Machine Learning in HEP data processing

李　刚

中国科学院高能物理研究所

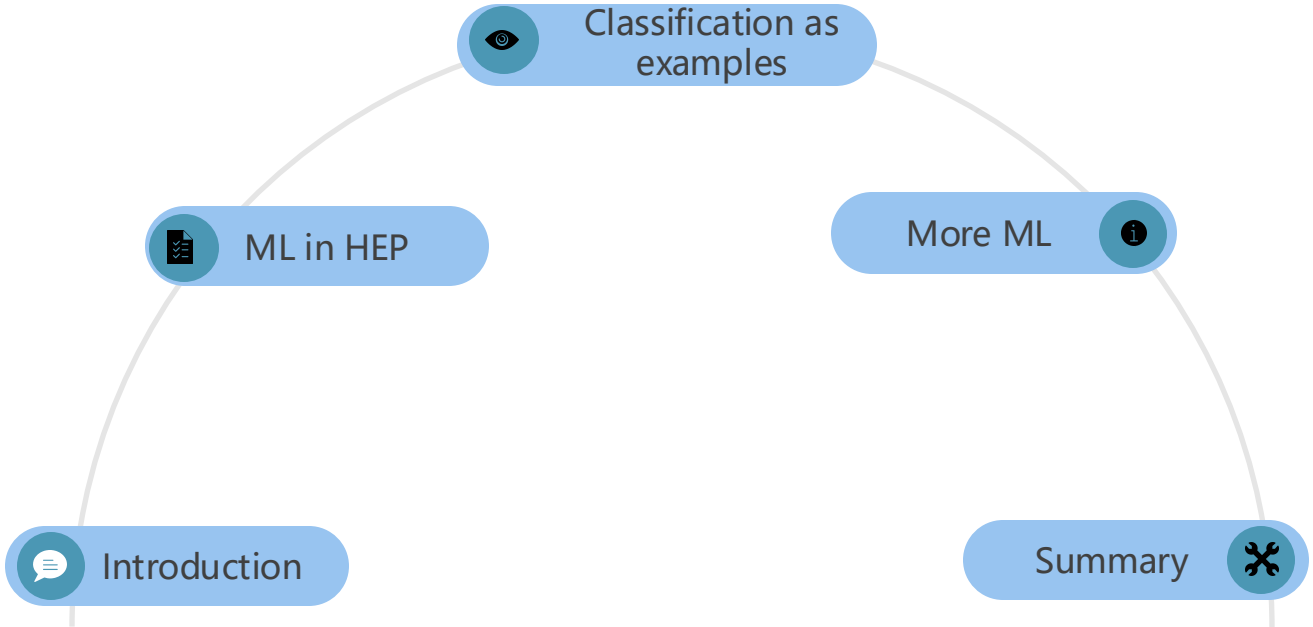
IHEP School of Computing 2024, 2024.08.21-23

北京延庆

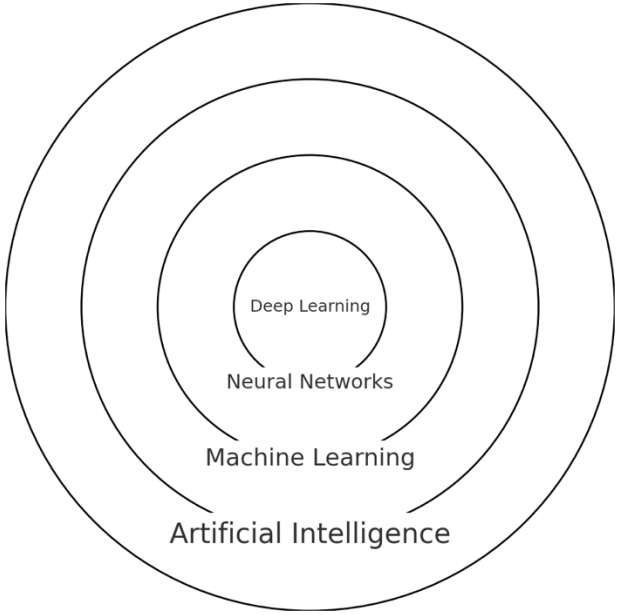
Disclaimers

- This is a very personal review, highly biased
- And mainly focusing on classification problems in offline data processing

Outline

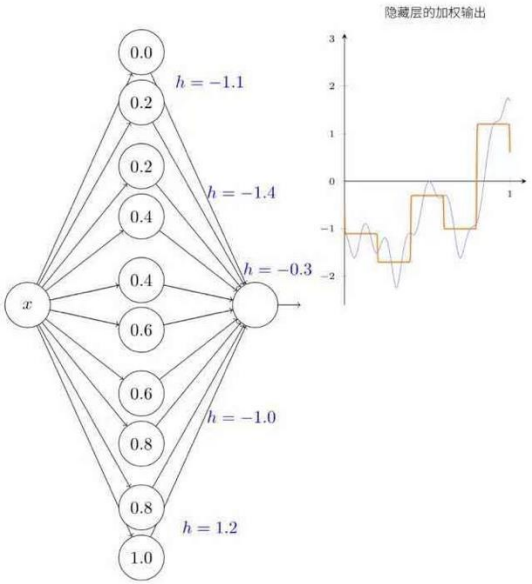

Classification as examples
ML in HEP
More ML
Introduction
Summary

What is Machine Learning ?

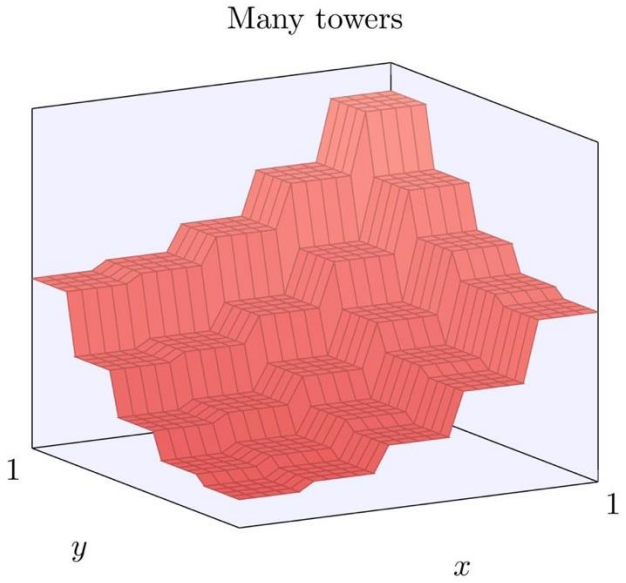


- ✓ Field of study that gives computers the ability to learn without being explicitly programmed
- ✓ A set of rules that allows systems to learn directly from examples, data and experience
- ✓ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E
- ✓ Machine learning is a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data or other outcomes of interest
- ✓ … …

事实 1: Neural network as universal function approximator



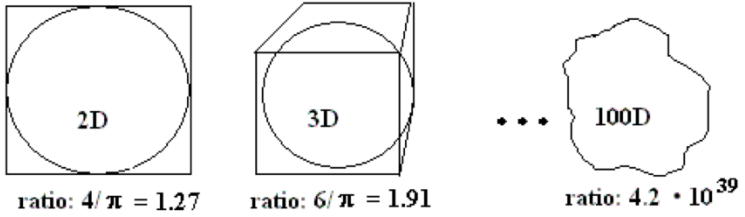
1D



2D

A notable fact about neural networks is that they can approximate a continuous function to any desired level of precision, provided that there are enough neurons in the hidden layers.

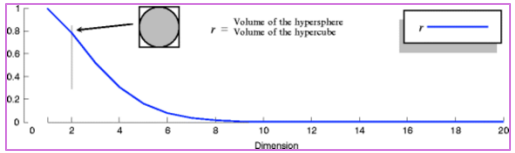
事实 2 : Curse of dimensionality

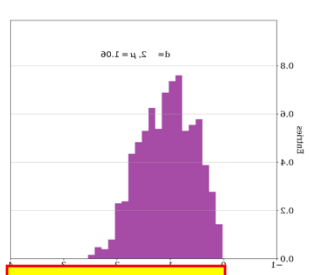


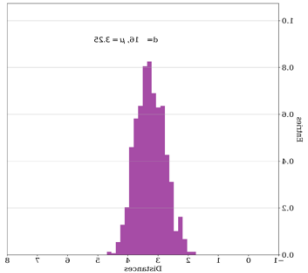
$$\frac{A_{circle}}{A_{square}} = \frac{\pi}{4} \text{ for } d = 2$$

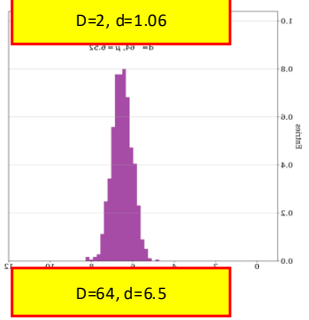
$$\frac{V_{sphere}}{V_{cube}} = \frac{\pi}{6} \text{ for } d = 3$$

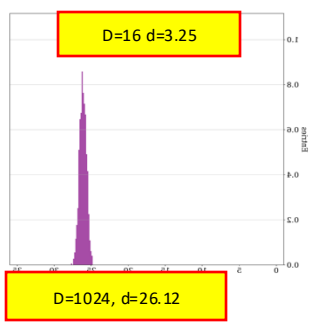
$$\frac{V_{hypersphere}}{V_{hypercube}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0 \text{ as } d \to \infty$$











- When D=1: 100 evenly distributed points can sample a unit interval with a distance no greater than 0.01;
- When D=10: it requires 10^{20} sampling points to achieve the same sampling rate.
- Almost all points in high-D are isolated

