ML at CEPC

李刚

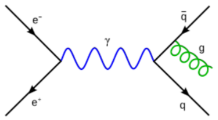
2023-02-28

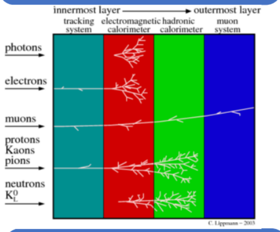
中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

CEPC

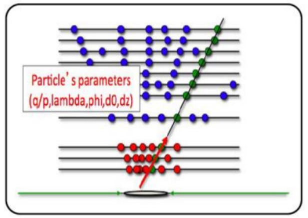
ML@HEP

nature → detector





event generator → simulation of detector response



reconstruction → event selection → Statistical inference

Classify signal/background

hits→ final state particle :
Cluster hits
Classify clusters as particles
Infer/regress properties

Hypothesis testing
Parameter estimation

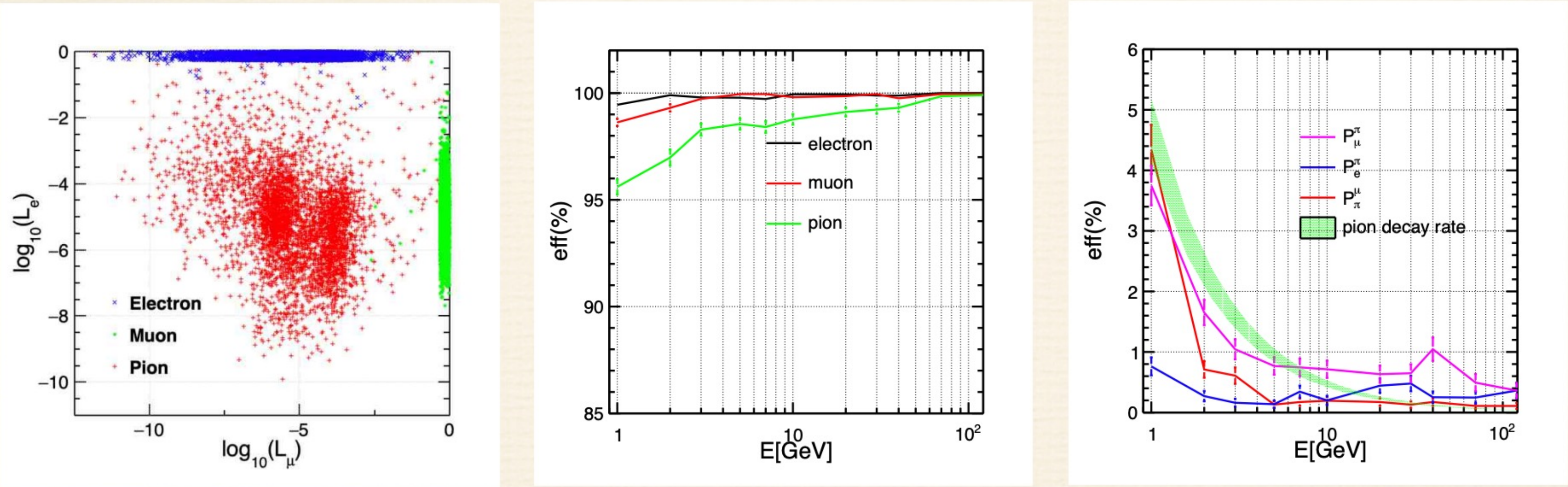
ML@CEPC

- Classification
  - PID
  - Jet flavor tagging
  - Event classification
- Pattern recognition
  - Using RNN to reconstruct peaks of primary ionization
- Background suppression + data compression
- Simulation

粒子分类

- TMVA + hand engineering features

- LICH uses TMVA methods to summarize 24 input variables into two likelihoods, corresponding to electrons and muons.
- The efficiency for electron and muon is higher than 99.5% (E>2 GeV). Pion efficiency ~ 98%.



Migration Matrix at 40GeV (LICH)

Type	e^- *like*	μ^- *like*	π^+ *like*
e^-	99.71 ± 0.08	< 0.07	0.21 ± 0.07
μ^-	< 0.07	99.87 ± 0.08	0.05 ± 0.05
π^+	0.14 ± 0.05	0.35 ± 0.08	99.26 ± 0.12

Migration Matrix for ALEPH PID ($> 2GeV$)(*Eur.Phys.J.C20:401-430,2001*)

Type	e^- *like*	μ^- *like*	π^+ *like*	undefined
e^-	99.57 ± 0.07	< 0.01	0.32 ± 0.0	0.09 ± 0.04
μ^-	< 0.01	99.11 ± 0.08	0.88 ± 0.08	0.01 ± 0.01
π^+	0.71 ± 0.04	0.72 ± 0.04	98.45 ± 0.06	0.12 ± 0.03

