ML @ Belle II @ DESY.

November 6th 2018, GPU Computing & Machine Learning @ DESY

Torben Ferber (torben.ferber@desy.de), Simon Wehle

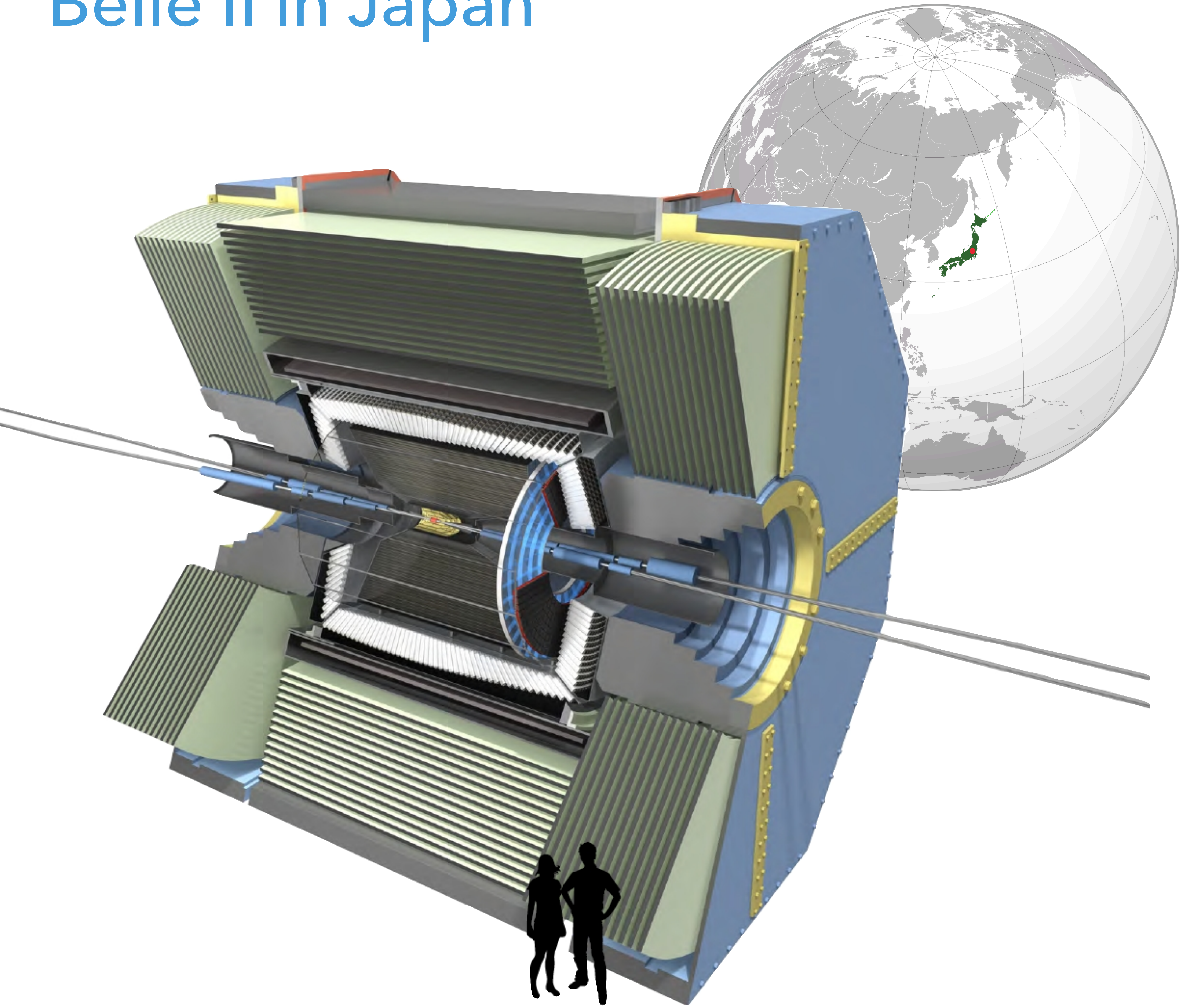
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Belle II

Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

DESY.

Belle II in Japan



- Intensity frontier flagship experiment: **30kHz event rate.**
- 750+ researchers from 30 countries. 100+ from Germany, ~20 from DESY (incl. 1 Helmholtz YIG and 1 Helmholtz W2).
- Precision physics and searches for (very) rare decays including Dark Matter.