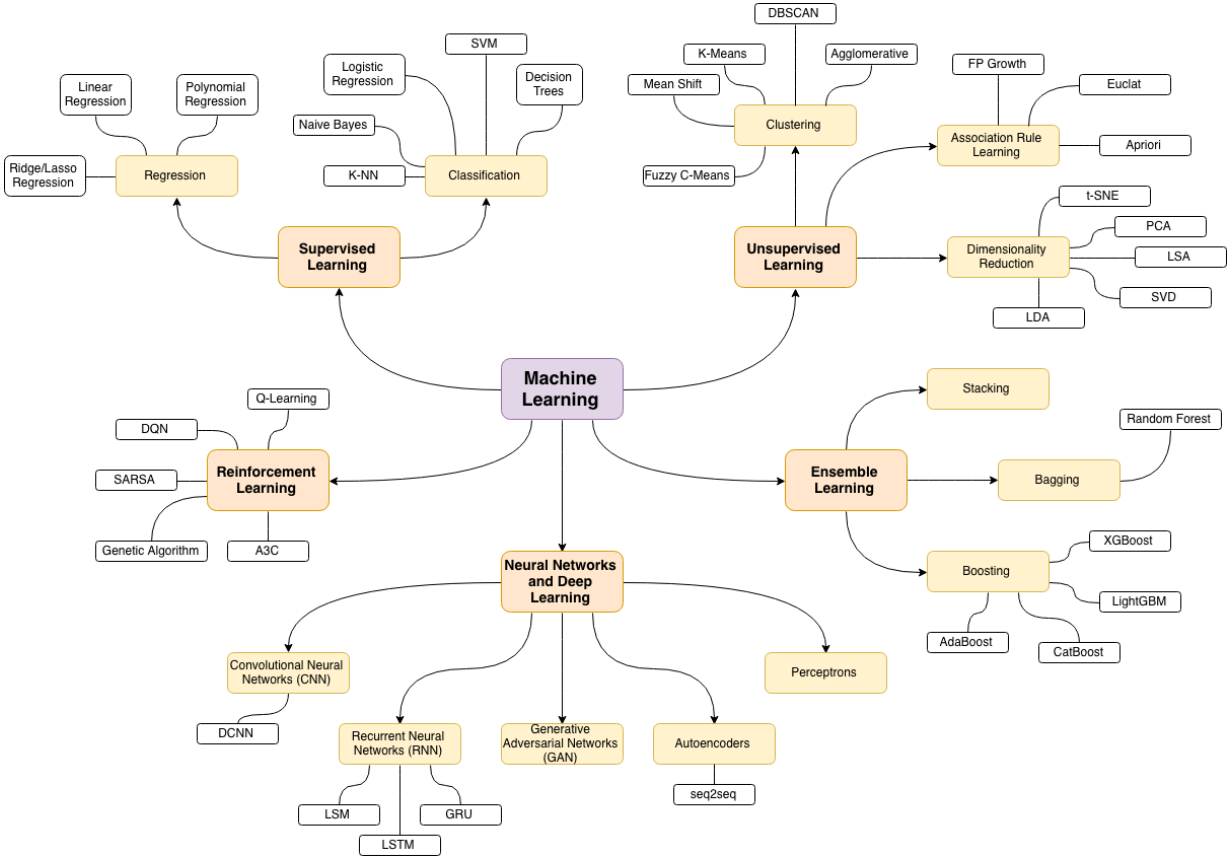
On one hand: fortunately most specific problems can be reduced in dimensionality!

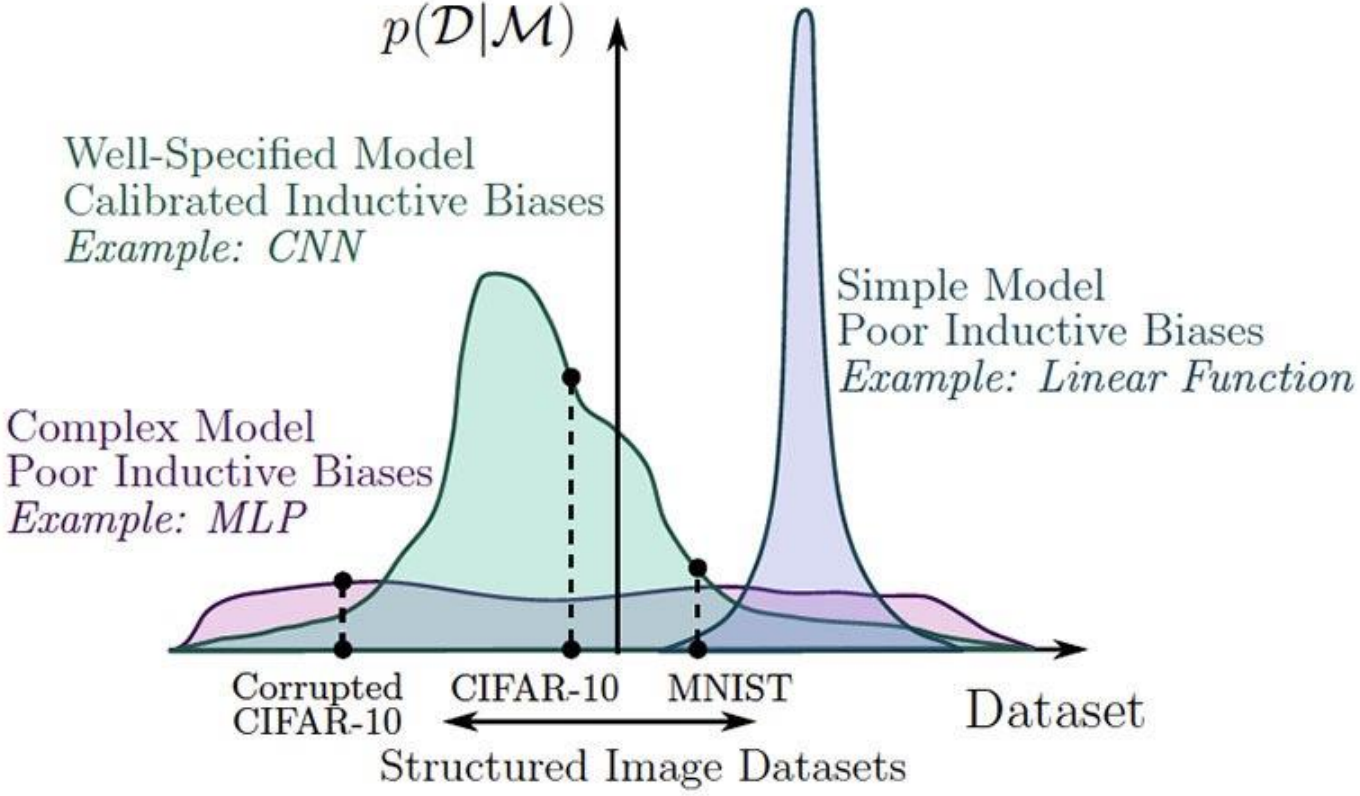
Neural networks have demonstrated their ability to effectively address the dimension problem!

事实 3: No free lunch theorem (http://www.no-free-lunch.org)

There is no single algorithm that is universally the best for all problems

Performance of a learning algorithm is problem-specific





Why do some models perform well on certain datasets? Inductive bias

DeepMind, et al, arXiv:1806.01261

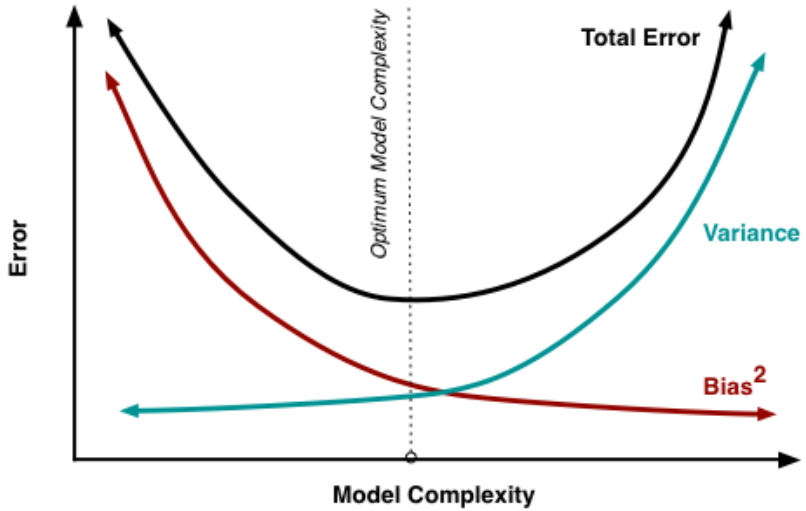
事实 4: bias-variance tradeoff

- 机器学习模型的目标是最小化预测误差。
- 预测误差可以分解为偏差、方差和不可减少的误差。
- Bias: 模型预测值与真实值之间的差异。高偏差导致模型在训练集上的表现不佳。
- Variance :模型在不同数据集上的预测结果的波动。高方差导致模型在新数据上的泛化能力差

$$\text{Total Error}^2 = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}^2$$

✓ 理解偏差和方差的关系对于构建有效模型至关重要。

✓ 持续调整模型以找到最佳平衡点。



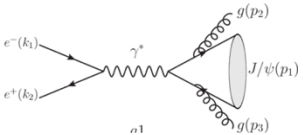
ML in HEP experiments

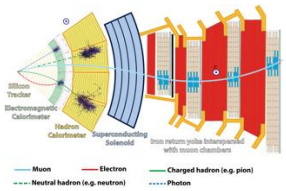
Noise suppression
(FPGA/ASIC)

