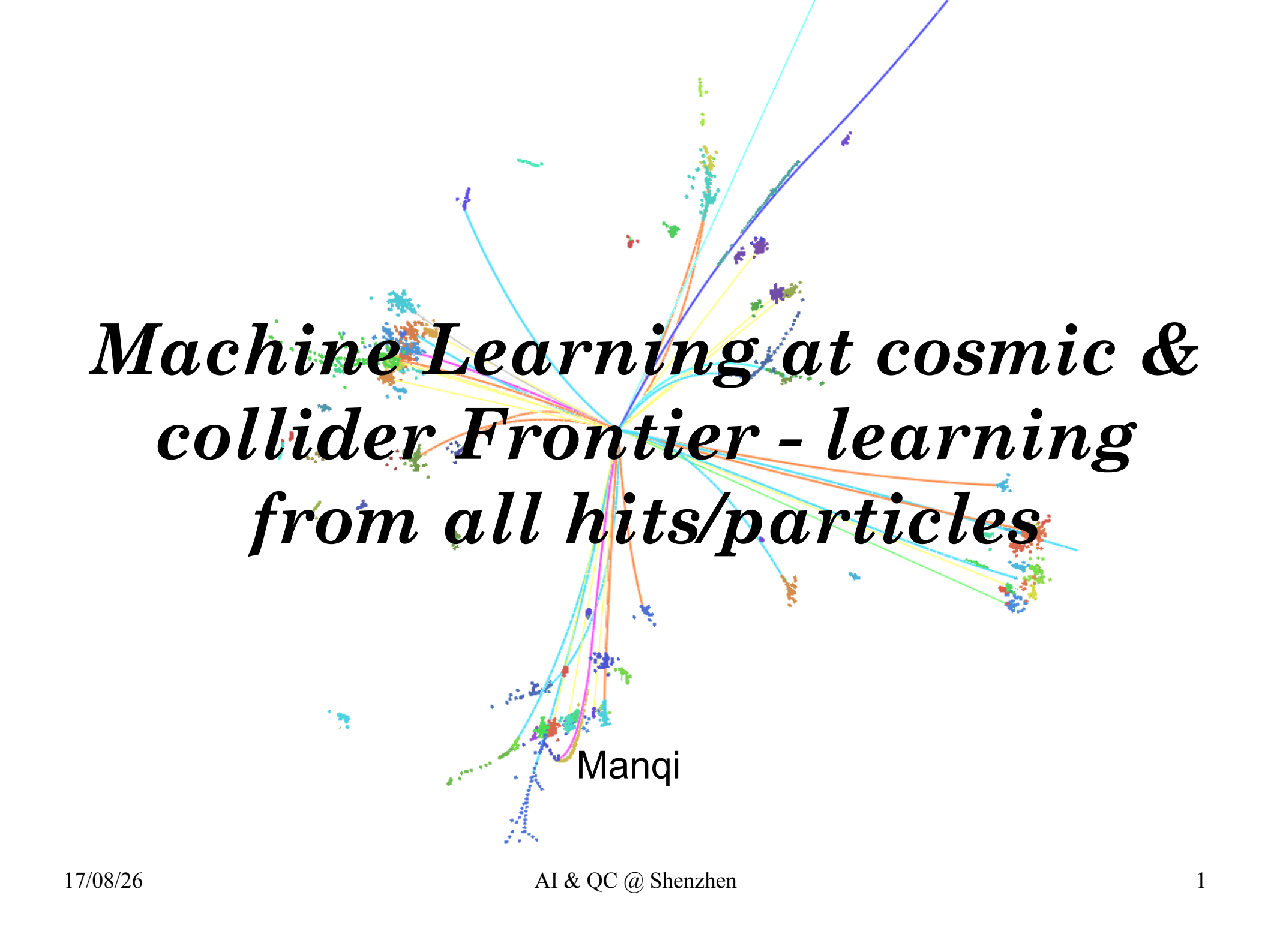




Machine Learning at cosmic & collider Frontier - learning from all hits/particles

Manqi

LHAASO

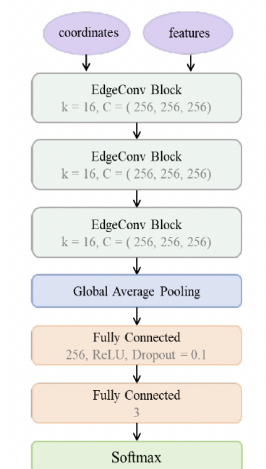
- **Location:** $29^{\circ}21'27.6''$ N, $100^{\circ}08'19.6''$ E
- **Altitude:** 4410 m
- **2021-07**

AI group @ LHAASO

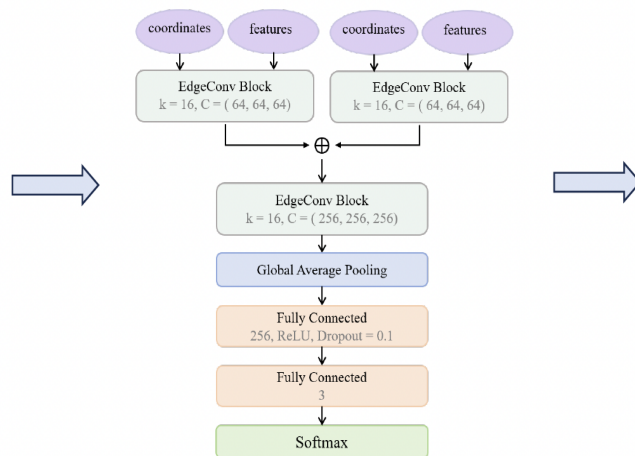


Learning from all hits

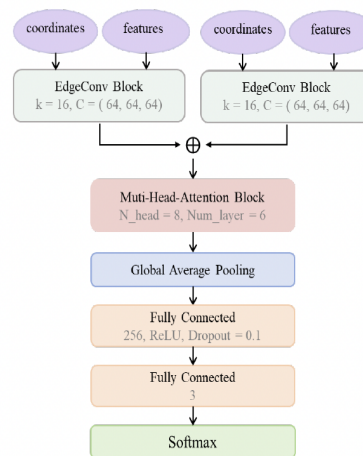
- Tasks:
 - Identification: Proton VS Gamma, Different species of ions...
 - Regression: Angular, energy, and central position...
- ML Models, evolution & optimization
 - i.e., Dual Stream Transformer@ KM2A, physics info. guided optimization



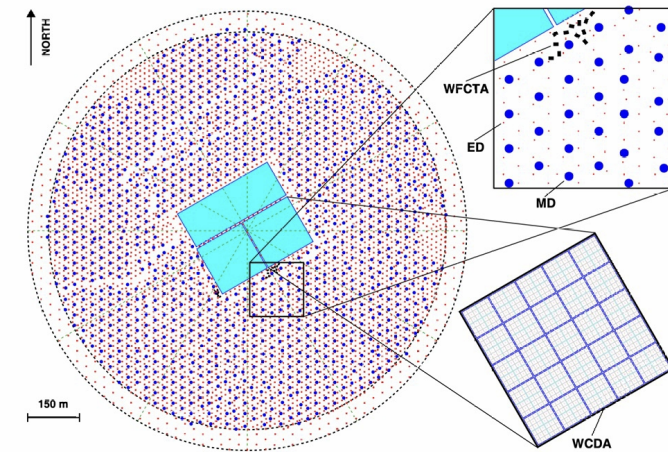
2025.10 ~ ParticleNet



2025.12 ~ Dual Stream ParticleNet

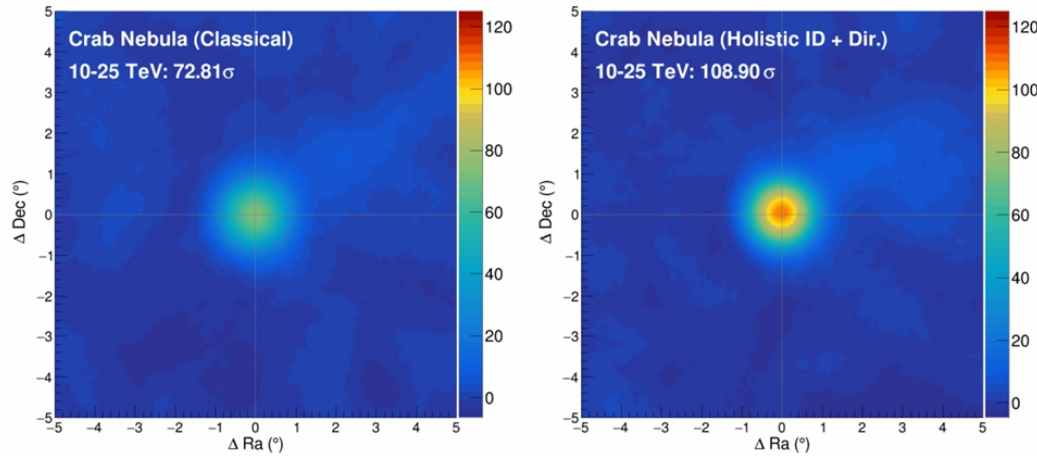


2026.3 ~ Dual Stream
GNN-Transformer

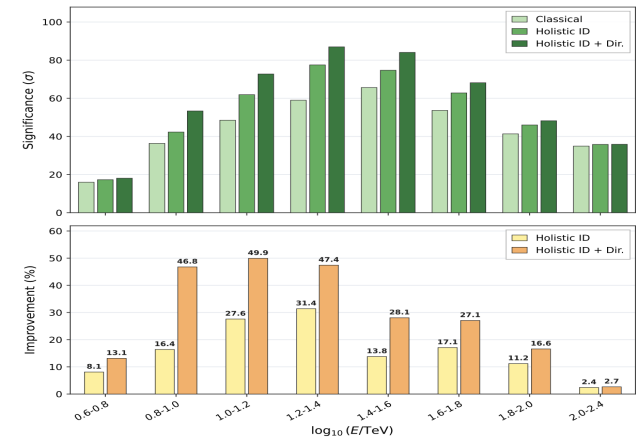
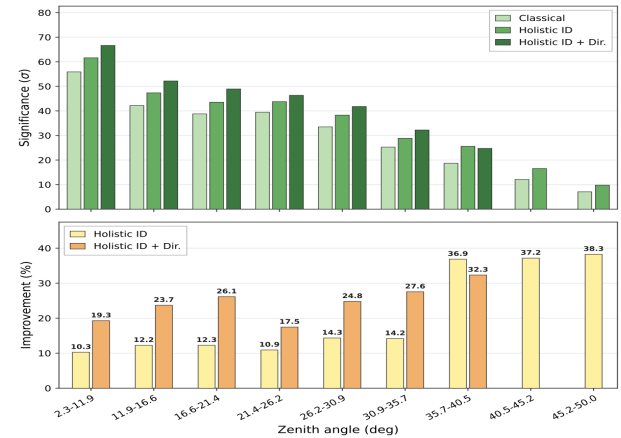
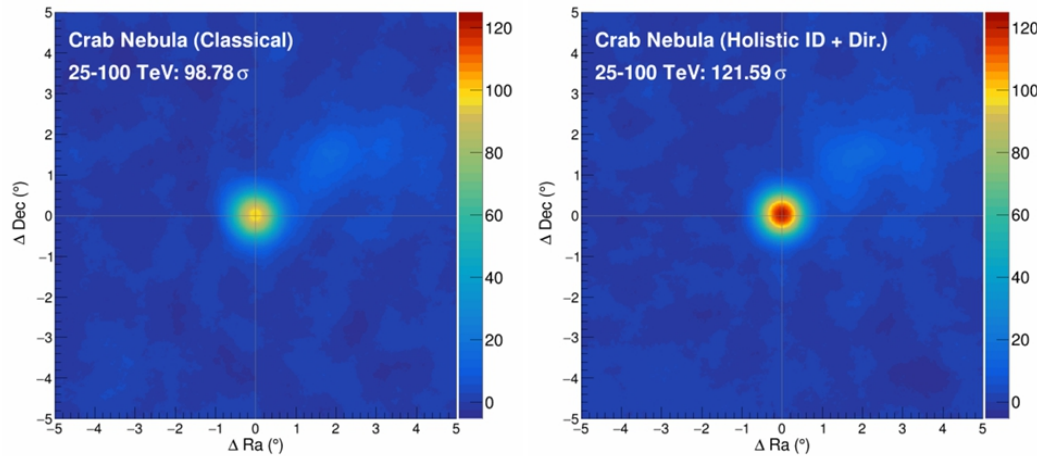


2

KM2A: ~ 1.5x @ Grab Nebula (2024)