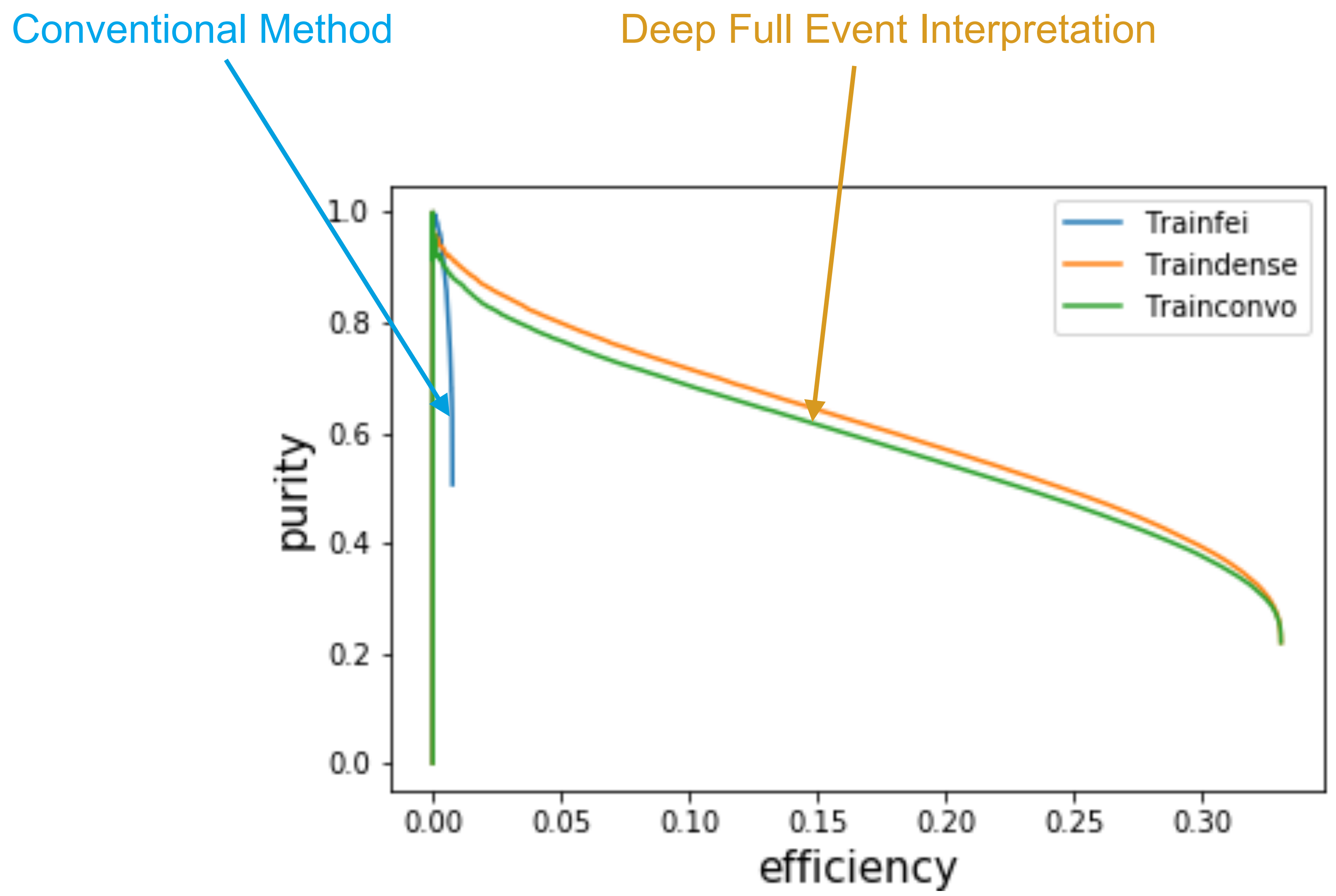
Data formats at Belle II

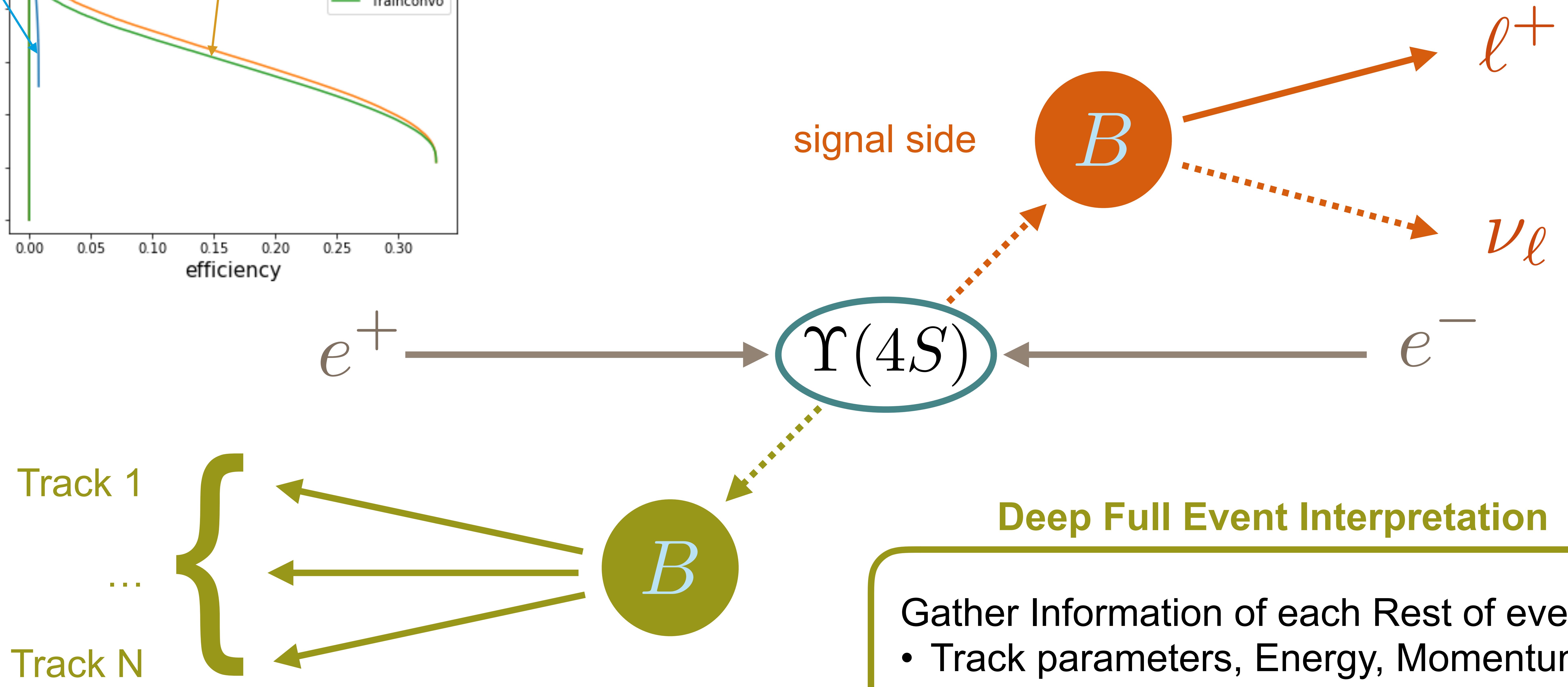
- Event reconstruction (C++, including online HLT) is performed on ROOT data.
  - Multiple ML packages (XGBoost, TMVA, Tensorflow) interfaced.
- User analysis (Python) high level output are (multiple) flat ROOT files. HDF support is planned.
- Final offline analysis either based on ROOT (C Macros) or Pandas/Numpy (via root_pandas, root_numpy or uproot). **Strong trend towards python at DESY.**

Overview Machine Learning at Belle II at DESY

- Physics analysis (NNs, BDTs)
  - Full Event Interpretation (FEI)
  - Adversarial approaches (bump hunts ↔ true mass, precision physics ↔ correlations)
- Electromagnetic calorimeter (NNs)
  - Energy and position reconstruction
  - Charged and Neutral Particle Identification (PID)
  - Calibration
  - Seedless clustering
  - Photon direction/displaced photons
- Tracking (BDTs)

Full Event Interpretation (FEI)

