# Machine Learning in HEP data processing

李　刚

中国科学院高能物理研究所

**Workshop on Super Tau-Charm Facility,**

**2024.07.7-10, Lanzhou**

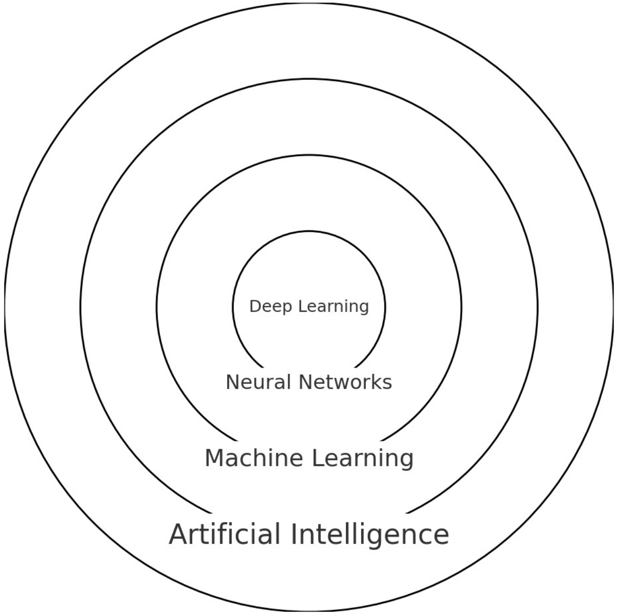
# Disclaimers

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

# Outline

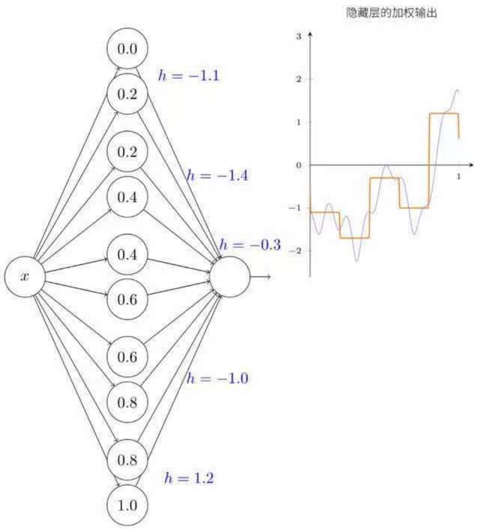
- Introduction
- ML in HEP
- Classification as examples
- More ML
- Summary

# What is Machine Learning ?

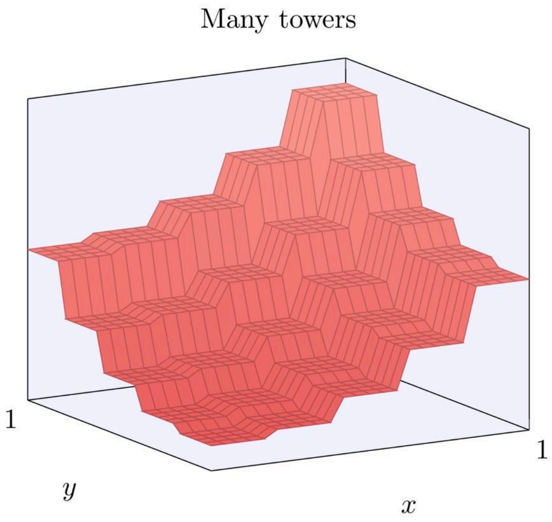



- ✓ Field of study that gives computers the ability to learn without being explicitly programmed
- ✓ A set of rules that allows systems to learn directly from examples, data and experience
- ✓ A computer program is said to learn from experience E with respect to some class of tasks T and performance 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
- ✓ … …

# Fact 1: Neural network as universal function approximator



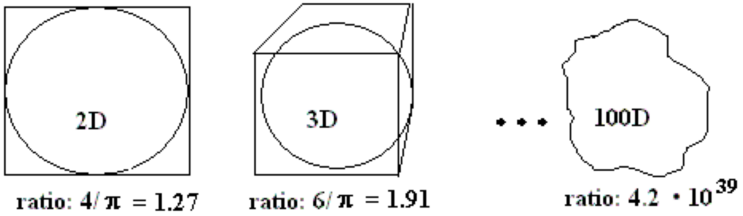

1D




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.

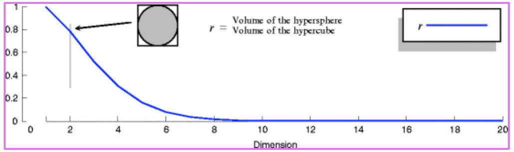
# Fact 2 : Curse of dimensionality

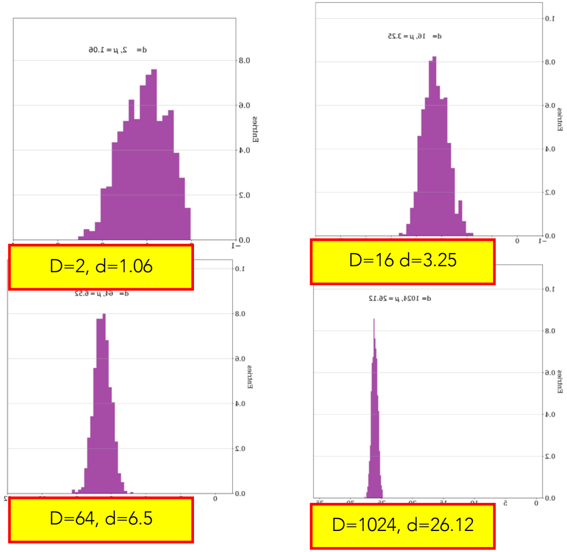



$$\frac{A_{circle}}{A_{square}} = \frac{\pi}{4} \text{ for } d = 2$$

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

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




- 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

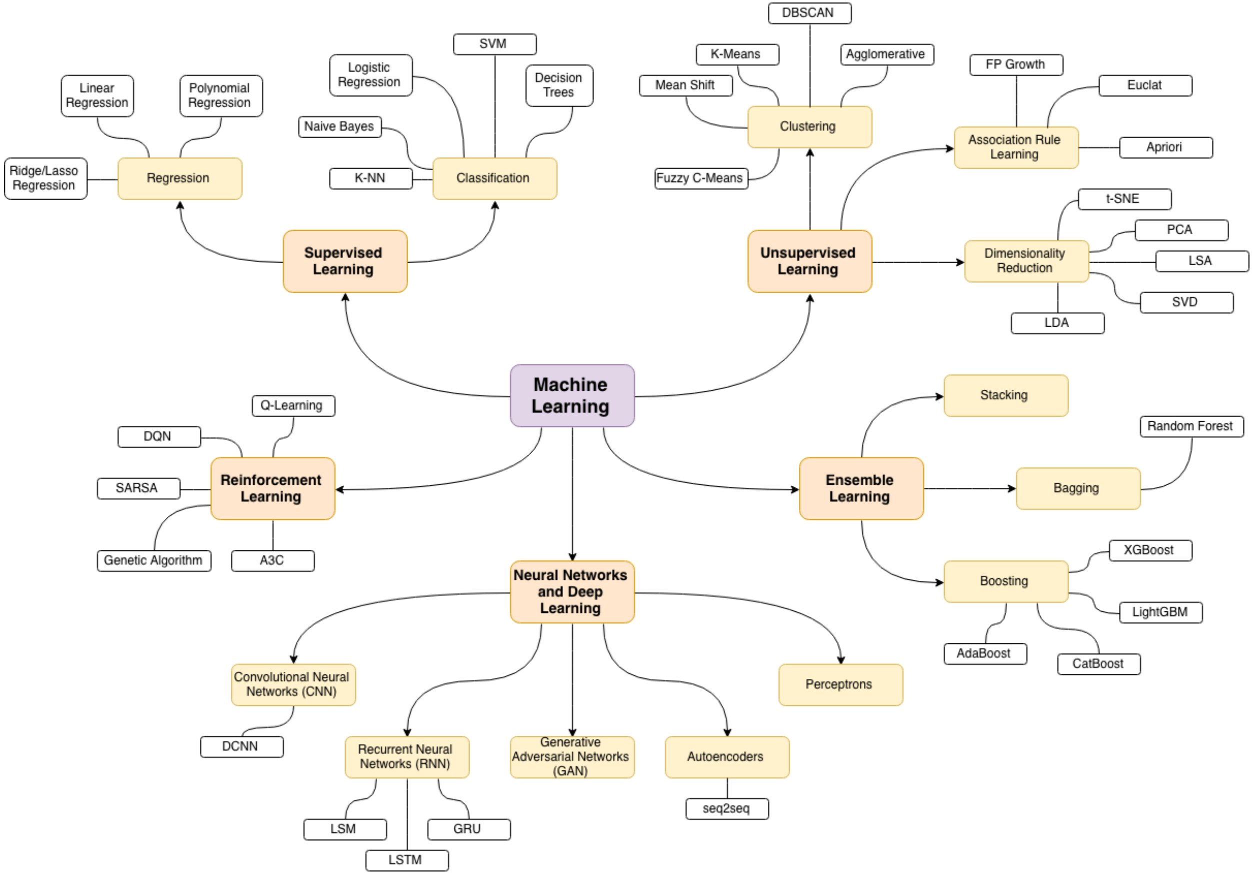
Fortunately most specific problems can be reduced in dimensionality!