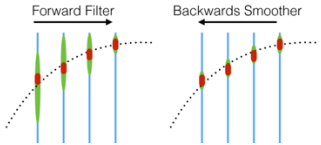
From Y.C. Zhai

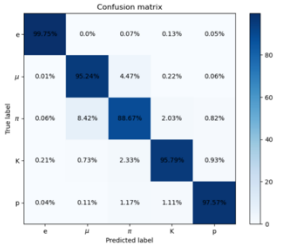
Nature

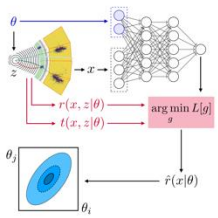
Detector











Reconstruction

Event selection

Statistical inference

Generators

Simulation

Complex processes

Detector response:
e.g. Diffusion model for showering

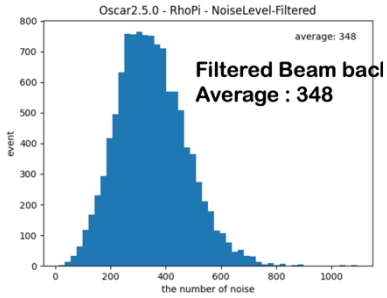
Hits →final state particle :

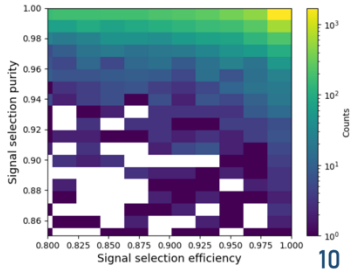
- Pattern recognition/clustering
- Regression

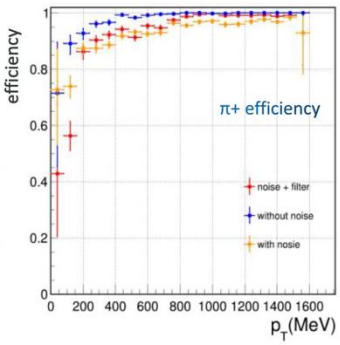
(Multi-)Classification

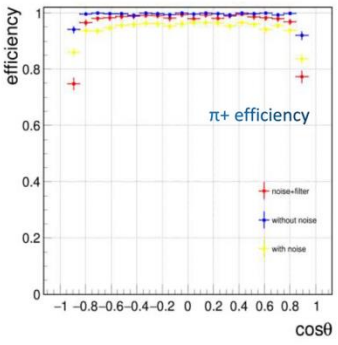
Likelihood free inference

X. Jia: CNN for tracking

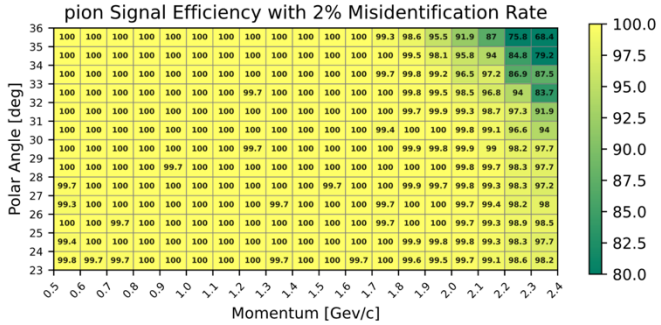




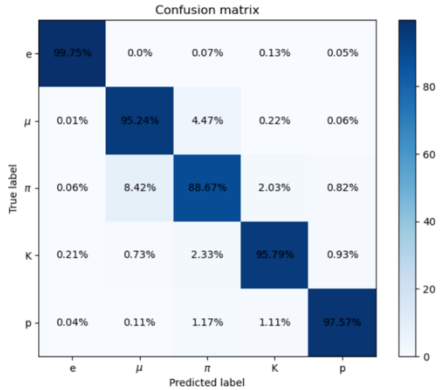




Z. Yao: CNN for PID



Y. Zhai: BDT for (global)PID

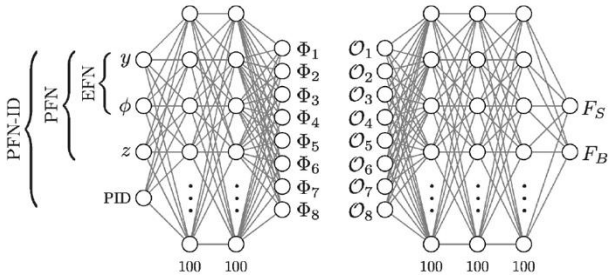


(Multi-)Classification problem

- Jet tagging/W tagger
- Event classification

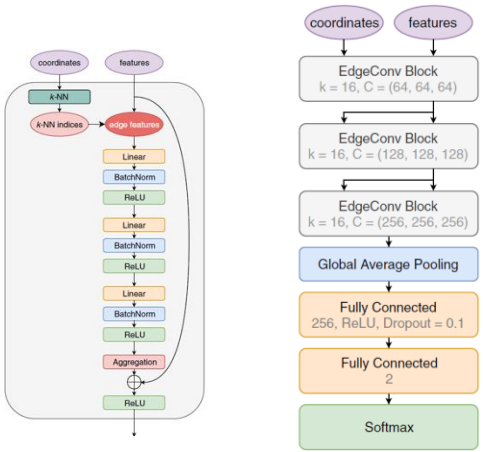
Algorithms used

Energy Flow Network(EFN) / Particle Flow Network(PFN)



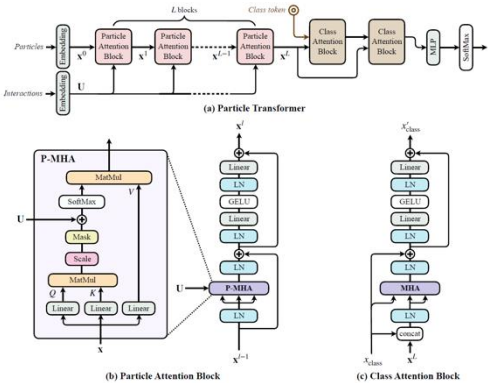
P. T. Komiske, E. M. Metodiev and J. Thaler [*JHEP01(2019)121*]

ParticleNet



H. Qu and L. Gouskos [*Phys.Rev.D 101 (2020) 5, 056019*]

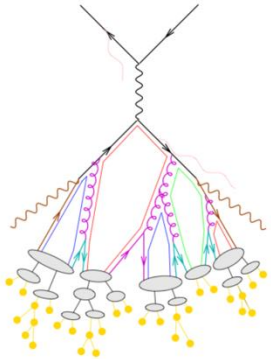
ParticleTransformers (ParT)

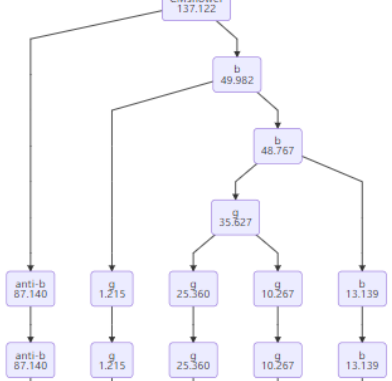


H. Qu , C. Li, S. Qian [*2202.03772*]

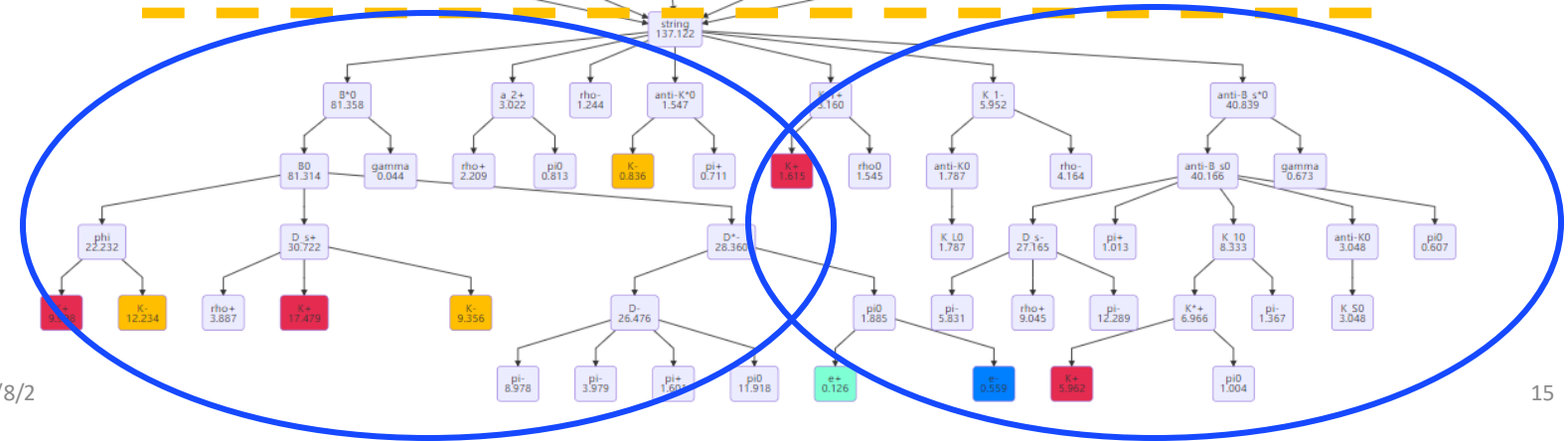
$e^+e^- \to b\bar{b}$

Hard process





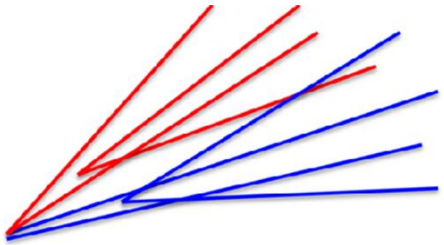
Fragmentation



Jet (flavor) tagging(单个喷注)