Dan Yu

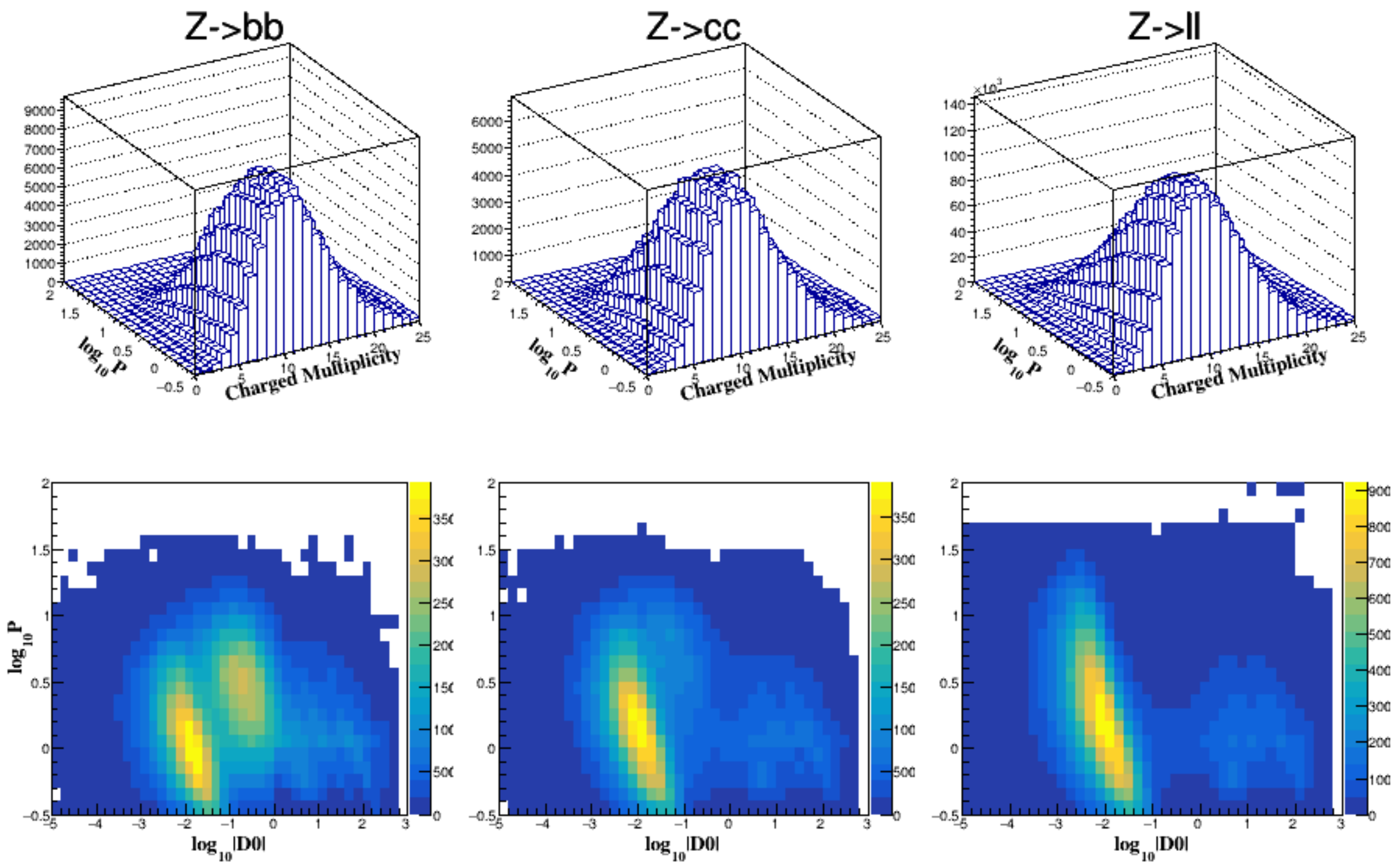
Jet 分类

ArXiv:2208.13503, submitted to EPJC

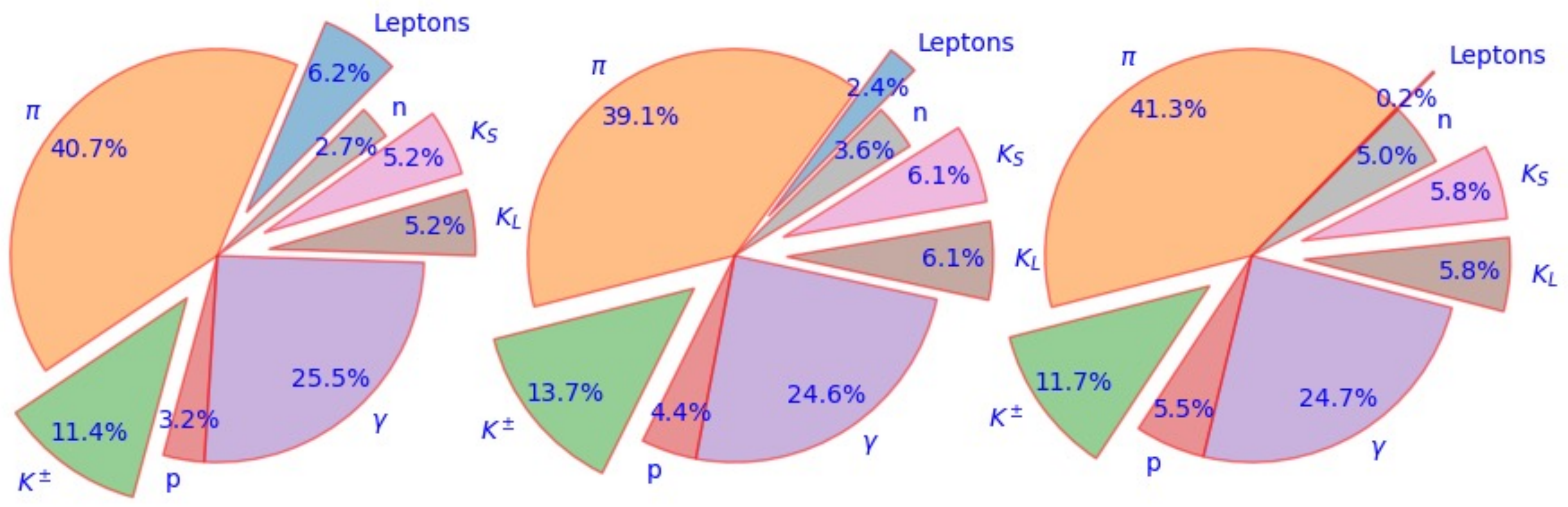
数据集

- 91 GeV
- Z → bb, cc, oo (uu,dd,ss)
- WHIZARD 产生/全模拟/重建
- Jet Clustering
- 每种样本 450k 事例（900k jets）

数据集中的 features



数据集中的 features



粒子种类特征

能看到的非常有限，更多的还需要算法去挖掘

不同算法结果比较（一）

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

>0.90 @ fast sim

不同算法结果比较（二）

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

简单估算c-tag：

sqrt(0.584/0.345)=1.3

统计误差减小 30%

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

PERSONAL RANK THE DIFFICULTNESS OF HIGGS ANALYSIS AT EE COLLIDERS

4 x 9 modes in this study, [5 production and 13 (9) decays modes in SM]

Prod/decay	cc	bb	μμ	ττ	γγ	gg	WW	ZZ	γZ	ee, uu,dd,ss
eeH (incl. Z fusion)	3	1	5	2	4	1	2	3	5	Not covered yet
μμH	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	
ννH (incl. W fusion)	5	1	3	2	3	5	4	2	4	

According to production rate, signal signature, backgrounds, complication of analysis, ...

