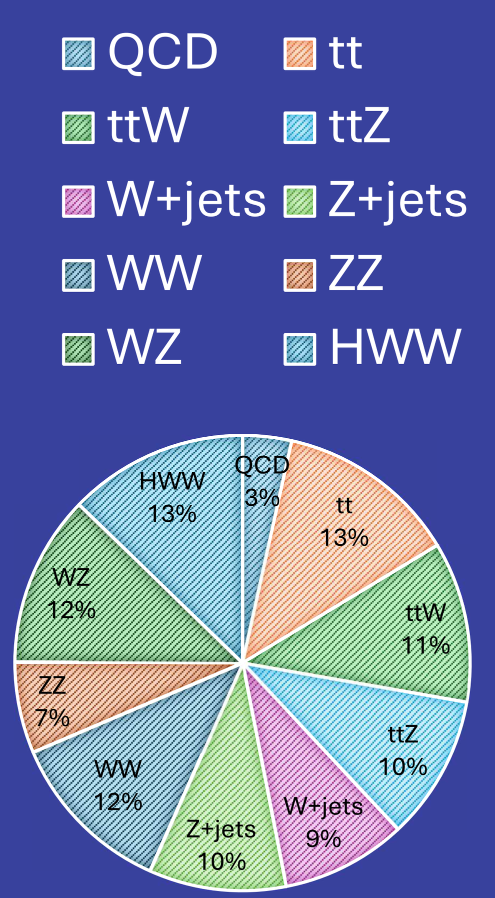
EveNet: Towards a Generalist Event Transformer for Unified Understanding and Generation of Collider Data

Ting-Hsiang Hsu, Bai-Hong Zhou, Yuan-Tang Chou, Qibin Liu, Wei-Po Wang, Yue Xu, Haoran Zhao,
Shu Li, Benjamin Nachman, Shih-Chieh Hsu, Vinicius Massami Mikuni, Yulei Zhang

Big HEP Data



Data generated using MC
- Madgraph, Pythia and Delphes

Big data for big model
- 10 SM processes: diverse topology
- 3B raw evt generated (500M used)

Hierarchy representation
- Constituent level *Point Cloud:*
  - Up to 18 particles per event
  - 7 features per particle: 4-momentum, PID and charge
- Event level *High-level Variables:*
  - Missing ET, HT, ST
  - Multiplicity and invariant mass

Large-scale Collider Dataset

Raw detector response

Trackers, Clusters

Constituents

Event-level Objects

Pre-Training

Unsupervised Training

Representation learning
- Masked prediction of 10%-100% particles
- Masking rate ramping up during training

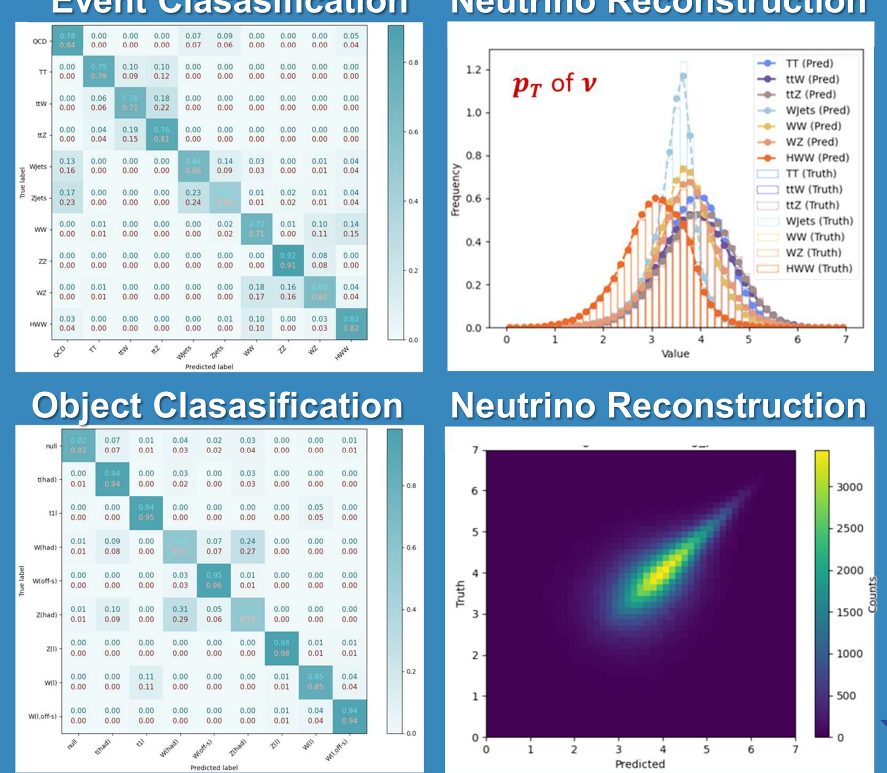
Learn intra-particle info and global event info
- Source of MET, multiplicity, invariant mass