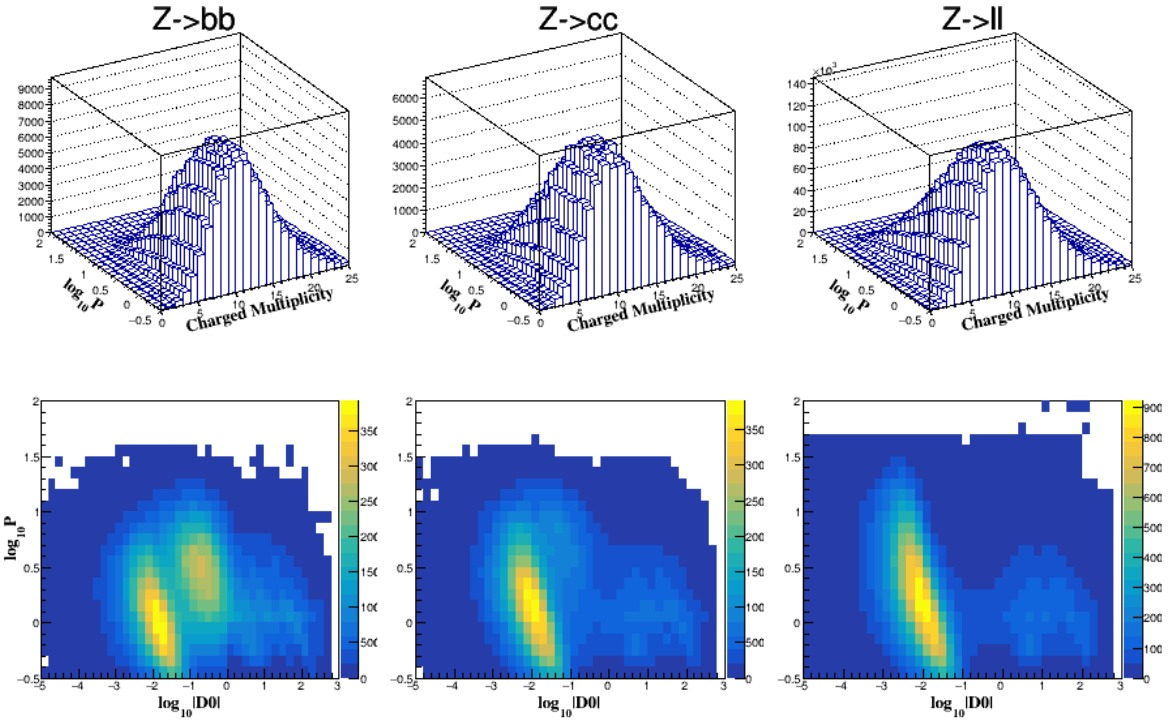
- 91 GeV
- Z → bb, cc, ll (uu,dd,ss)
- 450k events（900k jets）for each class

- Take particle level information a input

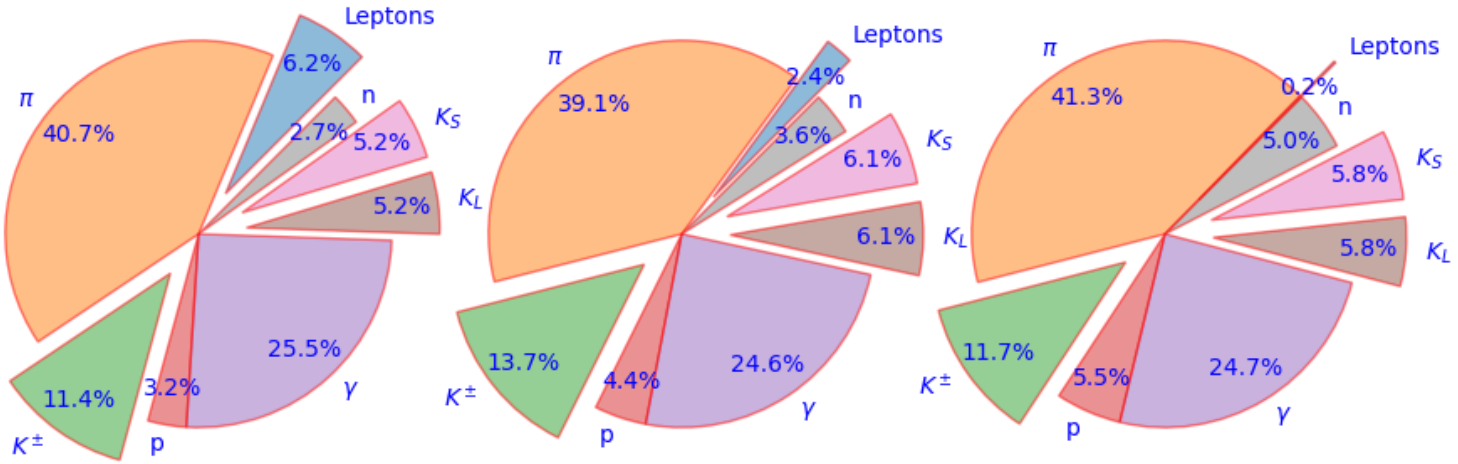


- 4-momenta
- d0/z0
- PID
- … …

Multiplicity, impact parameters



PID information



b c light

Weighted by momenta

Accuracy →

Algorithm	ParticleNet	PFN	DNN	BDT	GBDT	gcforest	XGBoost
Accuracy	0.872	0.850	0.788	0.776	0.794	0.785	0.801

Purity × efficiency →

tag	$\epsilon_S(\%)$	$\epsilon \times \rho$			
		LCFIPlus	XGBoost	ParticleNet	PFN
b	60	-	-	0.589	0.596
	70	-	-	0.694	0.689
	80	-	0.747	0.780	0.763
	90	0.72	0.713	0.810	0.752
	95	-	0.609	0.721	0.645
c	60	0.36	-	0.548	0.485
	70	-	-	0.589	0.497
	80	-	0.345	0.584	0.467
	90	-	0.292	0.516	0.402
	95	-	0.251	0.451	0.348

Take c-tagging as example

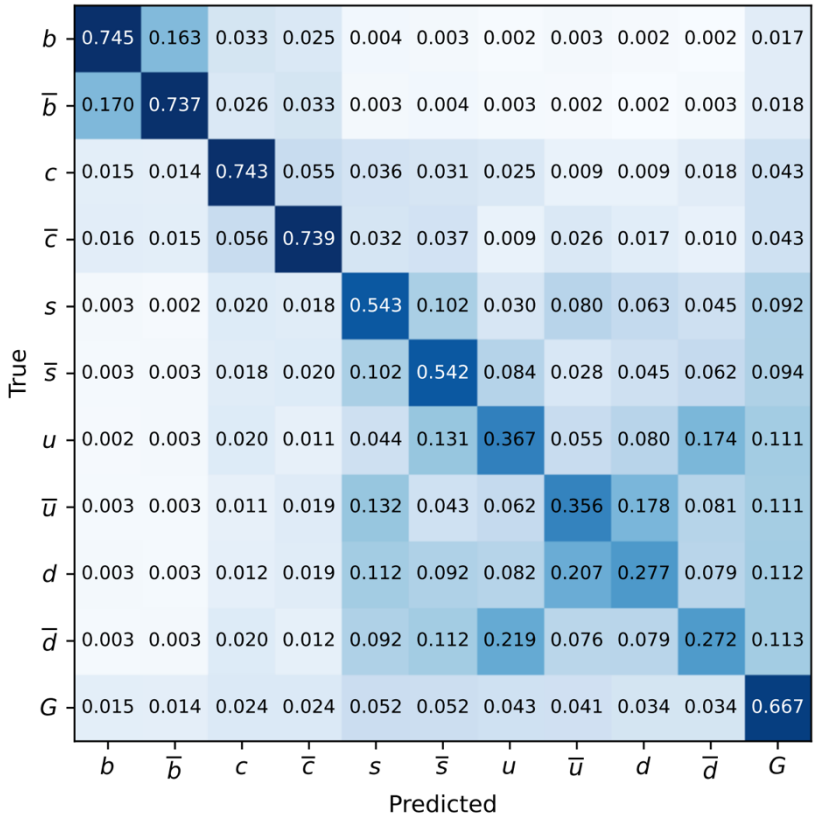
sqrt(0.584/0.345)=1.3

Statistical uncertainty: 30% ↓

$$\frac{1}{(\Delta\sigma_s)^2} = \frac{1}{\sigma_s}\mathcal{L}\epsilon_s\rho = \frac{1}{\sigma_s^2}S_{\text{tot}}\epsilon_s\rho$$

11 classes

Ambitious test by M. Ruan

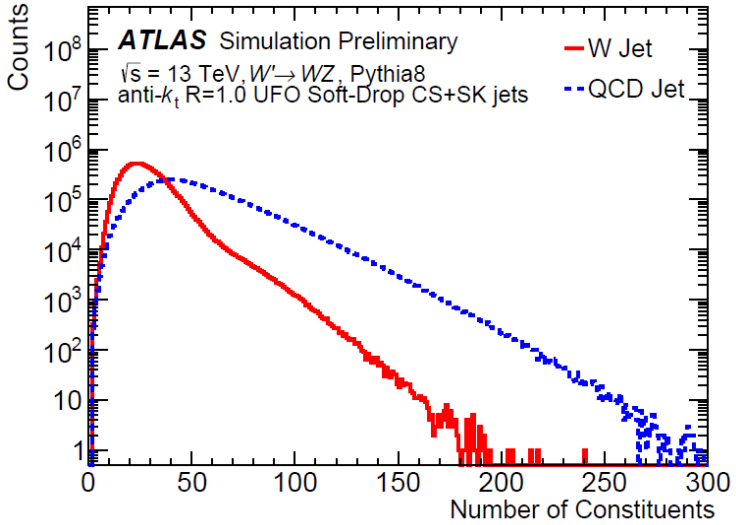


Phys. Rev. Lett. 132, 221802 (2024)

W Jet Taggers (ATLAS, by Shudong Wang)

(一对喷注)

- In this study, a maximum of 200 constituents are considered by all constituent-based taggers. Only a small portion of jets in the dataset have more than 200 constituents (less than 0.04%). As jet constituents are sorted by decreasing p_T, truncation eliminates the softest constituents of the jet.



