Models for Joint Analysis Optimization*, 2401.13536

P. Harris et al, *Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models* 2403.07066

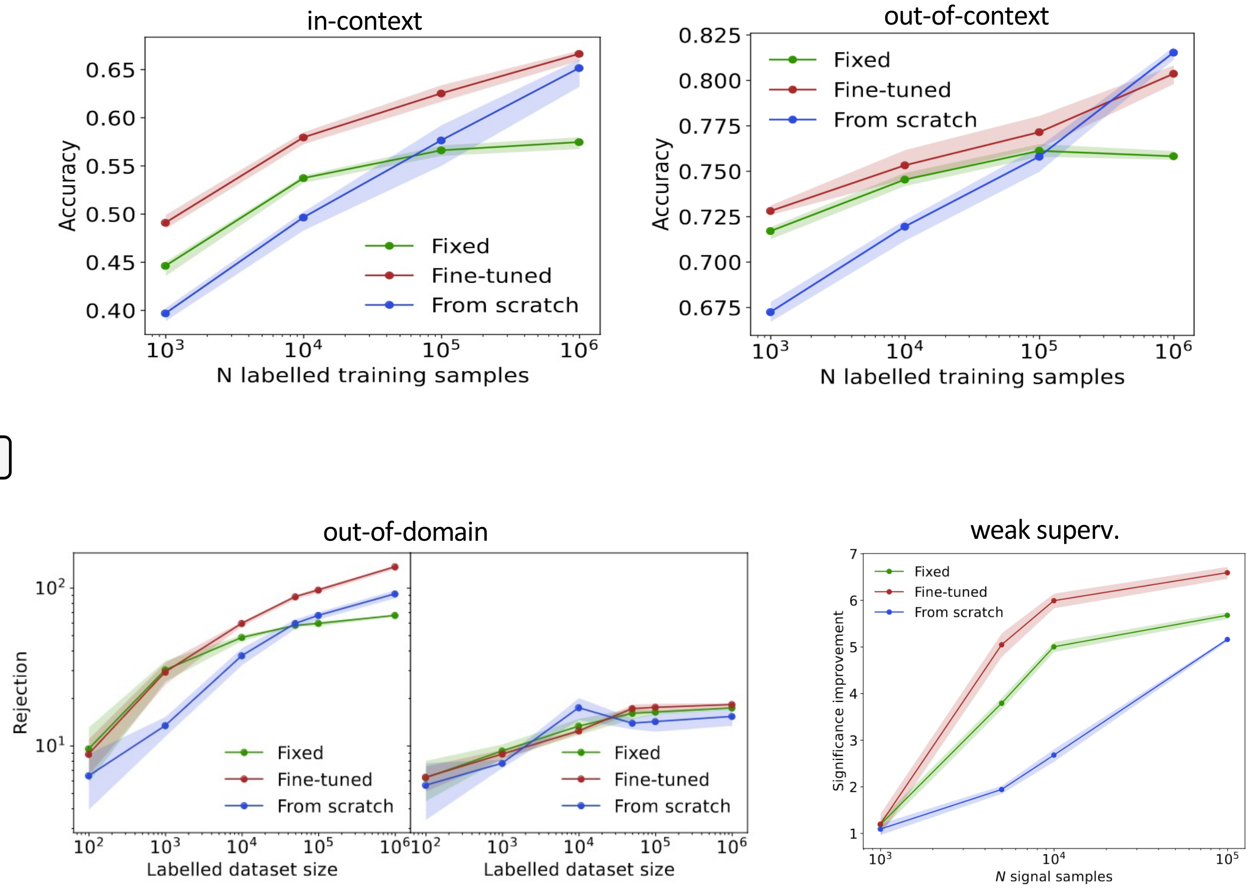
J. Birk et al, *OmniJet-α: The first cross-task foundation model for particle physics,* 2403.05618