Supervised Training

Task-specific knowledge learning
- Classification: separate SM processes
- Reconstruction: clustering, pairing, regression
- Generation: particle 4-momentum

Map general representation to specific HEP tasks

Physics Performance



Ready for HEP Tasks!

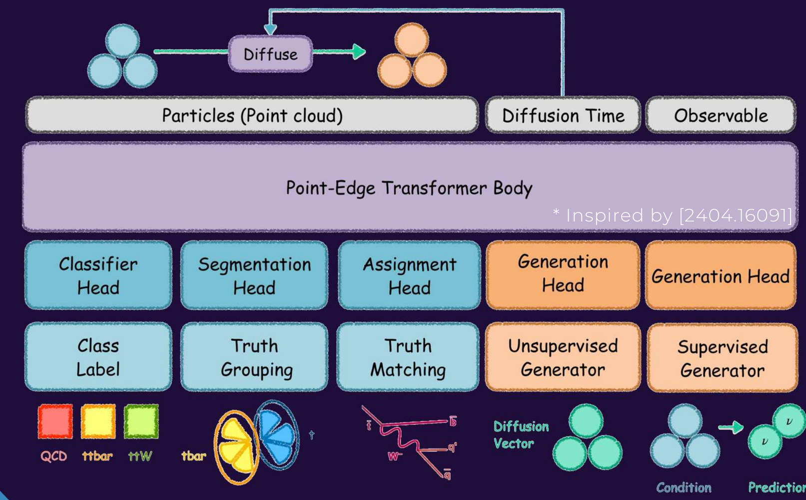
Training

Foundation Model

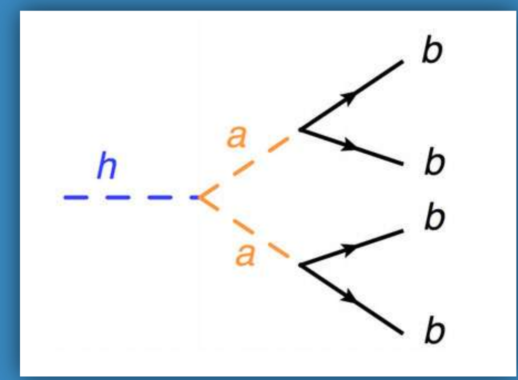
QCD
ttbar
ttW
tbar
Condition
Prediction

One Model for All

Foundational HEP model which learns physics and useful to various tasks



Exotic Higgs



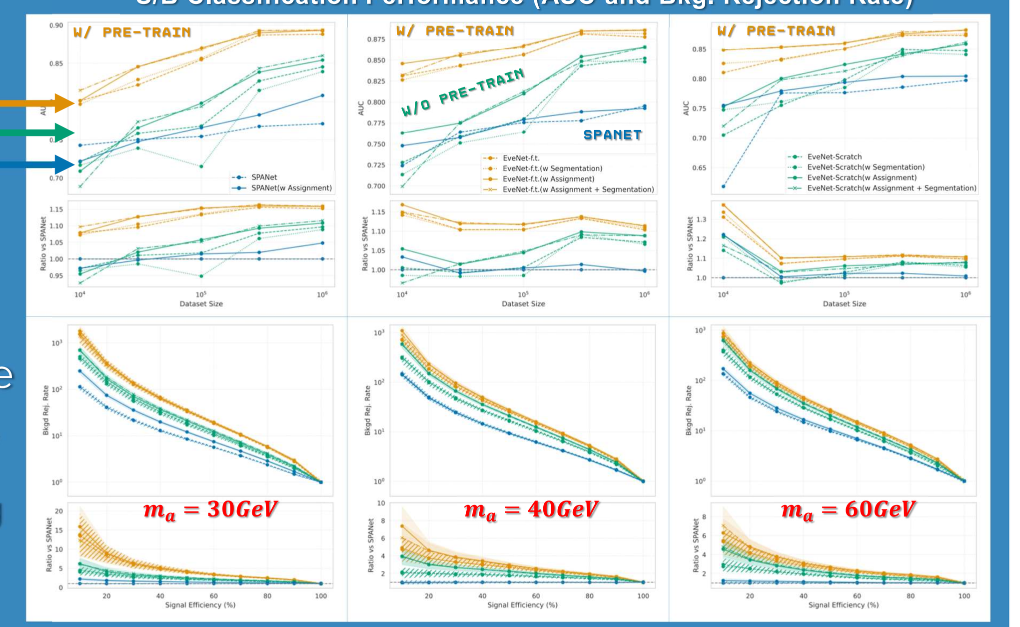
Tasks

Classification

Search for exotic Higgs $H \to aa \to 4b$
- Complex hadronic final states requires good b-tagging and jet pairing: benchmark various HEP tasks
- Varying scalar mass provides different kinematics : new signals unseen during pre-training
- Background dominated by QCD: heavy flavor jets and high multiplicity (>=4 jets with 0,1,2,3 b-jets)
- ML plays important role: successful S/B separation reply on efficient capture of local and global info.

Foundational model shows improvement
- *EveNet* outperforms dedicated model [SPANet, Commun Phys 7, 139 (2024)]
- Including assignment task (b-jet pairing) further improves the S/B separation
- Benefit from the pre-training and outperform the one w/o it: good performance generalized to new BSM signals which are unseen in the pre-training and make it useful for general HEP applications

S/B Classification Performance (AUC and Bkg. Rejection Rate)



Entanglement

Measurement of quantum entanglement of leptonic ttbar
- Leptonic channel $pp \to t\bar{t} \to b\bar{b}\ell\bar{\nu}\ell\nu$: large yield and clean signatures
- Top kinematics and spin correlation → spin density matrix of two-particle state