Distributions of the number of constituents in a large-R jet.

W Jet Taggers

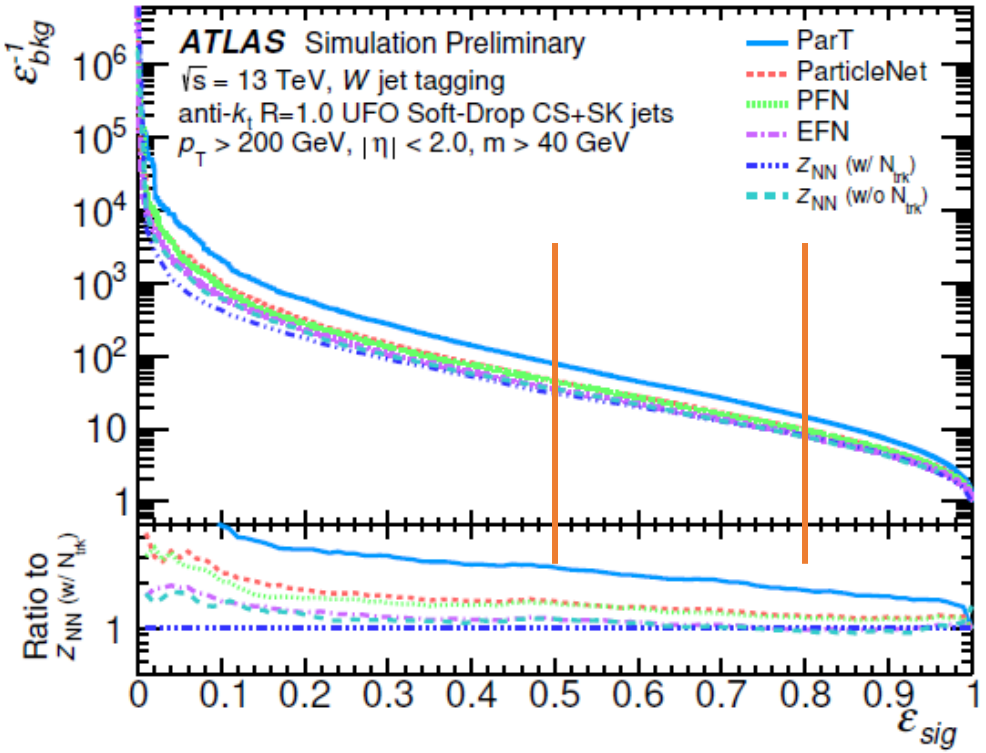
- Particle Flow Network(PFN)/Energy Flow Network(EFN)
  - Based on Deep Sets Theorem
  - *JHEP01(2019)121*
- ParticleNet
  - Customized graph neural network architecture for jet tagging with the point cloud approach
  - *Phys.Rev.D 101 (2020) 5, 056019*
- ParticleTransformer
  - Transformer designed for particle physics
  - *arxiv: 2202.03772*

Models	**Input variables**
EFN	$\Delta\eta$, $\Delta\phi$, $\ln p_{\mathrm{T}}$
PFN	$\Delta\eta$, $\Delta\phi$, $\ln p_{\mathrm{T}}$, $\ln E$, $\ln \frac{p_{\mathrm{T}}}{\sum_{jet} p_{\mathrm{T}}}$, $\ln \frac{E}{\sum_{jet} E}$, ΔR
ParticleNet	$\Delta\eta$, $\Delta\phi$, $\ln p_{\mathrm{T}}$, $\ln E$, $\ln \frac{p_{\mathrm{T}}}{\sum_{jet} p_{\mathrm{T}}}$, $\ln \frac{E}{\sum_{jet} E}$, ΔR
ParticleTransformer	$\Delta\eta$, $\Delta\phi$, $\ln p_{\mathrm{T}}$, $\ln E$, $\ln \frac{p_{\mathrm{T}}}{\sum_{jet} p_{\mathrm{T}}}$, $\ln \frac{E}{\sum_{jet} E}$, ΔR (E, p_x, p_y, p_z)

Tagger Performance

Calculated using samples with steeply falling pT spectra, i.e. both sig & bkg are weighted to have falling pT spectra.

For a signal efficiency of 0.5 (0.8) case, the background rejection of ParticleTransformer is about 1.8-2.8 (1.6-2.7) times better than the baseline tagger.



Tagger Performance

Model	AUC	ACC	ε_{bkg}^{-1} @ $\varepsilon_{sig} = 0.5$	ε_{bkg}^{-1} @ $\varepsilon_{sig} = 0.8$	# Params	Inference Time
EFN	0.920	0.835	35.1	7.95	56.73k	0.065 ms
PFN	0.931	0.853	44.7	9.50	57.13k	0.11 ms
ParticleNet	0.933	0.826	46.2	9.76	366.16k	0.36 ms
ParticleTransformer	0.951	0.880	77.9	14.6	2.14M	0.28 ms

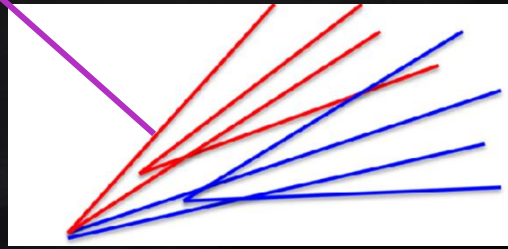
Table 3: The performance of each W jet tagger is measured with several metrics evaluated on the testing set.

Transformers the best

But the # of parameters is almost one order of magnitude larger

$e^+e^- \rightarrow ZH \rightarrow 4\ jets$

(一个事例喷注)



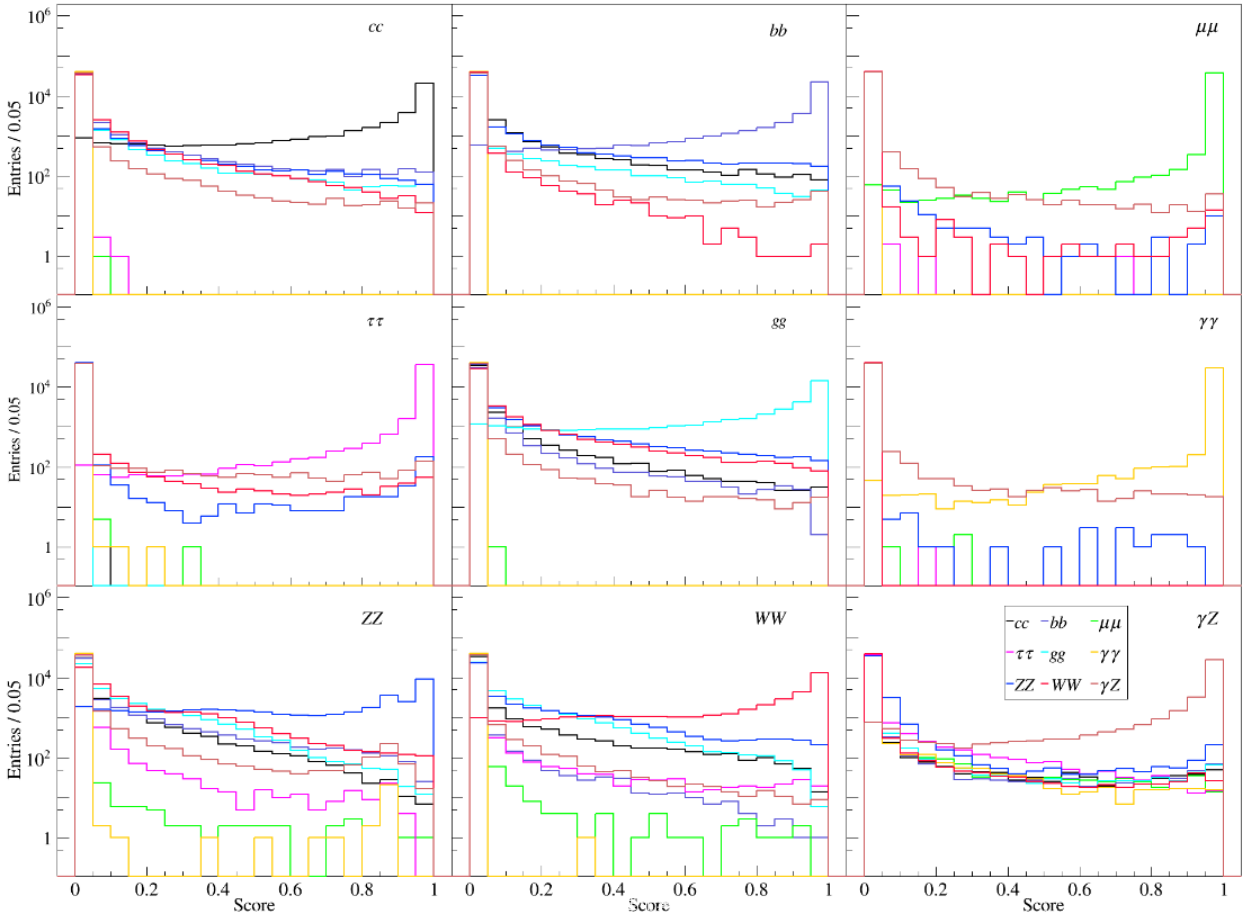
Many processes are selected simultaneously