Current estimation of Higgs precision

CEPC: 2205.08553

	240 GeV, 20 ab^{-1}		360 GeV, 1 ab^{-1}		
	ZH	vvH	ZH	vvH	eeH
any	0.26%		1.40%	\	\
H→bb	0.14%	1.59%	0.90%	1.10%	4.30%
H→cc	2.02%		8.80%	16%	20%
H→gg	0.81%		3.40%	4.50%	12%
H→WW	0.53%		2.80%	4.40%	6.50%
H→ZZ	4.17%		20%	21%	
$H \to \tau\tau$	0.42%		2.10%	4.20%	7.50%
$H \to \gamma\gamma$	3.02%		11%	16%	
$H \to \mu\mu$	6.36%		41%	57%	
$Br_{upper}(H \to inv.)$	0.07%		\	\	
$H \to Z\gamma$	8.50%		35%	\	
Width	1.65%		1.10%		

FCC-ee

$\sqrt{s}$ (GeV)	240		365	
Luminosity (ab^{-1})	5		1.5	
$\delta(\sigma BR)/\sigma BR$ (%)	HZ	$\nu\bar{\nu}$ H	HZ	$\nu\bar{\nu}$ H
$H \to$ any	±0.5		±0.9	
$H \to b\bar{b}$	±0.3	±3.1	±0.5	±0.9
$H \to c\bar{c}$	±2.2		±6.5	±10
$H \to gg$	±1.9		±3.5	±4.5
$H \to W^+W^-$	±1.2		±2.6	±3.0
$H \to ZZ$	±4.4		±12	±10
$H \to \tau\tau$	±0.9		±1.8	±8
$H \to \gamma\gamma$	±9.0		±18	±22
$H \to \mu^+\mu^-$	±19		±40	
$H \to$ invisible	< 0.3		< 0.6	

- Results of CEPC and FCC-ee based individual analysis
- Comparable precision

A lots of efforts

DATA SETS: $e^+e^- \rightarrow ZH, Z \rightarrow l^+l^-, qq$

- 400 k events for each Higgs decays :

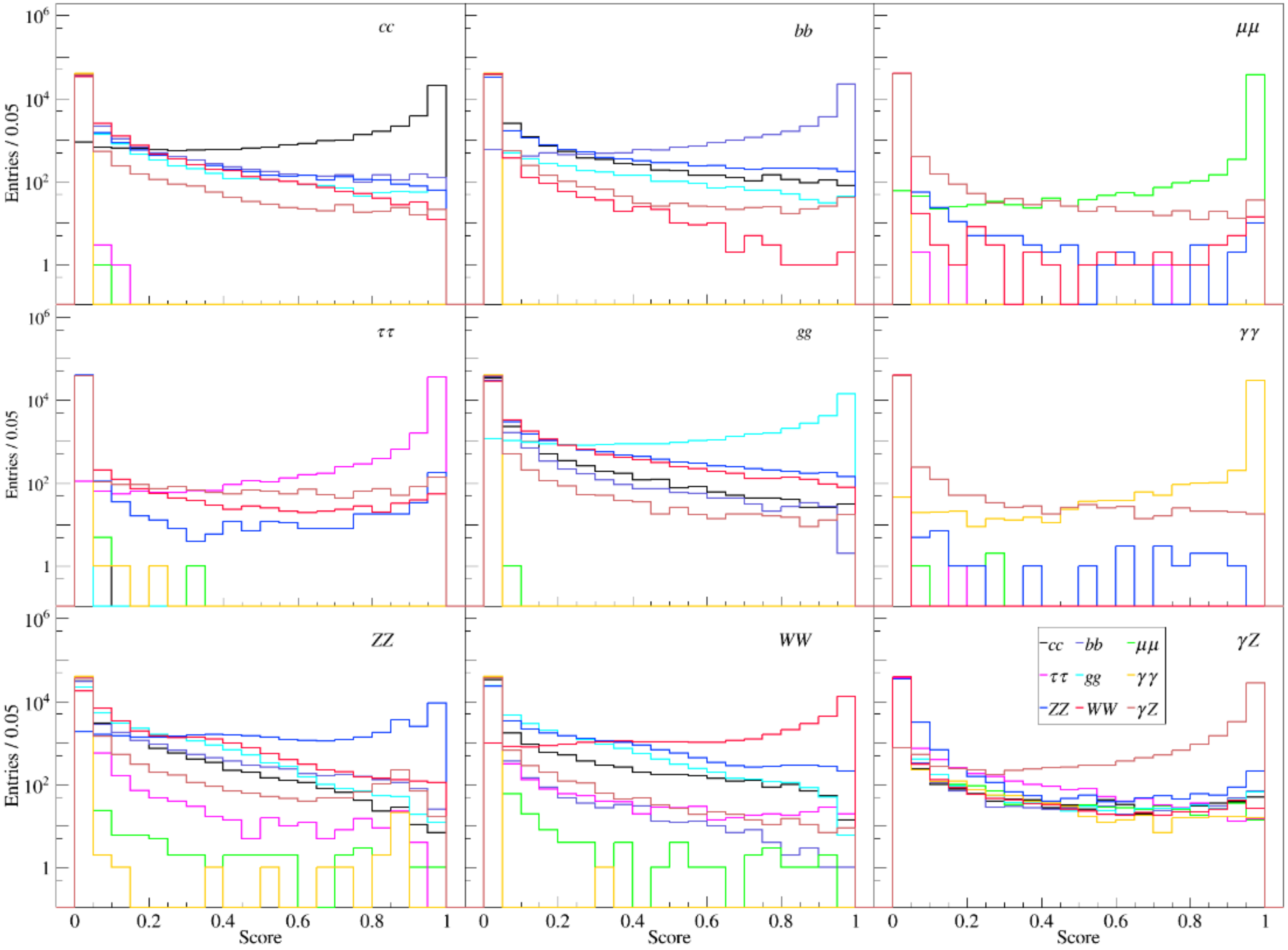
  $cc, bb, \mu\mu, \tau\tau, gg, \gamma\gamma, ZZ, WW, \gamma Z$

