HitPF

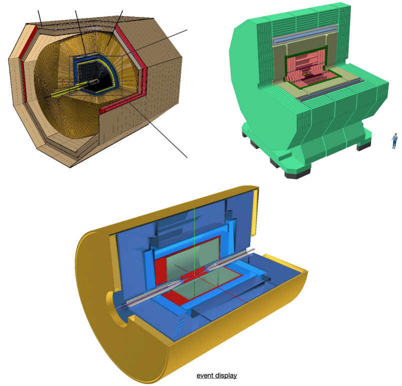
Dolores Garcia, Bohdan Durdar, Lena Herrmann, Gregor Krzmanc, Michele Selvaggi

Motivation

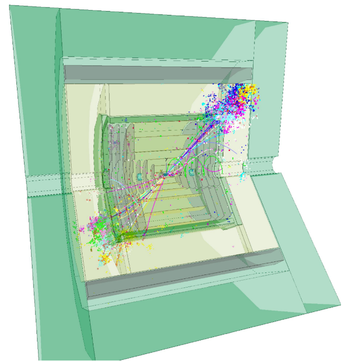
- FCC-ee has been recommended by the European Strategy as CERN's next flagship collider [1]
- FCC-ee will collide e+e- at Z, WW, ZH and $t\bar{t}$ thresholds with large luminosity
- Experiments aim to measure Higgs couplings, electroweak parameters, flavour observables with extemely high precision

Precisions depends on:

High performance detectors

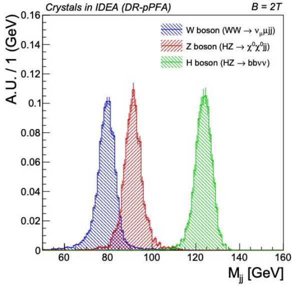


High quality reconstruction



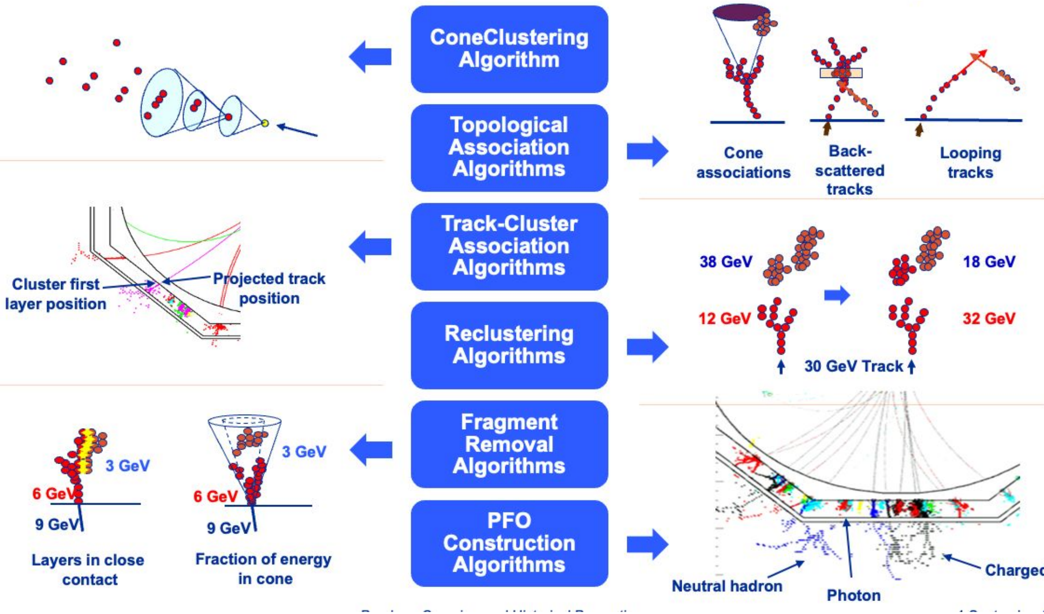
Best resolution translates into best physics sensitivity

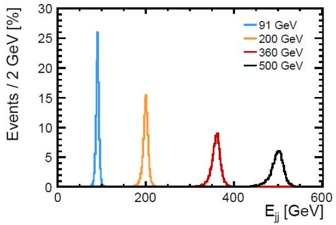
- Modern jet flavour tagging makes use of full jet per particle level data
- W/Z hadronic mass separation for H--> WW, ZZ
- Rare H→ $s\bar{s}$, H→ $c\bar{c}$ relative uncertainty scales with mass resolution



How is it tackled?

- Classically, multiple algorithms:
    - Tracking (one per sub-detector), clustering (ECAL/HCAL), and linking tracks and clusters
    - ALEPH, DELPHI, CMS, ATLAS, CLIC,…
- Pandora SDK (baseline FCC)
    - Series of distributed algorithms
    - Iterative clustering
    - Track cluster association and reclustering
    - Neutral identification
    - Relies on tuning, limiting flexibility during detector design phase (allows for geometry studies e.g transverse calo segmentation)

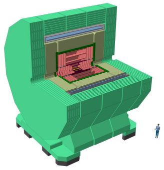




Talk by Mark Thomson at the FCC Particle Flow Workshop (Sept 2025)

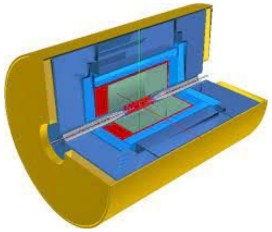
The need for an end-to-end approach: **Multiple detectors**

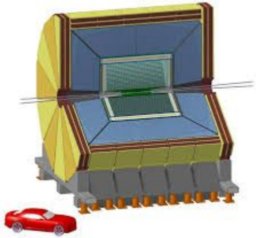
- Adaptable reconstruction is crucial for a systematic **design optimization**:
  - Parameter tuning is not straight forward; the detector will still change + different detectors
  - Costly conventional implementation



CLD
- Silicon pixel + silicon tracker
- ECAL Granular Silicon-Tungsten 5x5 mm^2 cells (15-17 %)
- Scintillator-steel, 44 layers, 30x30 mm^2
- Solenoid outside calo

IDEA
- Silicon pixel + drift chamber + silicon wrapper
- Dual-readout crystal
- Dual-readout fiber scintillating and cherenkov
- Solenoid between ECAL/HCAL





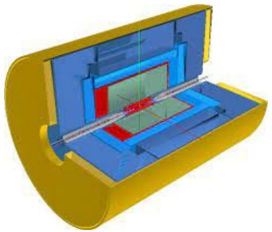
ALLEGRO
- Silicon pixel + drift chamber + silicon wrapper
- ECAL granular noble-liquid sampling LAr (8-10 %)
- Scintillating-tile HCAL
- Solenoid behind ECAL

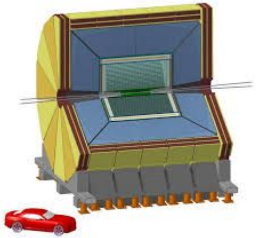
And others… ALFA, ILD

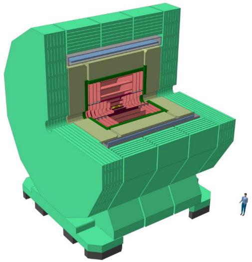
The need for an end-to-end approach: **Multiple detectors**

- ❑ The detector determines how the showers ‘look’:
  - ❑ What is the shape of an electron
    - o How wide (Molière radius)
    - o How long (longitudinal shower profiles)
  - ❑ How do hadrons shower in the HCAL, what is the shower geometry like, what are acceptable patterns?
  - ❑ What is the ‘correct’ threshold for a track-cluster separation to link them?

All these questions can be answered from data



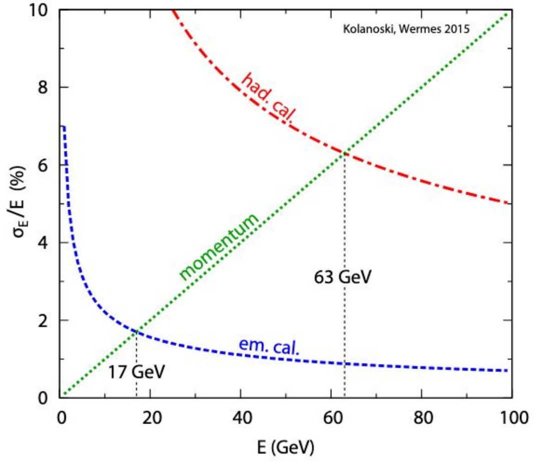




The need for an end-to-end approach

Increased performance?