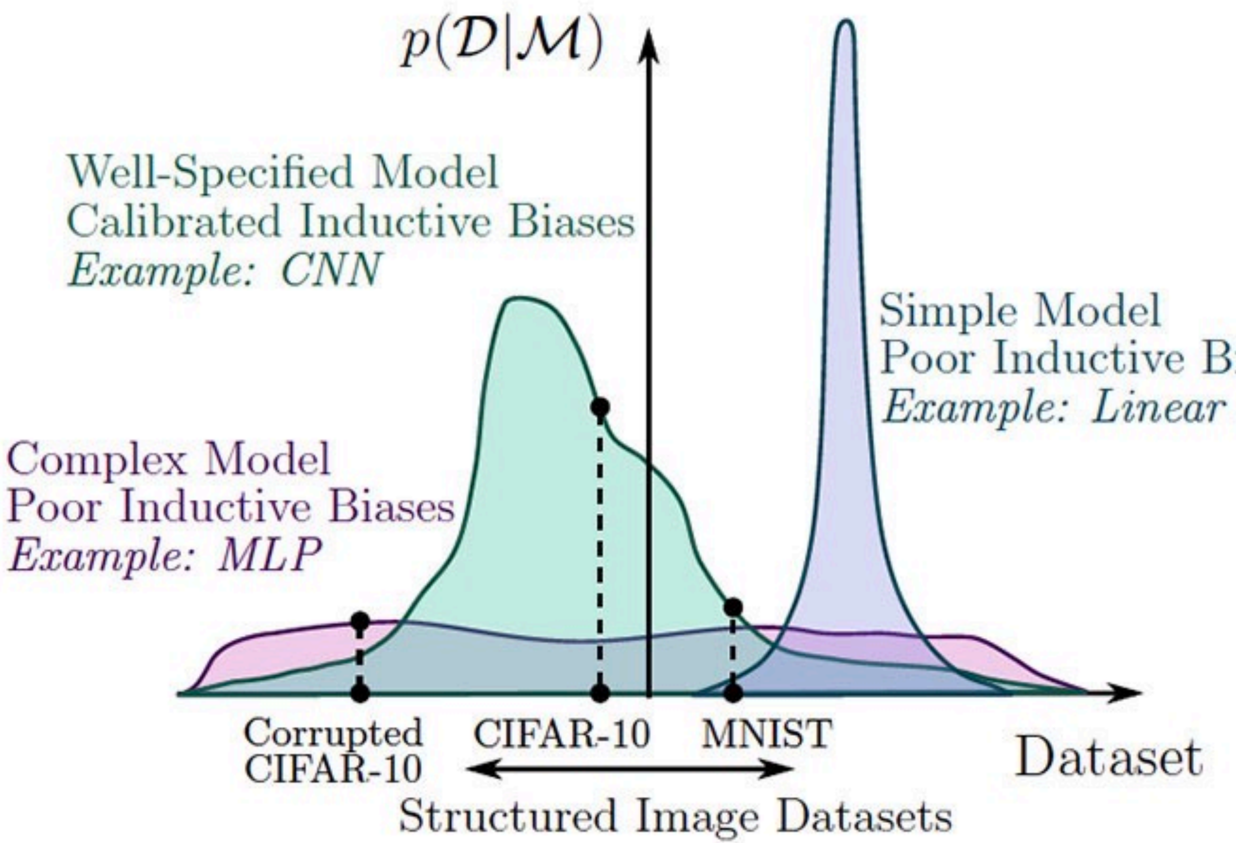
Neural networks have demonstrated their ability to effectively address the dimension problem!

Fact 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

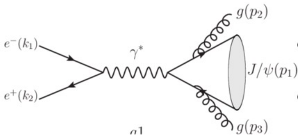


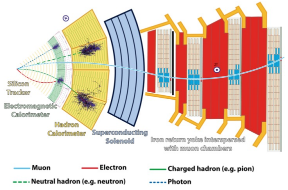
DeepMind, et al, arXiv:1806.01261

# ML in HEP experiments

Noise suppression
(FPGA/ASIC)

Nature → Detector

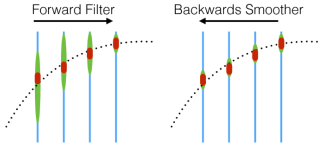


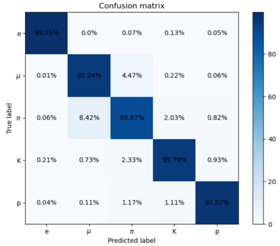

Generators → Simulation

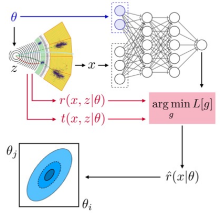
Complex processes

Detector response:
e.g. Diffusion model for showering




From Y.C. Zhai




Reconstruction → Event selection → Statistical inference

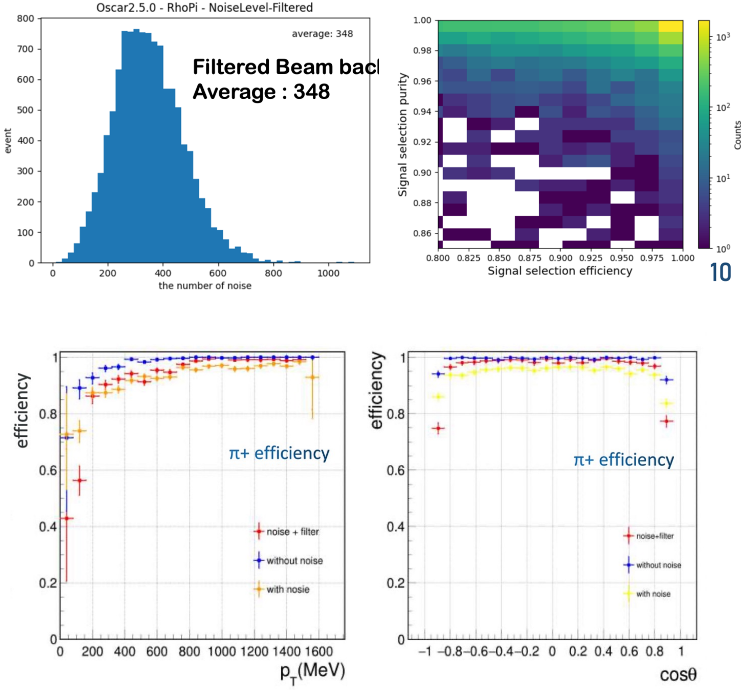
Hits →final state particle :