- Train: validation: test = 8:1:1
- Simple smearing fast simulation

Chinese Phys. C **46** 113001

Try eeH first

Probability distributions of each class

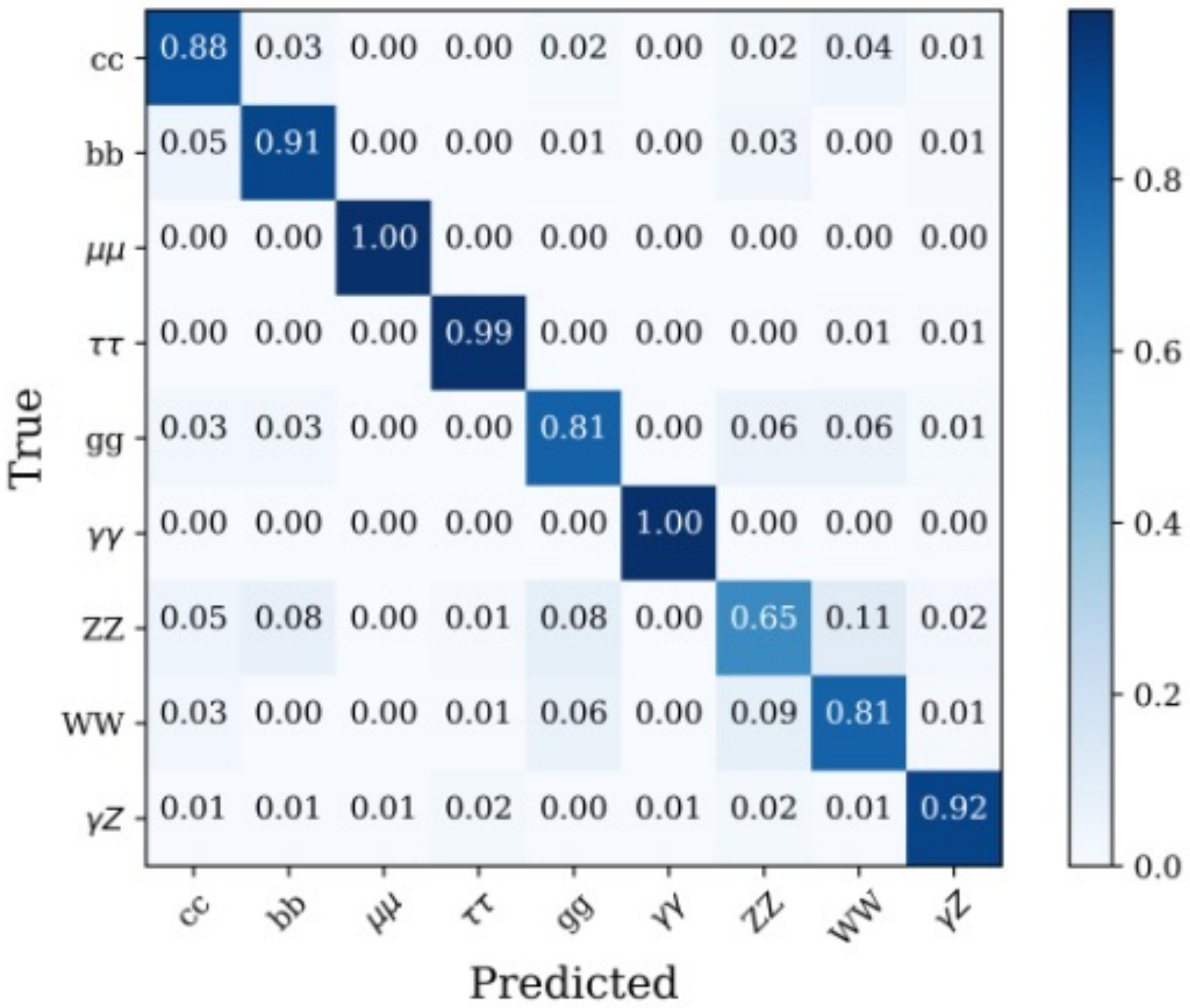


Try eeH first

Sufficiently good performance

Average Accuracy ~ 87%

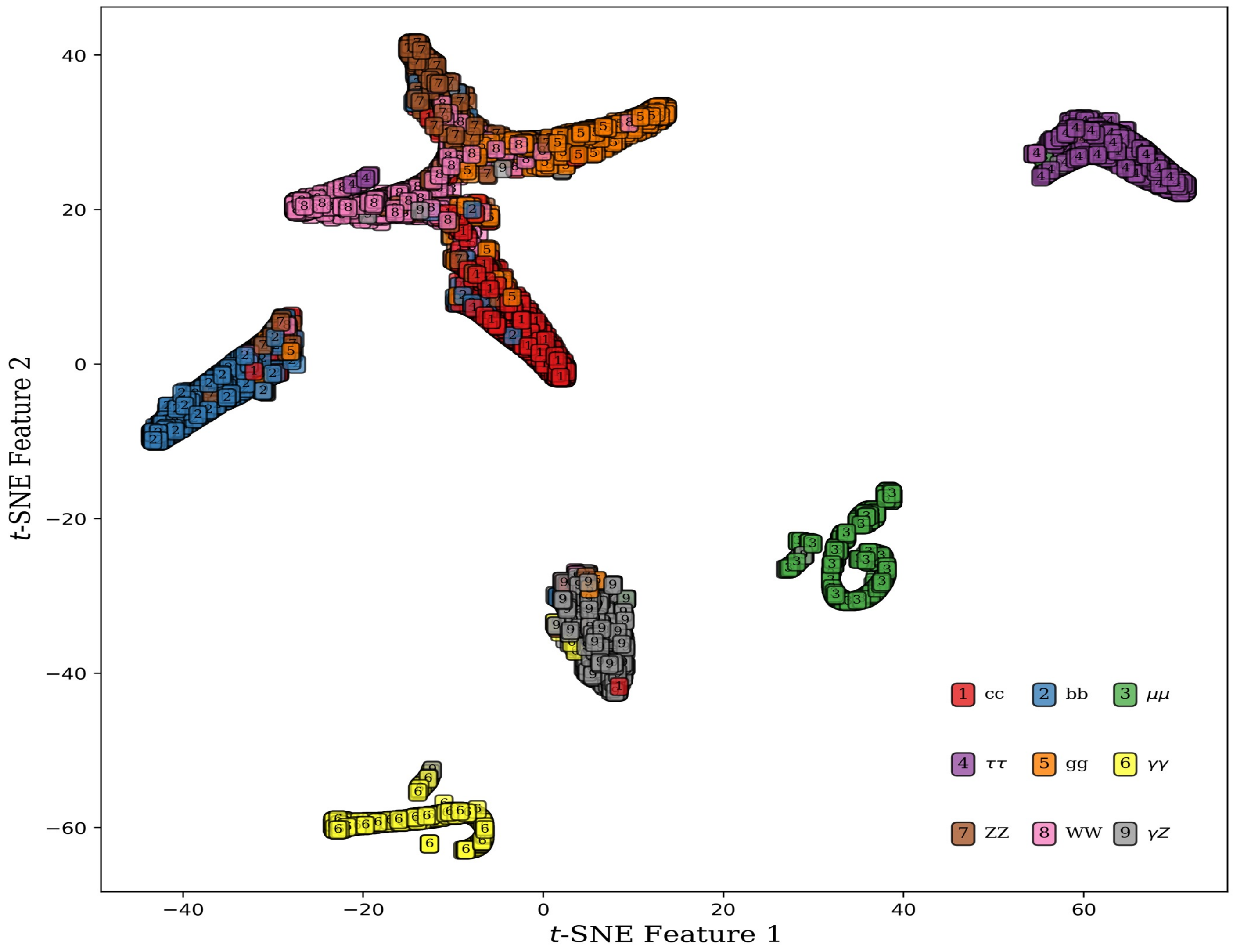
(11% for random guess)