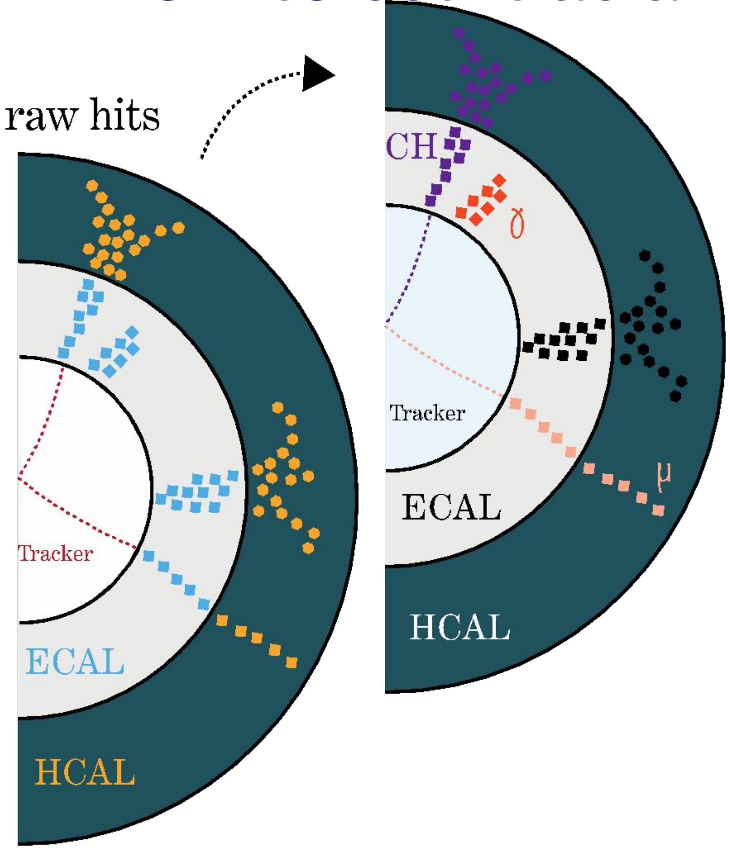
- Asymptotic resolution is 1-2% @100 GeV. Current performance 3-4% @100 GeV arXiv:0907.3577v1
- Key drivers of performance:
  - Tracking & calo cluster efficiency/purity
    - → Cluster merging: if two clusters are merged, with a track we lose resolution because we use the total calorimeter energy
    - → Shower splitting: double counting energy
  - Track cluster matching
  - Arbitration of track vs cluster energy



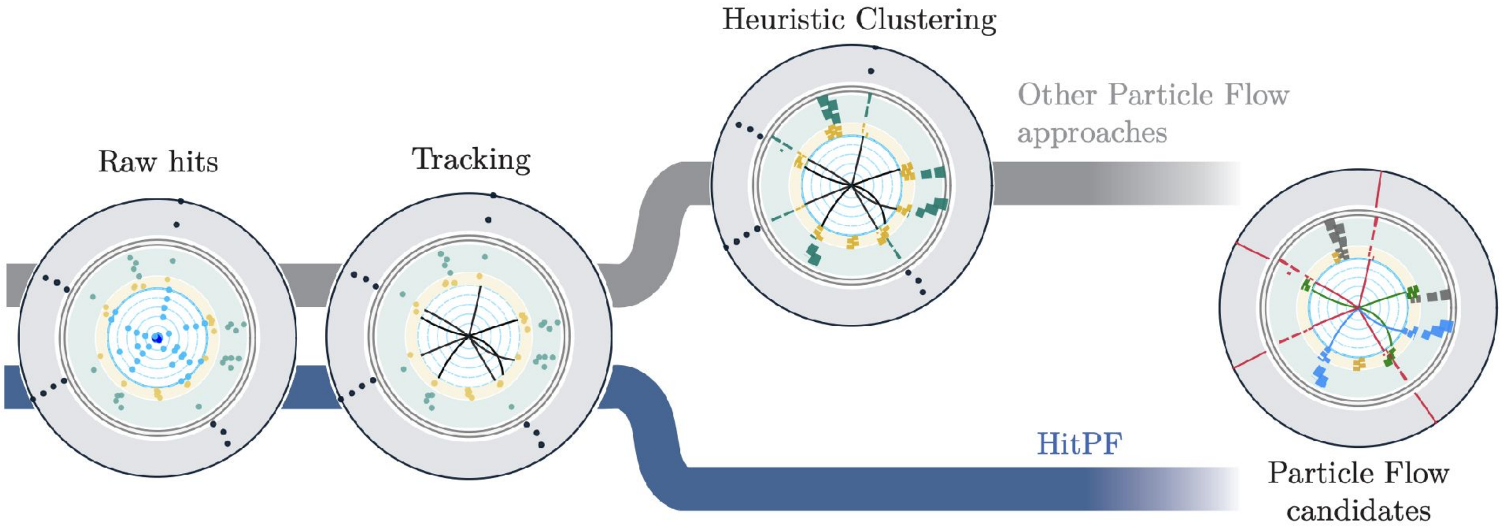
Average Jet Composition		Superior System
• 60% charged particles	→	tracker
• 30% photons	→	ECAL
• 10% neutral particles	→	HCAL

$$\sigma^2_{jet}(E_{vis}) = \sum_{i\in tr}\sigma^2_{tr}\left(E^{(i)}_{tr}\right) + \sum_{i\in\gamma}\sigma^2_{ECAL}\left(E^{(i)}_{ECAL}\right) + \sum_{i\in h^0}\sigma^2_{HCAL}\left(E^{(i)}_{h^0}\right)$$

How to set it as an ML problem

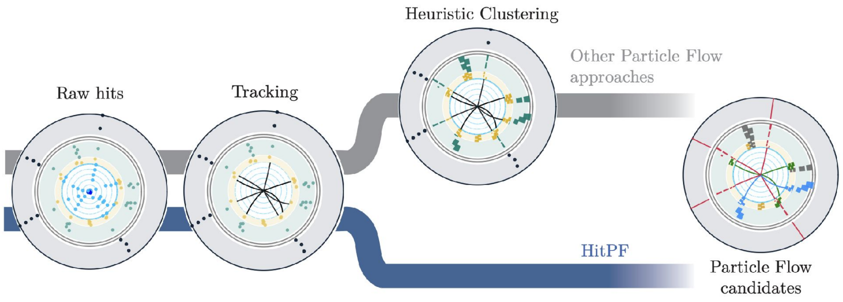


Hit-based approach to Particle Flow


Heuristic Clustering
Other Particle Flow approaches
Raw hits
Tracking
HitPF
Particle Flow candidates

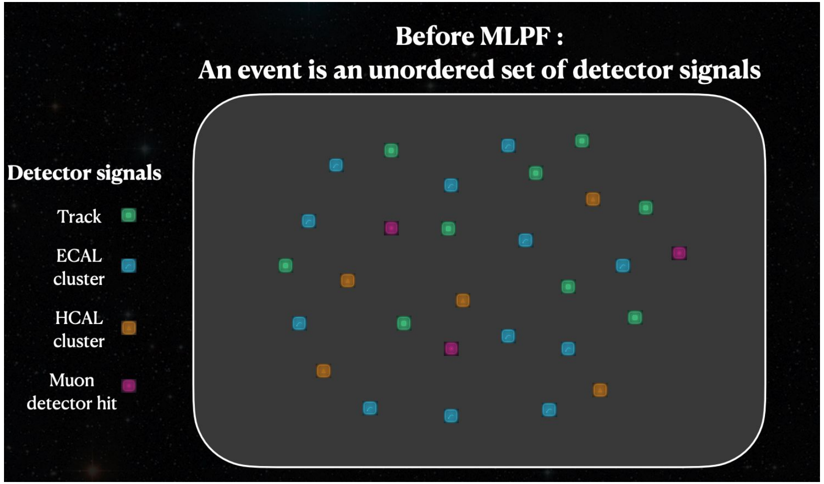
Comparison with other methods

Other PF approaches	**HitPF**
❑ Require heuristic clustering step + cluster track matching	❑ ML-based Clustering + track matching
❑ Complex parameter tuning for new detector	❑ Out of the box with multiple detector and hit geometries
❑ Requires geometry navigation	❑ Does not require the explicit detector geometry (navigation)

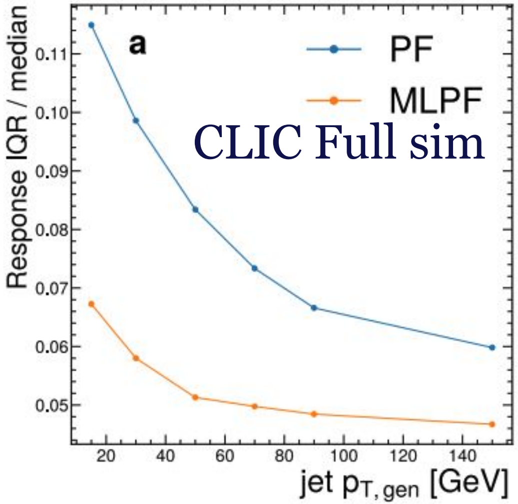


MLPF

- Cluster based approach:
  - Starts from pandora clusters and reconstructed tracks
  - In practice, very different problem
  - Improves the track-cluster linking and properties regression
  - 1 node in MLPF ~ 600 nodes in HitPF
  - Does not solve the geometry adaptation problem
  - New results for CMS

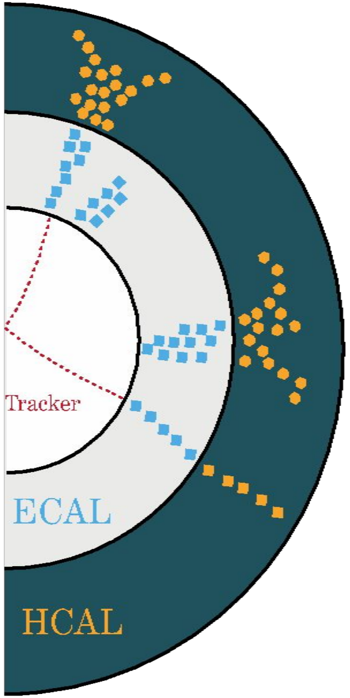


Farouk.M @ CERN Data Science Seminar



The workflow: Inputs

raw hits



Track 'hits:
Reference point in calorimeter (position)
p at vertex and calo
χ from track fit
Sub-detector ID