- Pattern recognition/clustering
- Regression

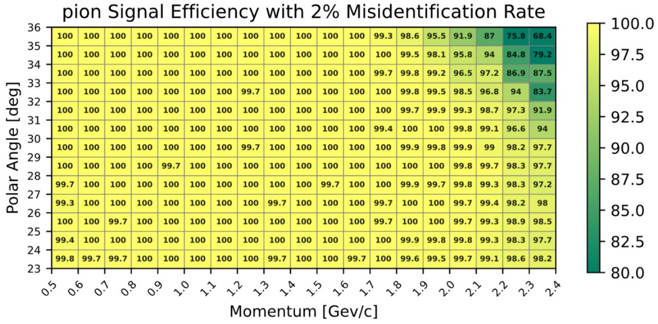
(Multi-)Classification

Likelihood free inference

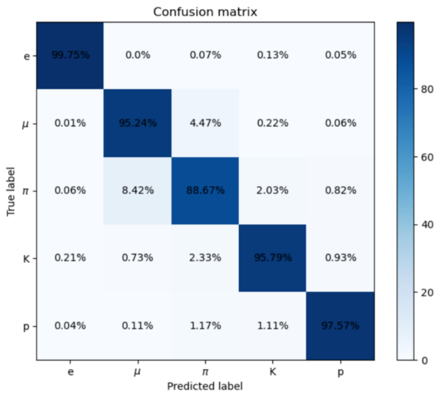
## X. Jia: CNN for tracking




## Z. Yao: CNN for PID




## Y. Zhai: BDT for (global)PID

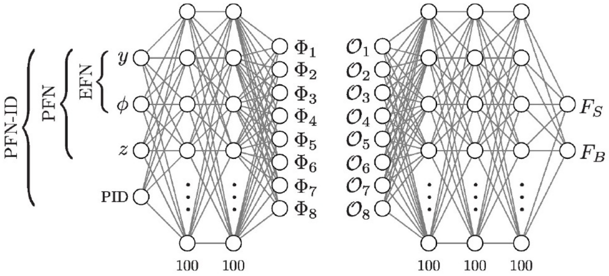


# (Multi-)Classification problem

- Jet tagging/W tagger
- Event classification

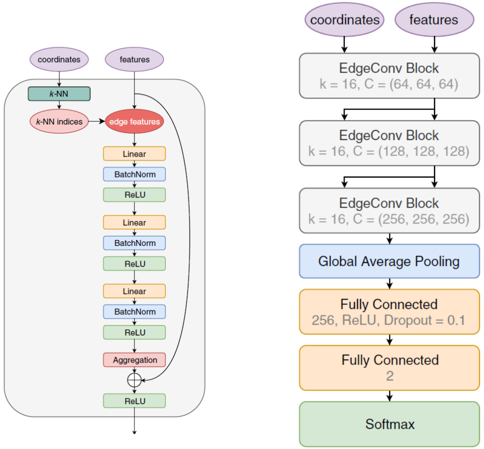
# Algorithms

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



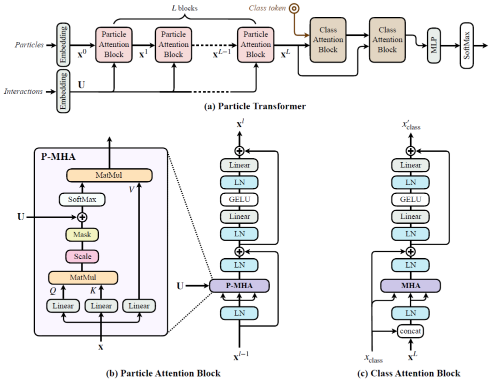

P. T. Komiske, E. M. Metodiev and J. Thaler [JHEP01(2019)121]

## ParticleNet




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

## ParticleTransformers (ParT)

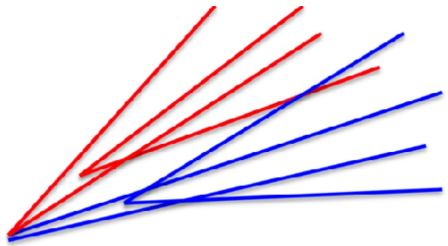



H. Qu , C. Li, S. Qian [2202.03772]

# Jet (flavor) tagging

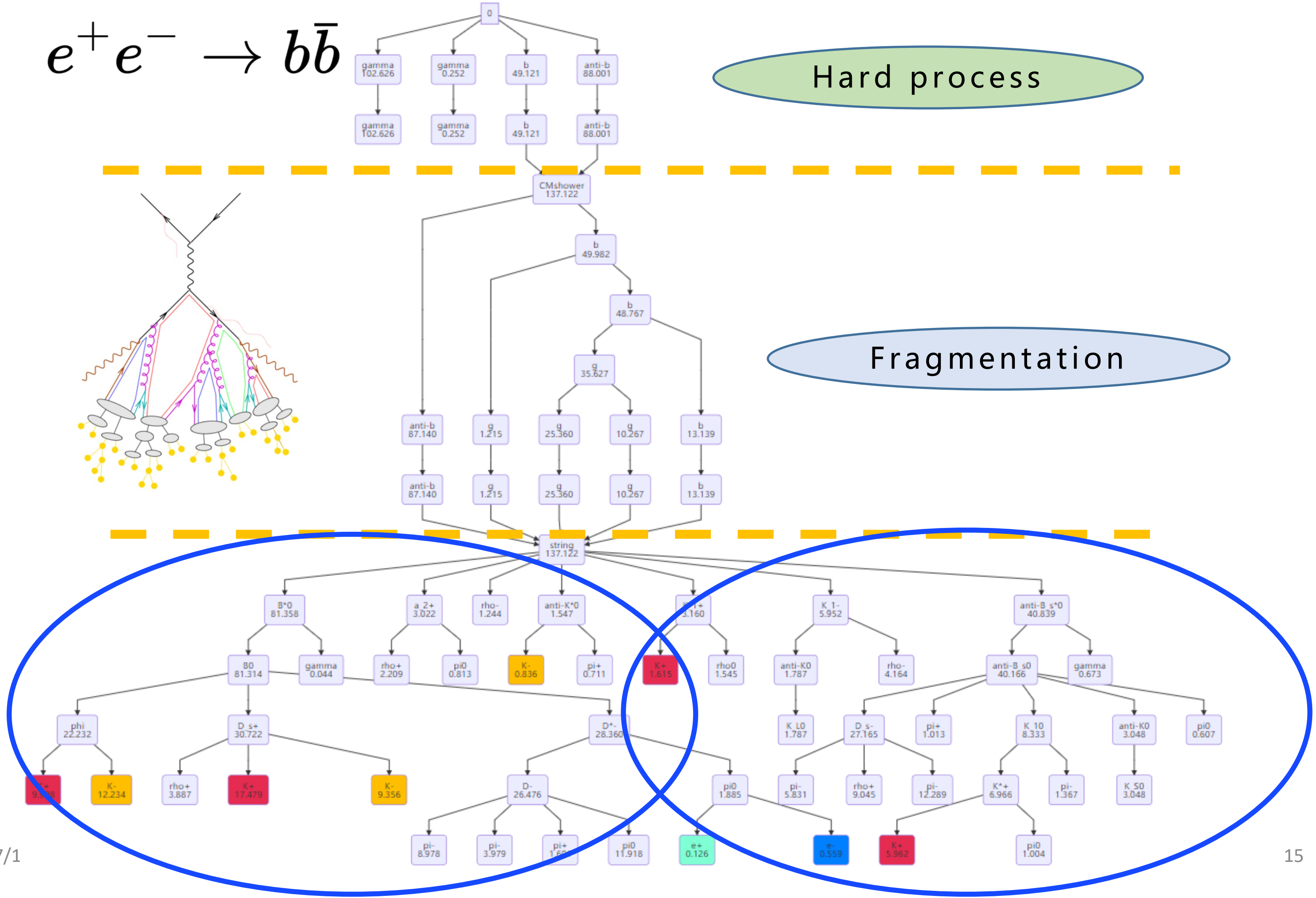
- 91 GeV
- Z → bb, cc, ll (uu,dd,ss)
- 450k events (900k jets) for each class