Foundation models in HEP

L. Heinrich et al, MPM (masking)



Datasets: JetClass; RODEM

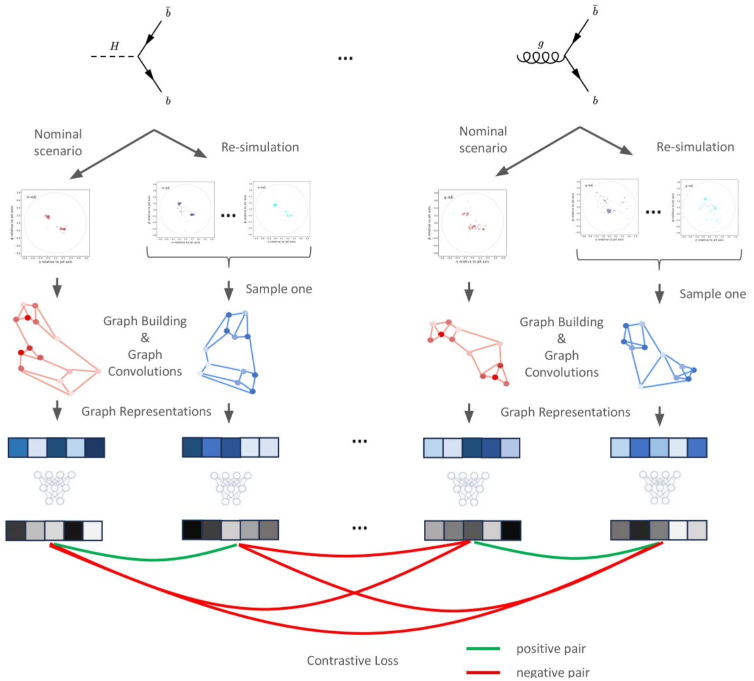


UniGe | MaLGa

Foundation models in HEP

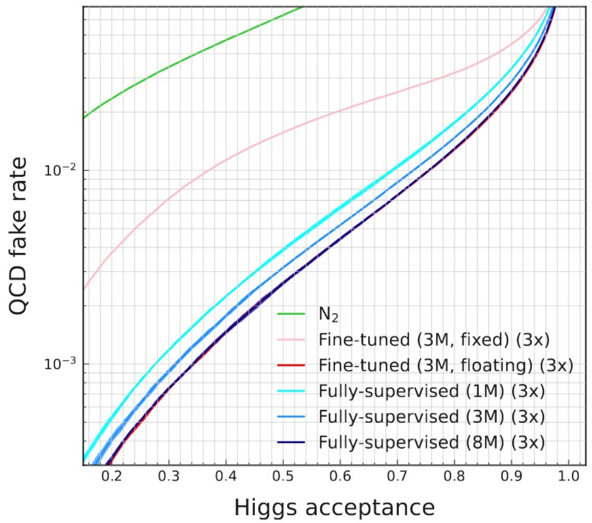
Harris et al, R3SL (self-supervision with contrastive learning)

Strategy: map a data point and its augmentation(s) to similar representations, while pushing different data points toward differing representations.



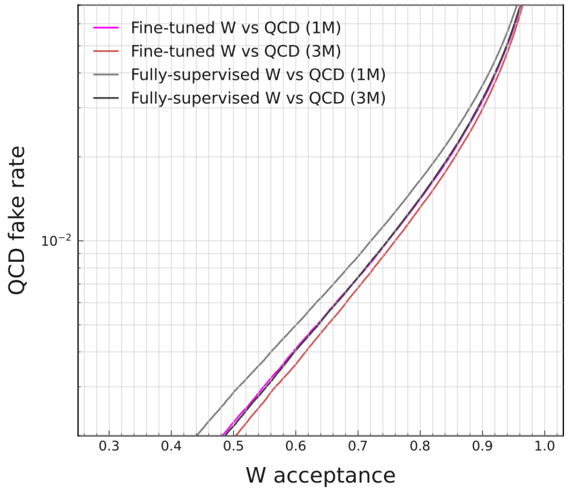
$$\mathcal{L} = -\log \frac{e^{s(z_i, z_i')/\tau}}{\sum_{i \neq j \in \text{minibatch}} \left[e^{s(z_i, z_j)/\tau} + e^{s(z_i, z_j')/\tau} \right]}$$

Higgs vs QCD jets (in-domain)



Higgs efficiency	0.3	0.5	0.7
N_2	29	16	8
Fine-tuned (3M, fixed)	140 ± 1	64 ± 0	39 ± 0
Fine-tuned (3M, floating)	$\mathbf{1325 \pm 31}$	380 ± 4	$\mathbf{135 \pm 1}$
Fully-supervised (1M)	834 ± 11	257 ± 6	96 ± 2
Fully-supervised (3M)	1070 ± 29	317 ± 3	115 ± 1
Fully-supervised (8M)	1312 ± 15	$\mathbf{381 \pm 7}$	134 ± 1

QCD vs W (out-of-domain)



W efficiency	0.3	0.5	0.7
Training setup	1/(QCD efficiency)		
Fine-tuned (1M, floating, all)	1589 ± 88	438 ± 9	134 ± 2
Fine-tuned (3M, floating, all)	$\mathbf{1928 \pm 31}$	$\mathbf{504 \pm 3}$	$\mathbf{147 \pm 1}$
Fully-supervised (1M, all)	1288 ± 102	357 ± 16	114 ± 2
Fully-supervised (3M, all)	1763 ± 18	459 ± 1	137 ± 2