Taking the one of the largest probability (ArgMax)

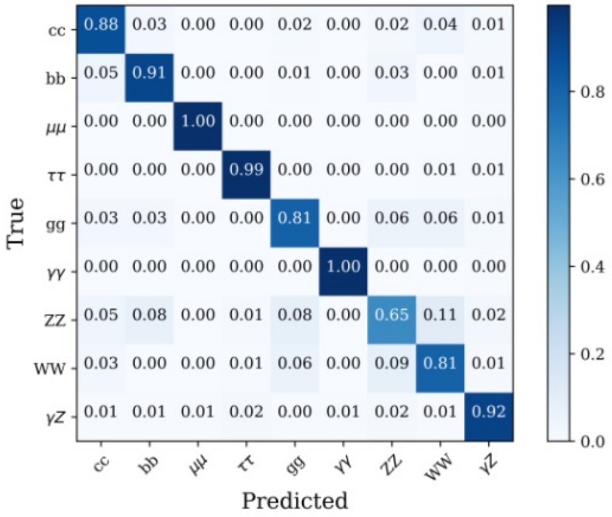
Dimension reduction tells us more

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

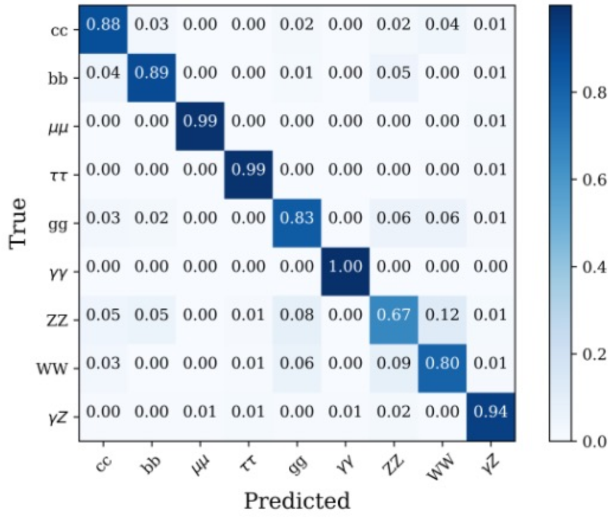


Dimensional reduction (t-SNE)