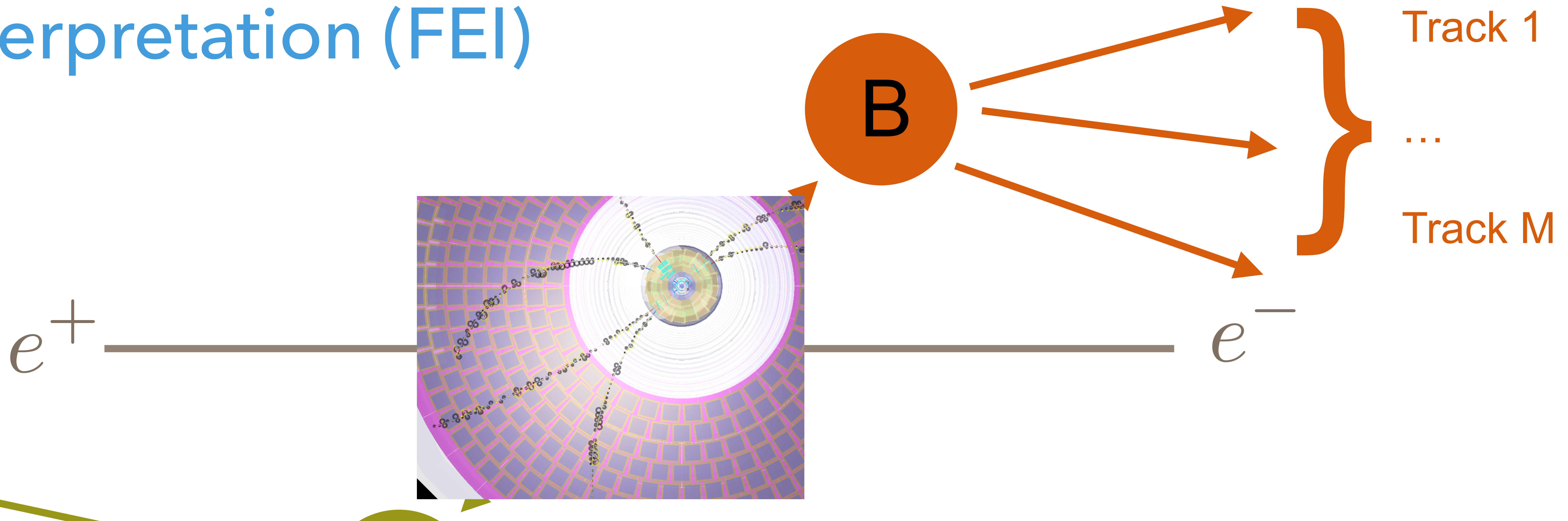


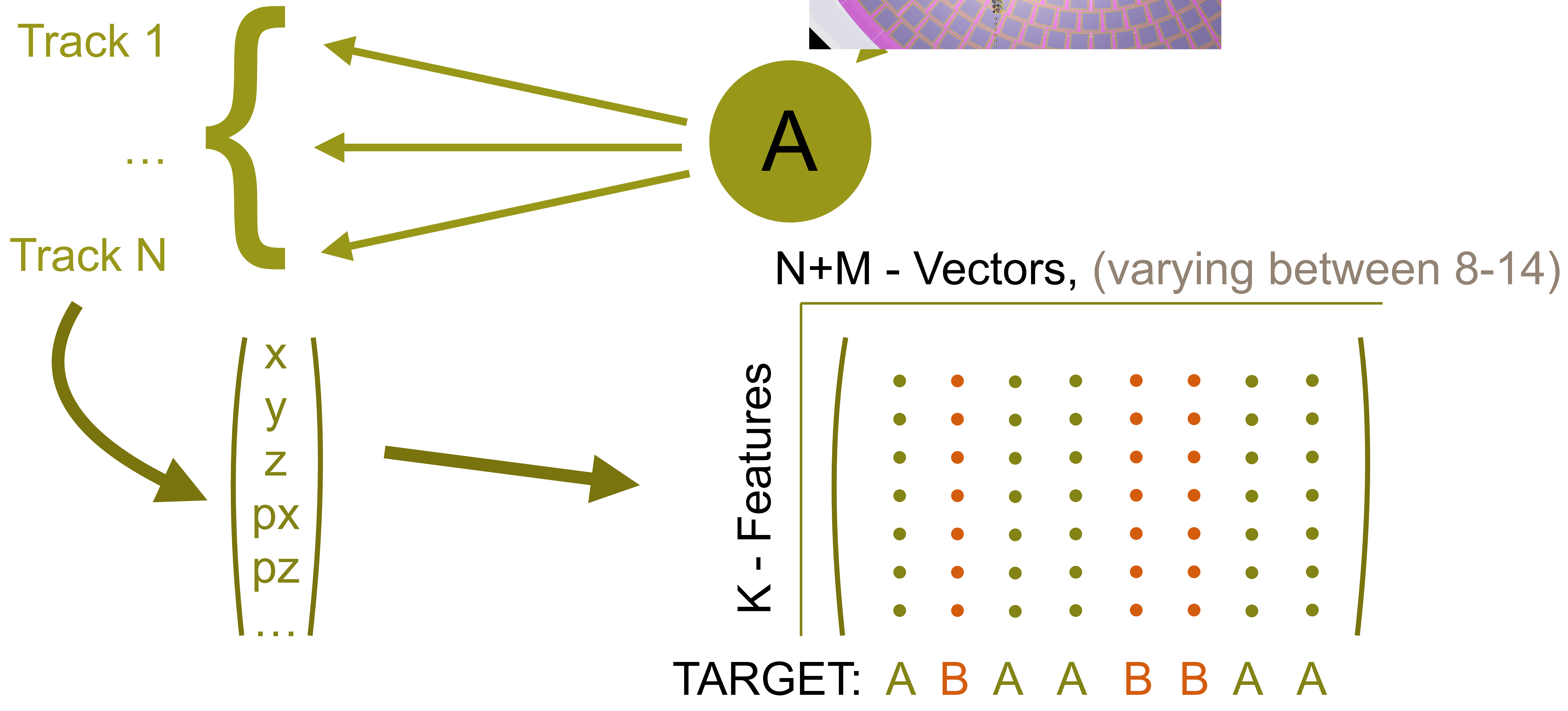


Deep Full Event Interpretation

Gather Information of each Rest of event Track