- Take particle level information a input

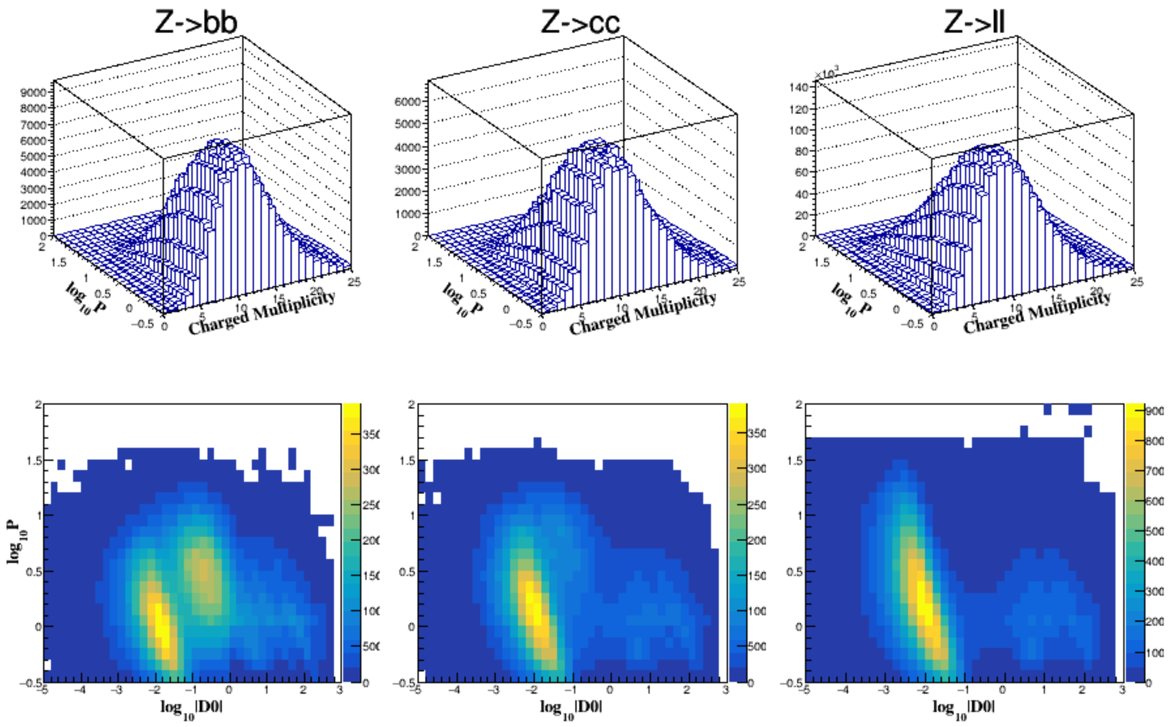


- ➢ 4-momenta
- ➢ d0/z0
- ➢ PID
- ➢ … …

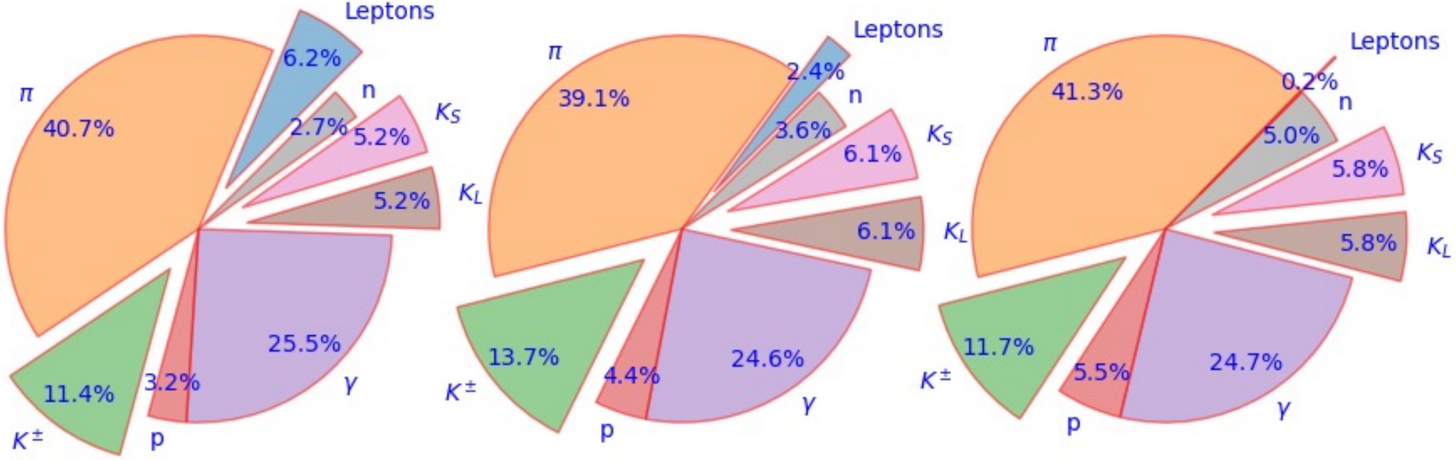
$$e^+e^- \to b\bar{b}$$


Hard process
Fragmentation
0
gamma 102.626
gamma 0.252
b 49.121
anti-b 88.001
CMshower 137.122
b 49.982
b 48.767
g 35.627
anti-b 87.140
g 1.215
g 25.360
g 10.267
b 13.139
string 137.122
B*0 81.358
a_2+ 3.022
rho- 1.244
anti-K*0 1.547
K_1- 5.952
anti-B_s*0 40.839
B0 81.314
gamma 0.044
rho+ 2.209
pi0 0.813
K- 0.836
pi+ 0.711
rho0 1.545
anti-K0 1.787
rho- 4.164
anti-B_s0 40.166
gamma 0.673
phi 22.232
D_s+ 30.722
D*- 28.360
K_L0 1.787
D_s- 27.165
pi+ 1.013
K_10 8.333
anti-K0 3.048
pi0 0.607
K- 12.234
rho+ 3.887
K+ 17.479
K- 9.356
D- 26.476
pi0 1.885
pi- 5.831
rho+ 9.045
pi- 12.289
K*+ 6.966
pi- 1.367
K_S0 3.048
pi- 8.978
pi- 3.979
pi0 11.918
e+ 0.126
e- 0.559
K+ 5.962
pi0 1.004

# 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 $\times$ 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

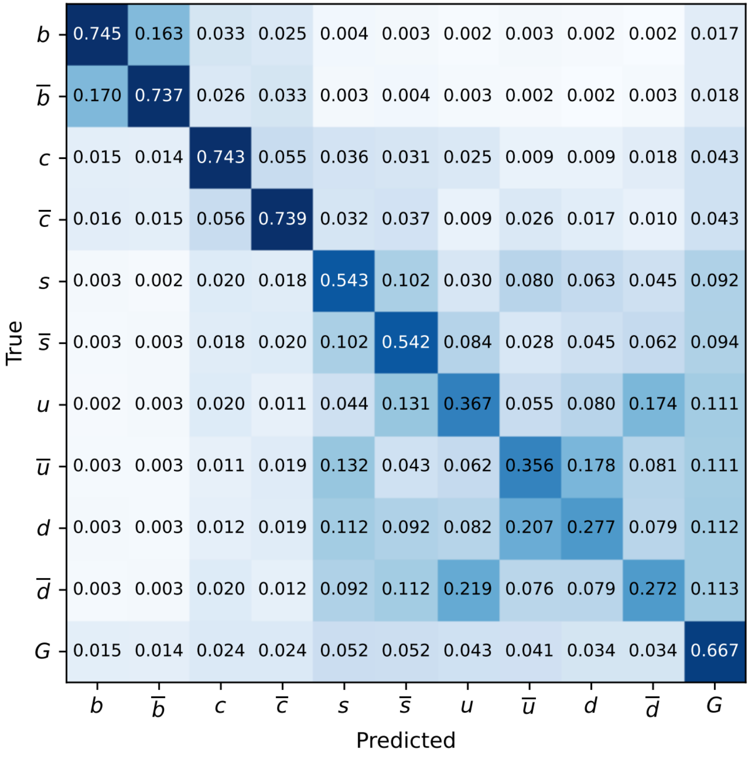
sqrt(0.584/0.345)=1.3

Statistical uncertainty: 30%

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

11 classes

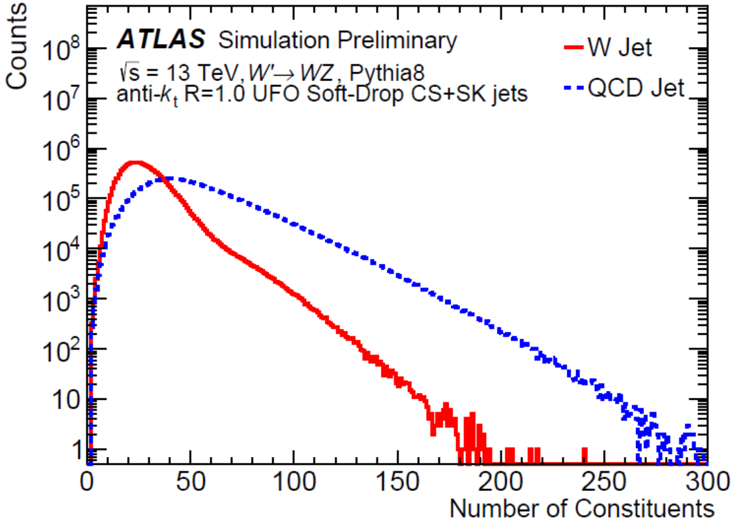
Ambitious test by M. Ruan




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

# $W$ Jet Taggers

- 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 $p_{\mathrm{T}}$, truncation eliminates the softest constituents of the jet.