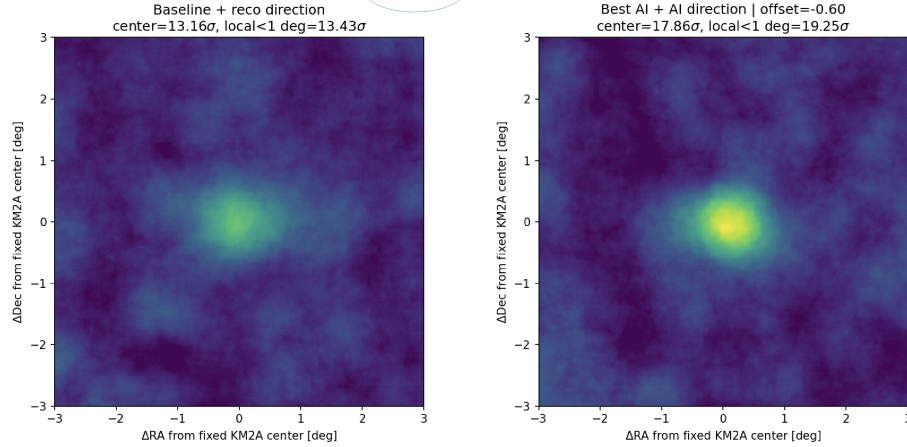
Credit: Yangchao, Yongfeng, etc



KM2A: more candles

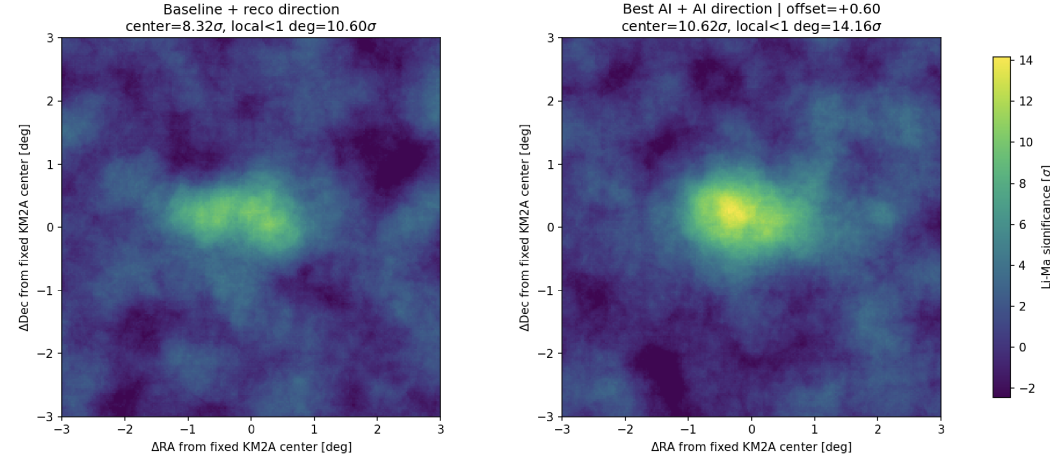
这个就是源的编号

1LHAASO J2108+5153u | 25-100TeV | shared color scale



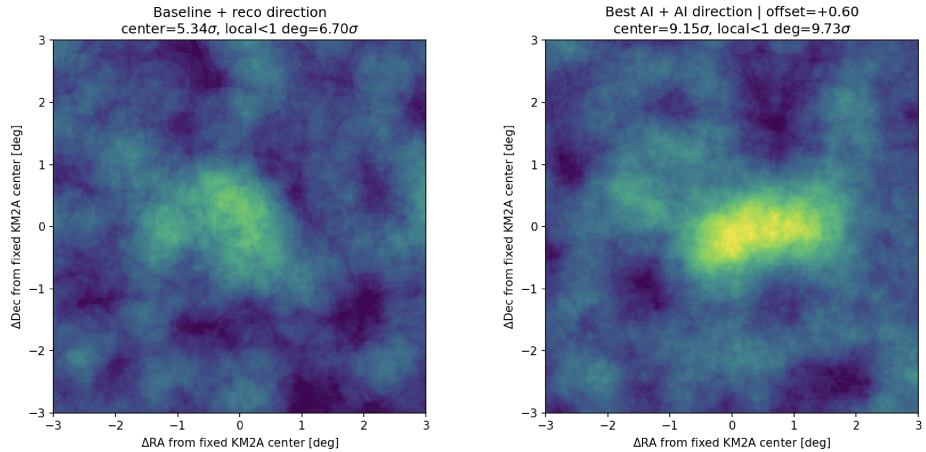
↑ 43% 天顶角范围: 23°—50°

1LHAASO J0056+6346u | 25-100TeV | shared color scale



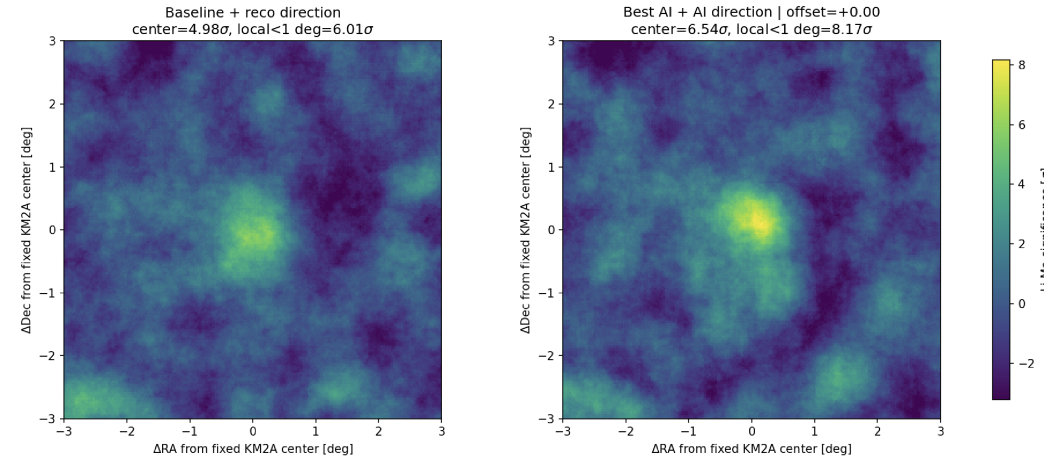
↑ 34% 天顶角范围: 34°—50°

1LHAASO J0007+7303u / CTA 1 | 25-100TeV | shared color scale



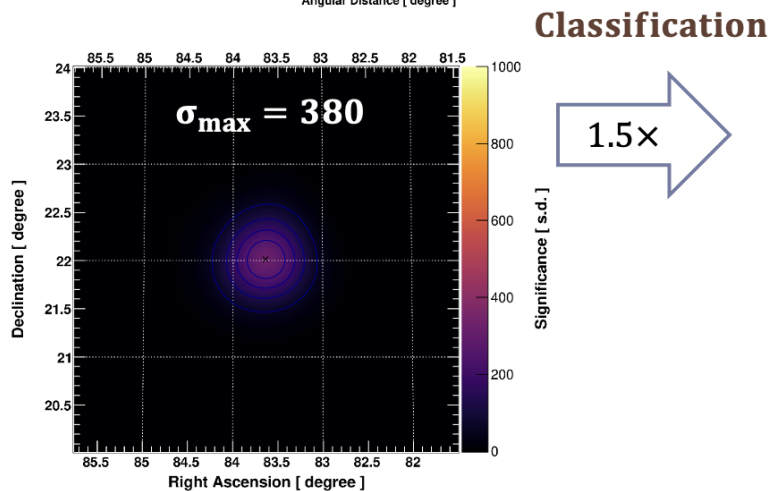
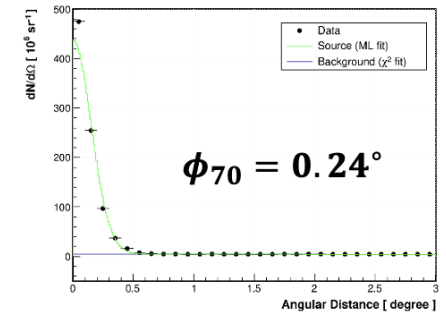
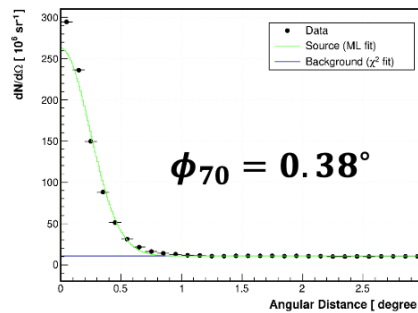
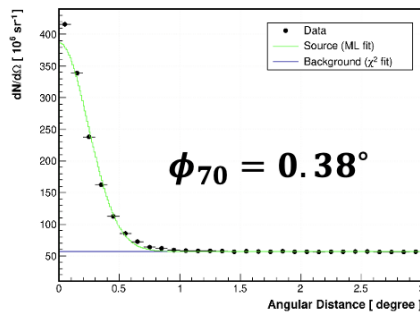
↑ 45% 天顶角范围: 44°—50° (也称为CTA 1)

1LHAASO J1740+0948u | 25-100TeV | shared color scale

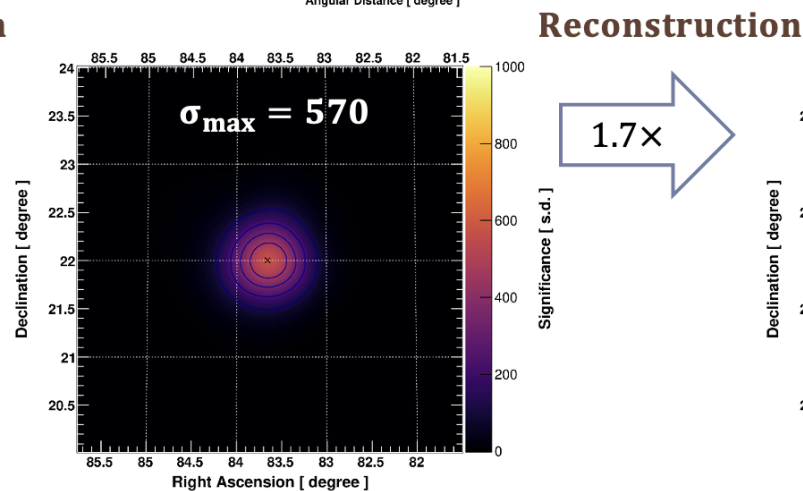


↑ 36% 天顶角范围: 20°—50°

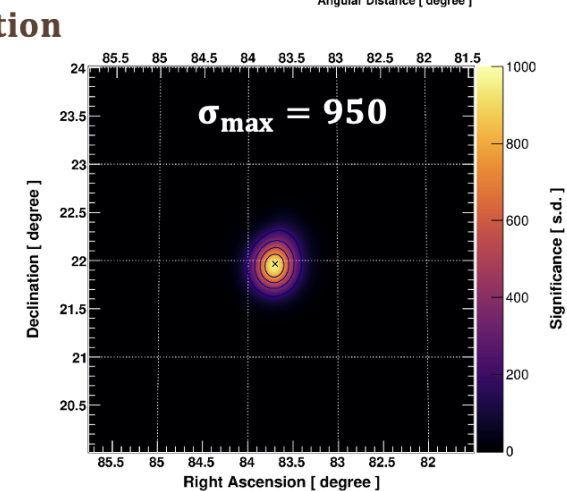
WCDA: $\sim 3\times$ @ Grab Nebula (2022)



1.5×

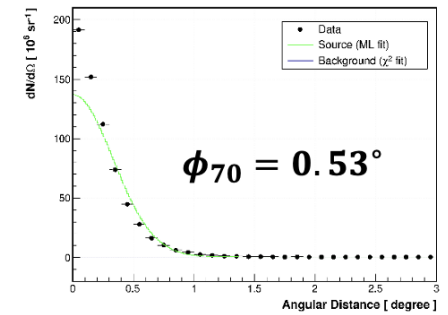
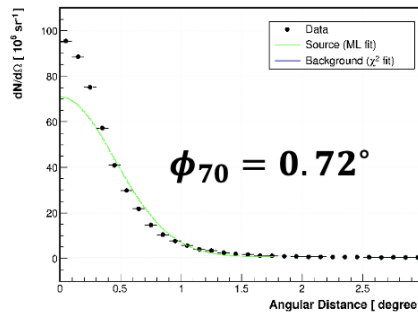
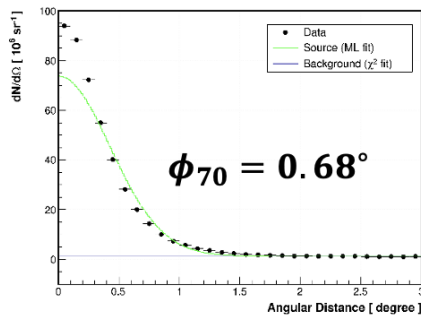
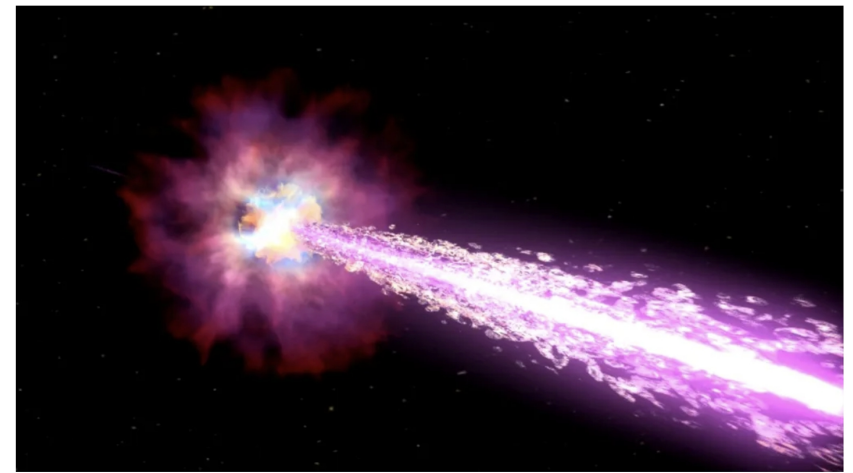


1.7×

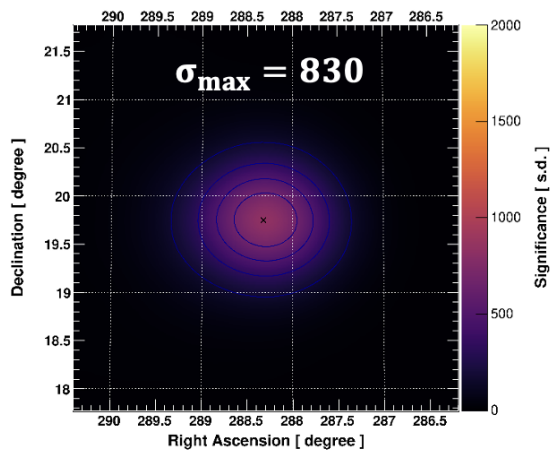


Credit: Shicong, Weiguo, etc

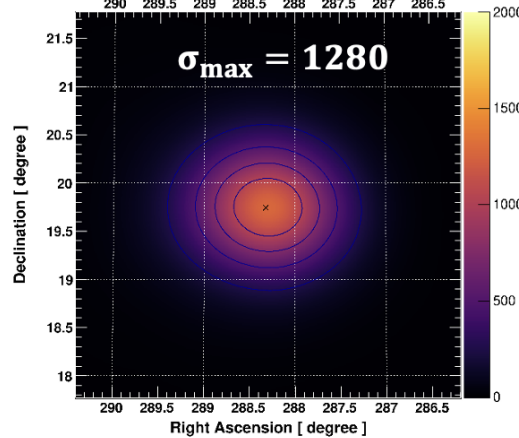
WCDA: $\sim 3\times$ @ GRB 221009A (2h)



Classification

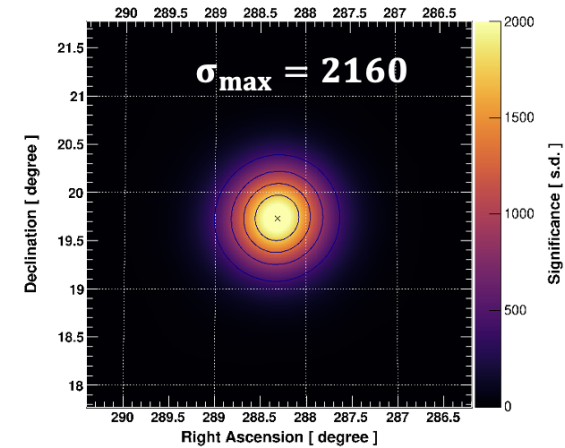


1.5×



Reconstruction

1.7×

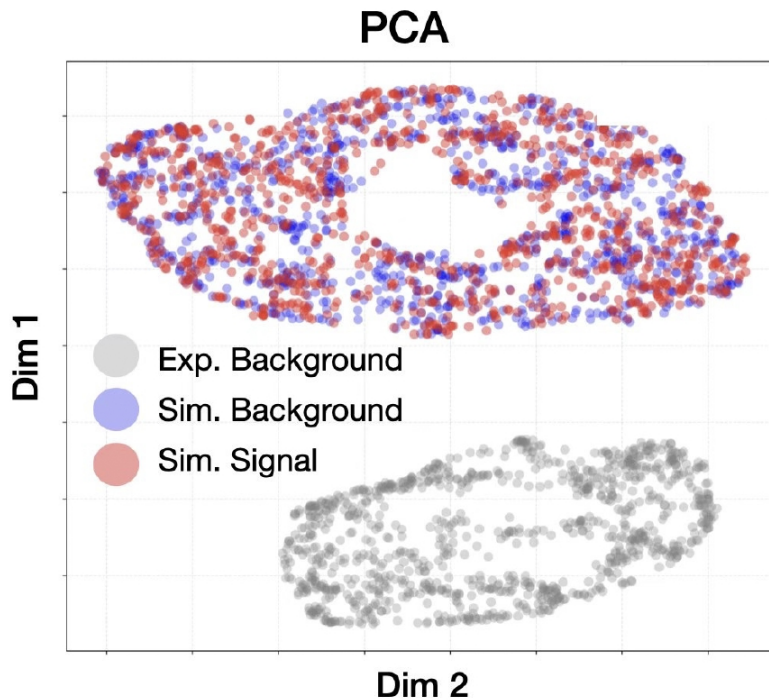


Credit: Shicong, Weiguo, etc

Current Bottleneck: Data-MC Discrepancy

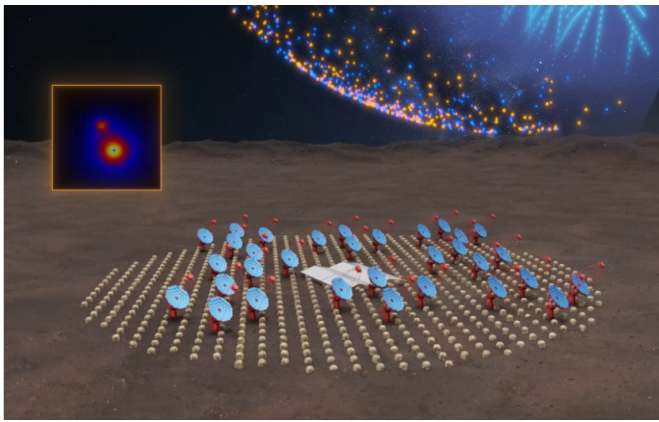
➤ Possible solutions:

- **Better Simulation:** Develop more accurate Simulation (from CORSIKA, to Geant4, to digitization), especially to model noises at different stages.
→ **Ongoing Efforts:** The AI-LHAASO group is actively preparing new MC samples with various combinations.
- **Better Architecture that less sensitive to the Discrepancy:** Self-supervised Learning (*to guide the attention of ML ...*)

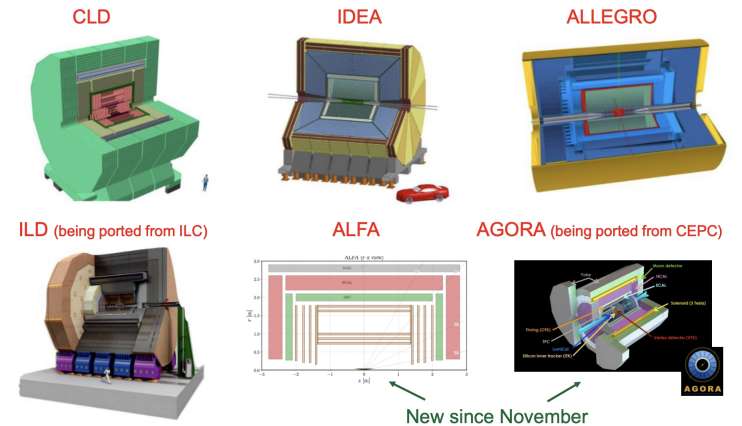
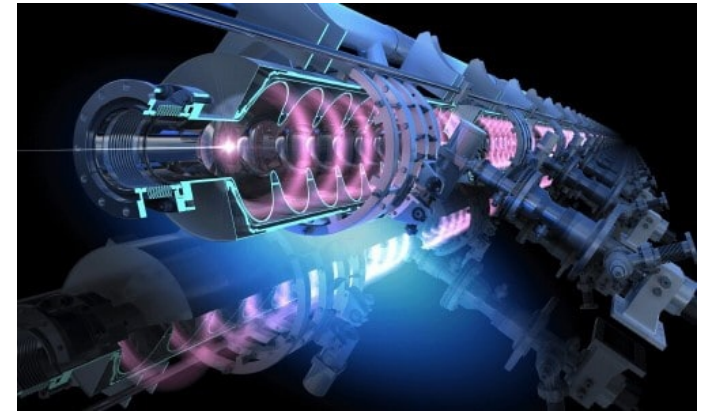
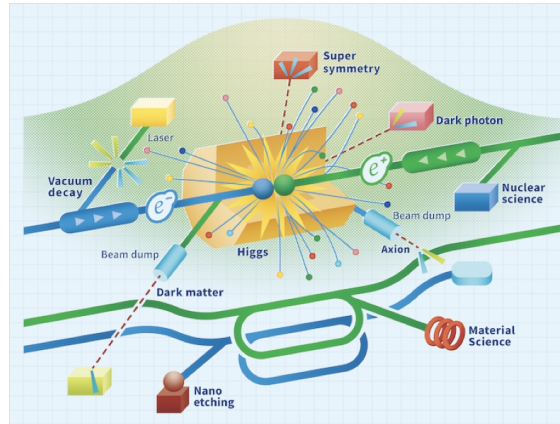
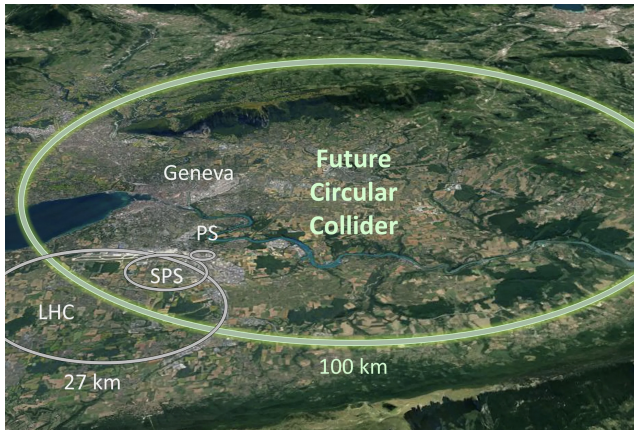


Next milestones

- More tasks, especially to address MC-data discrepancy
- Sky survey: increase the #sources by multifold
- LACT upgrade & Fusion of info. from different sub detectors
- Agents
- Synergies with more facilities
- AI interpretation
- ...