- Track parameters, Energy, Momentum
- Particle Identification Information

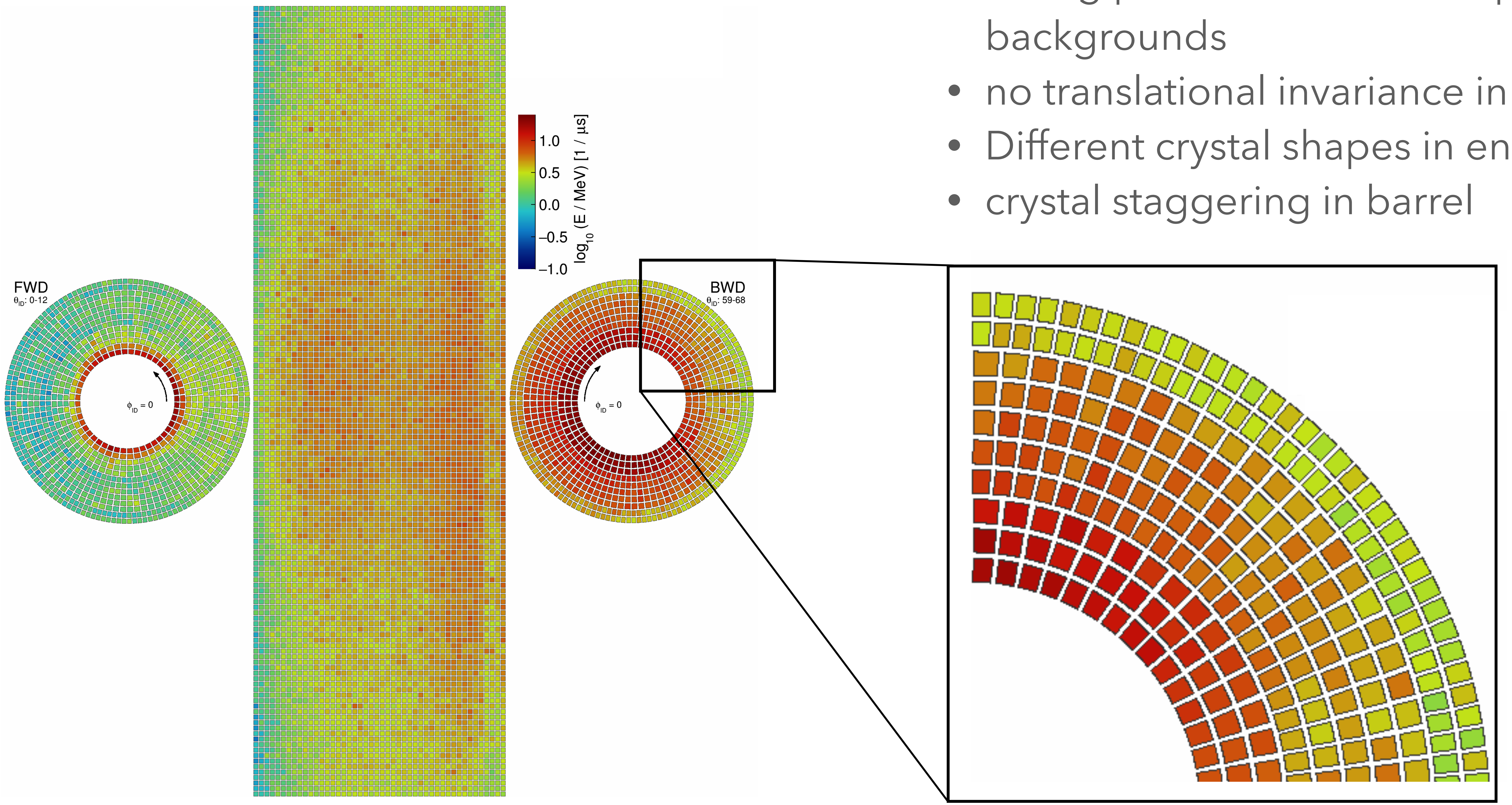
Full Event Interpretation (FEI)