ECAL hit
Position (x,y,z)
E energy
Sub-detector ID

HCAL hit
Position (x,y,z)
E energy
Sub-detector ID

Muon hit
Position (x,y,z)
E energy
Sub-detector ID

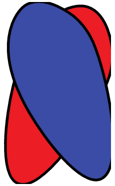
Other sub-detectors?
eg. ARC

pointcloud

Adapting ML methods to PF

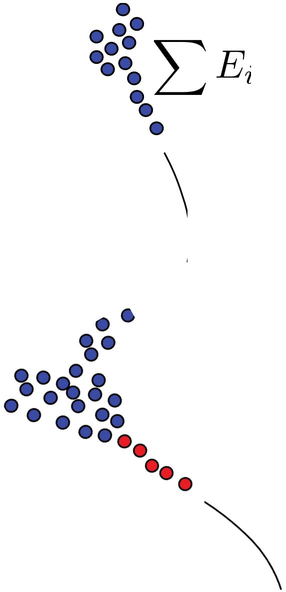
Inductive bias

Self- intersections
No equivalent in 3D ML dataset

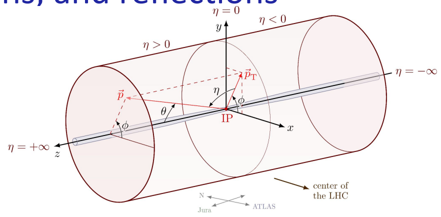


Multiple measurements in different subdetectors that 'add up' --> Optimal transport
No equivalent in 3D ML datasets (LiDAR scans of chairs do not add up to their mass)

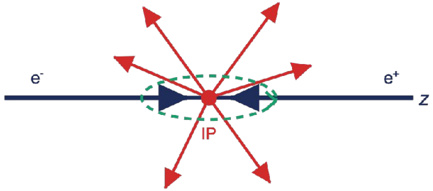
Need to preserve 'fine' structures in ECAL/tracker
Prevents voxelization and standard pulling



SO(2) + z mirror (endcaps)
~E(3) rotations, translations, and reflections



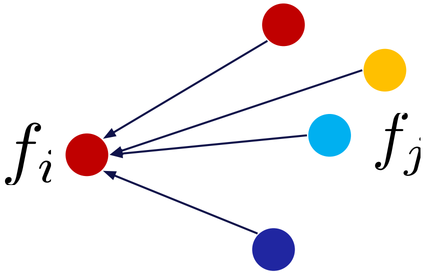
Beam direction
Preferred direction set by gravity



ML 3D data

Detector data

Model GNNs, transformers



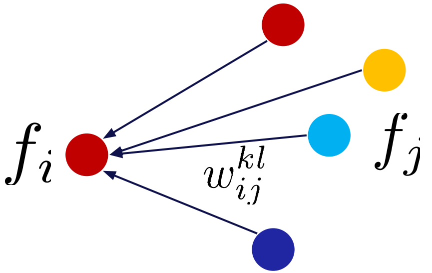
$$f_i' = \square_{j:(i,j)\in\mathcal{E}} h_{\Theta}(\mathbf{f}_i, \mathbf{f}_j).$$

Building block 1

$\square_{j:(i,j)\in\mathcal{E}}$ Aggregation operator

h_{Θ} Non-linear function with learnable parameters

Model GNNs, transformers



$$f_i' = \square_{j:(i,j)\in\mathcal{E}} h_\Theta(\mathbf{f}_i, \mathbf{f}_j).$$

Aggregation operator

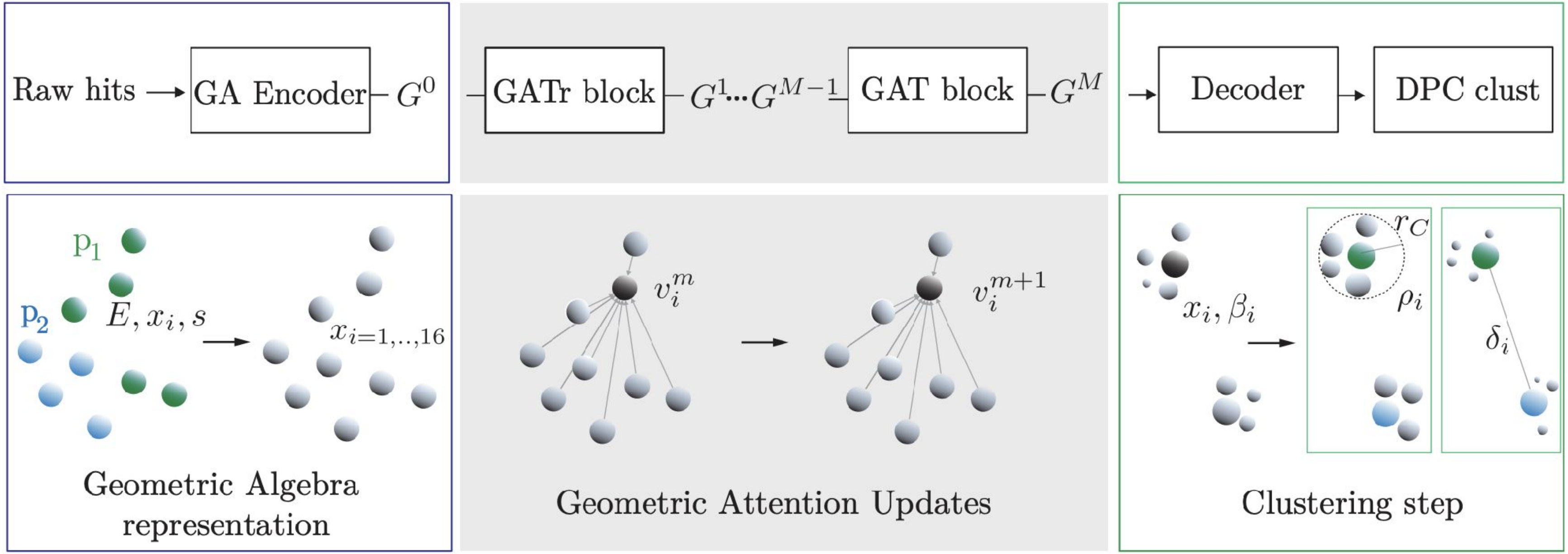
$$\square_{j:(i,j)\in\mathcal{E}} \longrightarrow \quad f_i' = \sum_{j:(i,j)\in\mathcal{E}}$$

Non-linear function with learnable parameters

$$h_\Theta \longrightarrow \quad h_\Theta(\mathbf{f}_i^l, \mathbf{f}_j^l) = w_{ij}^{kl}\, \mathbf{V}^{kl}\mathbf{f}_j^l,$$

$$w_{ij}^{kl} = \mathrm{softmax}_j\left(\frac{(\mathbf{Q}_q^{kl}\mathbf{f}_i^l)^\top(\mathbf{K}_k\mathbf{f}_j^l)}{\sqrt{d}}\right)$$

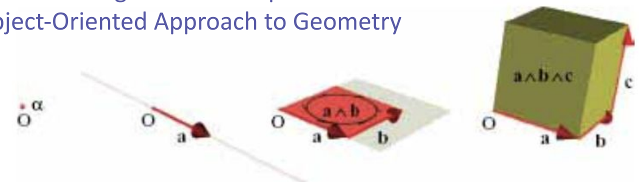
Geometric Algebra Transformer

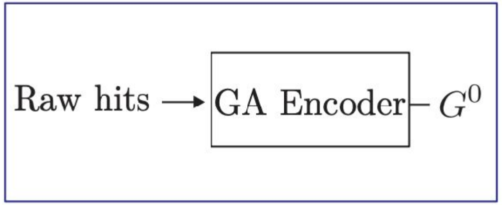


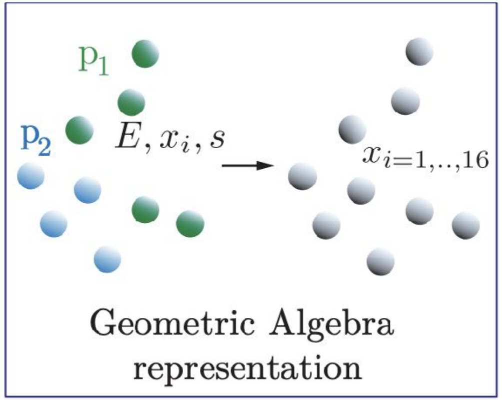
GATr [1]