Higgs factories



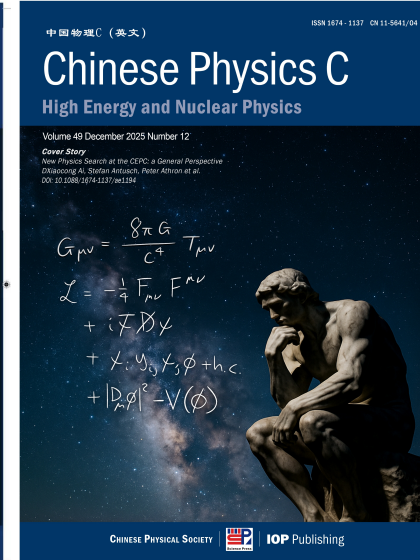
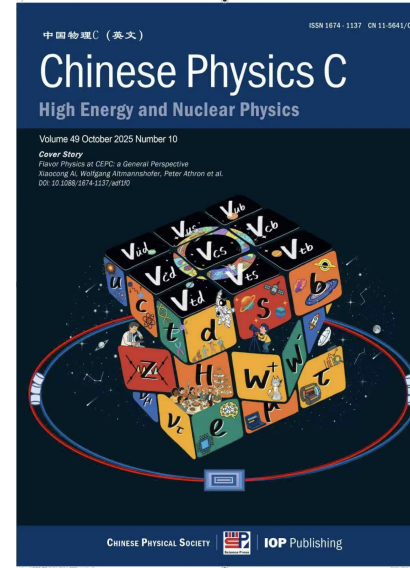
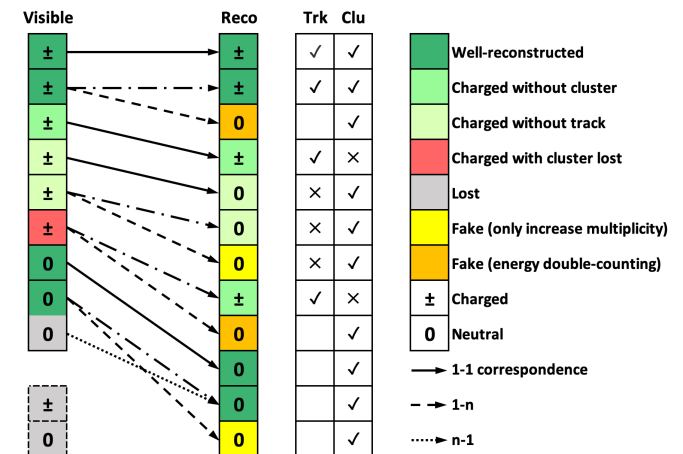
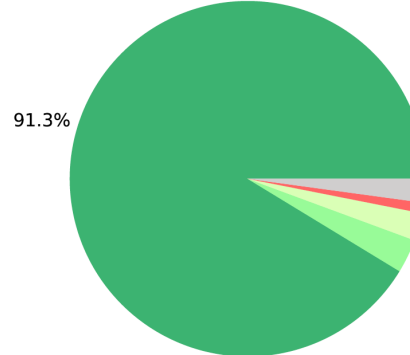
Physics metrics

- CEPC Physics White papers:
- Higgs: 2019, CPC
- ESPPU input: 2019 ArXiv
- Snowmass input: 2022 ArXiv
- **Flavor physics: 2025, CPC**
- **New physics: 2025 CPC**
- **Current focus**

- **EW White paper**
- **QCD studies**
- Reco & Det. Optimization
- ...

- *Boost horizon by orders of magnitudes*

*Address/answers profound mysteries:
mass/matter origin, nature of Dark Matter, Phase
Transitions, evolution and fate of the Universe...*



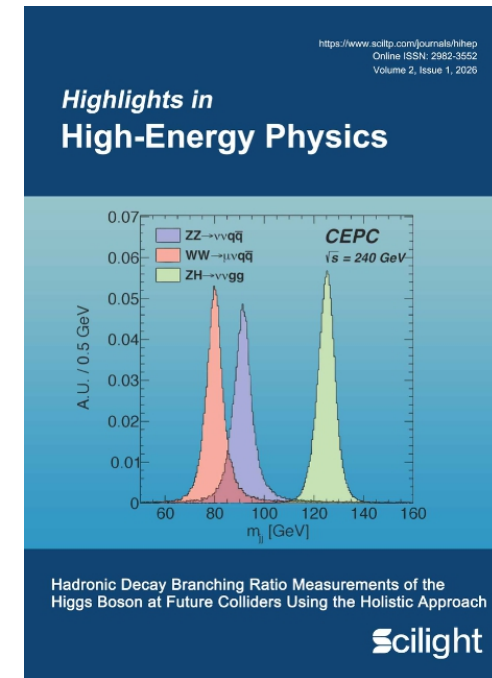
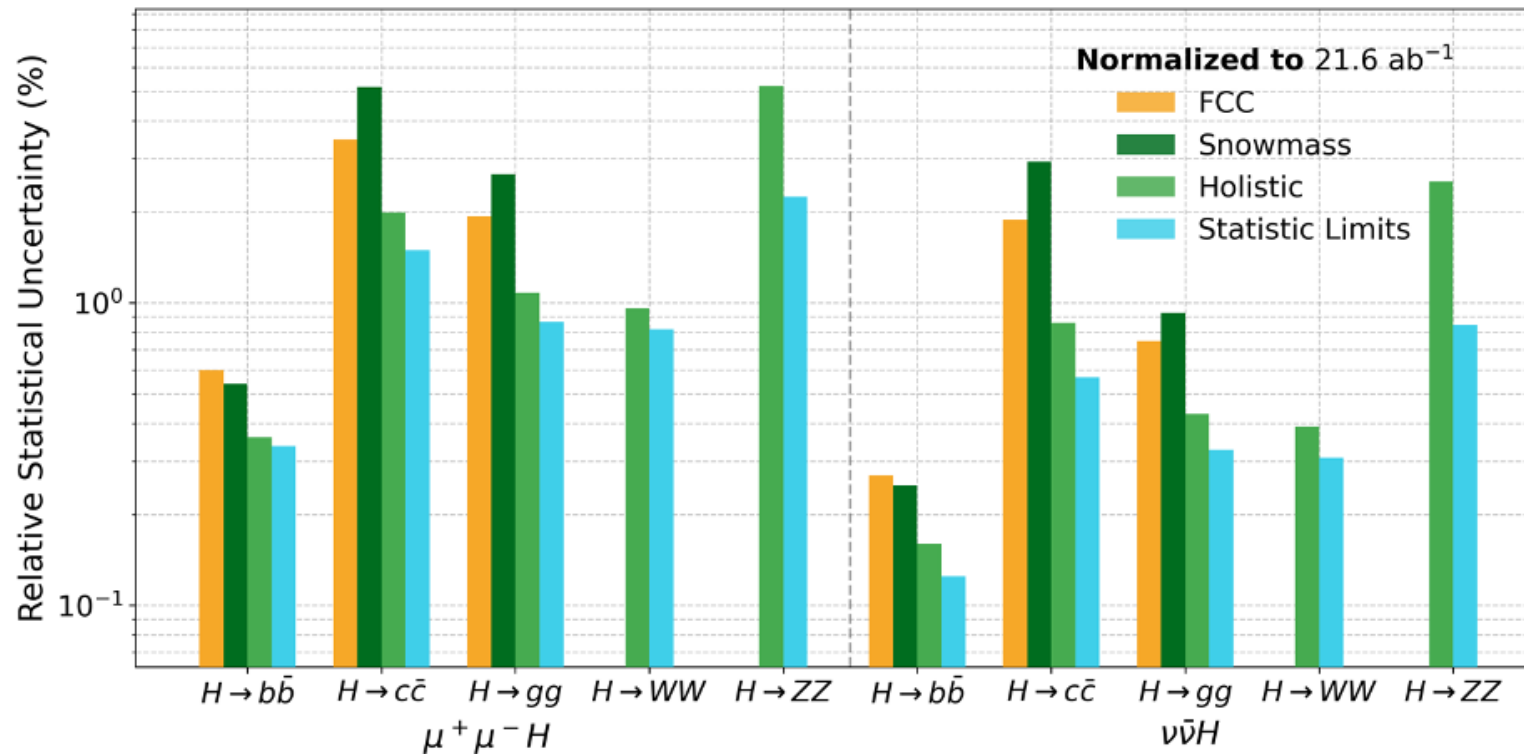
Holistic analysis: identify event type

Truth \ Predicted	$H \rightarrow b\bar{b}$	$H \rightarrow c\bar{c}$	$H \rightarrow gg$	$H \rightarrow WW^*$	$H \rightarrow ZZ^*$	ZZ_{sl}
$H \rightarrow b\bar{b}$	93.36%	1.14%	1.00%	0.09%	4.31%	0.10%
$H \rightarrow c\bar{c}$	0.94%	90.03%	2.68%	3.38%	2.88%	0.10%
$H \rightarrow gg$	1.89%	2.79%	84.40%	5.93%	4.94%	0.05%
$H \rightarrow WW^*$	0.06%	2.04%	5.82%	82.15%	9.87%	0.05%
$H \rightarrow ZZ^*$	3.67%	2.95%	6.26%	18.73%	68.32%	0.07%
ZZ_{sl}	0.23%	0.21%	0.17%	0.06%	0.08%	99.25%

Truth \ Predicted	$H \rightarrow b\bar{b}$	$H \rightarrow c\bar{c}$	$H \rightarrow gg$	$H \rightarrow WW^*$	$H \rightarrow ZZ^*$	ZZ	WW	$q\bar{q}$
$H \rightarrow b\bar{b}$	93.24%	1.08%	1.07%	0.09%	3.97%	0.01%	0.44%	0.10%
$H \rightarrow c\bar{c}$	0.75%	90.29%	2.54%	3.10%	2.69%	0.04%	0.52%	0.07%
$H \rightarrow gg$	1.82%	2.79%	84.45%	5.72%	4.93%	0.01%	0.25%	0.03%
$H \rightarrow WW^*$	0.05%	2.05%	5.74%	82.20%	9.71%	0.04%	0.18%	0.02%
$H \rightarrow ZZ^*$	3.75%	3.03%	6.20%	17.86%	68.81%	0.03%	0.27%	0.03%
ZZ	0.01%	0.05%	0.01%	0.05%	0.03%	99.33%	0.49%	0.04%
WW	0.84%	0.71%	0.31%	0.12%	0.22%	0.30%	96.84%	0.67%
$q\bar{q}$	0.12%	0.07%	0.02%	0.01%	0.03%	0.02%	0.25%	99.48%

- To identify Higgs (full hadronic) signals

@ Higgs measurements

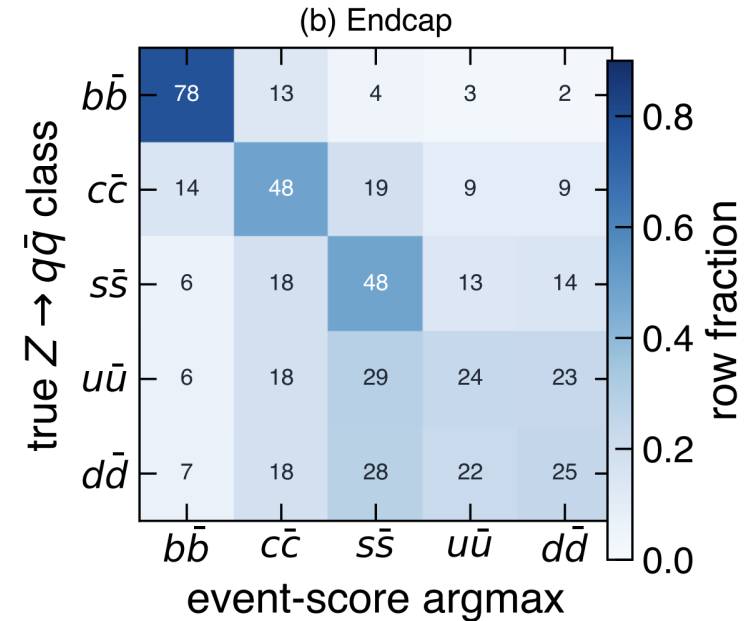
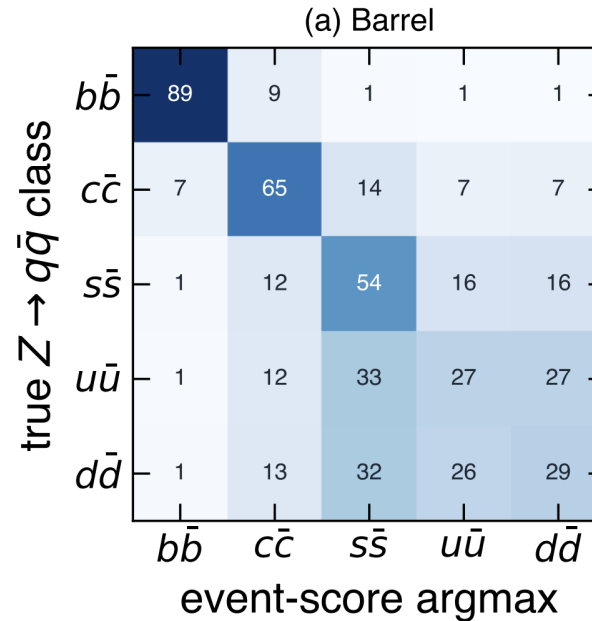
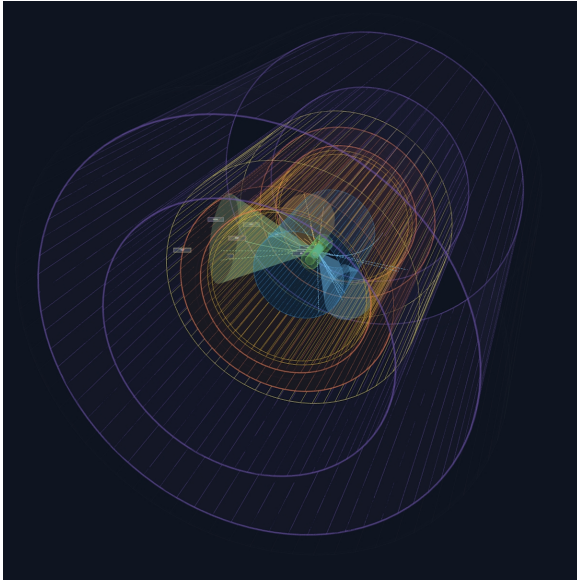


[Link to paper](#)

Holistic approach: improves the accuracy ~3 times, approach the statistic limit for $H \rightarrow b\bar{b}$...