- Non-separable state → quantum entanglement (QE)

Reconstruction of top: vital and challenging task
- Reconstruction of 2ν from MET and pairing of top decay products
- *EveNet* enables multi-ν prediction and objects assignment with single model
- Notable gain from pre-training: foundation model learns priori knowledge

Tasks

Reconstruction

Labeled Events [M]	Top Reconstruction Efficiency [%]		
(Training)	w/o pre-train	w/ pre-train	Improvement [%]
3.6	69.71	71.02	1.88
2.5	68.28	70.78	3.66
1.0	66.64	70.60	5.94
0.4	39.38	70.56	79.18

Uncertainty of QE measurement improved by *EveNet*

QE observable D constructed using spin correlation [Eur. Phys. J. C (2022) 82:285]
- $D = -C_{kk} - C_{rr} - C_{nn} > 1$: sufficient condition for QE

Rel. uncertainty reduced to 1% level and pre-training shows good improvement

MC Events [M]	Rel. Uncertainty: $\epsilon/(D-1)$ [%]		
(Training)	w/o pre-train	w/ pre-train	Improvement [%]
3.6	1.37	1.24	9.78
2.5	1.45	1.23	15.21
1.0	1.54	1.23	20.06
0.4	1.89	1.24	34.45

Re-Discover $\Upsilon \to \mu\mu$

Tasks

Generation

Anomaly Detection to re-discover $\Upsilon \to \mu\mu$

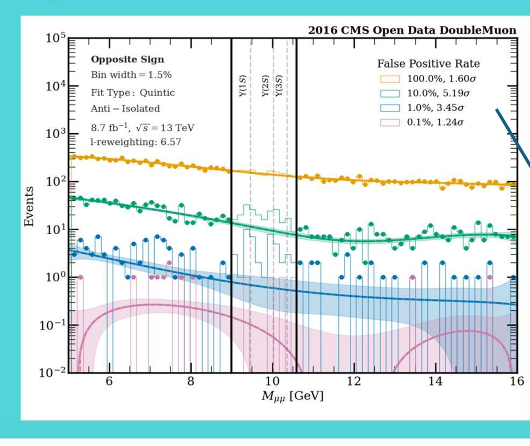
Inspired from [2502.14036] using CMS open data

Signal (SR) & Sideband (SB) defined on $m_{\mu\mu}$: SR in [9, 10.6] GeV

Background modeled with data-driven: [2109.00546]
- *EveNet* learns from SB to generate bkg. in SR

Weak supervised classification: [1805.02664]
- BDT trained in SR to separate generated bkg. v.s. data

Significance extraction:
- Event counting after cut or reweight w/ classifier score



Cuts on weak classifier, trained using EveNet generated bkg.

Tsung-Dao Lee Institute
NERSC
Berkeley Lab
Stanford
SLAC
University of Washington