- Currently use deep FC NN
- Probably better use recurrent NN (however order of the vectors has no information)
- Segmentation approaches?

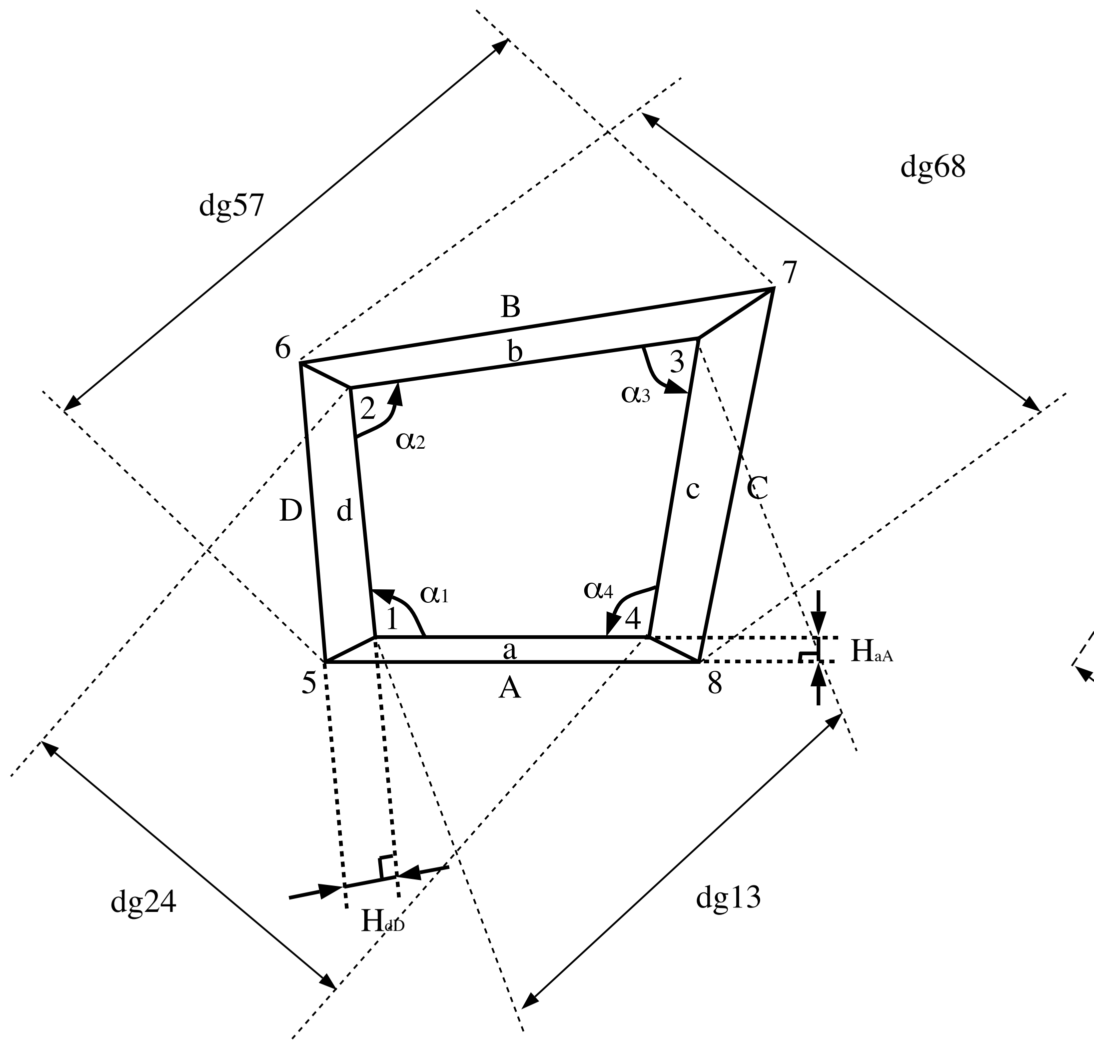
Electromagnetic calorimeter (ECL)