Prod/decay	cc	bb	mm	ττ	gg	gg	WW	ZZ	aZ	ee, uu,dd,ss
eeH	3	1	5	2	4	1	2	3	5	Not covered yet
mmH	3	1	5	2	4	1	2	3	5	
ττH	3	1	5	2	4	1	2	3	5	
qqH	4	1	2	1	2	5	5	5	3	
nnH	5	1	3	2	3	5	4	2	4	

Consider: psi(2S) → pi+ pi- J/psi, J/psi → various processes

Try eeH first

Probability distributions of each class

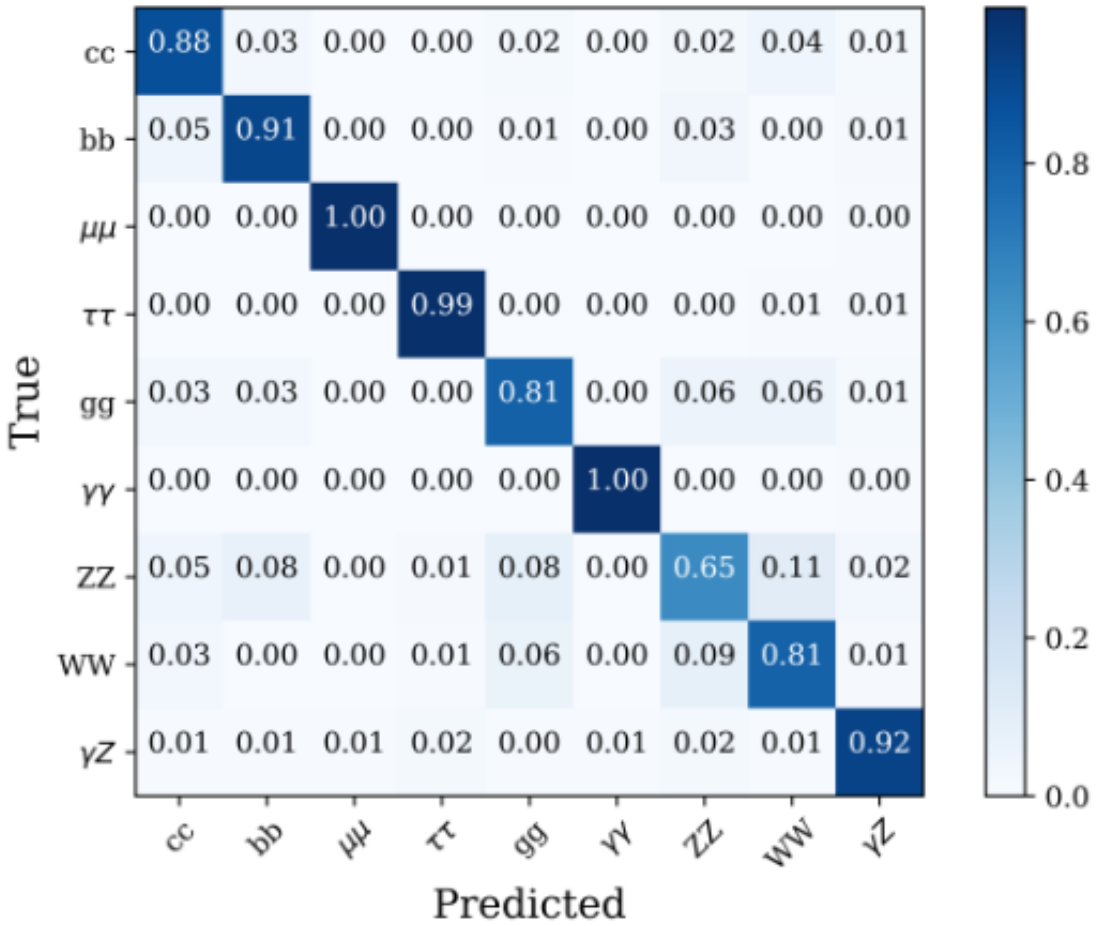


Try eeH first

Sufficiently good performance

Average Accuracy ~ 87%

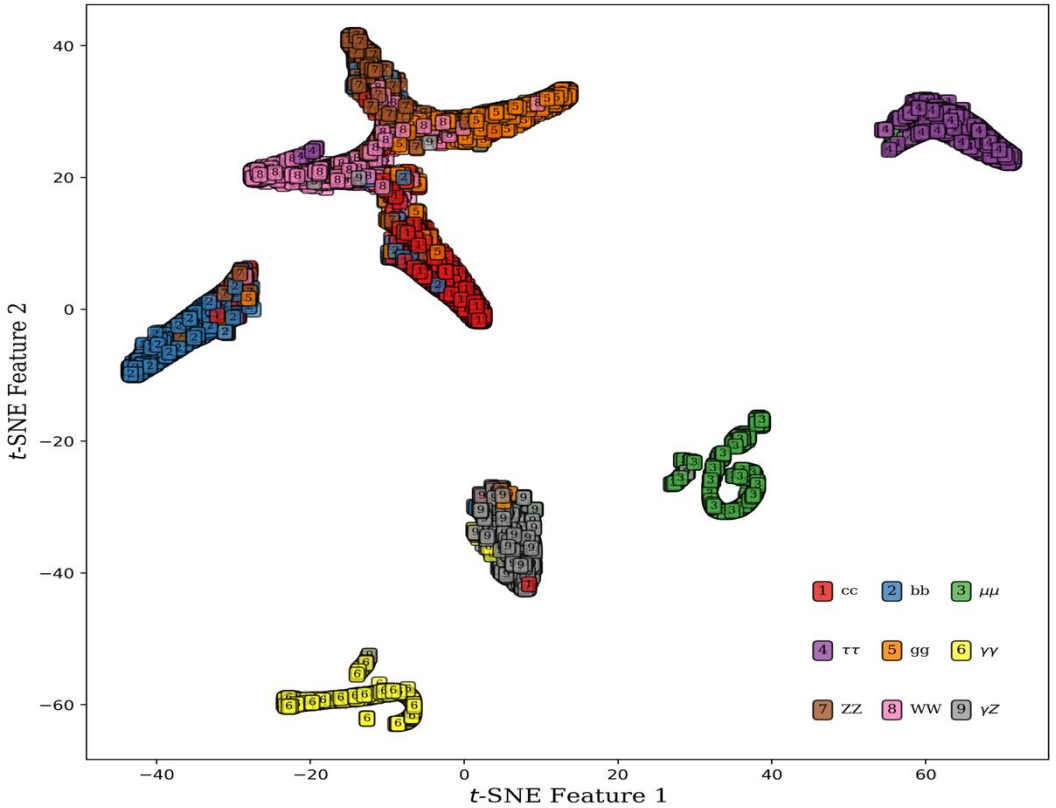
(11% for random guess)



Taking the one has largest probability (ArgMax)

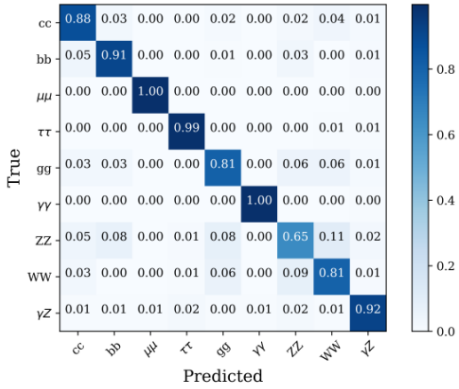
Dimension reduction tells us more

- ✓ μμ, γγ, ττ well classified as expected
- ✓ bb and γZ also good
- ✓ cc, gg, WW, and ZZ fake each other, but under control

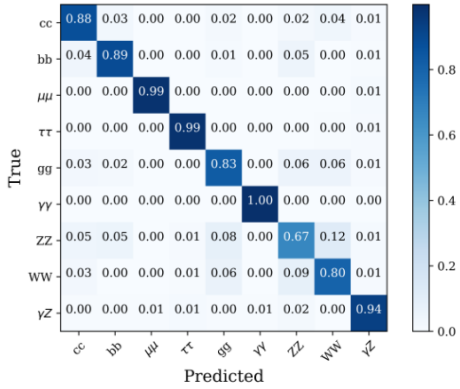


Dimensional reduction (t-SNE)