SM background included

LEP Legacy data: 1 min at CEPC/FCC



Training Events



Data: DELPHI 1994–1995, 40 pb^{-1} at the Z pole, $\sim 2\text{M}$ events



Key methods:

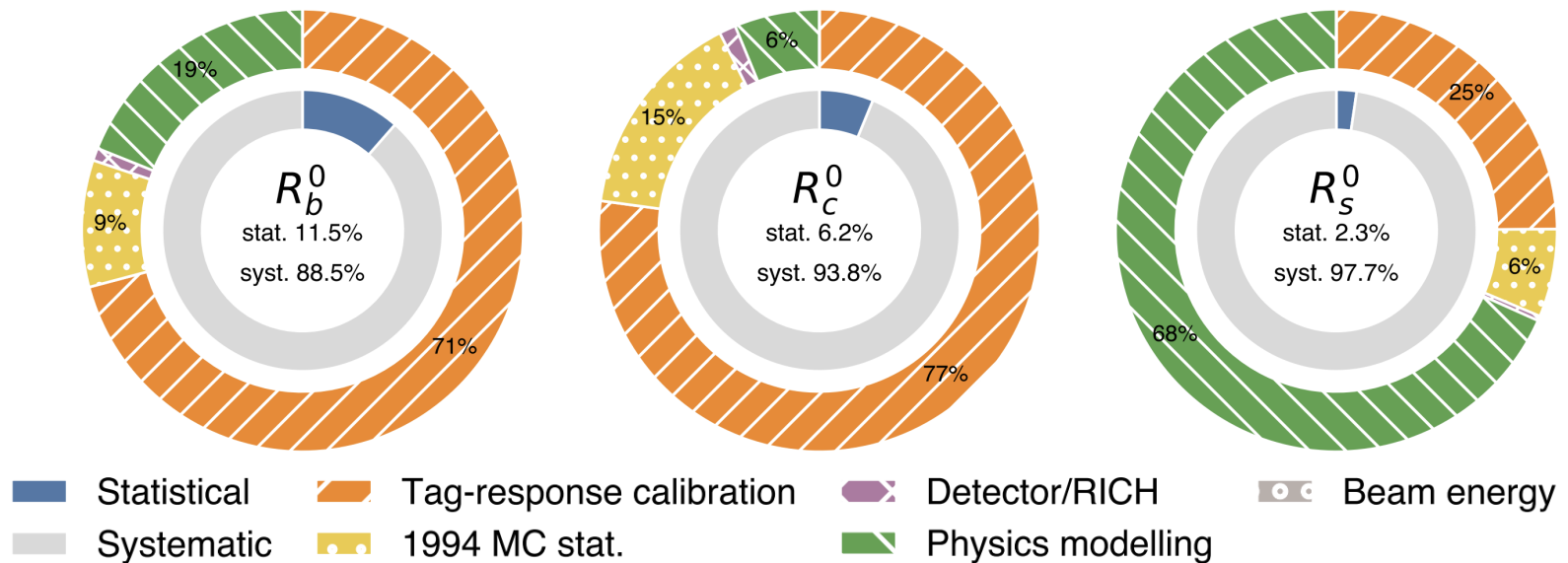
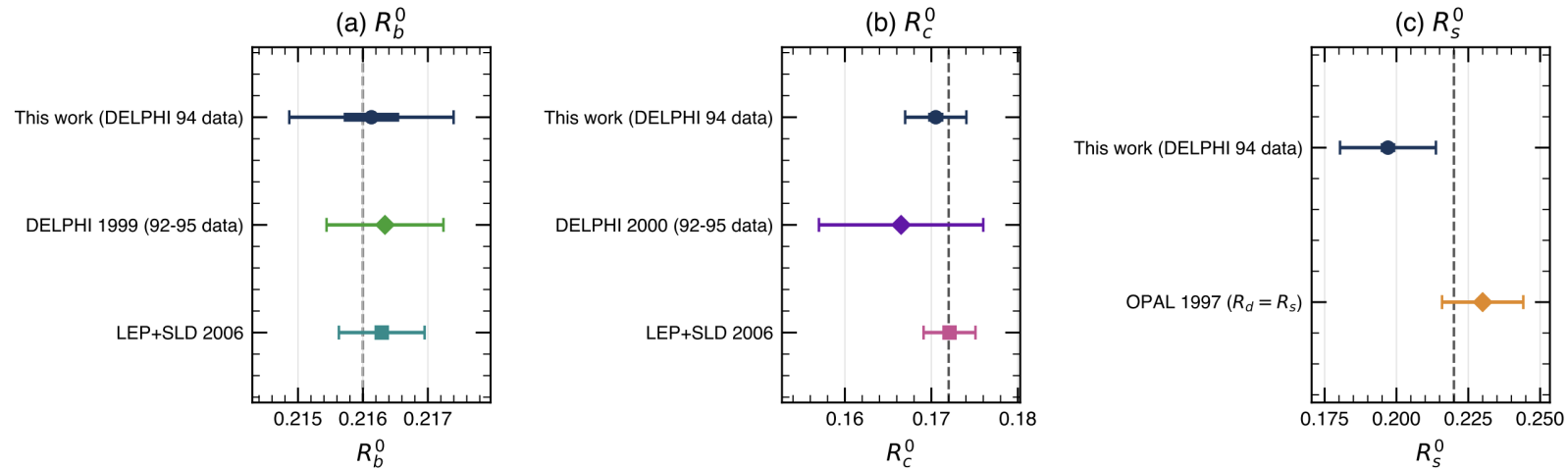
- ParticleNet 5-flavor tagger
- Cut-optimization criteria: minimum pre-fit data/MC χ^2
- 8-bin categorization: barrel vs endcap, b/c/s-enriched vs other
- Likelihood profiling of R_b , R_c , R_s simultaneously



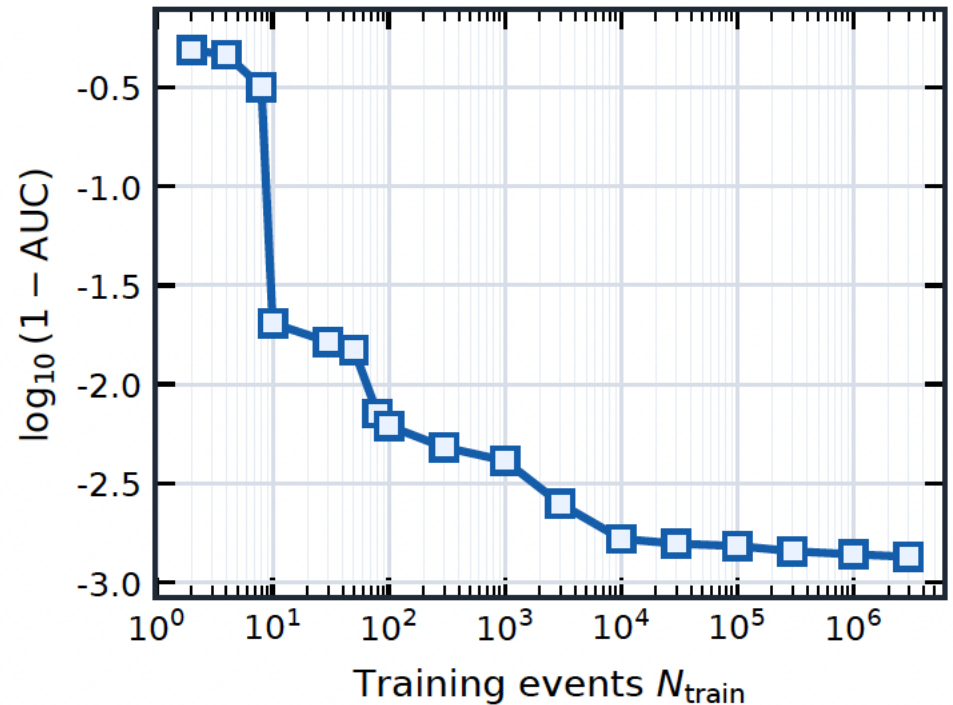
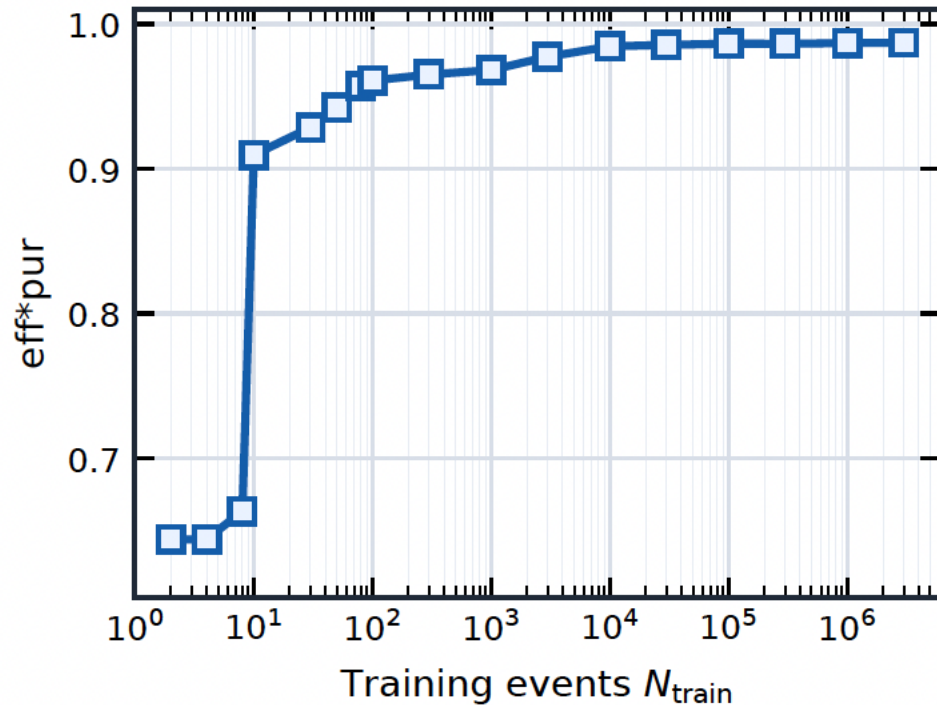
Systematics: cut threshold, jet tagging, fragmentation, detector response, etc..



LEP Legacy data: Preliminary



Scaling @ LHAASO Pid

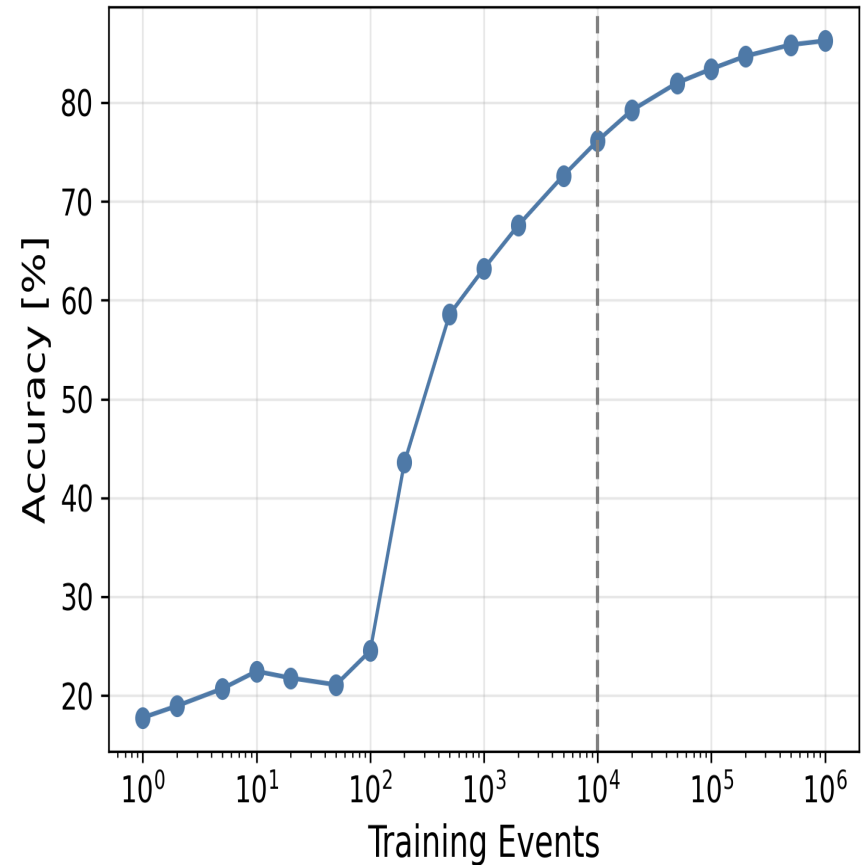


Model could distinguish Proton from Photon, after ~ 10 events training...

Difference between EM/Had showers is much more significant than hadronic events/jets

Scaling: @ event id

Truth \ Predicted	$H \rightarrow b\bar{b}$	$H \rightarrow c\bar{c}$	$H \rightarrow gg$	$H \rightarrow WW^*$	$H \rightarrow ZZ^*$	ZZ_{sl}
$H \rightarrow b\bar{b}$	93.36%	1.14%	1.00%	0.09%	4.31%	0.10%
$H \rightarrow c\bar{c}$	0.94%	90.03%	2.68%	3.38%	2.88%	0.10%
$H \rightarrow gg$	1.89%	2.79%	84.40%	5.93%	4.94%	0.05%
$H \rightarrow WW^*$	0.06%	2.04%	5.82%	82.15%	9.87%	0.05%
$H \rightarrow ZZ^*$	3.67%	2.95%	6.26%	18.73%	68.32%	0.07%
ZZ_{sl}	0.23%	0.21%	0.17%	0.06%	0.08%	99.25%

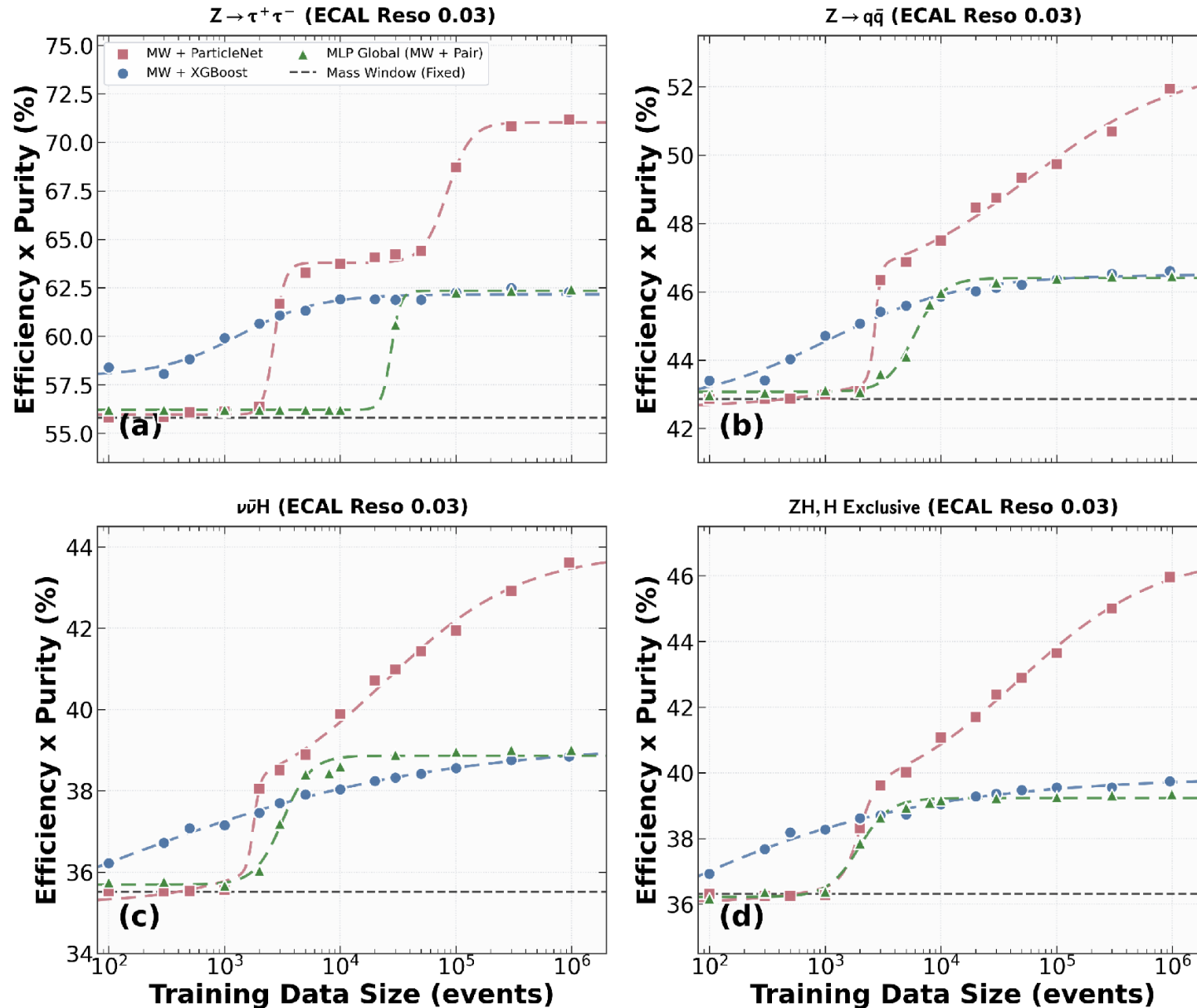


S-Curve: the emerging and saturation of the performance...

What determines the critical points?

What determines the ultimate performance & converging speed?