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

# $W$ Jet Taggers (ATLAS, by Shudong Wang)

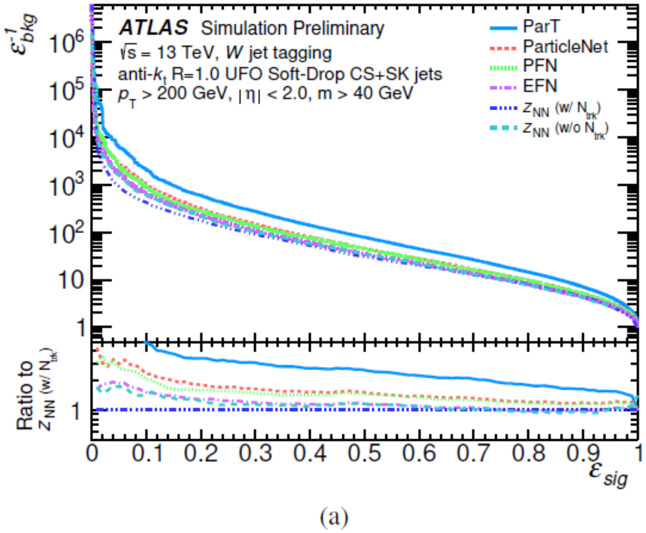
- Particle Flow Network(PFN)/Energy Flow Network(EFN)
  - Based on Deep Sets Theorem
  - *JHEP01(2019)121*
- ParticleNet
  - Customized graph neural network architecture for jet tagging with the point cloud approach
  - *Phys.Rev.D 101 (2020) 5, 056019*
- ParticleTransformer
  - Transformer designed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

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

# Tagger Performance

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.

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



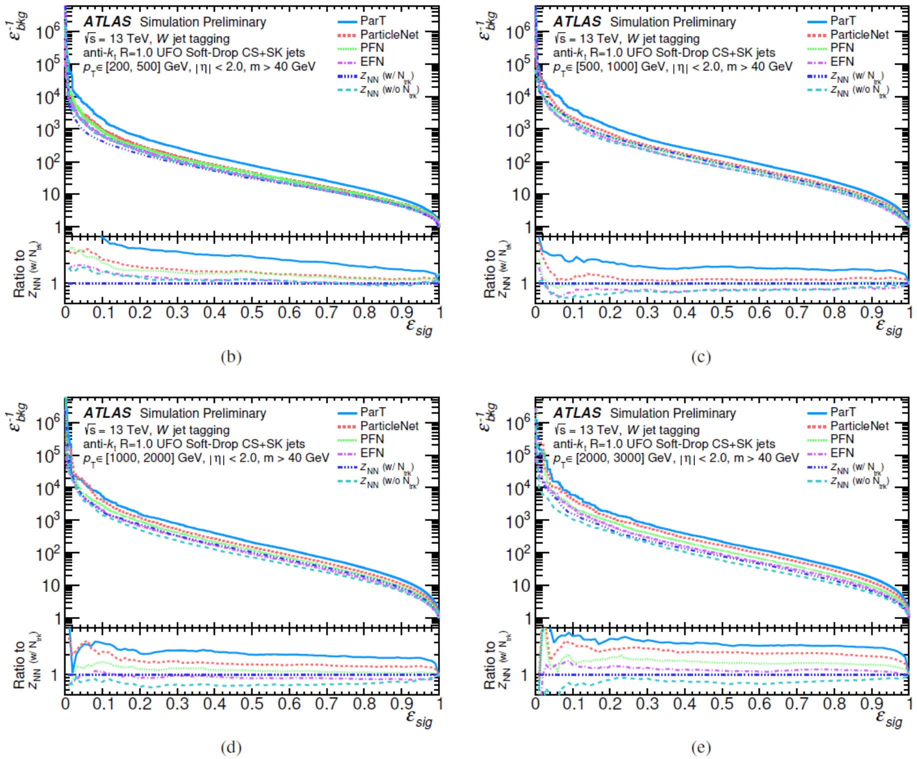




Figure 3: The QCD jets background rejection ($\varepsilon^{-1}_{bkg}$) versus the $W$-jets signal efficiency ($\varepsilon_{sig}$) for all the taggers studied. All of the constituent-based taggers studied surpass the performance of the high-level-feature-based tagger (noted as $z_{NN}$ in the figure) in the previous study [52].

# Tagger Performance

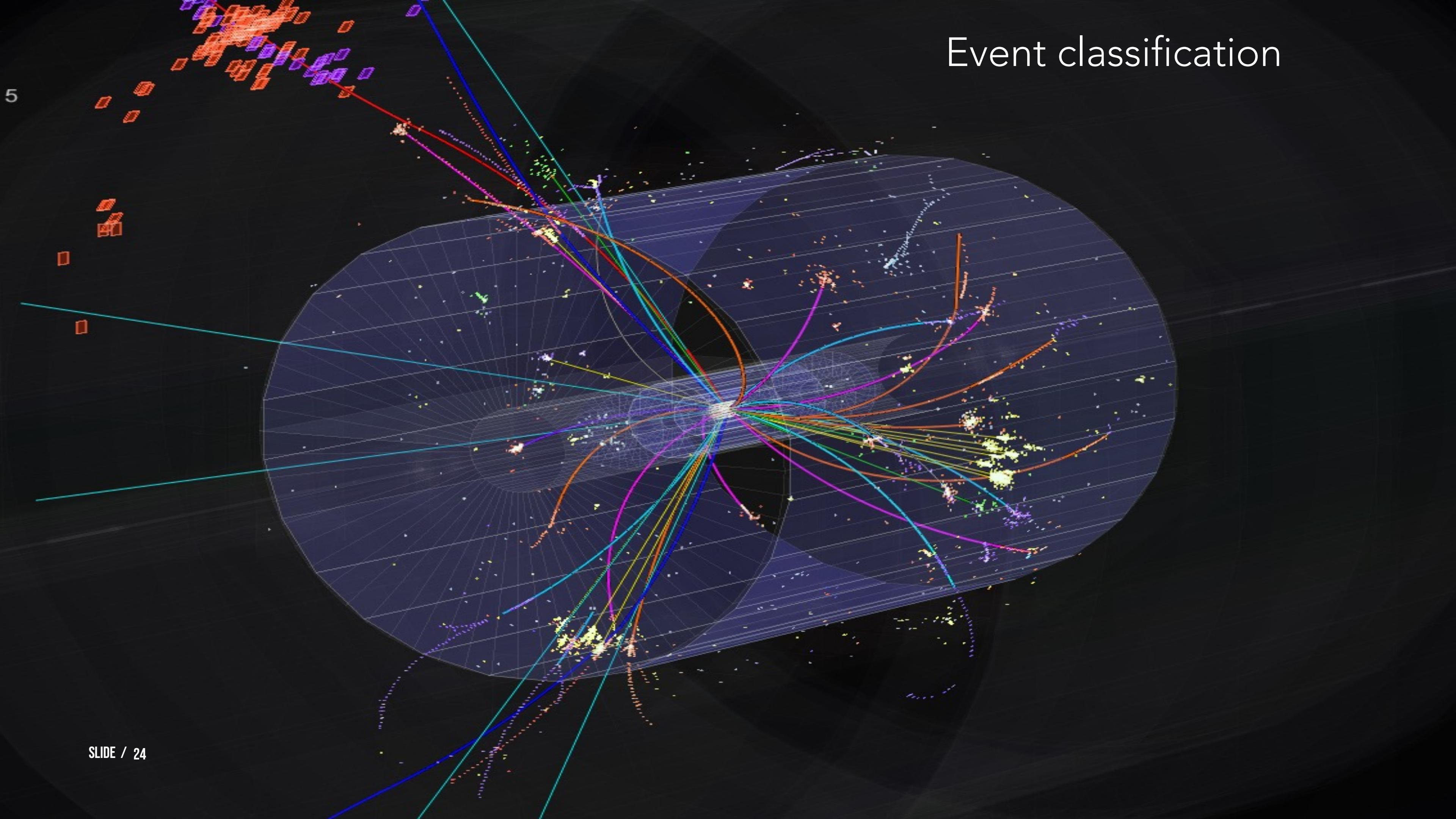
| Model | AUC | ACC | $\varepsilon_{bkg}^{-1}$ @ $\varepsilon_{sig} = 0.5$ | $\varepsilon_{bkg}^{-1}$ @ $\varepsilon_{sig} = 0.8$ | # Params | Inference Time |
|---|---|---|---|---|---|---|
| EFN | 0.920 | 0.835 | 35.1 | 7.95 | 56.73k | 0.065 ms |
| PFN | 0.931 | 0.853 | 44.7 | 9.50 | 57.13k | 0.11 ms |
| 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