Multivectors

Geometric Algebra for Computer Science : An Object-Oriented Approach to Geometry



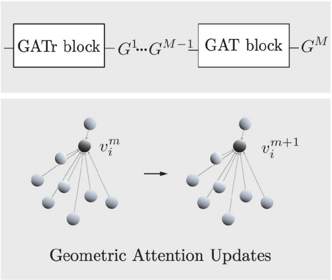


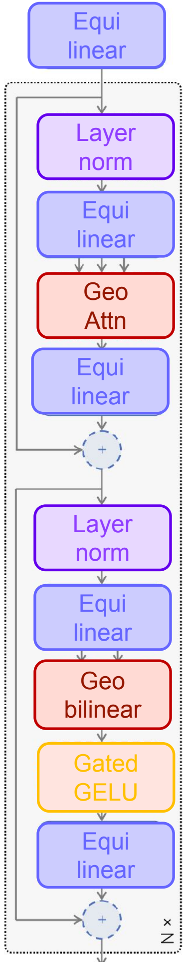


- Representation of multiple geometric objects of different grades
- Outer product a ∧ b (oriented span of a and b)
- Projective algebra (adds displacement from the origin)
- 3D object representation:
  - Planes are represented as vectors (normal + displacement)
  - Lines are intersection of two planes (bivectors)
  - Points are intersections of three planes (trivectors)
- Multivectors:
  - $x = x_s + x_1e_1 + x_2e_2 + x_3e_3 + x_{12}e_1 e_2 + x_{13}e_1 e_3 + x_{23}e_2 e_3 + x_{123}e_1e_2e_3$, x_i real coefficients
  - Built with the geometric product (metric)

Object / operator	Scalar 1	Vector e_0	e_i	Bivector e_{0i}	e_{ij}	Trivector e_{0ij}	e_{123}	PS e_{0123}
Scalar $\lambda \in \mathbb{R}$	λ	0	0	0	0	0	0	0
Plane w/ normal $n \in \mathbb{R}^3$, origin shift $d \in \mathbb{R}$	0	d	n	0	0	0	0	0
Line w/ direction $n \in \mathbb{R}^3$, orthogonal shift $s \in \mathbb{R}^3$	0	0	0	s	n	0	0	0
Point $p \in \mathbb{R}^3$	0	0	0	0	0	p	1	0
Pseudoscalar $\mu \in \mathbb{R}$	0	0	0	0	0	0	0	μ
Reflection through plane w/ normal $n \in \mathbb{R}^3$, origin shift $d \in \mathbb{R}$	0	d	n	0	0	0	0	0
Translation $t \in \mathbb{R}^3$	1	0	0	$\frac{1}{2}t$	0	0	0	0
Rotation expressed as quaternion $q \in \mathbb{R}^4$	q_0	0	0	0	q_i	0	0	0
Point reflection through $p \in \mathbb{R}^3$	0	0	0	0	0	p	1	0

Geometric Algebra Transformer





Equi linear

K-blade projections, no grade mixing

$$\phi(x) = \sum_{k=0}^{d+1} w_k \langle x \rangle_k + \sum_{k=0}^{d} v_k e_0 \langle x \rangle_k$$

Geo bilinear

Geometric product x,y --> xy= ⟨x,y⟩+x∧y --> grade mixing

Join product x,y --> (x* ∧ y*)*

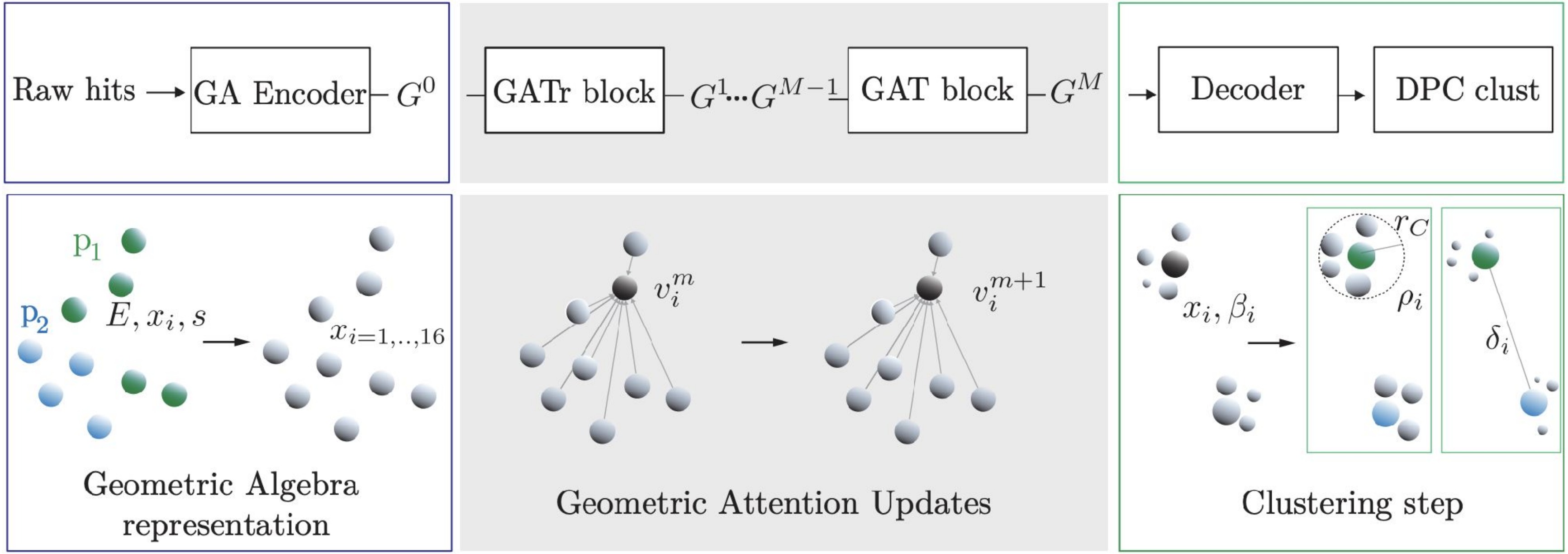
$$\mathrm{Geometric}(x, y; z) = \mathrm{Concatenate}_{\mathrm{channels}}\big(xy,\ \mathrm{EquiJoin}(x, y; z)\big)$$

Geo Attn

Attention mechanism as described previously but with inner product of the algebra

$$\mathrm{Attention}(q, k, v)_{i'c'} = \sum_i \mathrm{Softmax}_i \left(\frac{\sum_c \langle q_{i'c}, k_{ic} \rangle}{\sqrt{8 n_c}} \right) v_{ic'}$$

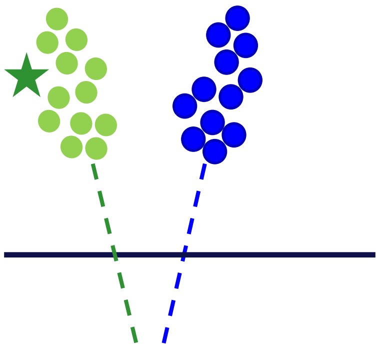
Geometric Algebra Transformer



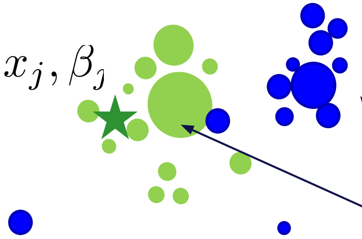
GATr [1]

Loss: object condensation

Inputs (K=2)



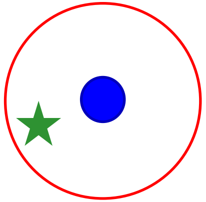
Embedding space



$$x_j, \beta_j$$

$$\check{V}_k(x) = \|x - x_\alpha\|^2 q_{\alpha k},$$

$$\hat{V}_k(x) = \max\big(0,\ \underline{1 - \|x - x_\alpha\|}\big)\, q_{\alpha k}$$

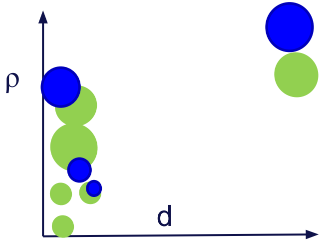


$$L_V = \frac{1}{N} \sum_{j=1}^{N} q_j \sum_{k=1}^{K} \Big(M_{jk}\, \check{V}_k(x_j) + (1 - M_{jk})\, \hat{V}_k(x_j) \Big)$$

$$q_{\alpha k} = \max_i q_i M_{ik},$$

[Object condensation @Jan K.]