Grouping: from leaves to the tree: pi0 reco. (Preliminary)



Scaling: not a summary

Meta question	Task	1 st Critical Point	2 nd Critical Point - Saturation	Architectures
Classification	Jet origin Id @ HF	300 – 2k	1 M	PN/ParT - LLM
	Event id @ HF	100	10 k – 1 M	PN
	Pid @ LHASSO KM2A	10	100 – 1 M	Multi-head PN, etc
	...			
Regression	Hadron Energy @ HF Calo	7 k	100 k	PN
	...			
Grouping	CSI @ HF	1000	300 k	ParT, Mamba, etc
	...			
...				

Perspective: exist certain principles, for certain architecture to be optimal at certain meta-question

A lot more examples ... to be accumulated & analyzed... from & beyond HEP

Remark: 2nd critical point & saturation needs better specification, in fact, S-Curve is only 1st order description...

Quantification depends on figure of merit

Saturation of performance: standard & optimization determined by requirements

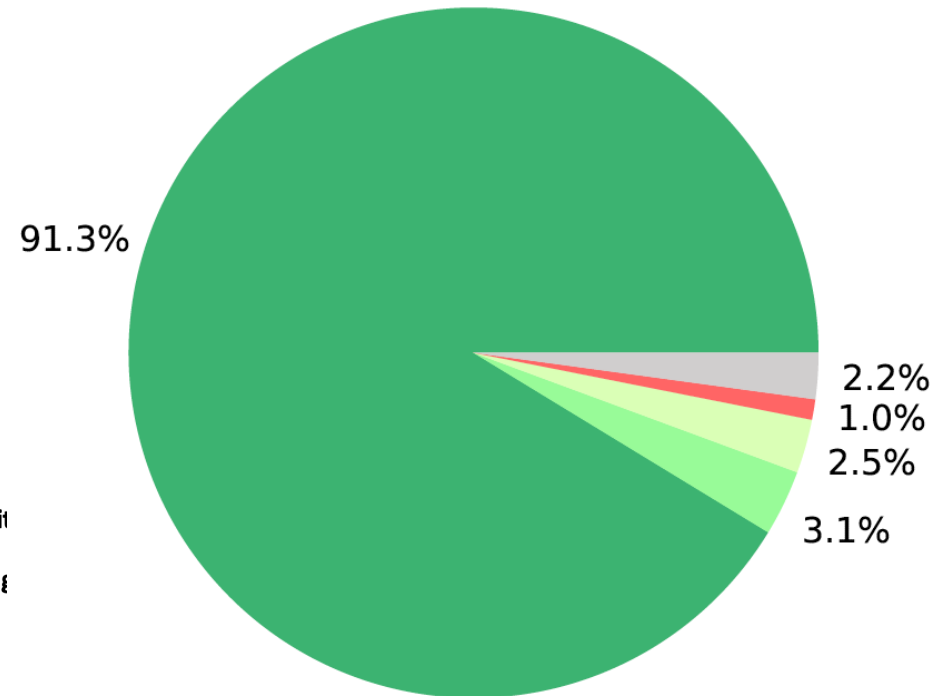
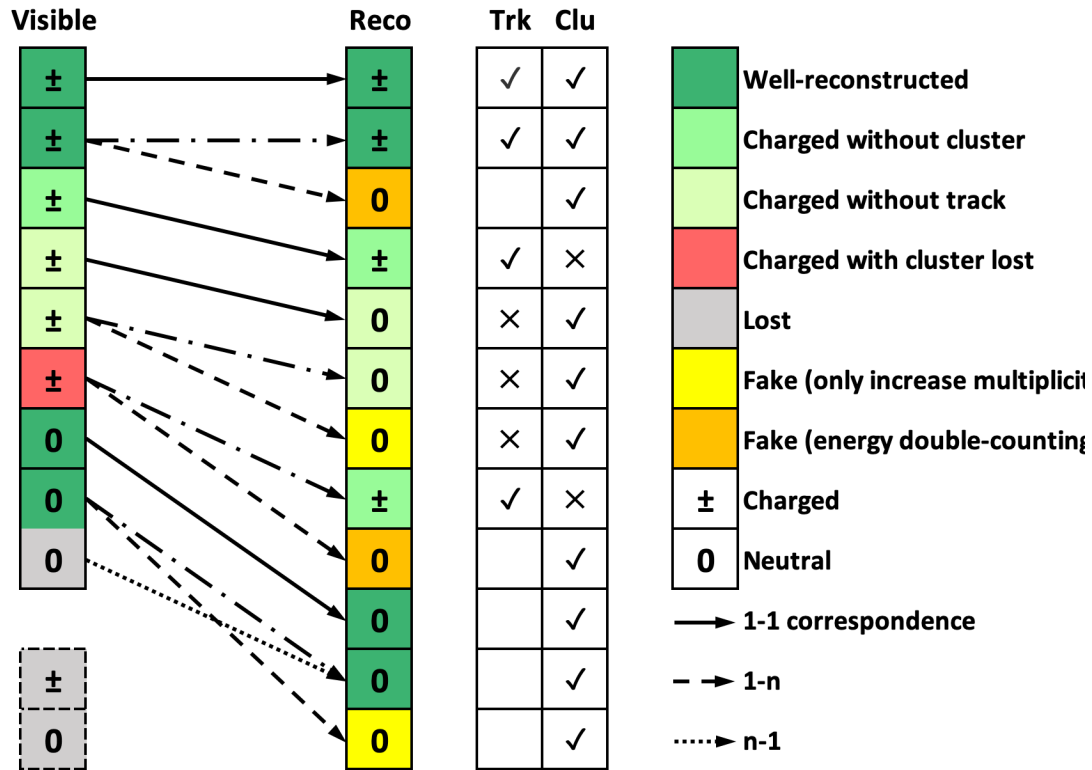
Summary

- Machine Learning, boost the effective luminosity/yields by ~ 1 order of magnitude
 - LHAASO: Significance of crab nebula (standard candle) & GRB $\times 1.5 - 3$ times
 - Higgs factories: accuracies of core Higgs measurements: $\times 2-4$ times
 - LEP legacy data: $\sim \times 4$ times on R_c and 1st direct measurement on R_s ([Preliminary](#))
- **Scaling**: quantify the emerge & saturation of performance, characteristic to ML models
- Future:
 - Novel methods to address MC-data discrepancy & systematic control
 - Fusion of Agent + ML + X...
 - AI interpretability: what AI perceived, beyond conventional methods
 - ...

$3*3 + 1$, *Why?*

Thank you

1-1 correspondence reconstruction



Computer Physics Communications 314 (2025) 109661



Contents lists available at ScienceDirect
 Computer Physics Communications
 journal homepage: www.elsevier.com/locate/cpc



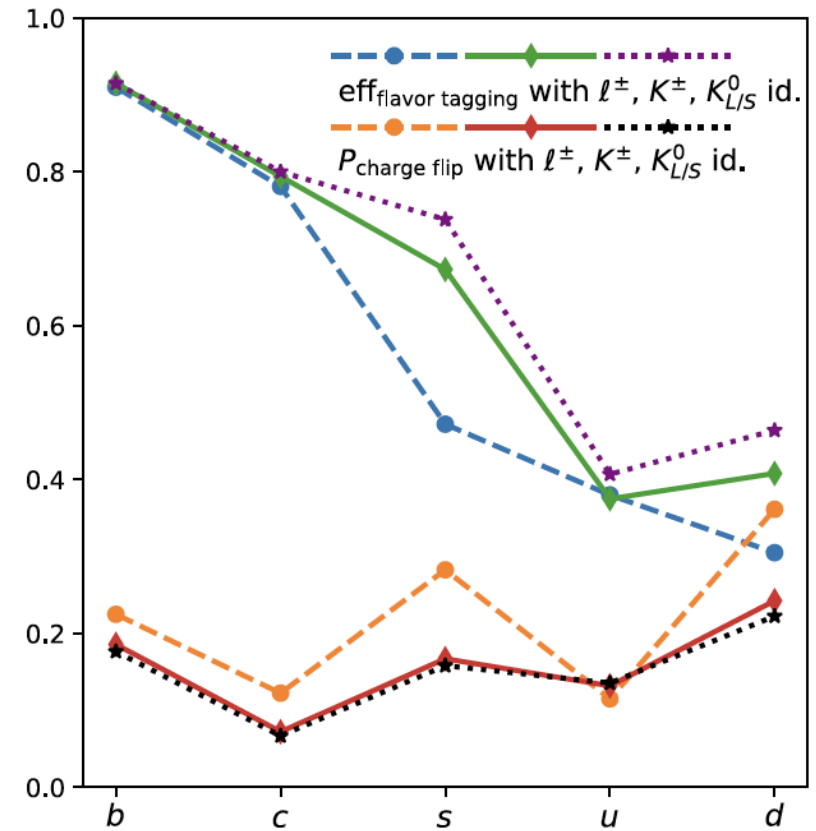
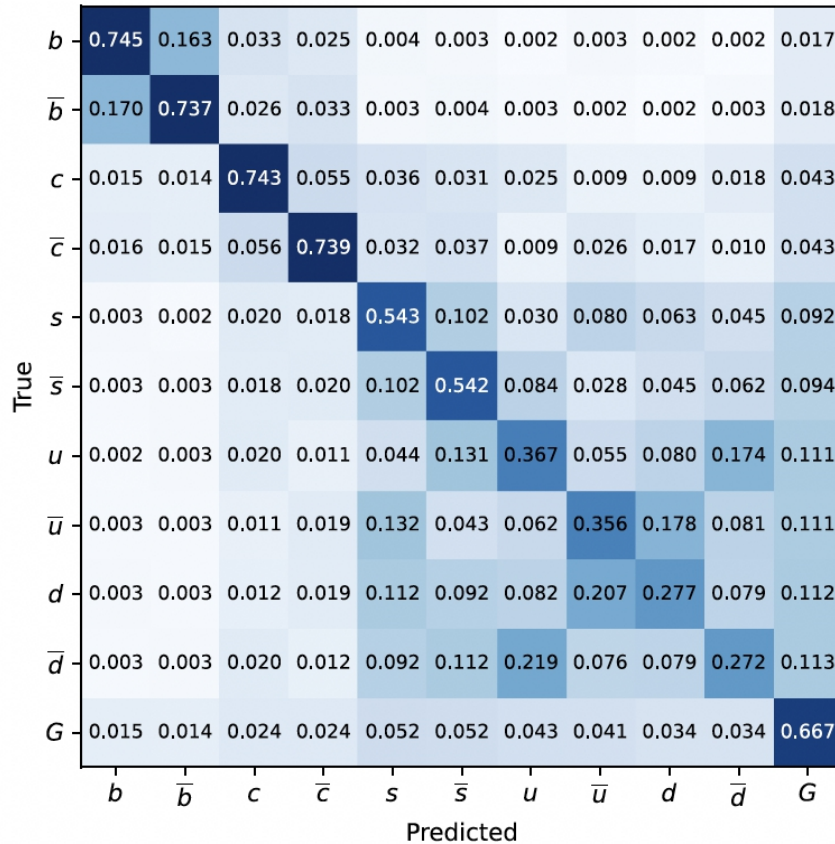
Computational Physics

One-to-one correspondence reconstruction at the electron-positron Higgs factory

Yuxin Wang^{a,b}, Hao Liang^{a,c,d}, Yongfeng Zhu^e, Yuzhi Che^{a,f}, Xin Xia^{a,c}, Huilin Qu^g, Chen Zhou^e, Xuai Zhuang^{h,i}, Manqi Ruan^{a,c,e}

<https://arxiv.org/abs/2411.06939>

Holistic Reco: Jet origin id

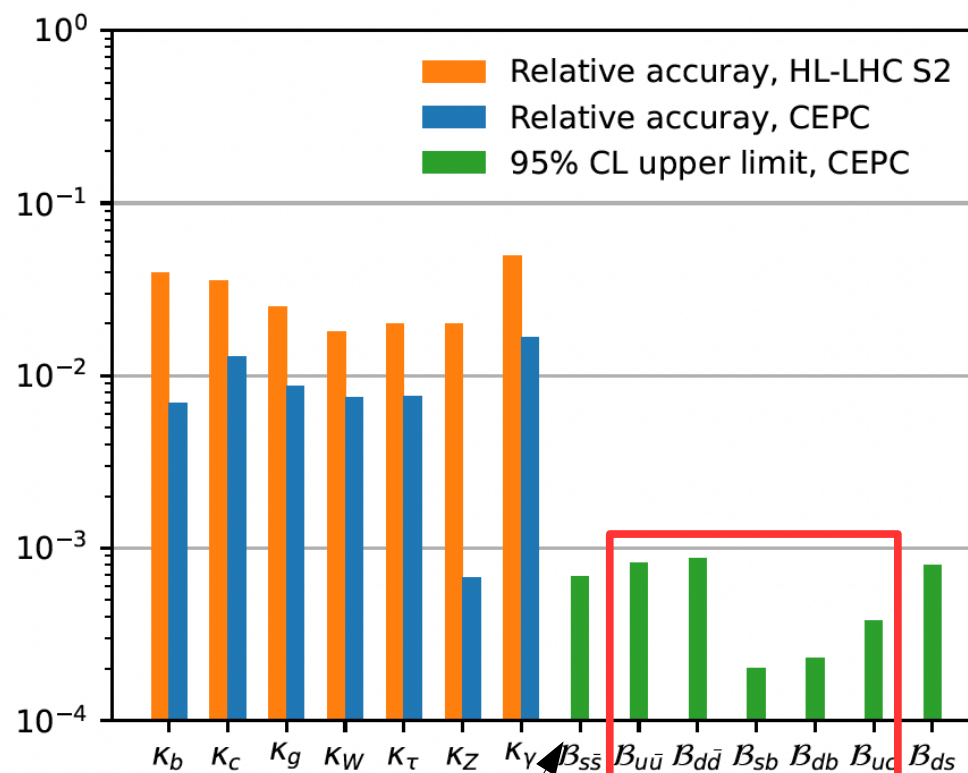
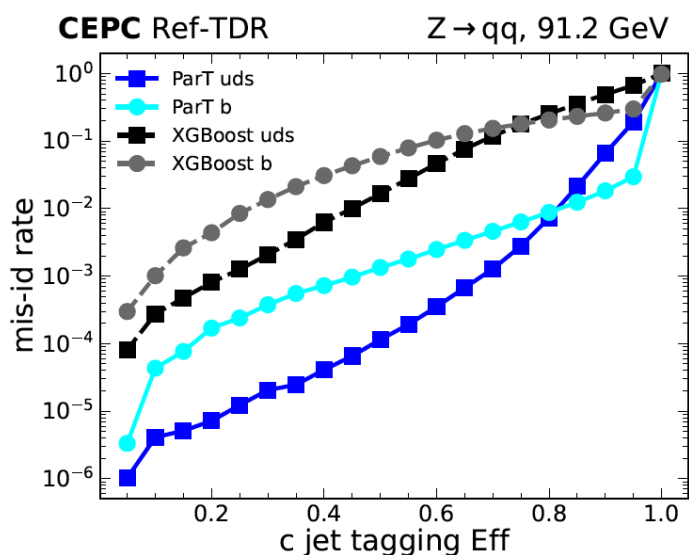
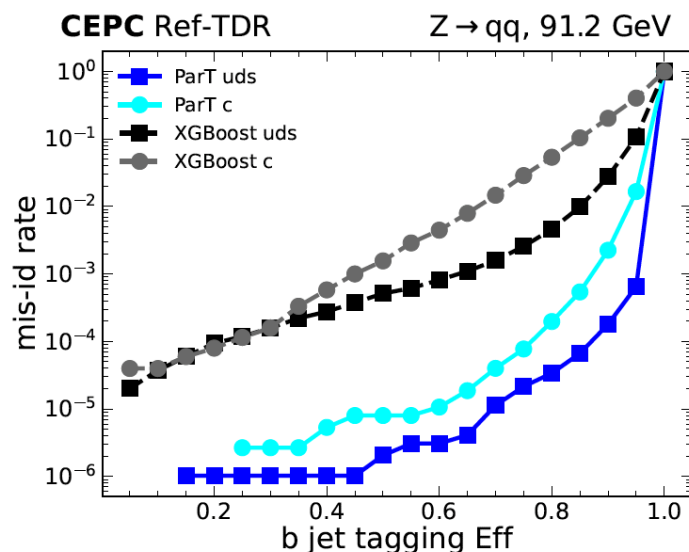


- 11 categories (5 quarks + 5 anti quarks + gluon) identification, realized at Full Simulated di-jet events at CEPC CDR baseline with **Arbor + ParticleNet**
- Published in PRL 132, 221802 (2024). Comment from the referee: *"demonstrate the world-leading performance of tagger", "a "game changer" and opens new horizons for precision flavor studies at all future experiments."*