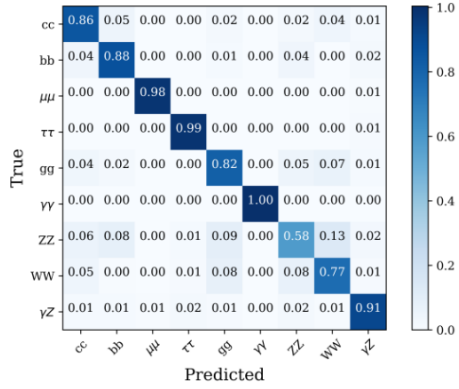
All 4 production modes



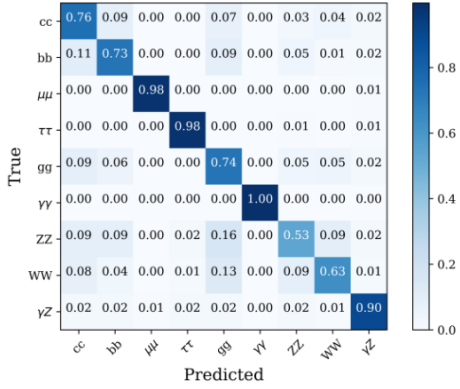
eeH



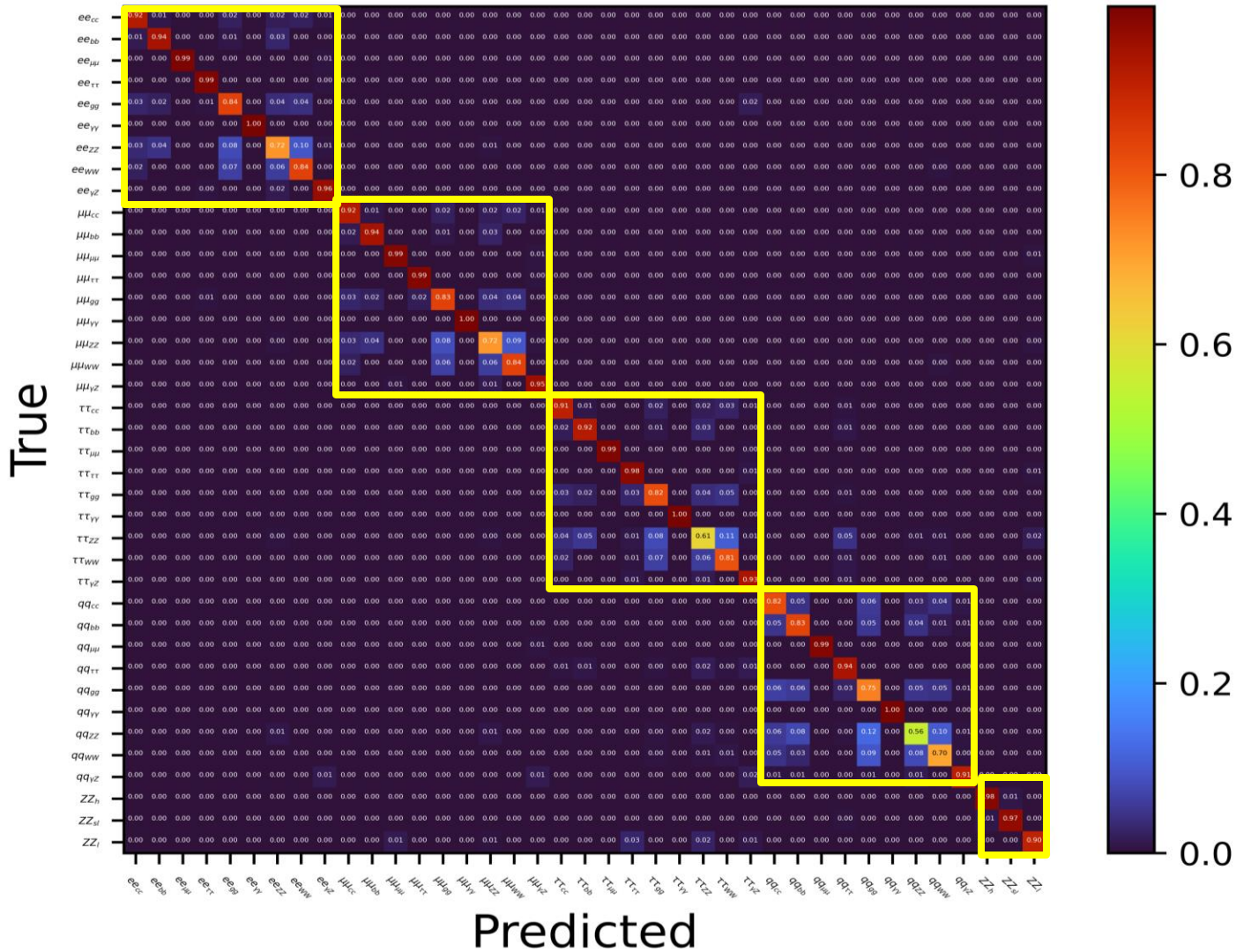
$\mu\mu$H



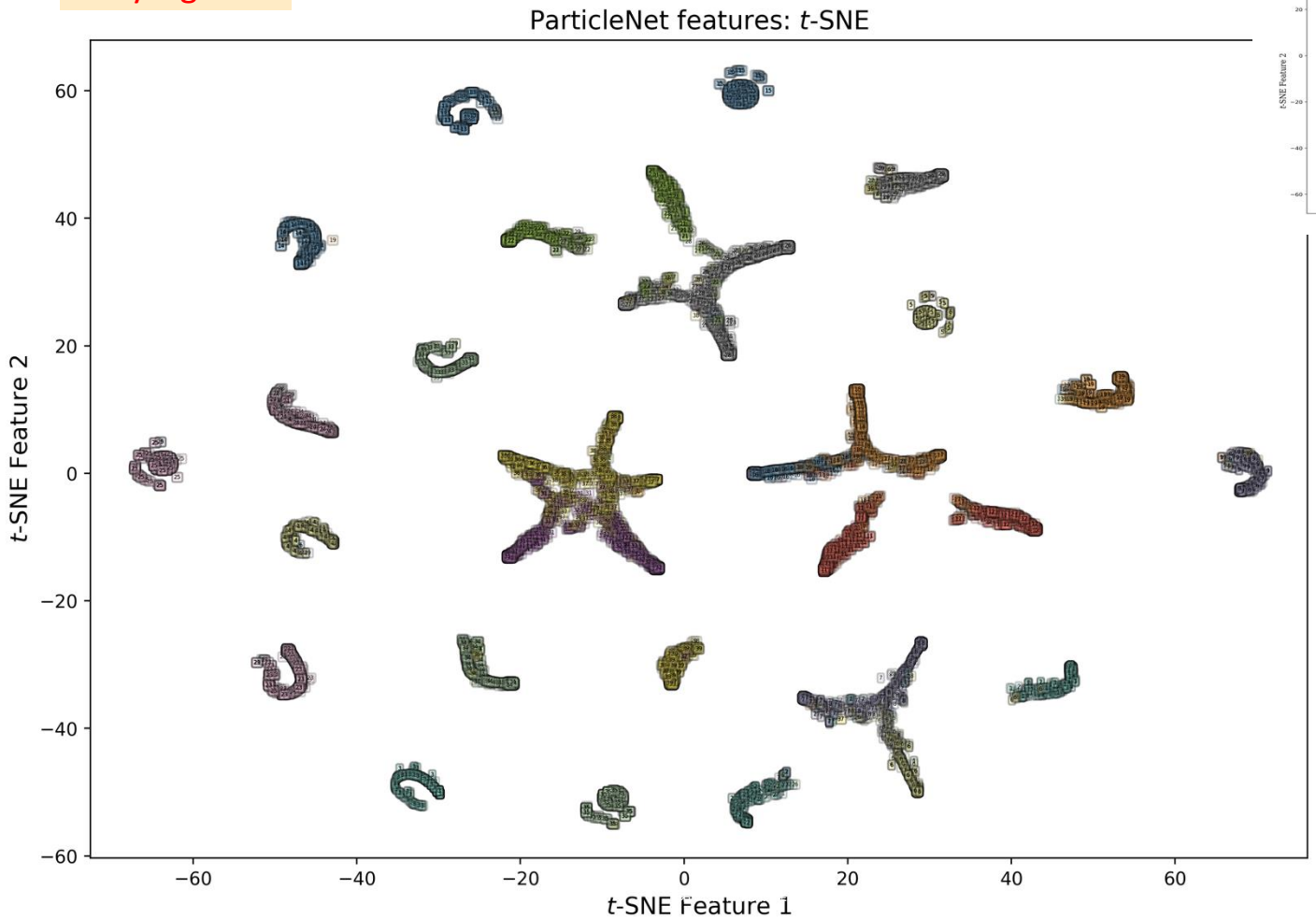
$\tau\tau$H

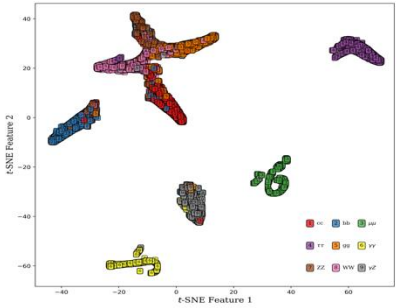


qqH


True
Predicted
0.8
0.6
0.4
0.2
0.0

Only signals







(LHC 上的“事例”)

Based on arXiv:2405.12972

[Github] [Google Colab]

Sophon meets LHC:
Accelerating resonance discovery via signature-oriented pre-training

Congqiao Li (李聪乔), *Peking University*

based on the work with our colleagues in the CMS Collaboration:

Antonios Agapitos[1], Jovin Drews[2], Javier Duarte[3], Dawei Fu[1], Leyun Gao[1], Raghav Kansal[3], Gregor Kasieczka[2], Louis Moureaux[2], Huilin Qu[4], Cristina Mantilla Suarez[5], Qiang Li[1]

1) Peking U. 2) Hamburg U. 3) UC San Diego 4) CERN 5) FNAL

also thanks* Yuzhe Zhao[1] *for his contribution

Quantum Computing and Machine Learning Workshop 2024, Changchun

6 August, 2024

Statistical property of multi-class classifier

→ Statistical theory shows that:

A **multi-class** classifier with minimum **cross-entropy loss estimates the probability ratios** on the input classes:

$$g_i(\mathbf{x}) = \frac{p(\text{class} = i \,|\, \mathbf{x})}{\sum_{j=1}^{N_{\text{out}}} p(\text{class} = j \,|\, \mathbf{x})}$$

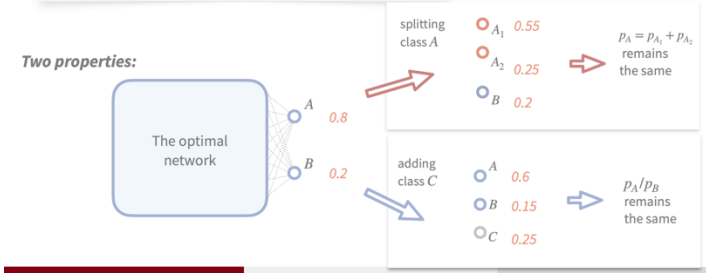
The key question in this context

Does the model's capacity still enable us to reach the best achievable performance in existing tasks?