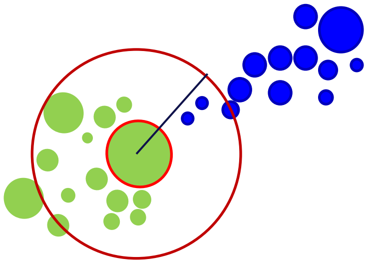
Clustering in embedding space



Two methods

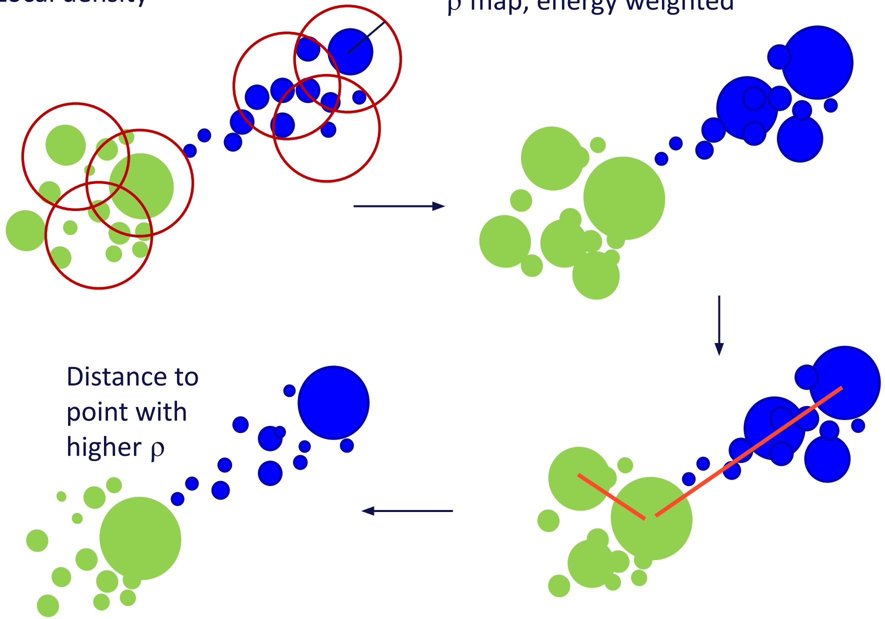
Beta based clustering [1]

x_j, β_j



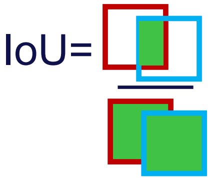
- Two parameters, β_{min}, distance threshold
- Depends on good training of β distribution

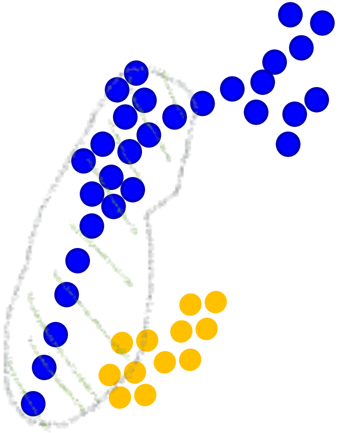
Density Peak clustering [2]



Matching

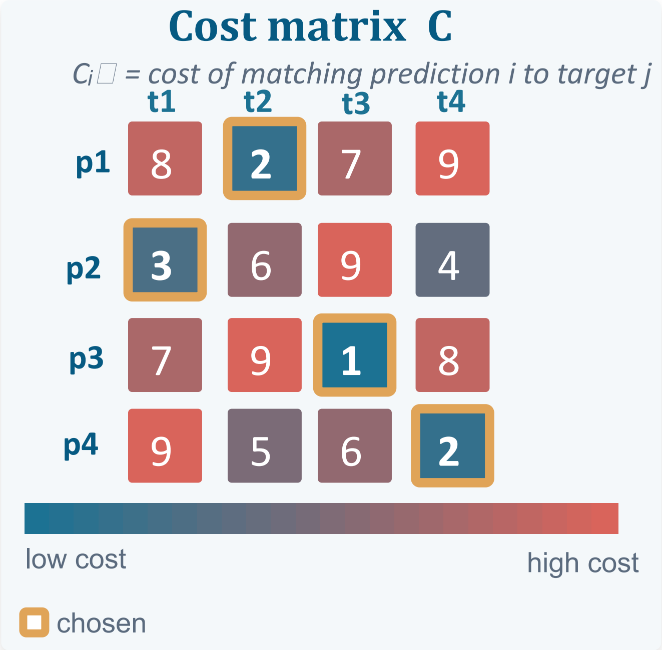
- We use a **Hungarian matching** to match a reco particle to a target particle with the intersection over union cost
- Builds a cost matrix
- One-to-one constraint
- Cost minimization





Matching

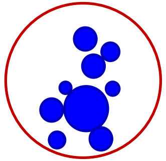
- We use a **Hungarian matching** to match a reco particle to a target particle with the intersection over union cost
- Builds a cost matrix
- One-to-one constraint
- Cost minimization



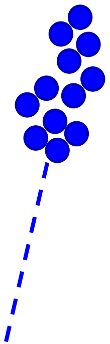
$$\min_{X\in\{0,1\}^{N\times N}} \sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}\, X_{ij} \quad \text{s.t.}$$

$$\sum_j X_{ij} = 1 \ \ \forall i, \qquad \sum_i X_{ij} = 1 \ \ \forall j.$$

Regression & Classification



Clustering
proposed PF

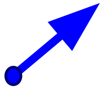


x_i, E_i, χ, ID

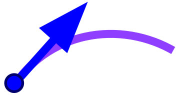
Per PF candidate property determination

$\vec{p}/|p|$

Neutral
Direction of energy-weighted average of hits



Charged
Direction from track state at IP



Charged
Energy measurement based on tracking

Energy

PID

CLUSTER EMBEDDINGS (GATr)
HIGH-LEVEL FEATURES

$E_{cluster}, \frac{E_{ECAL}}{E_{cluster}}, \frac{E_{HCAL}}{E_{cluster}}$, dispersion

# of hits

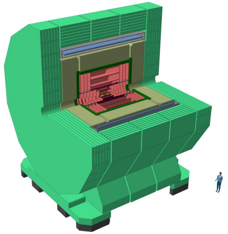
# of tracks, $p_{track}, \chi^2_{track}$

# of muon hits, E_{muon}

MLP Energy — neutral energy

MLP PID — CH, e, μ, NH, γ

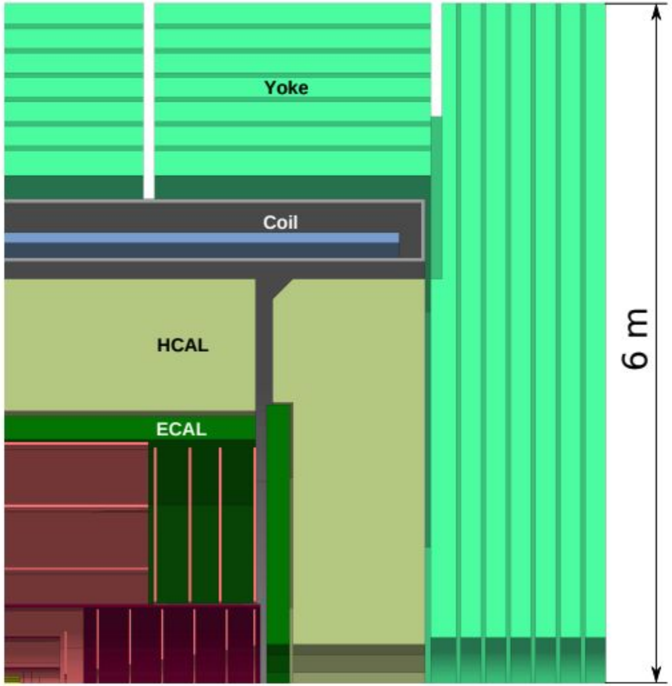
CLD



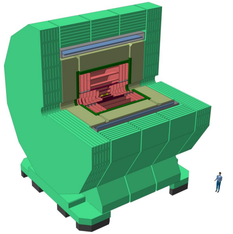
Full simulation

- ❑ Implemented in DD4hep
- ❑ Full simulation in Key4hep
- ❑ Comparison againts Pandora tunned for CLD (CLD_o2_v05)

- ❑ $Z \rightarrow q\overline{q}, q = (u, d, s)$ 91 GeV
- ❑ Jet like events:
  - ❑ Close particles
  - ❑ High particle multiplicity



ILD



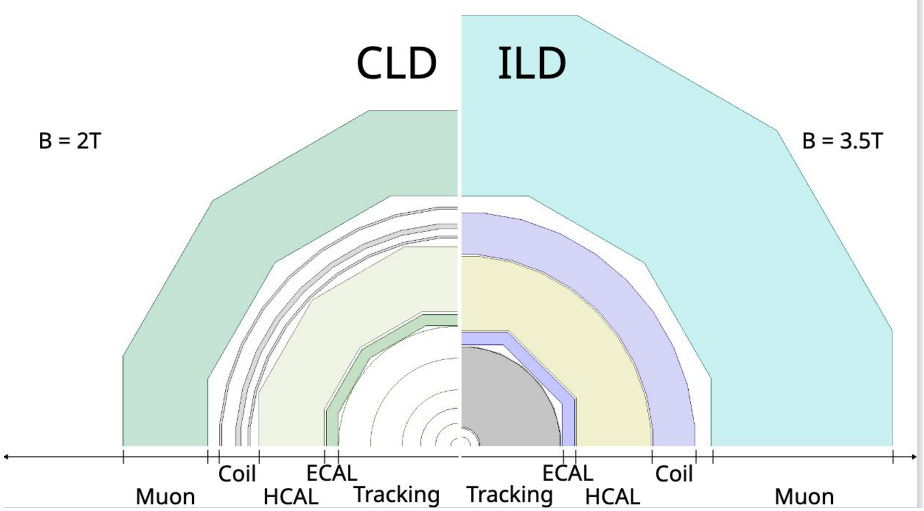
Full simulation

- Implemented in DD4hep
- Full simulation in Key4hep
- Comparison againts Pandora tunned for ILD_l5_o1_v02

- $Z \rightarrow q\overline{q}, q = (u, d, s)$ 91 GeV
- Jet like events:
  - Close particles
  - High particle multiplicity