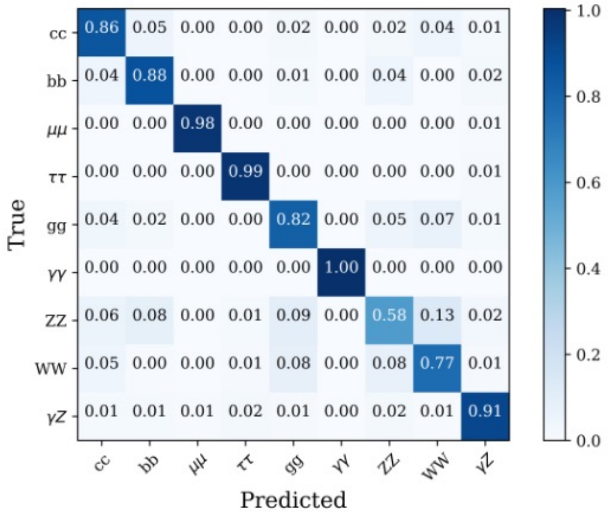
All 4 production modes



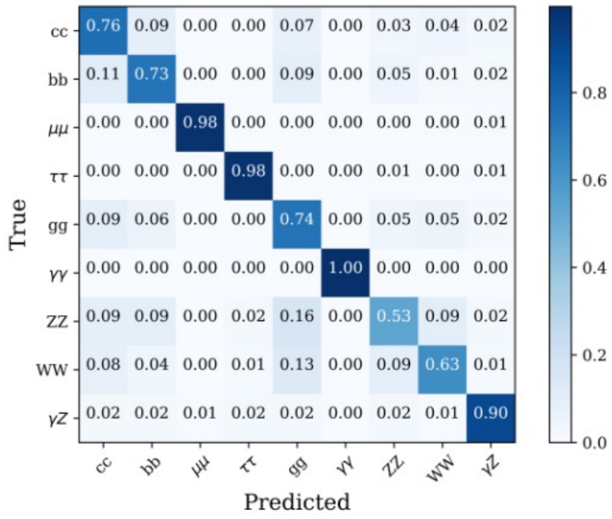
eeH



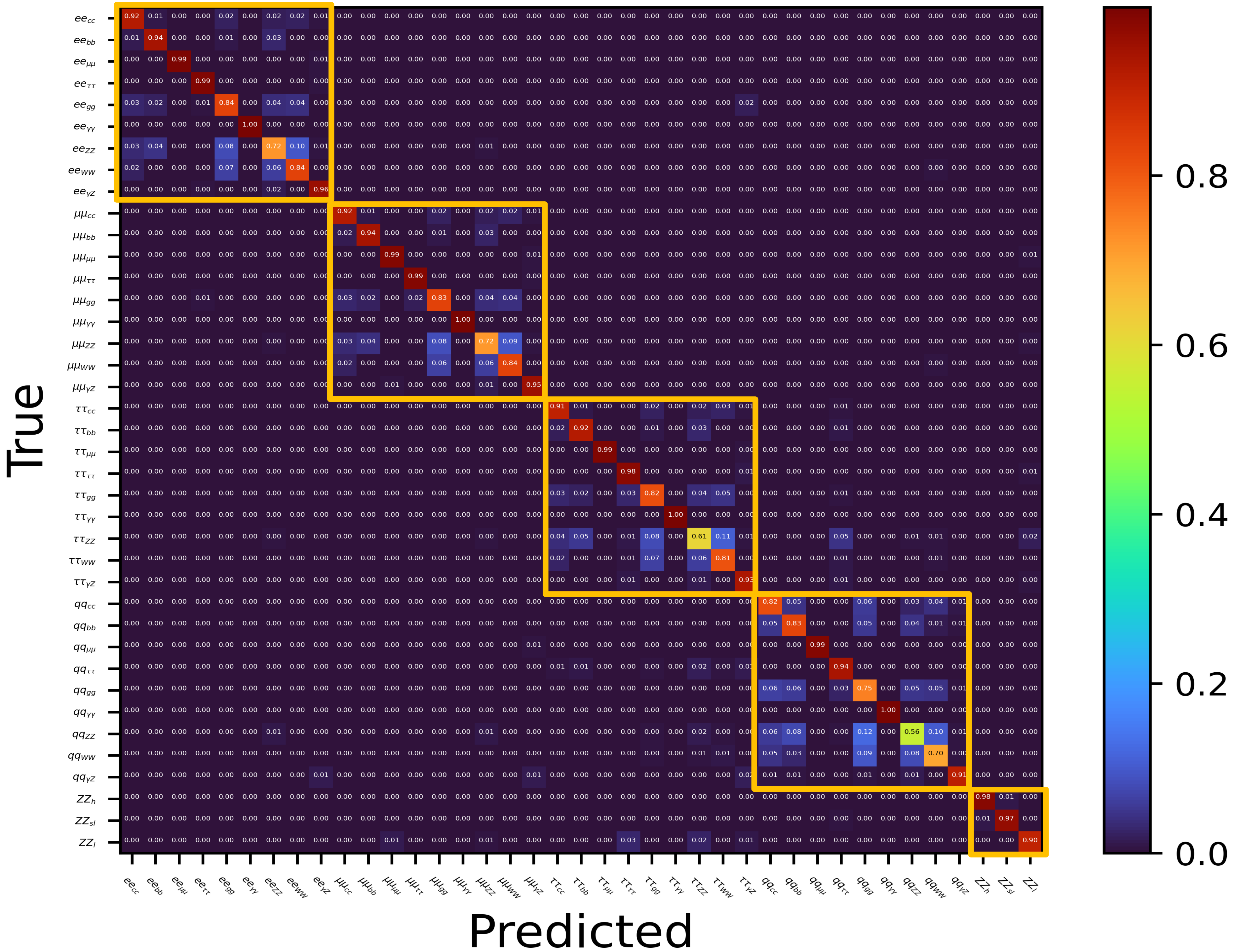
$\mu\mu$H

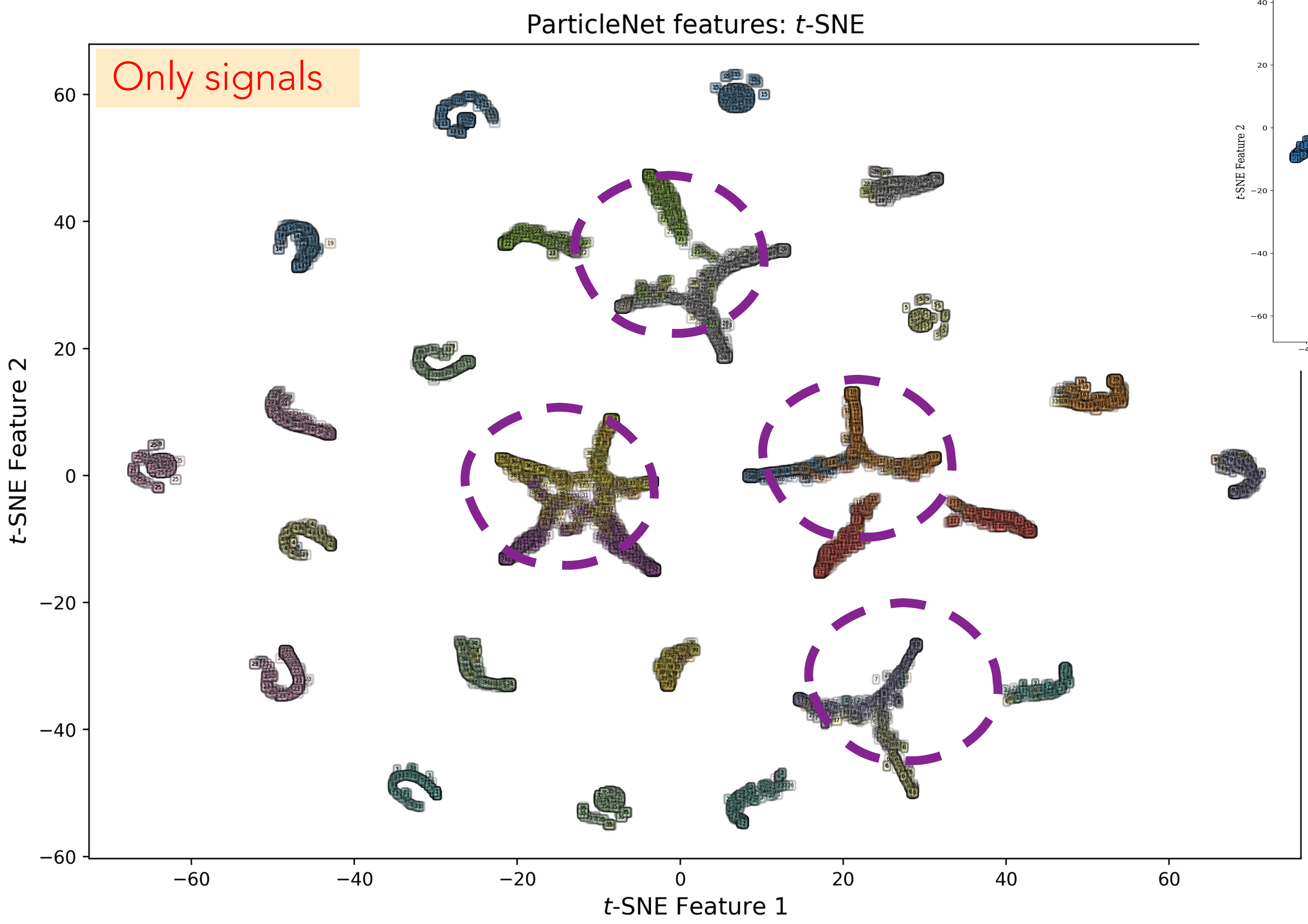


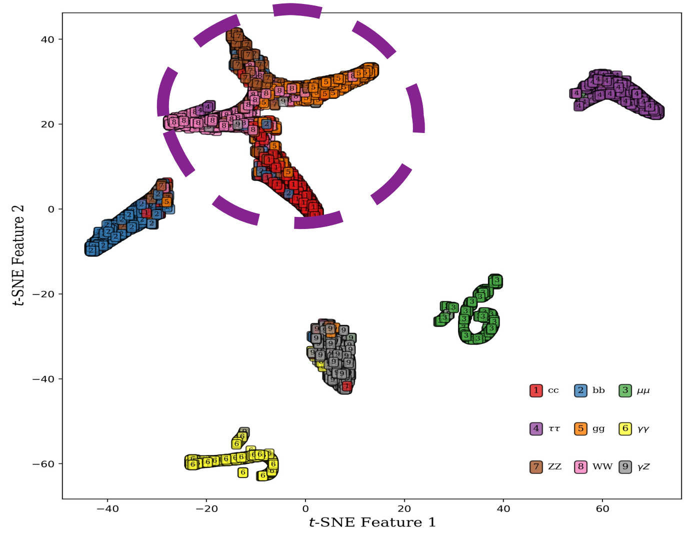
$\tau\tau$H



qqH


True
Predicted
0.8
0.6
0.4
0.2
0.0



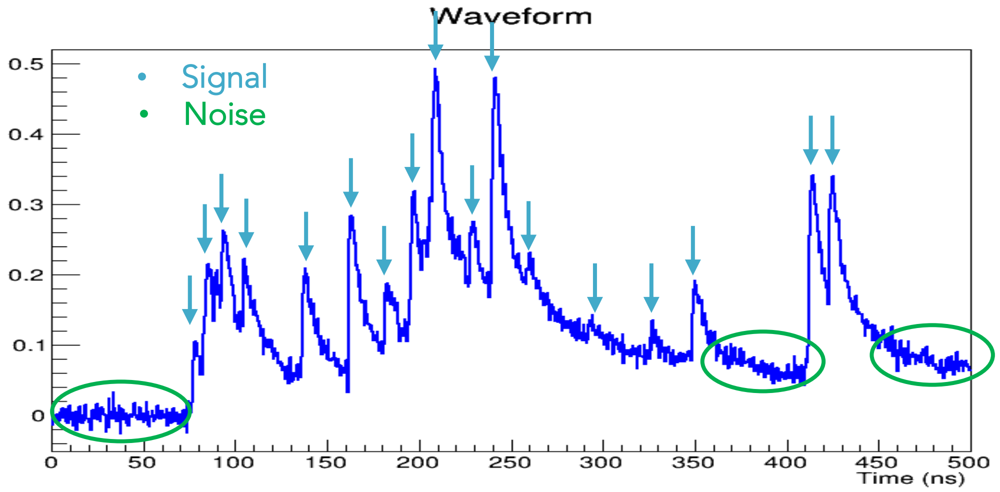


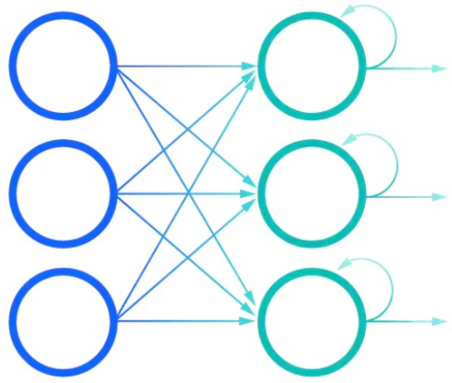
Will add more backgrounds, more statistics, …

一个时序重建问题的例子

By 赵光， et al

- Peak detection of waveforms from the DC
- Supervised-classification: “signal” and “noise”





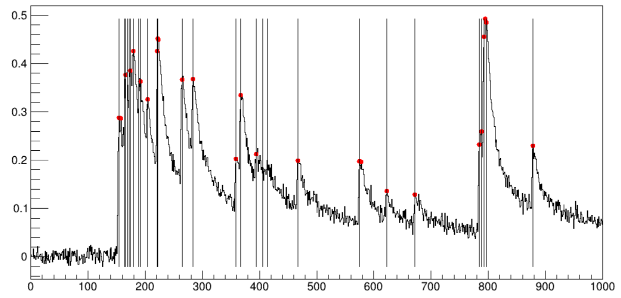
<u>Recurrent Neural Network (RNN)</u>:

- “Memory” structure: internal loops over sequence elements
- Powerful to handle time-sequences

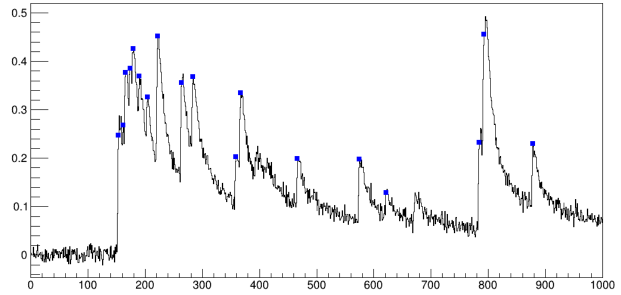