Our result will show: Yes.

hence it contains all the information the ideal $N(N-1)$ binary classifiers can do

Two properties:



Introducing Sophon

arXiv:2405.12972

https://github.com/jet-universe/sophon

→ We explore this possibility in the CMS experiment first, and also in a recent pheno work:

- **Signature-Oriented Pre-training for Heavy-resonant ObservatioN**
- the model is based on Particle Transformer (**ParT**) architecture ← *[H.Qu, CL, S.Qian. ICML 2022]* *The current "state-of-the-art" on large-scale dataset (o(100M), close to experimental setup)*
- a pre-trained model on a comprehensive dataset: **JetClass-II**
  - ***finely categorized labels:***

Major types	Index range	Label names
Resonant jets: $X \to 2$ prong	0–14	bb, cc, ss, qq, bc, cs, bq, cq, sq, gg, ee, $\mu\mu$, $\tau_h e$, $\tau_h \mu$, $\tau_h \tau_h$
Resonant jets: $X \to 3$ or 4 prong	15–160	$bbbb$, $bbcc$, $bbss$, $bbqq$, $bbgg$, $bbee$, $bb\mu\mu$, $bb\tau_h e$, $bb\tau_h \mu$, $bb\tau_h \tau_h$, bbb, bbc, bbs, bbq, bbg, bbe, $bb\mu$, $cccc$, $ccss$, $ccqq$, $ccgg$, $ccee$, $cc\mu\mu$, $cc\tau_h \tau_e$, $cc\tau_h \tau_\mu$, $cc\tau_h \tau_h$, ccb, ccc, ccs, ccq, ccg, cce, $cc\mu$, $ssss$, $ssqq$, $ssgg$, $ssee$, $ss\mu\mu$, $ss\tau_h \tau_e$, $ss\tau_h \tau_\mu$, $ss\tau_h \tau_h$, ssb, ssc, sss, ssq, ssg, sse, $ss\mu$, $qqqq$, $qqgg$, $qqee$, $qq\mu\mu$, $qq\tau_h \tau_e$, $qq\tau_h \tau_\mu$, $qq\tau_h \tau_h$, qqb, qqc, qqs, qqq, qqg, qqe, $qq\mu$, $gggg$, $ggee$, $gg\mu\mu$, $gg\tau_h \tau_e$, $gg\tau_h \tau_\mu$, $gg\tau_h \tau_h$, ggb, ggc, ggs, ggq, ggg, gge, $gg\mu$, bbe, cce, see, qee, gee, $b\mu\mu$, $c\mu\mu$, $s\mu\mu$, $q\mu\mu$, $g\mu\mu$, $b\tau_h \tau_e$, $c\tau_h \tau_e$, $s\tau_h \tau_e$, $q\tau_h \tau_e$, $g\tau_h \tau_e$, $b\tau_h \tau_\mu$, $c\tau_h \tau_\mu$, $s\tau_h \tau_\mu$, $q\tau_h \tau_\mu$, $g\tau_h \tau_\mu$, $b\tau_h \tau_h$, $c\tau_h \tau_h$, $s\tau_h \tau_h$, $q\tau_h \tau_h$, $g\tau_h \tau_h$, qqb, $qqqc$, $qqqs$, $bbcq$, $ccbs$, $ccbq$, $ccsq$, $sscq$, $qqbc$, $qqbs$, $qqcs$, $bcsq$, bcs, bcq, bsq, csq, $bc\epsilon\nu$, $cse\nu$, $bqe\nu$, $cqe\nu$, $sqe\nu$, $qqe\nu$, $bc\mu\nu$, $cs\mu\nu$, $bq\mu\nu$, $cq\mu\nu$, $sq\mu\nu$, $qq\mu\nu$, $bc\tau_e\nu$, $cs\tau_e\nu$, $bq\tau_e\nu$, $cq\tau_e\nu$, $sq\tau_e\nu$, $qq\tau_e\nu$, $bc\tau_\mu\nu$, $cs\tau_\mu\nu$, $bq\tau_\mu\nu$, $cq\tau_\mu\nu$, $sq\tau_\mu\nu$, $qq\tau_\mu\nu$, $bc\tau_h\nu$, $cs\tau_h\nu$, $bq\tau_h\nu$, $cq\tau_h\nu$, $sq\tau_h\nu$, $qq\tau_h\nu$
QCD jets	161–187	$bbccss$, $bbccs$, $bbcc$, $bbcss$, $bbcs$, bbc, $bbss$, bbs, bb, $bccss$, $bccs$, bcc, $bcss$, bcs, bc, bss, bs, b, $ccss$, ccs, cc, css, cs, c, ss, s, others

resulting to 4-prong or 3-prong

All final states!

Sophon: close to real experimental performance?

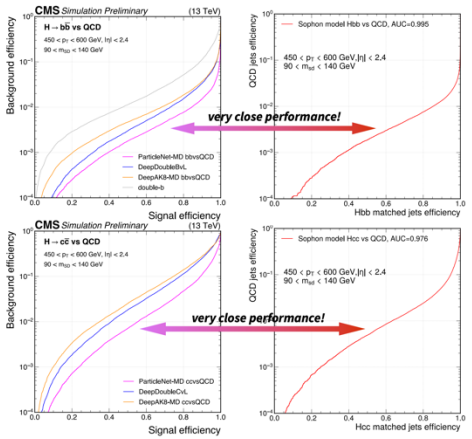
Thanks Yuzhe for his input

Take recent **CMS performance plots** as an example

Benchmark performance on SM H→bb̄/cc̄ jets vs QCD background jets

- Demonstrate that **applying "Sophon" on Delphes dataset for pheno study** is pretty realistic

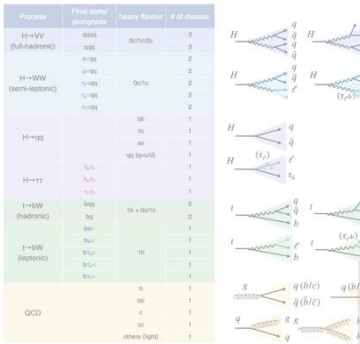
CMS results *CMS-PAS-BTV-22-001*



Sophon results *(on Delphes dataset)*

Highlight: CMS's Global ParT tagger

CMS-PAS-HIG-23-012

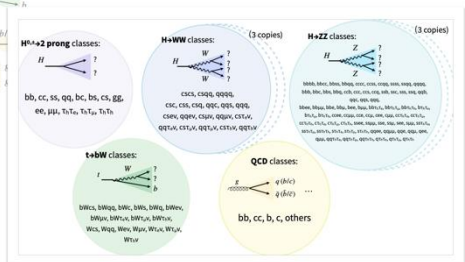


A global large-R mass-decorrelated tagger for **37-category classification**

- First time identifying the H→WW→4q signature with a jet tagger
- set a strong limit to κ_{2V} in the search of HH→bbVV signal

and we plan to update the CMS tagger under Sophon's philosophy

please stay tuned!



More?

Diffusion models

Large Language Models

… …

Used as copilot-like assistant: Dr. Sai

Used to HEP data directly: tokenization is a key

How to represent a HEP events? Tokenization

Feature engineering

Some mathematical methods? Such as fox-wolfram moments

$$H_l \equiv \left(\frac{4\pi}{2l+1}\right) \sum_{m=-l}^{+l} \left| \sum_i Y_l^m(\Omega_i) \frac{|\vec{p}_i|}{\sqrt{s}} \right|^2$$

$$= \sum_{i,j} \frac{|\vec{p}_i||\vec{p}_j|}{s} P_l(\cos\varphi_{ij}),$$

Autoencoder?

作业：实战内容（借用李聪乔的劳动）

- Particle Transformer 用作分类
- https://gitee.com/colizz/qcml-2024-tutorial
- lxslcaaa.ihep.ac.cn → lxloginaaa.ihep.ac.cn （需要修改的地方）
  - ✓chap1 是视频教学：看看就行
  - ✓chap2 是 general 的展示：需要科学上网，了解下就行
  - ✓chap3 我实测可以跑（CPU）：复现结果

Summary

- Machine learning is statistical learning (NFL)
- Machine learning is useful (~~CoD~~): high dimensional HEP data
- Machine learning method with proper bias is helpful and easy to explain
- Machine learning methods can be applied to almost all aspects of HEP experiments.
- LLM demonstrated astonishing capabilities, which are worth exploring from two aspects:

➢ Use LLMs as language-based assistants

➢ Employing LLMs to directly to process data: how to represent HEP events is the key