# Event classification



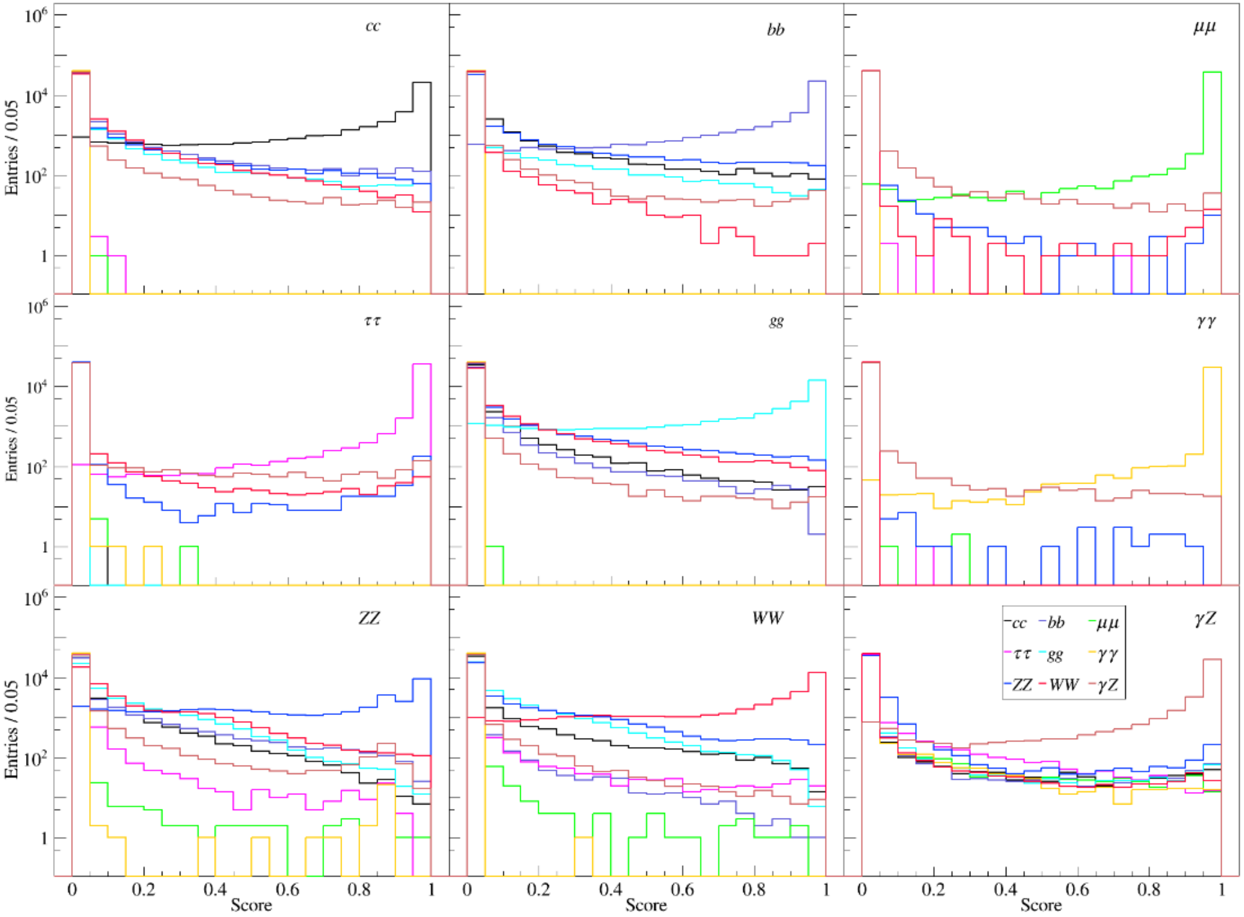
# Many processes are selected simultaneously

| 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

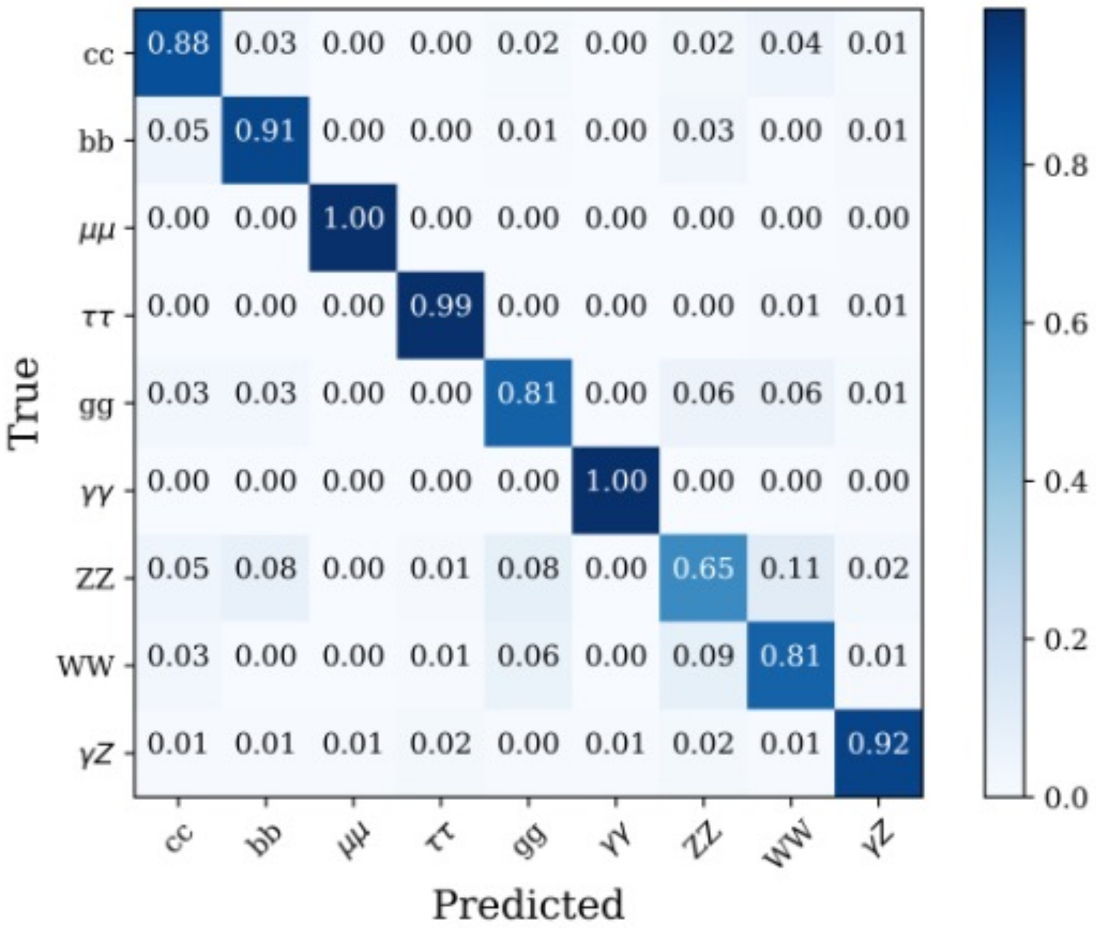


## Try eeH first

Sufficiently good performance

Average Accuracy ~ 87%

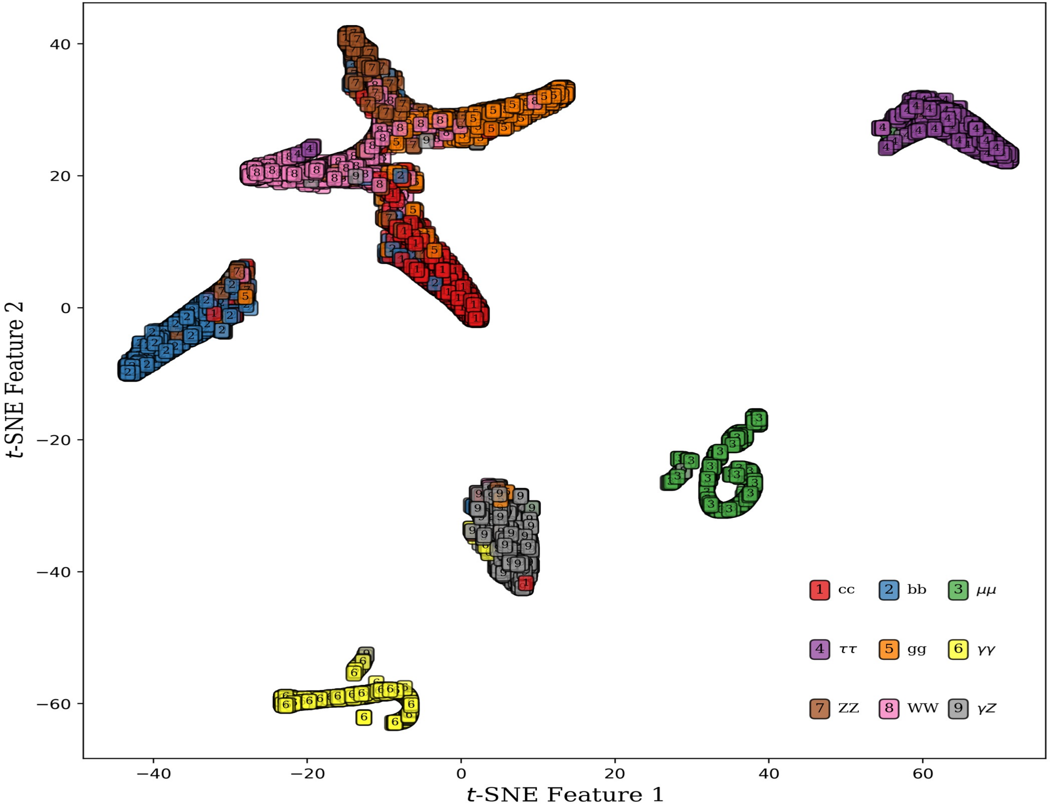
(11% for random guess)