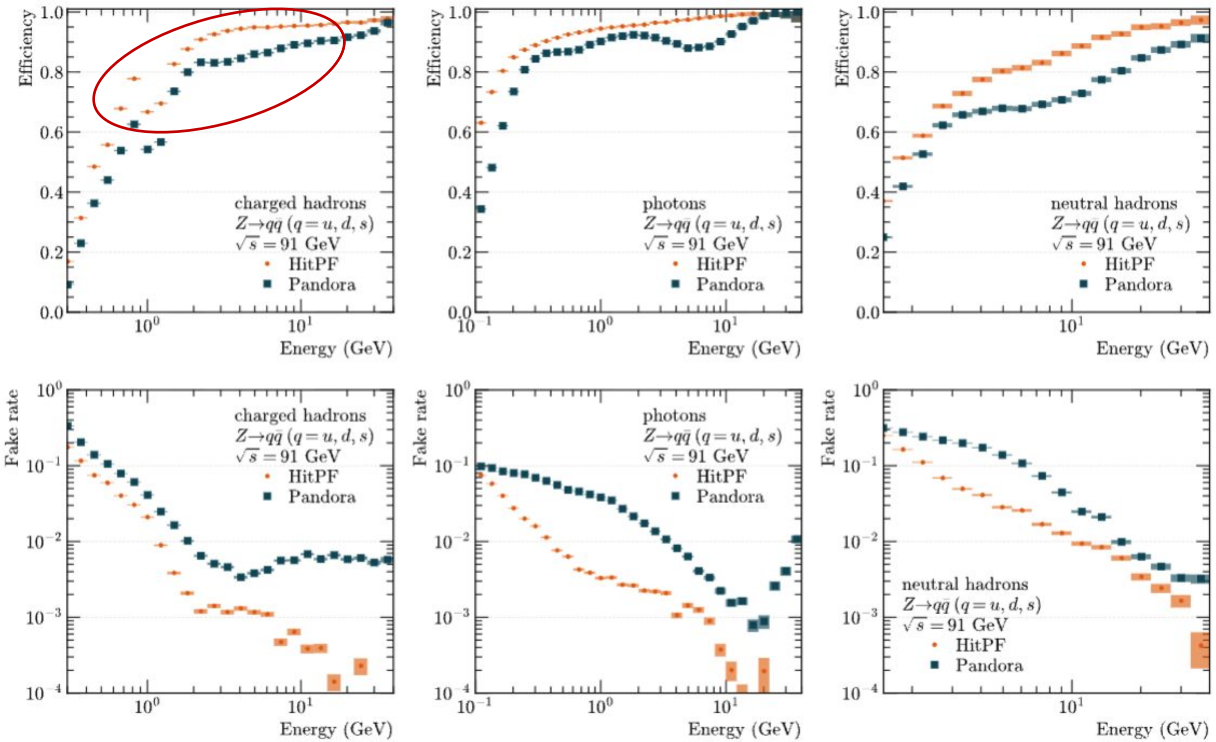
Trained on 1M events

Evaluated on 100k events



PID: Efficiency and fake rate

Charged Hadron   Photons   Neutral Hadrons



CH:

- Up to 15% higher eff.
- Fake rate $\mathcal{O}(2)$ smaller

→Eff. limited by track reco.

Photons:

- ~5% higher eff. at high E
- Improvement by up to factor 2 for low E
- Fake rate up to ~ $\mathcal{O}(1)$

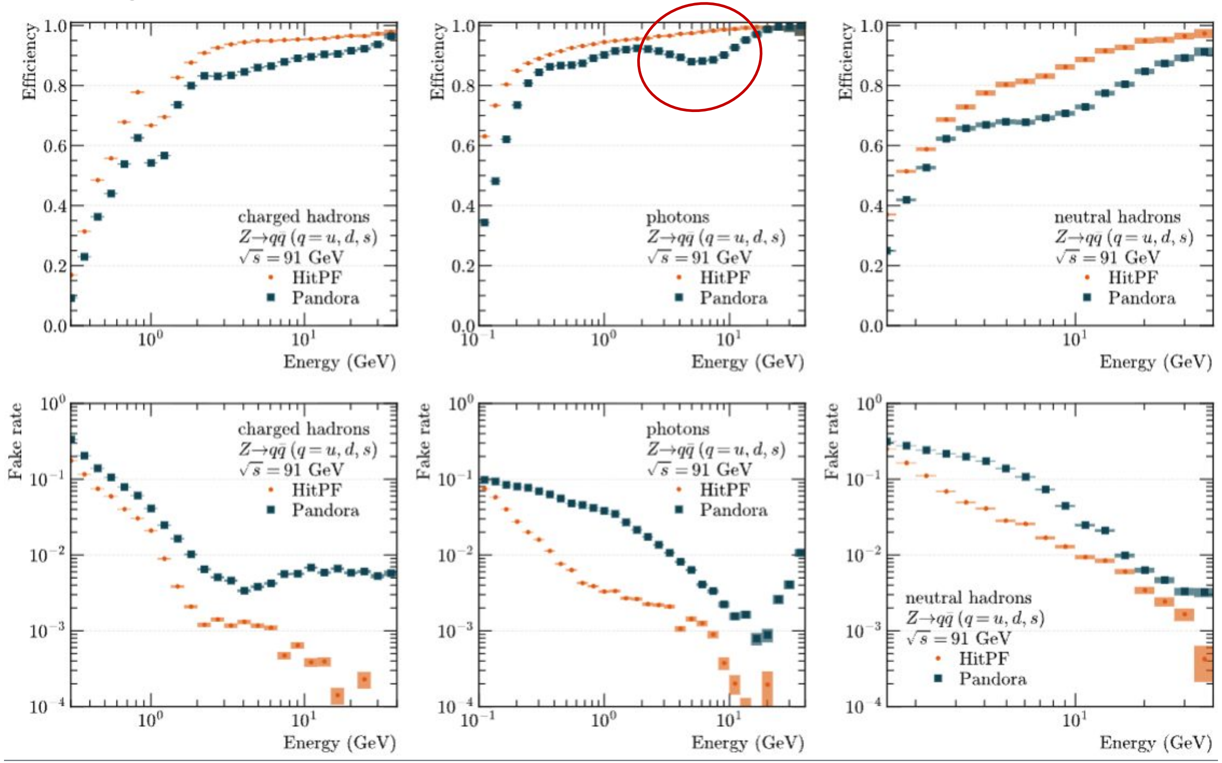
NH:

- Up to 15% higher eff. with Fake rate reduced by almost factor 5 in this region

→ Better clustering (no fake clusters from splitting)

PID: Efficiency and fake rate

Charged Hadron | Photons | Neutral Hadrons



CH:

- Up to 15% higher eff.
- Fake rate $\mathcal{O}(2)$ smaller

→Eff. limited by track reco.

Photons:

- ~5% higher eff. at high E
- Improvement by up to factor 2 for low E
- Fake rate up to ~ $\mathcal{O}(1)$

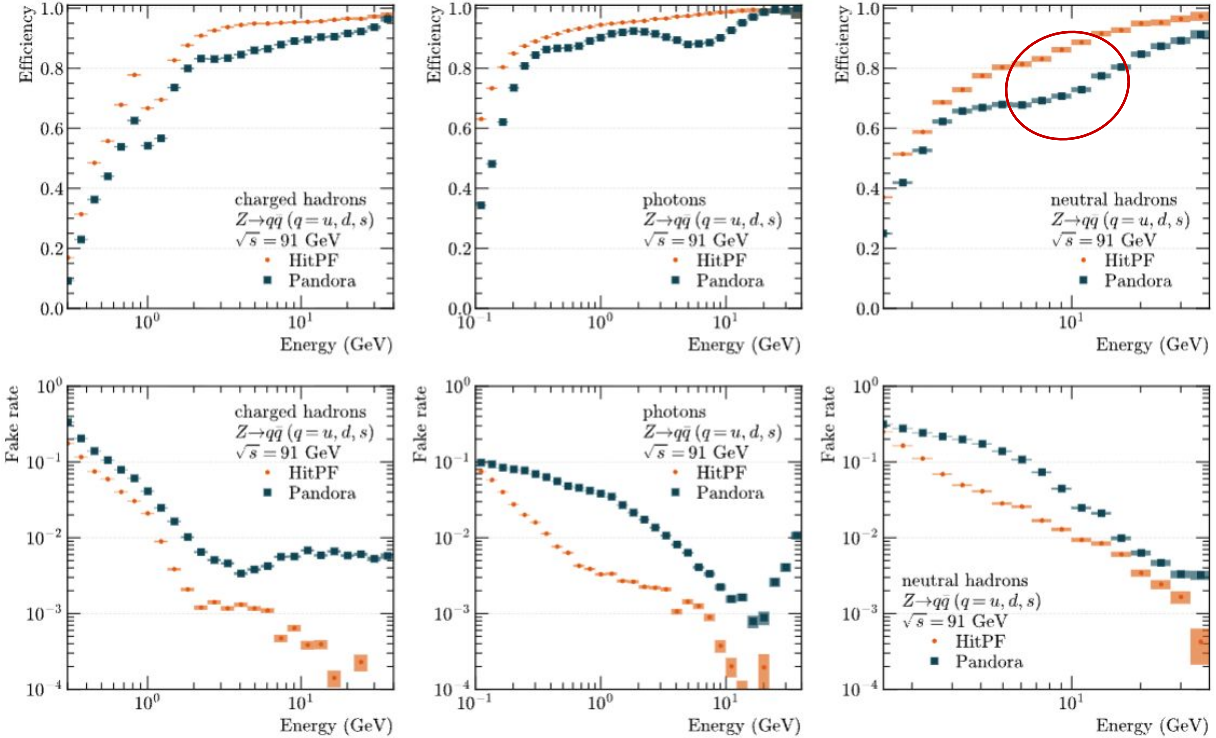
NH:

- Up to 15% higher eff. with Fake rate reduced by almost factor 5 in this region

→ Better clustering (no fake clusters from splitting)

PID: Efficiency and fake rate

Charged Hadron Photons Neutral Hadrons



CH:

- Up to 15% higher eff.
- Fake rate $\mathcal{O}(2)$ smaller

→Eff. limited by track reco.