DL RESULTS AND COMPARE TO TRADITIONAL ALGORITHM

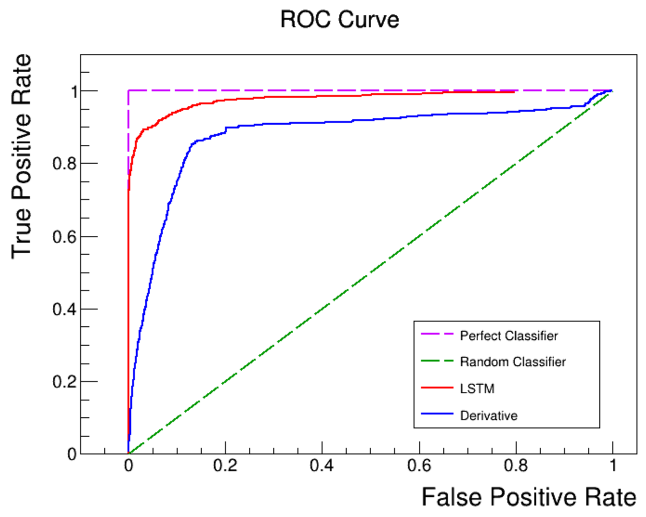
Black lines: MC truth times

RNN (LSTM) 24/28 detected



Derivative 18/28 detected





RNN (LSTM) is much more powerful than the derivative for the peak finding problem

Thanks to Zhao Guang

Intelligent Readout of Pixel Sensors

By 卢云鹏

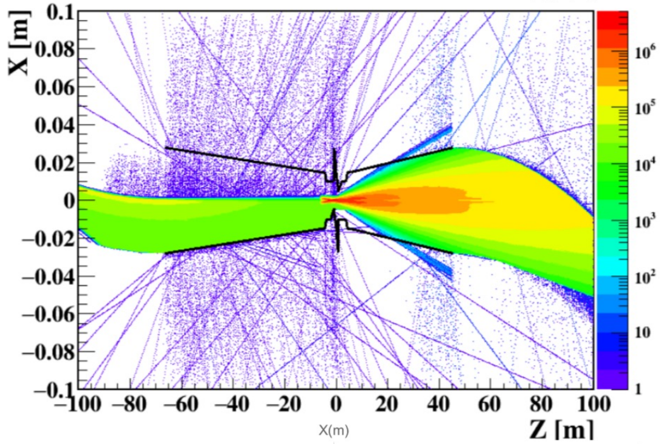
Challenges in the Vertex detector

- Data rate > Gbps / pixel chip, while power consumption limited < 50 mW/cm^2
  - 10 MHz particle hits / cm^2 at Z pole → 10 MHz * 3 pixels / cluster * 4 cm^2 / chip * 32 bit = 3.84 Gbps
  - High speed data link are always the hot spot of pixel chip
- The Neural Network was explored for possible solutions:
  - Data compression algorithm
  - Background suppression method