<https://arxiv.org/abs/2310.03440>

<https://arxiv.org/abs/2309.13231>

Impact on Physics

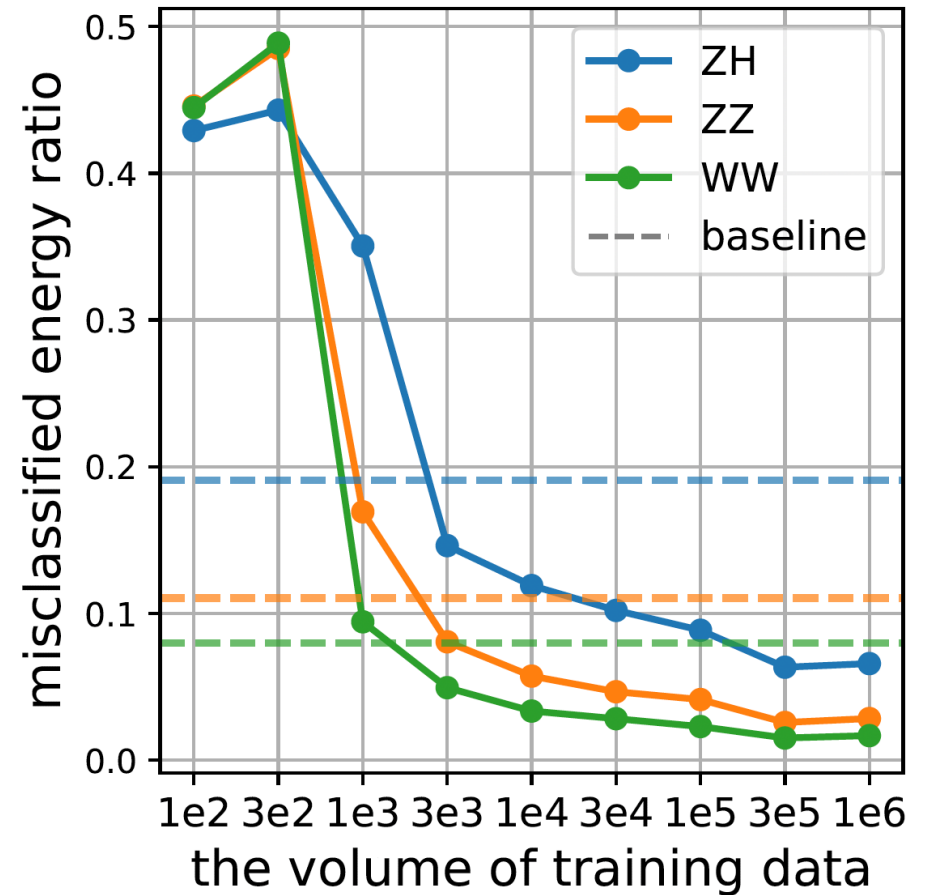
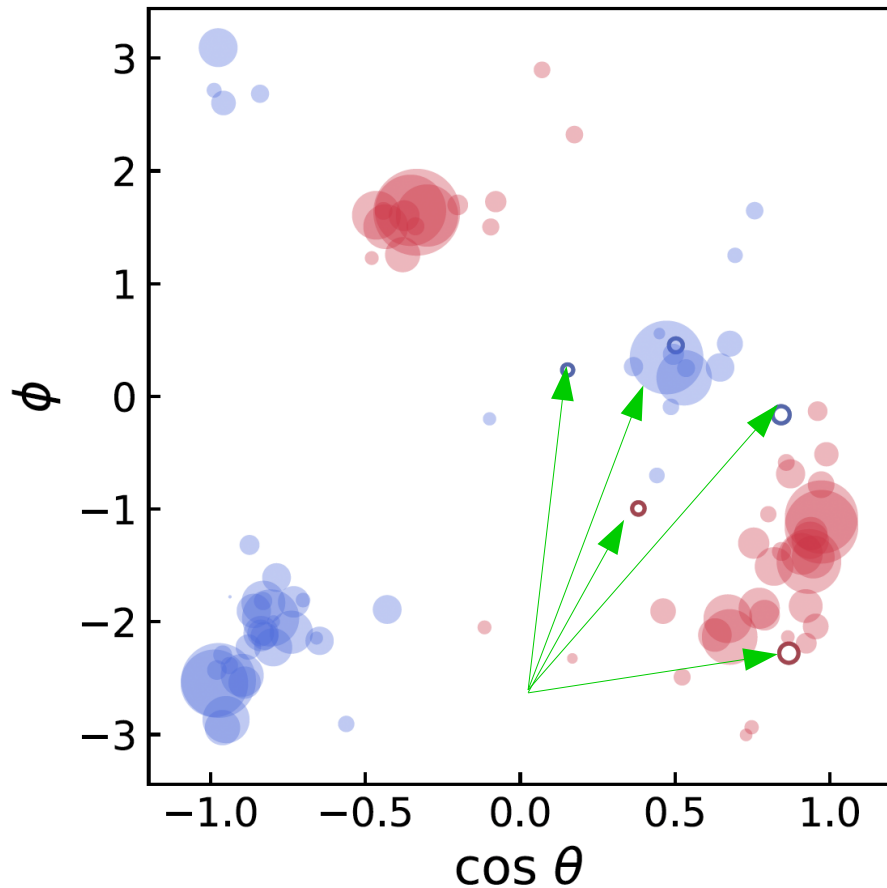


Improved by ~3 times

Improved by 1-2 orders of magnitudes
~ 2 folds improvement at $H \rightarrow cc$ & V_{cb}

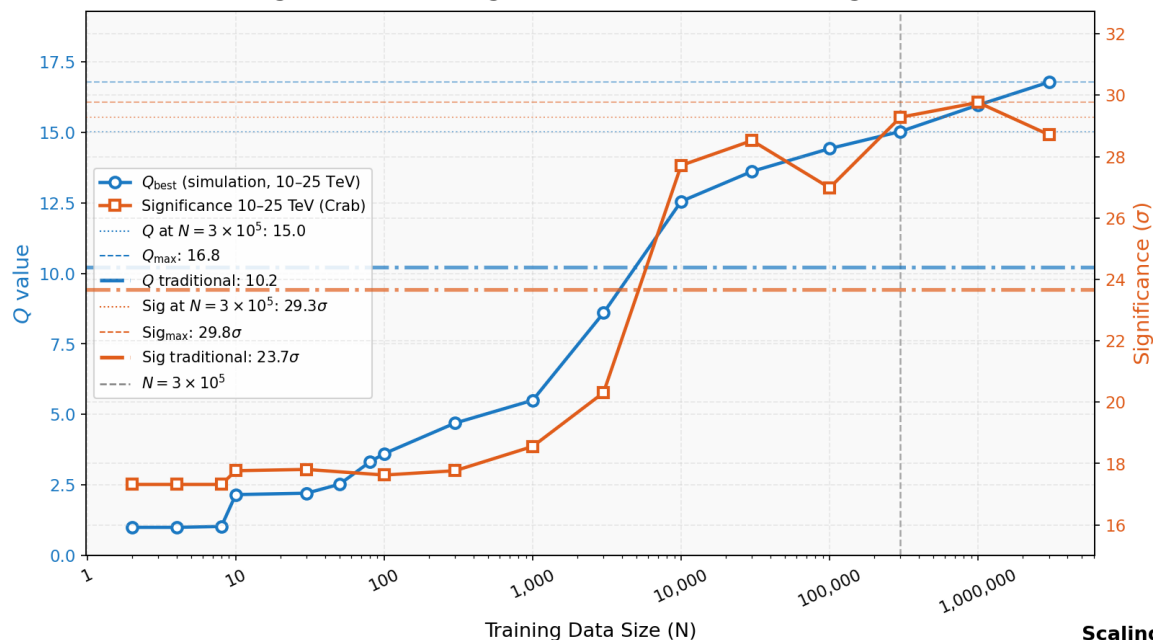
Presumably... firstly quantified

Grouping: CSI

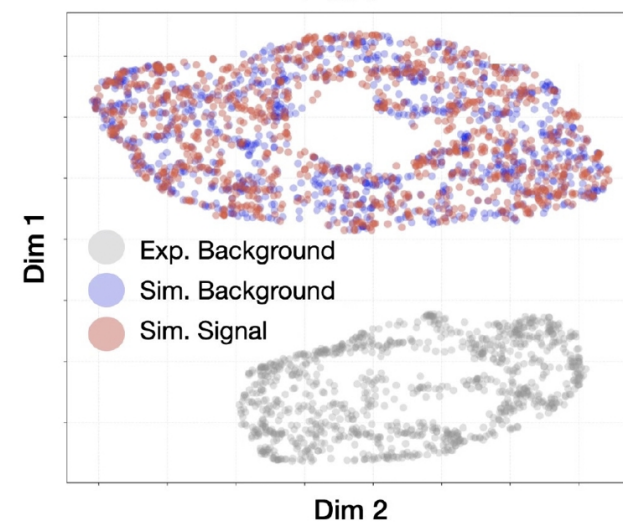


Reduced by 3-5 folds compared to conventional (jet clustering + matching)

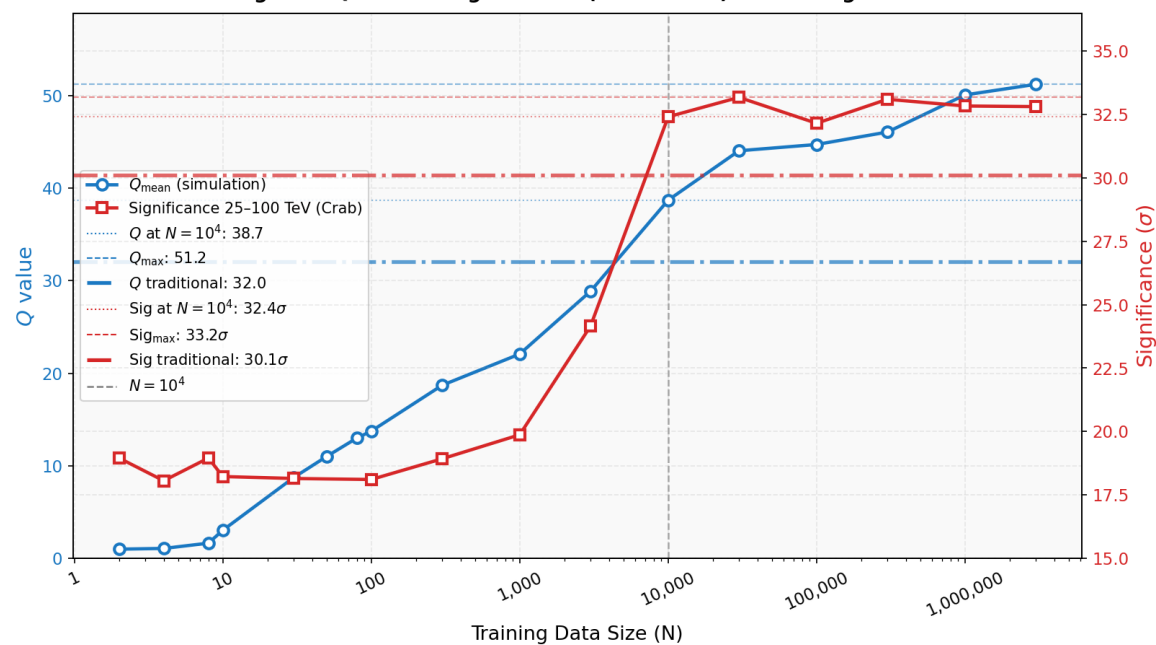
Scaling Law: Q value & Significance (10-25 TeV) vs Training Data Size



PCA



Scaling Law: Q value & Significance (25-100 TeV) vs Training Data Size

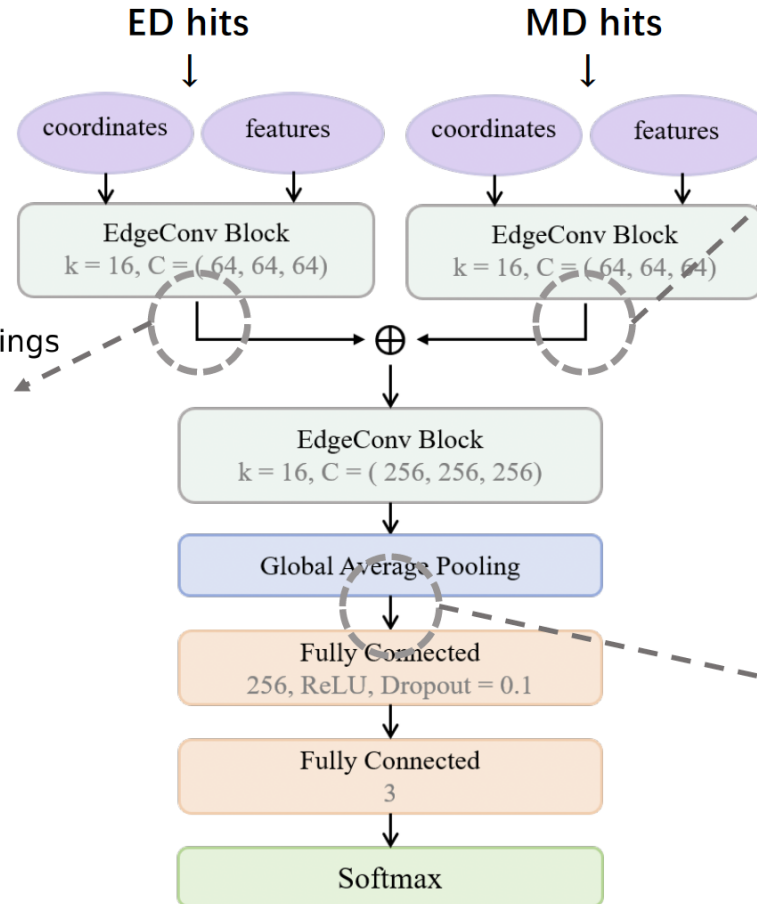
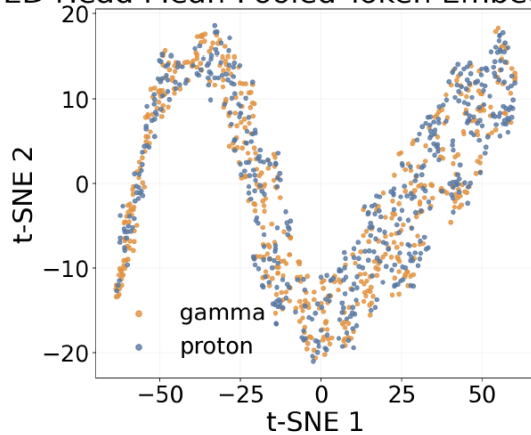


Preliminary...

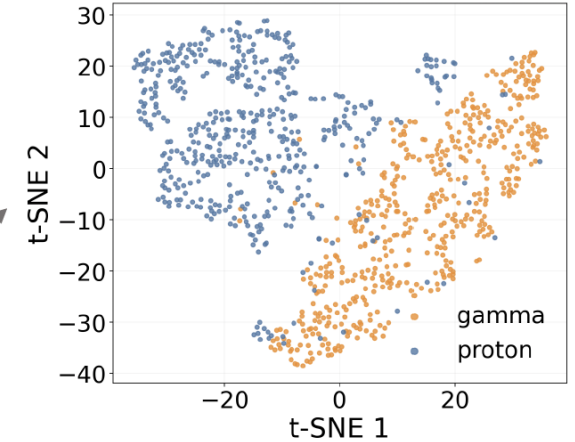
Feature Visualization

- Dimensionality-reduced visualization of high-level features extracted from different model depths.