Photons:

- ~5% higher eff. at high E
- Improvement by up to factor 2 for low E
- Fake rate up to ~ $\mathcal{O}(1)$

NH:

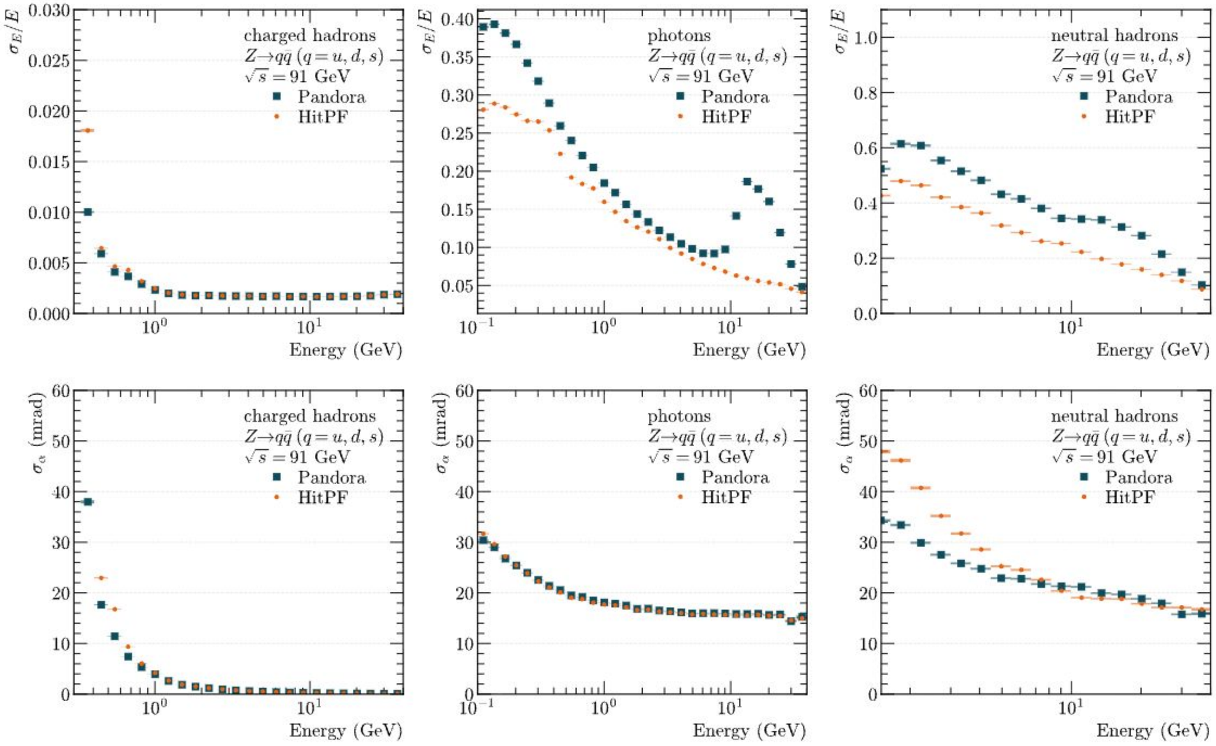
- Up to 15% higher eff. with Fake rate reduced by almost factor 5 in this region

→ Better clustering (no fake clusters from splitting)

Energy and angular resolution

Energy Resolution: half-width of 16th-84th percentile normalized to median
Angular resolution: 68th percentile of the opening angle between rec and true

Charged Hadron | Photon | Neutral Hadron



CH:

- Comparable by construction
- solely tracker information used

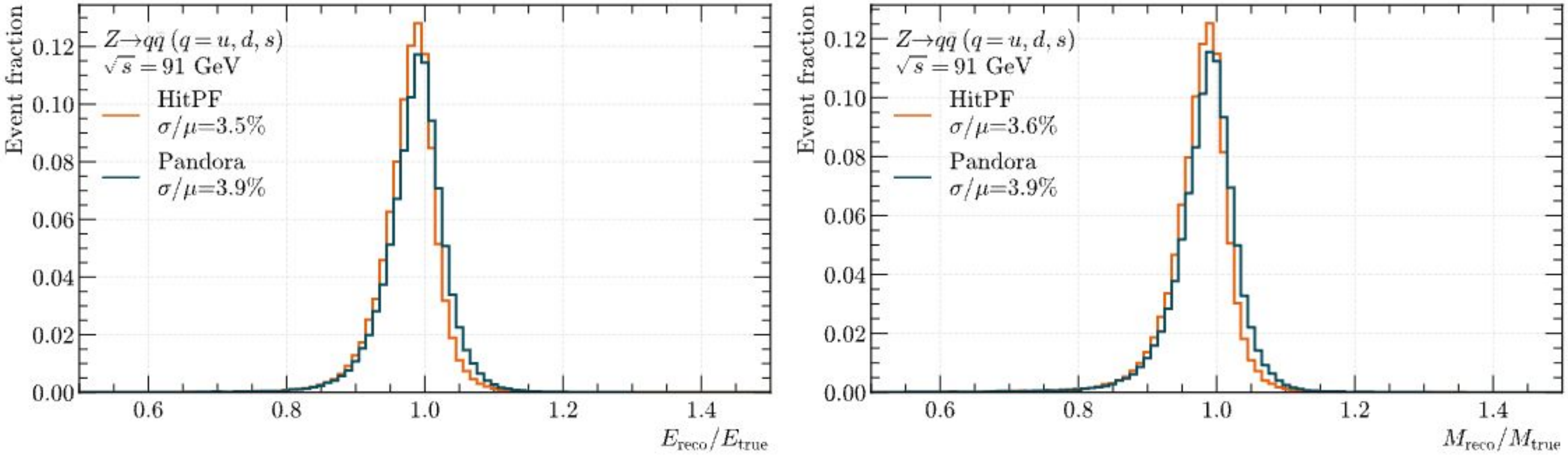
NH:

- Largest gain in energy/ angular resolution
- Less shower merging, better hit capture

Visible mass and energy

- ❑ Good complementary metric to per-particle metrics
- ❑ Improved mass resolution due to:
  - ❑ Higher efficiency in low energy neutrals
  - ❑ Improved energy resolution
  - ❑ Better track cluster-linking

7.7% relative improvement of HitPF compared to Pandora



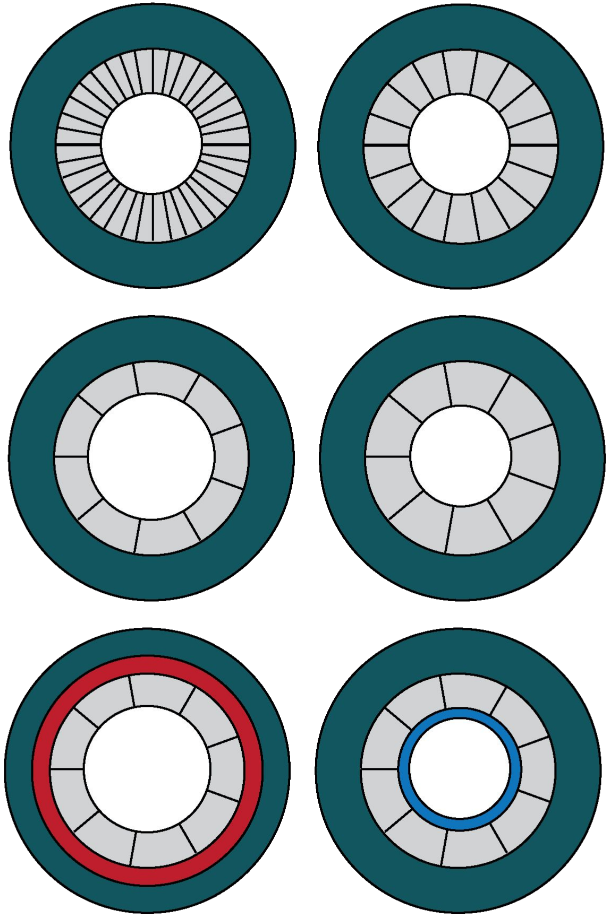
Isolated particles

		Efficiency		Fake rate		σ_E/E [%]		σ_α [mrad]	
process		HitPF	Pandora	HitPF	Pandora	HitPF	Pandora	HitPF	Pandora
e^+e^-	(t-ch.)	0.95	0.93	1.5×10^{-4}	3.3×10^{-3}	8.0	8.1	0.1	0.1
$\gamma\gamma$	(t-ch.)	1.00	1.00	1.7×10^{-4}	8.3×10^{-5}	3.6	3.0	11.7	11.7
$\mu^+\mu^-$	(s-ch.)	0.99	0.96	5.4×10^{-5}	3.6×10^{-3}	0.2	0.2	0.1	0.1
e^+e^-	(s-ch.)	0.97	0.95	5.6×10^{-5}	1.4×10^{-3}	6.5	6.3	0.1	0.1

What does this give access to?