Background suppression

- Hit rate dominated by the radiative background for the CEPC vertex detector
- A pattern recognition module can be integrated into the pixel chip
  - Local hit pattern can be classified by a neural network
  - Algorithm developed with simulation data
  - Parameter reconfigurable based on the chip position and experimental data
- Data can be processed at hit level, a simple network is essential for low power operation

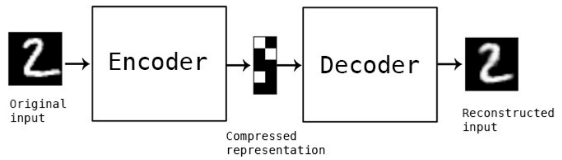
K. Li



Synchrotron radiation background

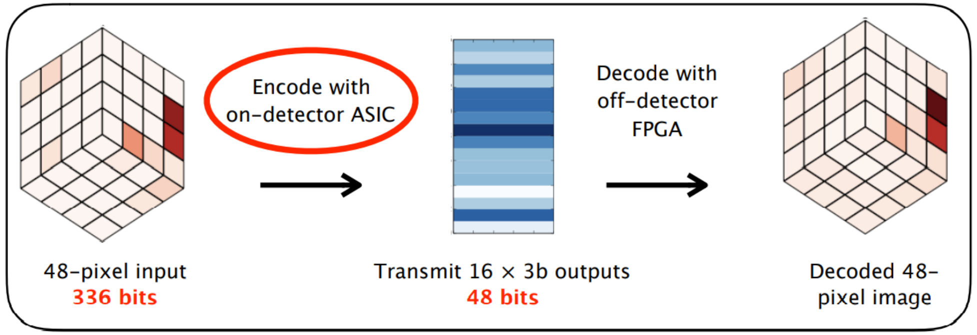
Autoencoder Neural Networks

- Compression algorithm, data-specific, lossy and **learned automatically**
  - https://blog.keras.io/building-autoencoders-in-keras.html
  - Being investigated by the High-Granularity Calorimeter Group
- Also considered for the data compression of CEPC vertex detector



  - Encoder on chip, and decoder in the back-end electronics or data processing software
  - Need to deal with much more channels and different data patterns

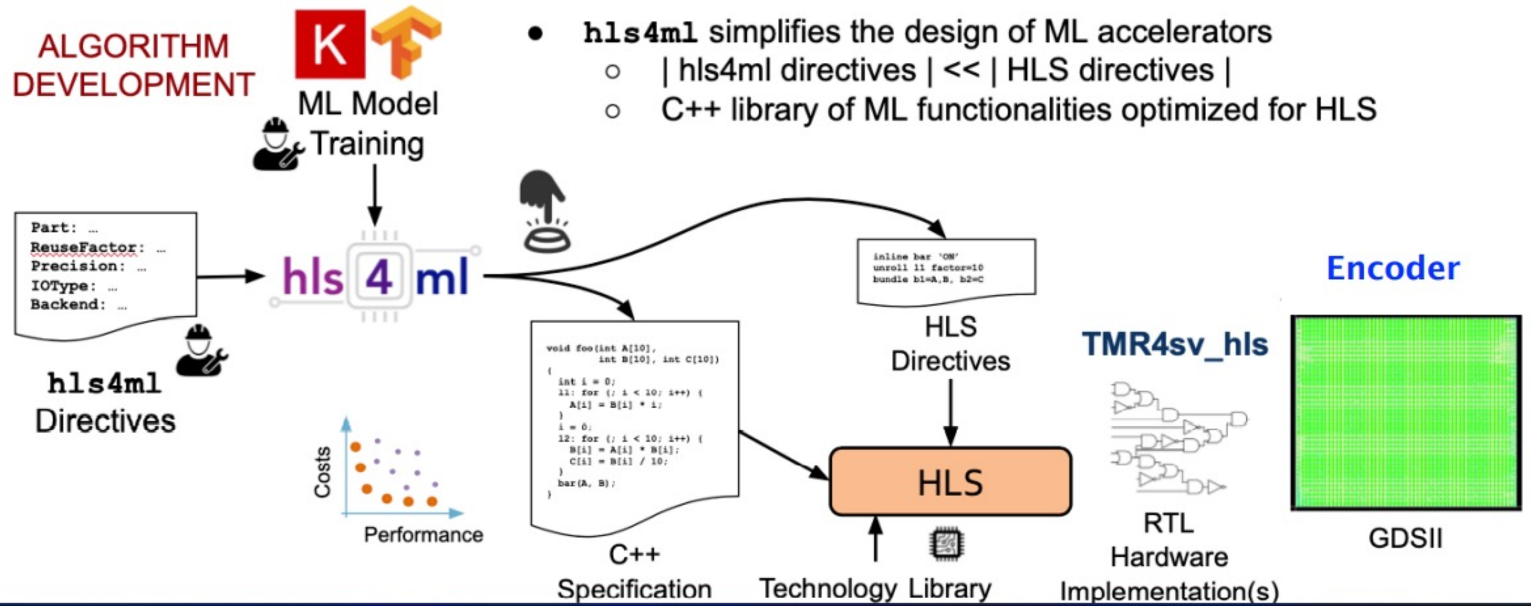
HGCAL 8" Module
Each trigger Cell consists of 3*3 sensors



Physics driven hardware co-design

Rapid prototyping and optimization of network achieved through

- **QKeras** : network development with **quantization-aware training** and physics simulation
- **hls4ml** : neural network description (h5 file e.g.) ➔ HLS-compliant C++ format
- **Catapult HLS** : C++ ➔ RTL
- **TMR4sv_hls** : Automated TMR for System Verilog



[Design of a reconfigurable autoencoder algorithm for detector front-end ASICs]
Giuseppe Di Guglielmo
2020/11/30, Fast Machine Learning for Science Workshop

小结与计划

- 用 ML-aided E2E 分析实现 CEPC 探测器的快速优化迭代
- Jet energy resolution , jet charge
- Peaking finding
- Background suppression + data compression
- Knowledge embedded ML in calorimeter reconstruction and simulation …