Taking the one has largest probability (ArgMax)

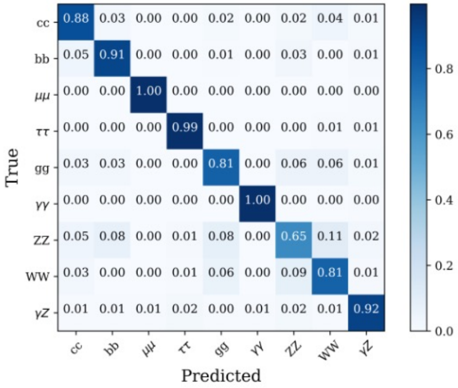
## Dimension reduction tells us more

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

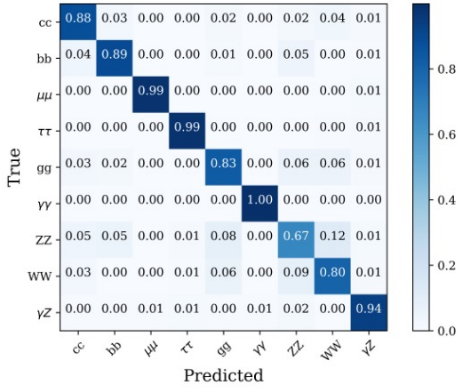



Dimensional reduction ( t-SNE )