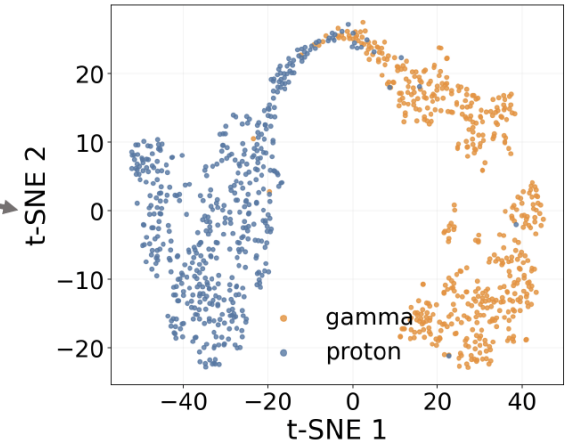
ED Head Mean-Pooled Token Embeddings



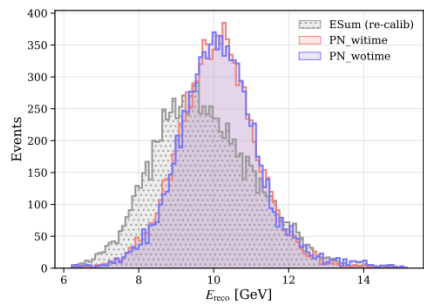
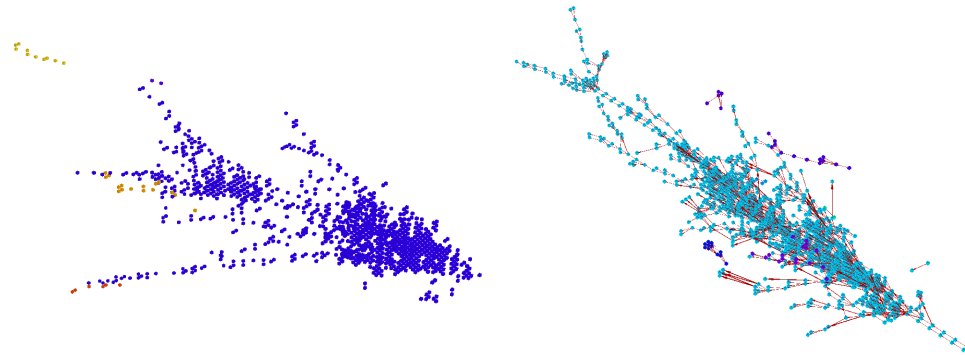
MD Head Mean-Pooled Token Embeddings



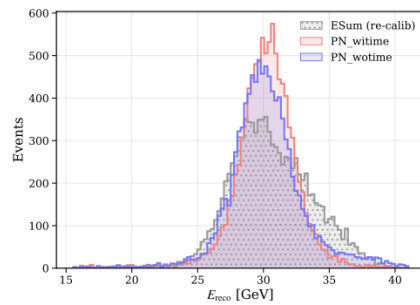
Fused Pre-Classification Embeddings



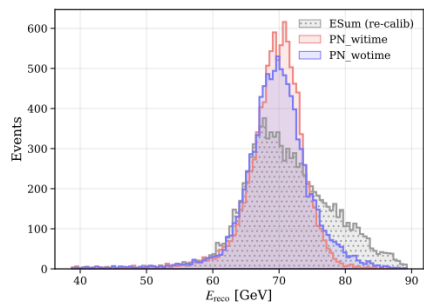
Regression: hadron energy measurement



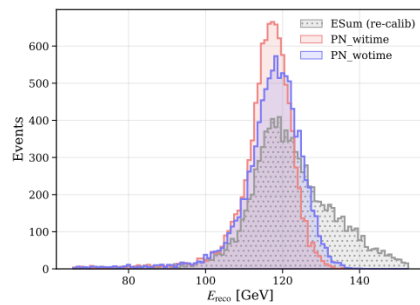
(a)



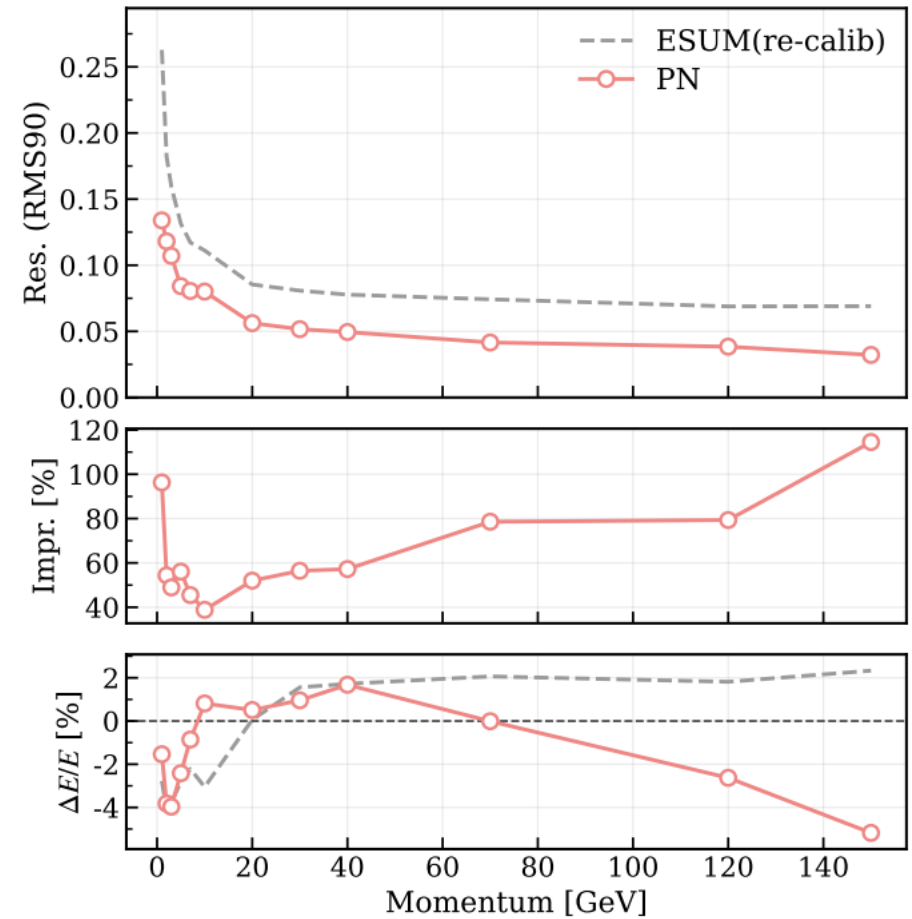
(b)



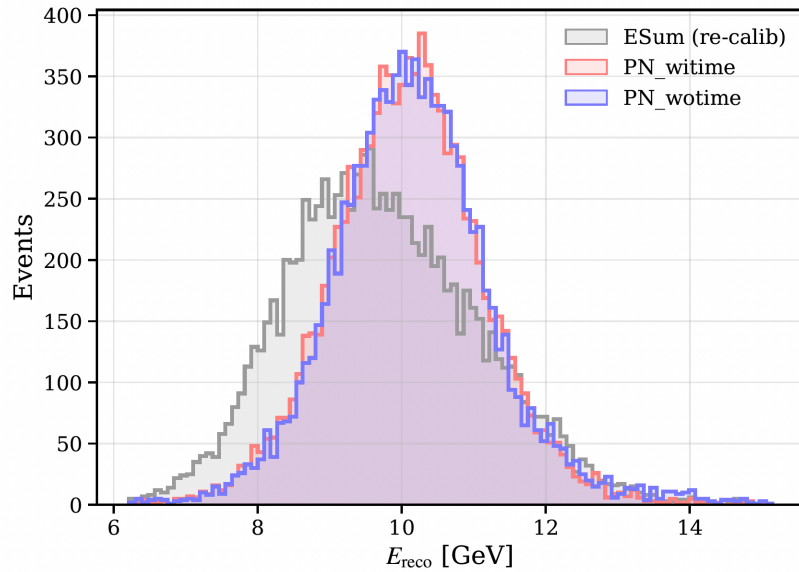
(c)



(d)



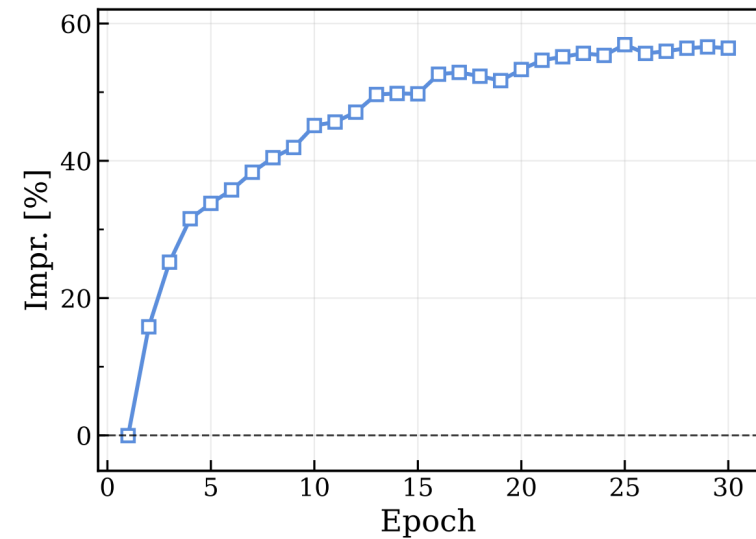
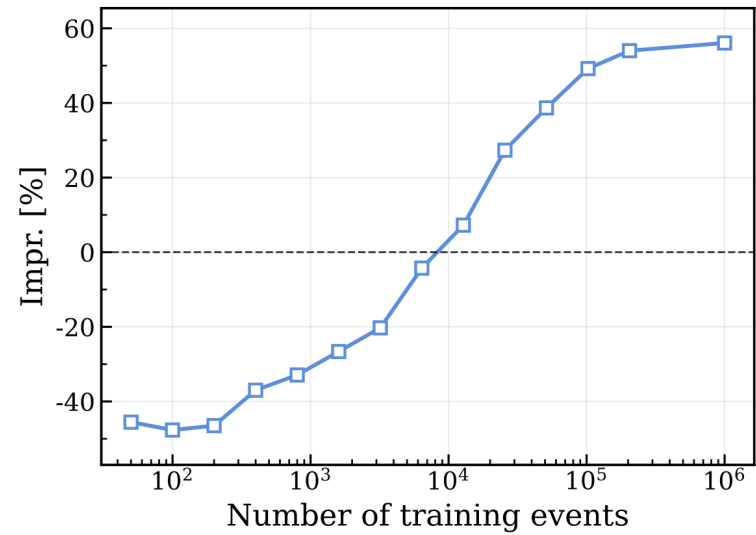
@ Regression



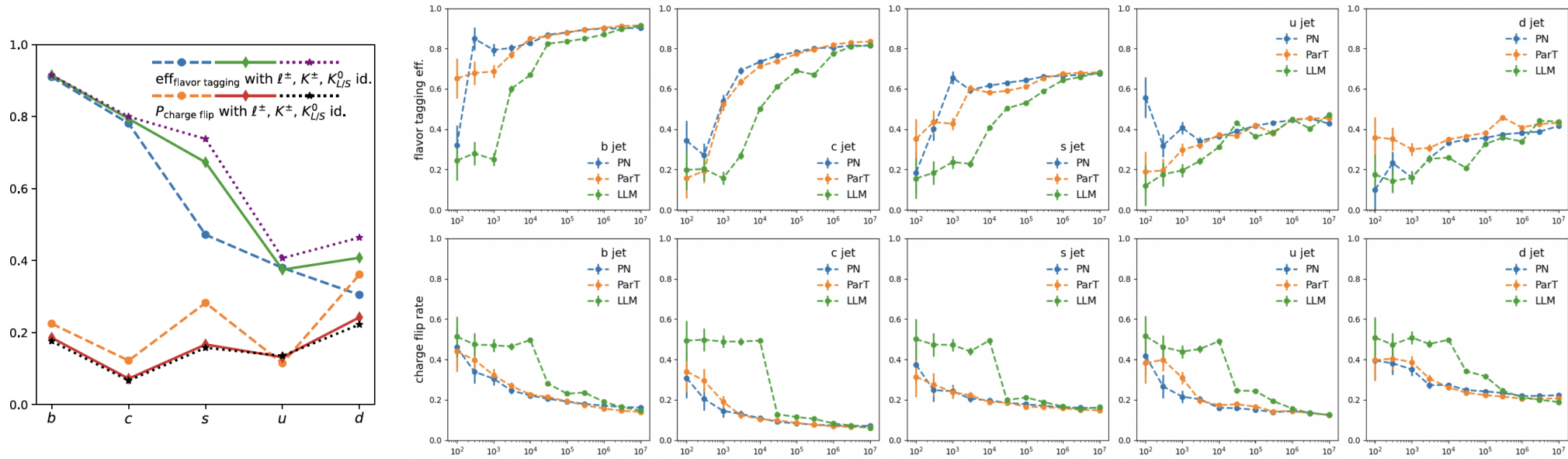
10 GeV Pion

~

7% relative energy resolution



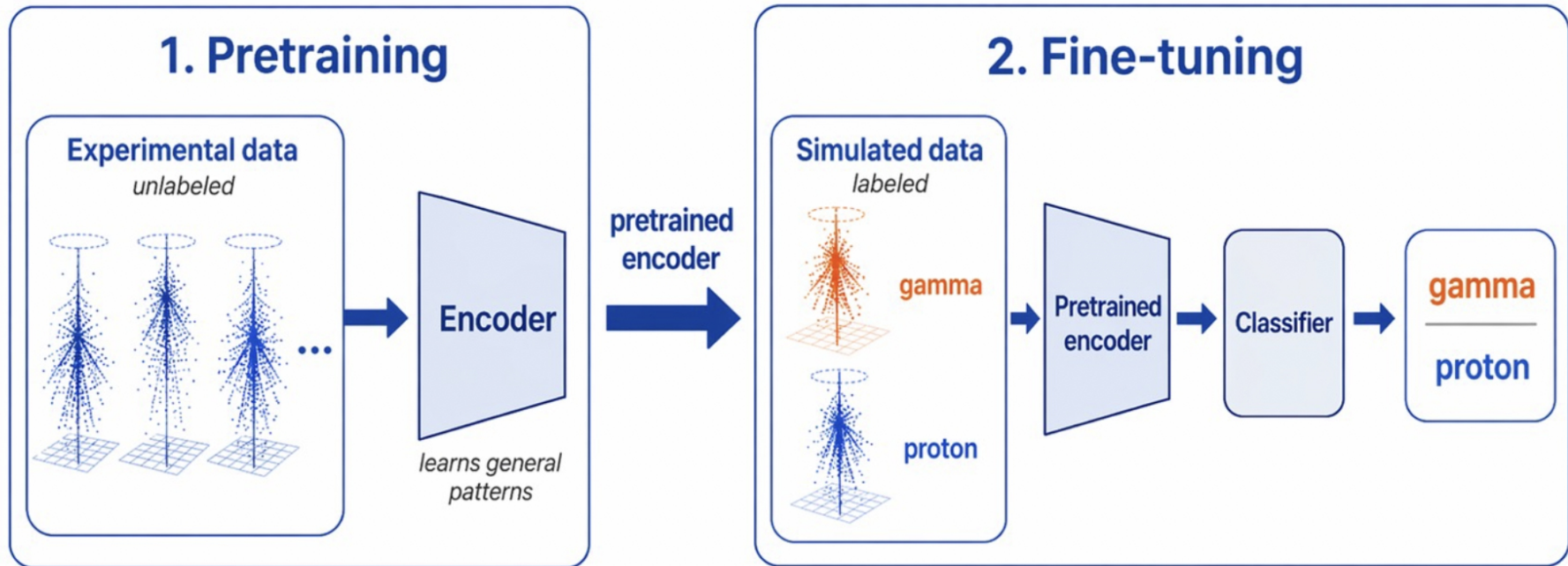
@ Jol with different Architectures...



- Comparable result with different scaling behavior
- Para. Numbers: PN 360k, ParT 2.4M, BINBBT(Large Language Base Model) 150 M
- More details at: <https://arxiv.org/pdf/2412.00129>