Full simulation physics studies

- ❑ Impact of beam background
- ❑ Effect of timing information in clustering
- ❑ Impact of low momentum tracks reco on PF
- ❑ Impact of acceptance
- ❑ Tau reconstruction
- ❑ Physics validation
    - Retrain jet taggers with HitPF input
    - Deploy for physics analysis in CLD

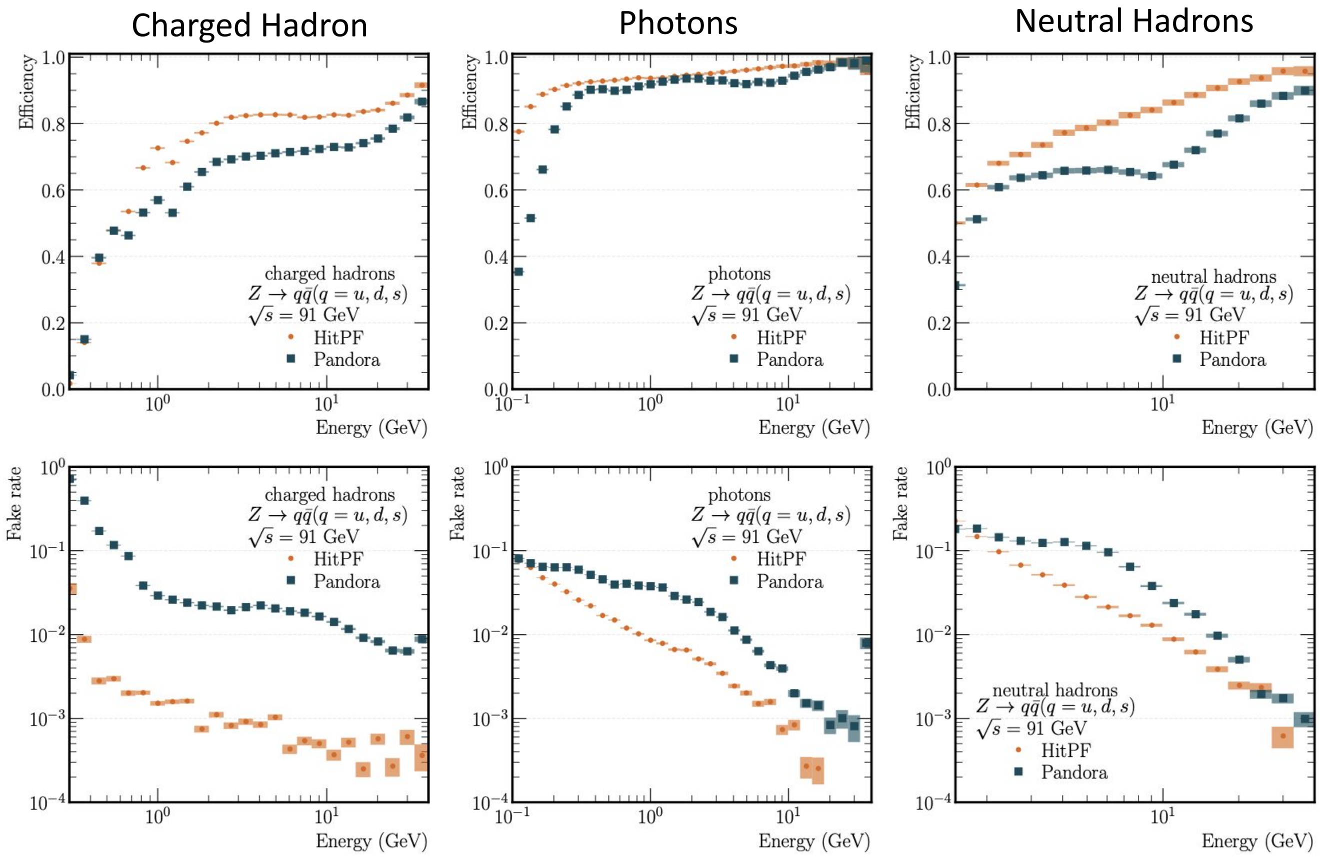


Summary

- Developments in ML allow for improved data representation for hit-based particle flow
- Improved performance in PF compared to classical approaches across all metrics: efficiency in complex hadronic topologies, neutral hadron efficiency, fakes and visible mass resolution (improvements by 24% CLD- 7.7% ILD)
- New geometric models could further improve performance, but development is needed as the properties of HEP are not the same as classic 3D datasets

Thank you for your attention.

PID: Efficiency and fake rate



CH:
- Up to 20% higher eff.
- Fake rate $\mathcal{O}(2)$ smaller

→Eff. limited by track reco.

Photons:
- ~5% higher eff. at high E
- Improvement by up to factor 2 for low E
- Fake rate up to ~ $\mathcal{O}(1)$

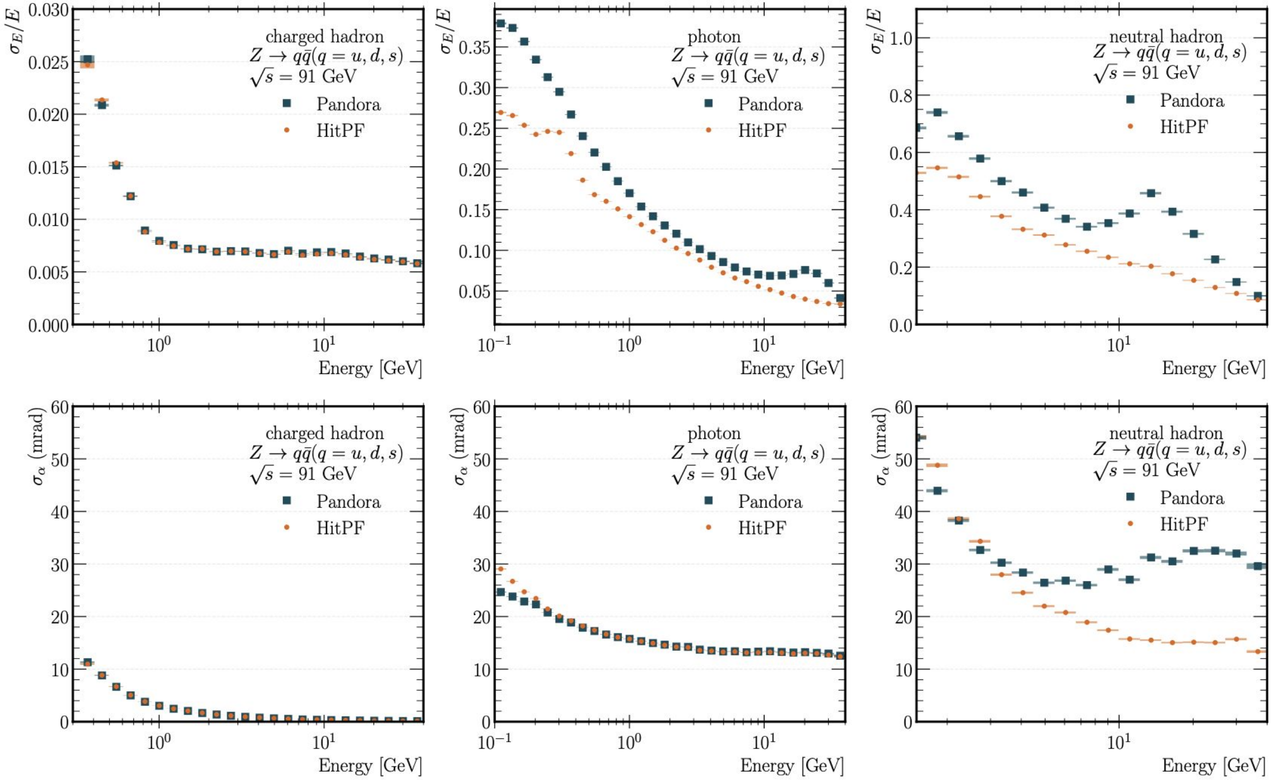
NH:
- Up to 20% higher eff. with Fake rate reduced by almost factor 5 in this region

→ Better clustering (no fake clusters from splitting)

Energy and angular resolution

Energy Resolution: half-width of 16th-84th percentile normalized to median
Angular resolution: 68th percentile of the opening angle between rec and true

Charged Hadron Photon Neutral Hadron



CH:

- Comparable by construction
- solely tracker information used

NH:

- Largest gain in energy/ angular resolution
- Less shower merging, better hit capture