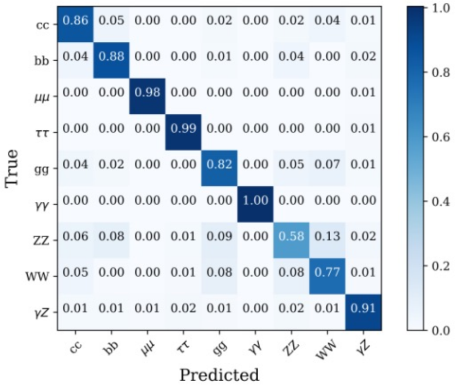
# All 4 production modes



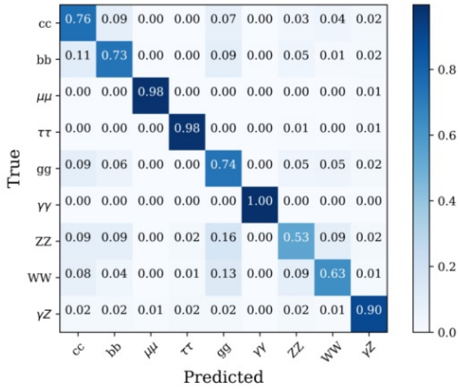

eeH



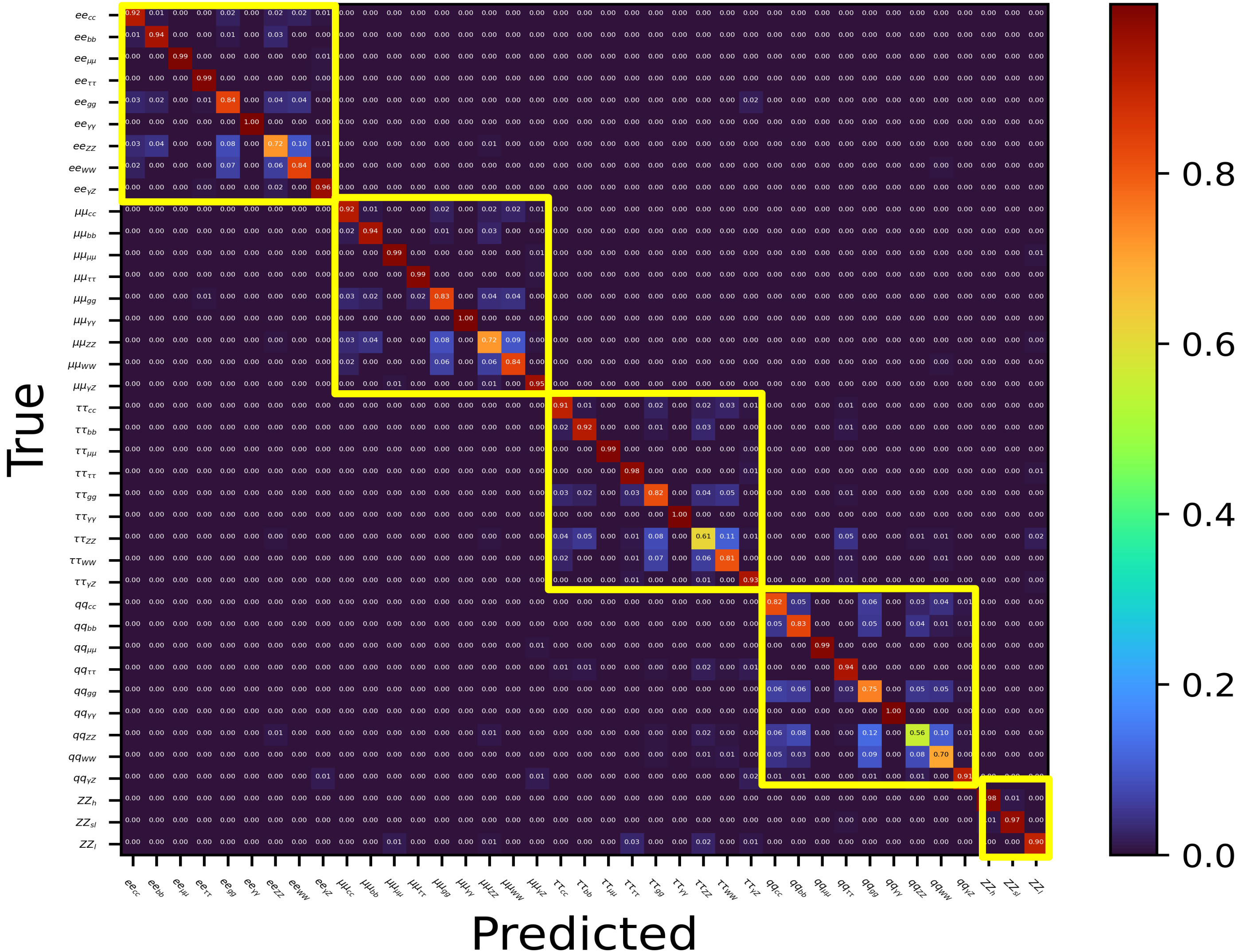

$\mu\mu$H



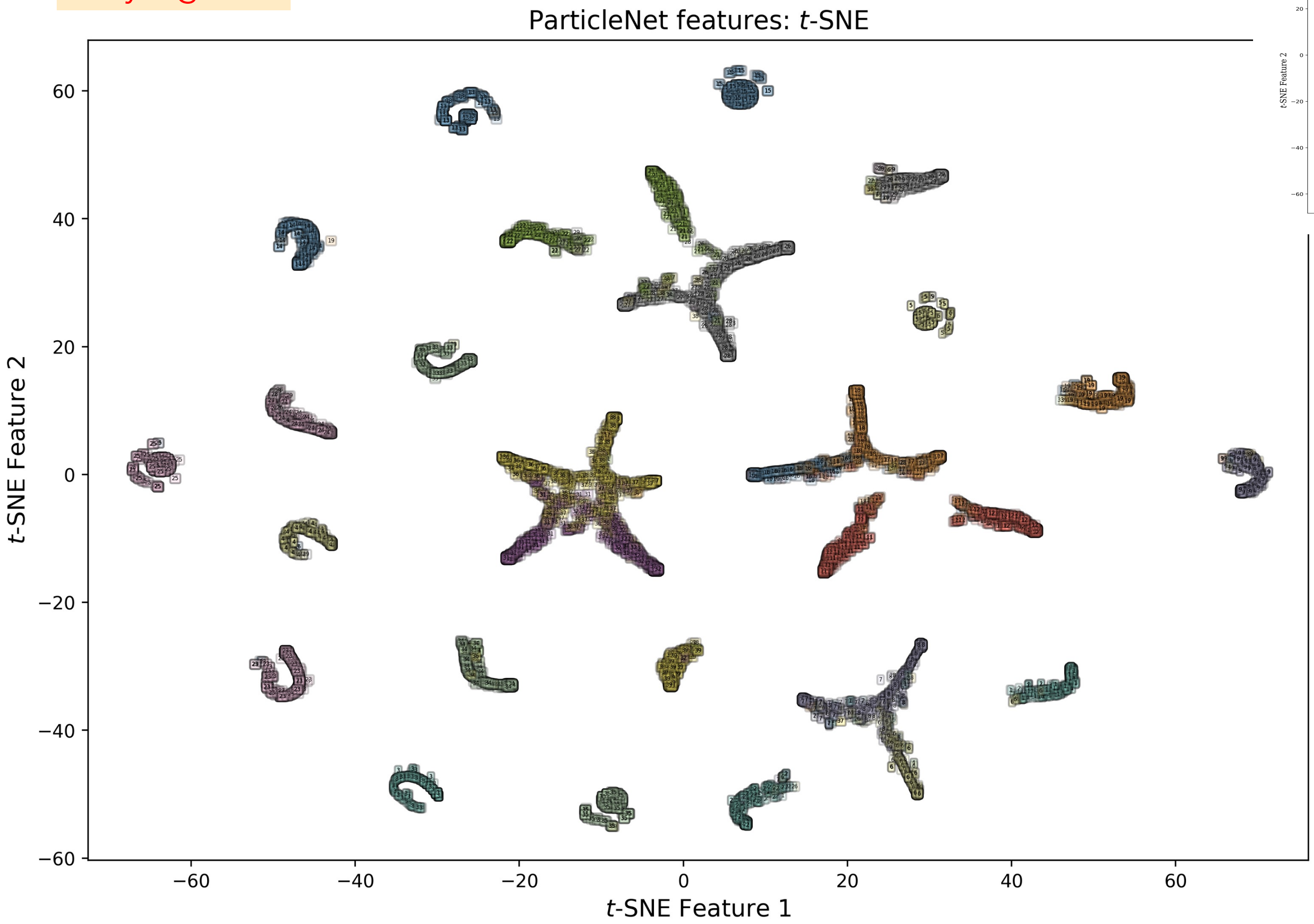

$\tau\tau$H

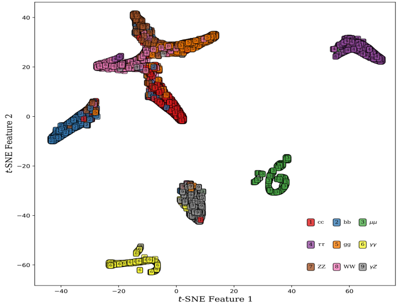



qqH


True
Predicted
0.8
0.6
0.4
0.2
0.0

Only signals



# More?

Large Language Models

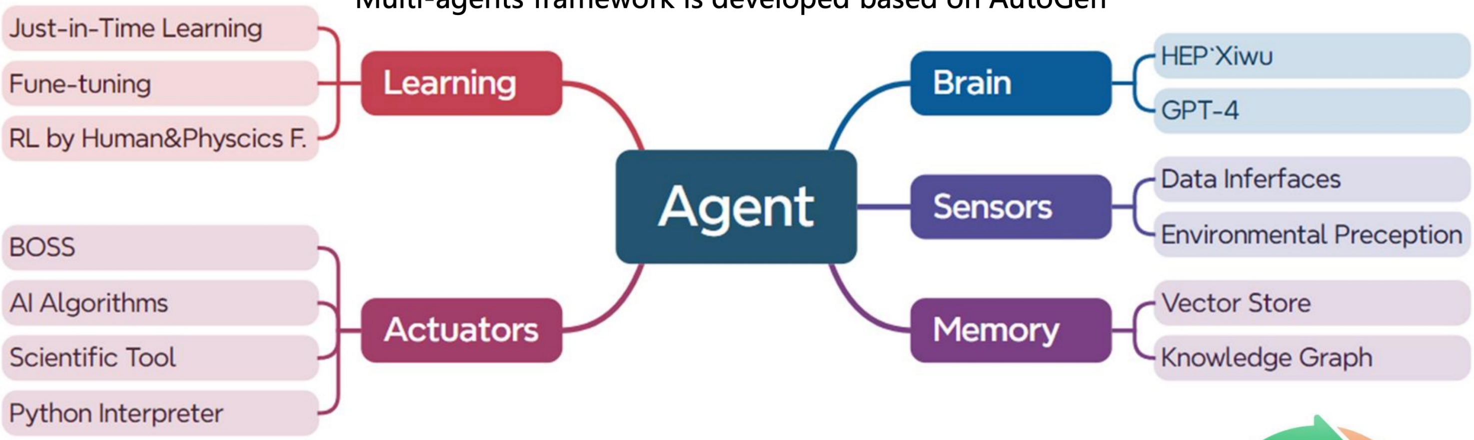
Used as copilot-like assistant: Dr. Sai

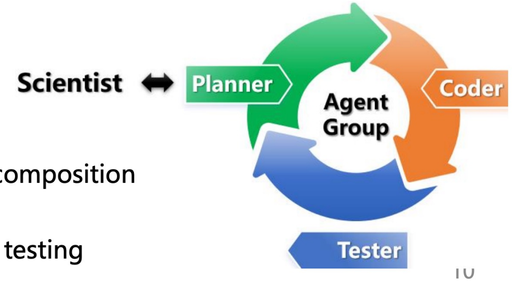
Used to HEP data directly: tokenization is a key

# Ke Li 's talk

## Dr. Sai

Multi-agents framework is developed based on AutoGen




Key of this project:

**make the results from AI more reliable**

- New architecture
- Good quality data
- In-the-fly validation and test

Main Agents:

- Planner: Planning and tasks decomposition
- Coder: Write BOSS code
- Tester: Using scientific tools for testing

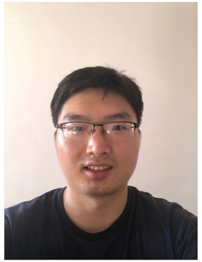
STCF workshop 2024 @ LZU 10

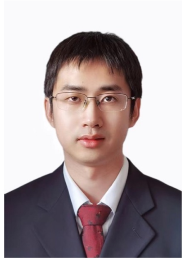
A-I-HEP Forum

Institute of High Energy Physics

A-I-HEP FORUM

“赛博士”科研智能体





| | |
|---|---|
| Speake: | 李科，张易于 |
| Host: | 李刚 |
| Time: | 10:00, July 15 2024 |
| Location: | C305 Main building |
| Indico: | https://indico.ihep.ac.cn/event/22949/ |
| Zoom ID: | 699 9017 4174 |
| Password: | 548072 |

如果你错过了沈阳的和昨天的报告 … …

可以听下周一的 lectures

# 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?

# Summary

- Machine learning is statistical learning (NFL)
- Machine learning is useful  (~~CoD~~): high dimensional HEP data
- Machine learning method with proper bias is powerful 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 – Ke Li's talk
  - ➢ Employing LLMs to directly to process  data: how to represent HEP events is the key

# 广告： 机器学习和量子计算的 workshop，日程含 2 天tutorial

## Quantum Computing and Machine Learning Workshop 2024

2024年8月3日至8日
Asia/Shanghai 时区

输入您的搜索词

- 概览
- 科学议程
- Committees
- 征集摘要
- 日程表
- 报告列表
- 注册
- 参会人名单
- Accommodation and Meeting Venue
- Travel Information
- Zoom Connection
- Previous Workshop

To promote the application of quantum computing and machine learning in high-energy theoretical and experimental physics, we will hold a workshop on quantum computing and machine learning at Jilin University, Changchun, China. Researchers from domestic and international fields related to quantum computing and machine learning are sincerely invited to exchange ideas and discuss on the application of quantum computing algorithms, machine learning, hardware advances, and the use of development platforms.

Look forward to meeting you in Changchun!

**Registration fee:**

- 2000CNY for regular attendee; 1000CNY for student
- No registration fee for online attendee
- Registration site: Herun Hometown Hotel (长春和润记忆酒店)
- Registration time: 9:00-17:00, August 5

**开始** 2024年8月3日 上午9:00
**结束** 2024年8月8日 下午5:00
Asia/Shanghai

会议通知.pdf


Contact
weiminsong@jlu.edu.cn
zhoumiao@jlu.edu.cn
xuhj@ihep.ac.cn


**征集摘要已开放**
您可以提交摘要供审核。
提交新摘要

**注册**
您已注册此会议。
看详情 ›

# 欢迎注册！