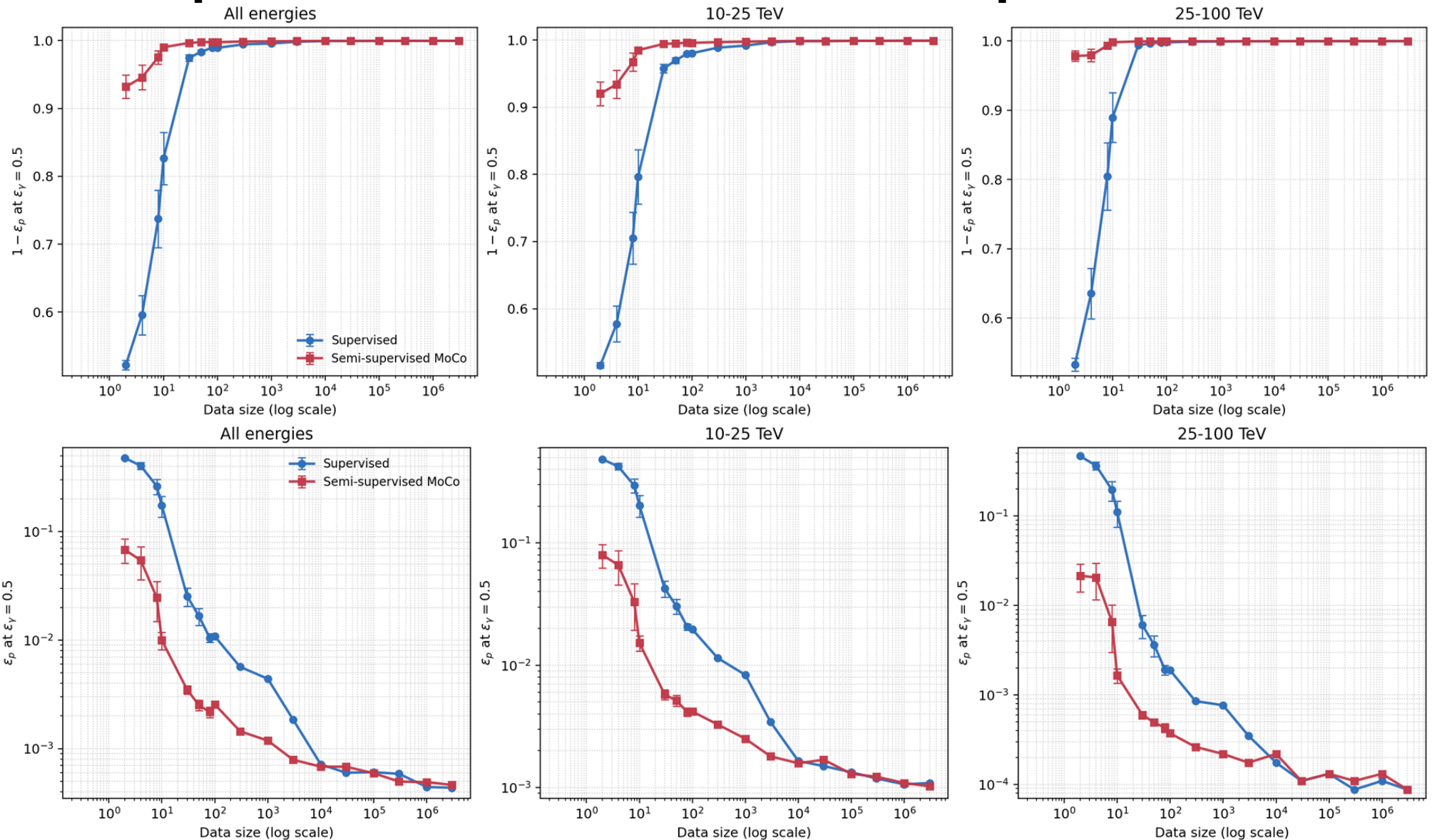
Semi-supervised Learning

*Encoder: Classical ParticleNet

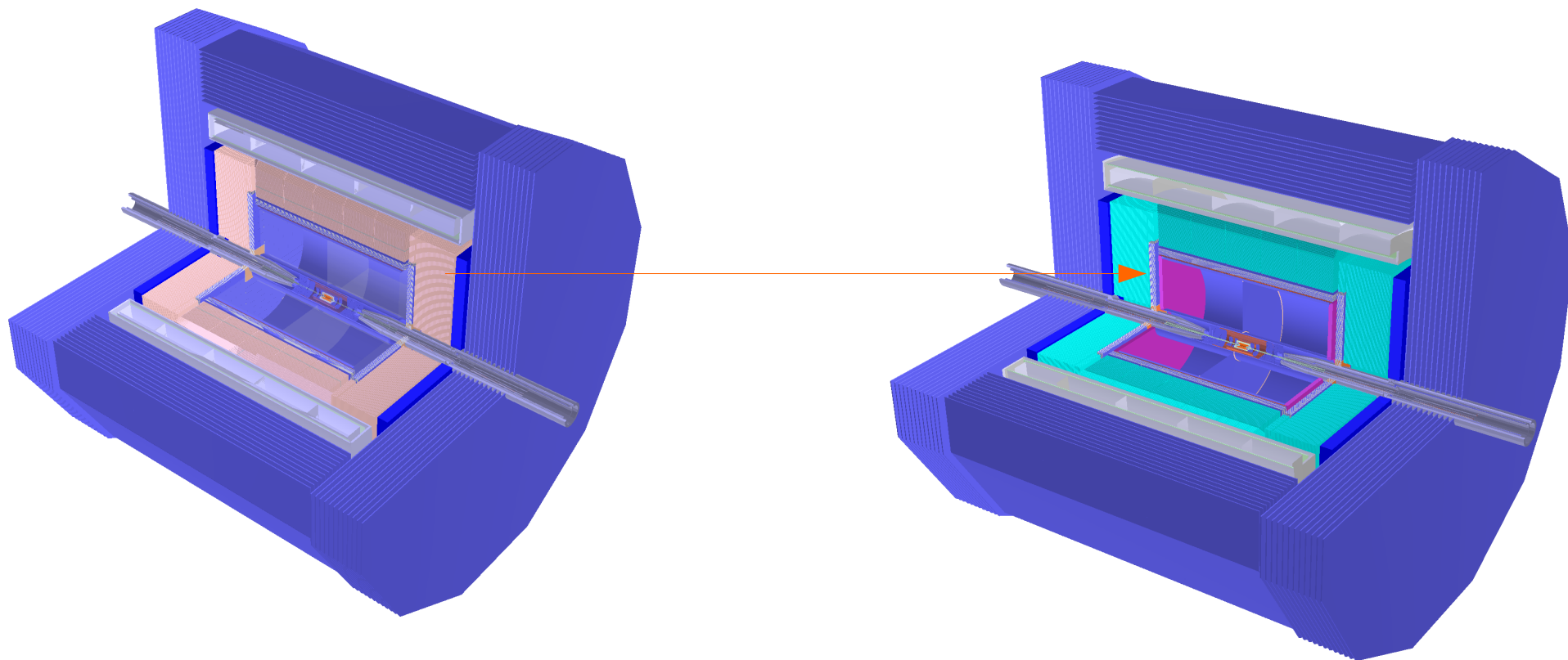


- **Abundant Experimental Data:** We possess an extensive amount of real experimental data.
- **Representation Learning:** Pre-training enables the encoder to effectively capture **robust representations** from this data.

Supervised VS. Semi-supervised...



From CDR baseline to AURORA



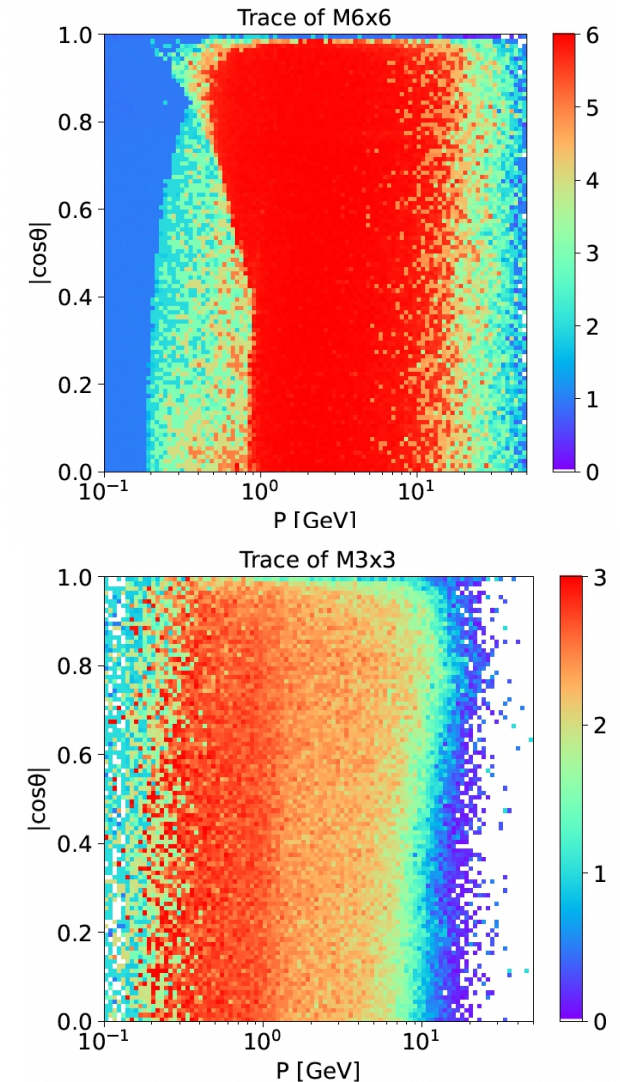
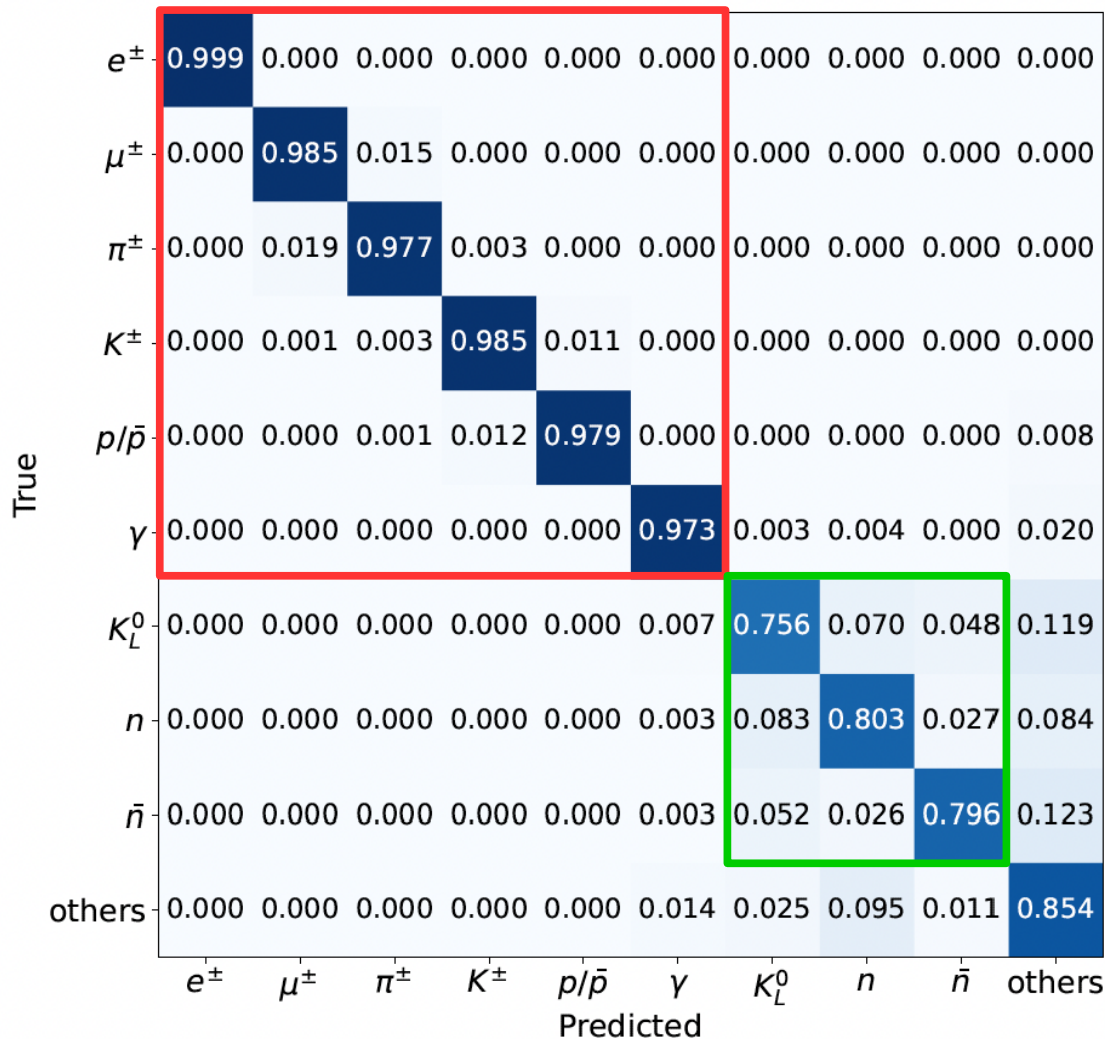
CEPC CDR baseline: TPC + SiW ECAL + RPC HCAL, BMR of 3.8%

Confusion & Acceptance & Intrinsic Resolution: 50%, 25%, 25%

Key contributions: Fragments (leading confusion) & Intrinsic HCAL resolution

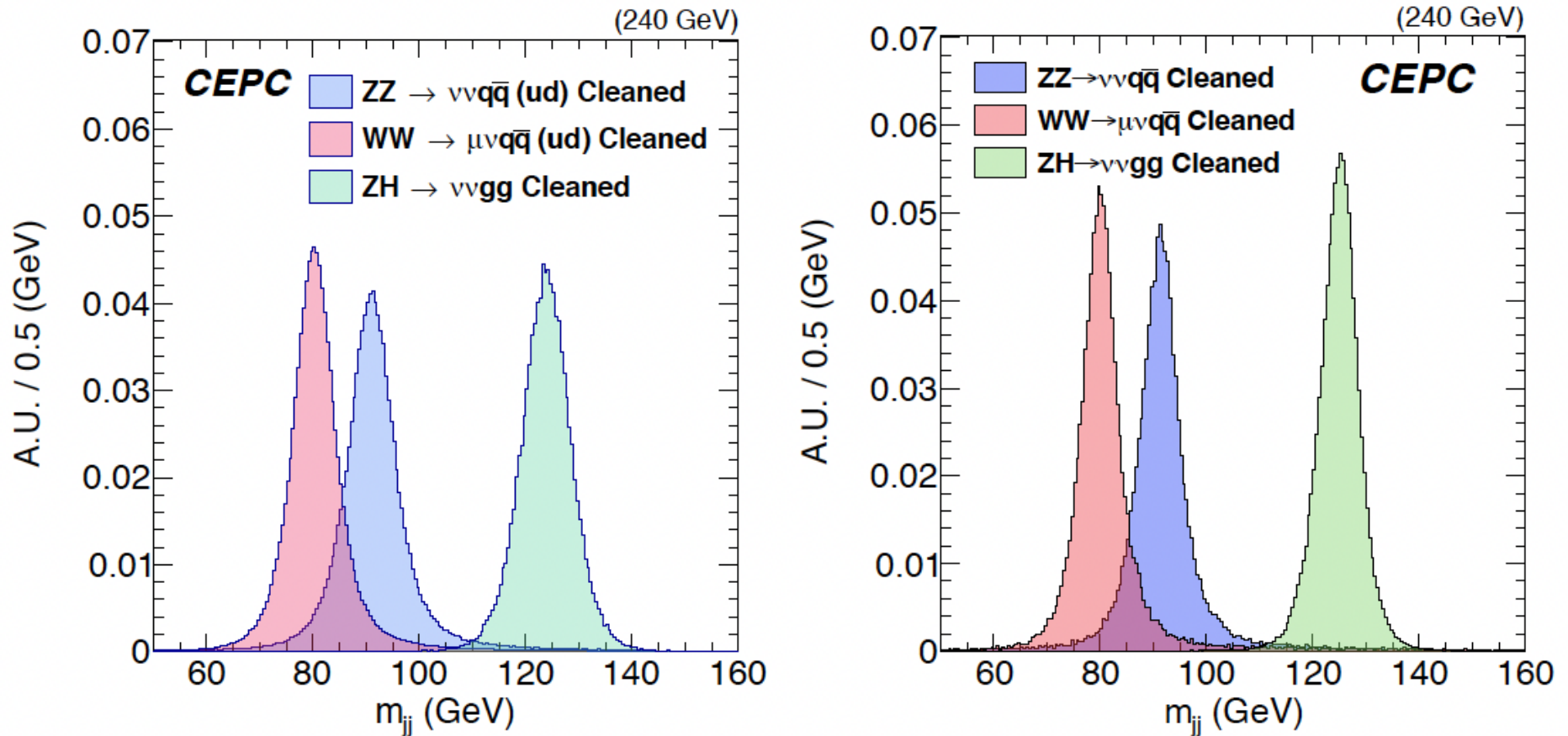
AURORA: DHFAL->GSHFAL, + 5d calo with 100 ps/Hit time resolution, BMR of 3.4%

Pid: differential performance



Neutral Hadron ID: 5d Calorimetry with ToF capability ($\delta t \sim 100$ ps/hit)

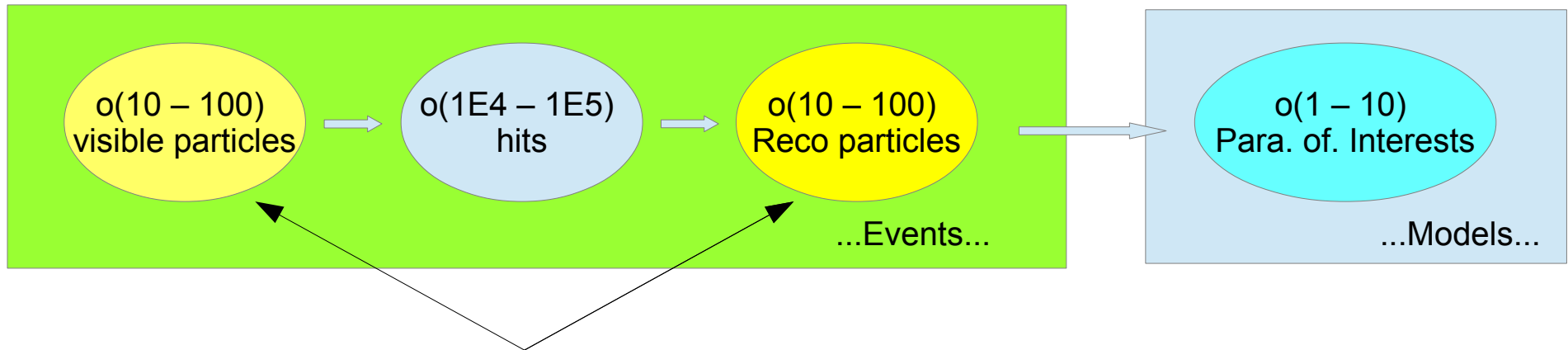
BMR of 2.7% reached



Detector change (usage of high density scintillating glass HCAL): BMR 3.7% $\rightarrow$ 3.4%;

AI enhanced reconstruction: 3.4% $\rightarrow$ 2.7%.

Learning from all particles/hits



- Ultimate goal for detector design & performance: **1-1 correspondence**
- **Holistic approach using**
 - All reco particles
 - Physics object reconstruction
 - Identifying the nature of physics events: Signal & Backgrounds
 - From leaves to the tree
 - All detector hits
 - Energy estimation for hadron
 - Pid at LHAASO

Impact on physics benchmarks...

BMR $\sim 4\%$, essential
For $H \rightarrow \text{inv}$, tautau, ...
(measurement with
 qqH , $H \rightarrow \text{non hadronic}$)

As well as Higgs width