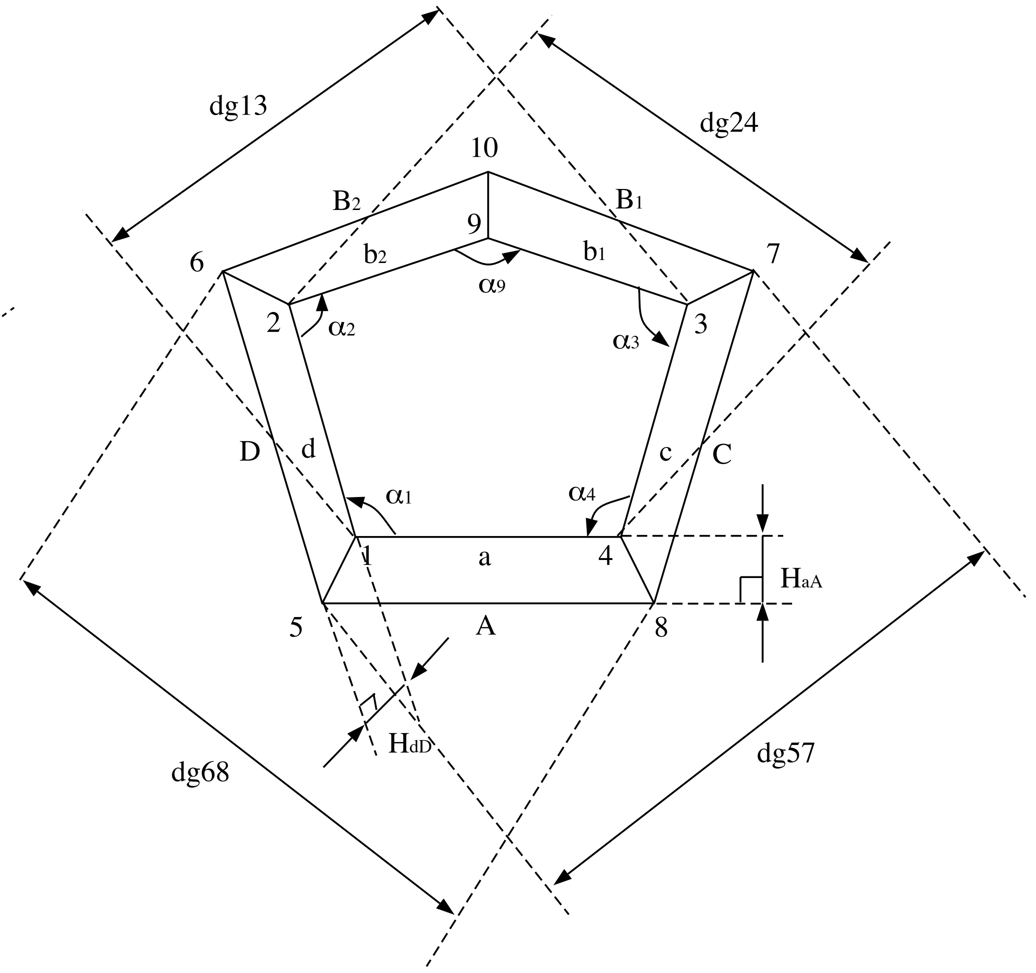


Challenges:

- strong position and time dependent backgrounds
- no translational invariance in endcaps
- Different crystal shapes in endcaps
- crystal staggering in barrel

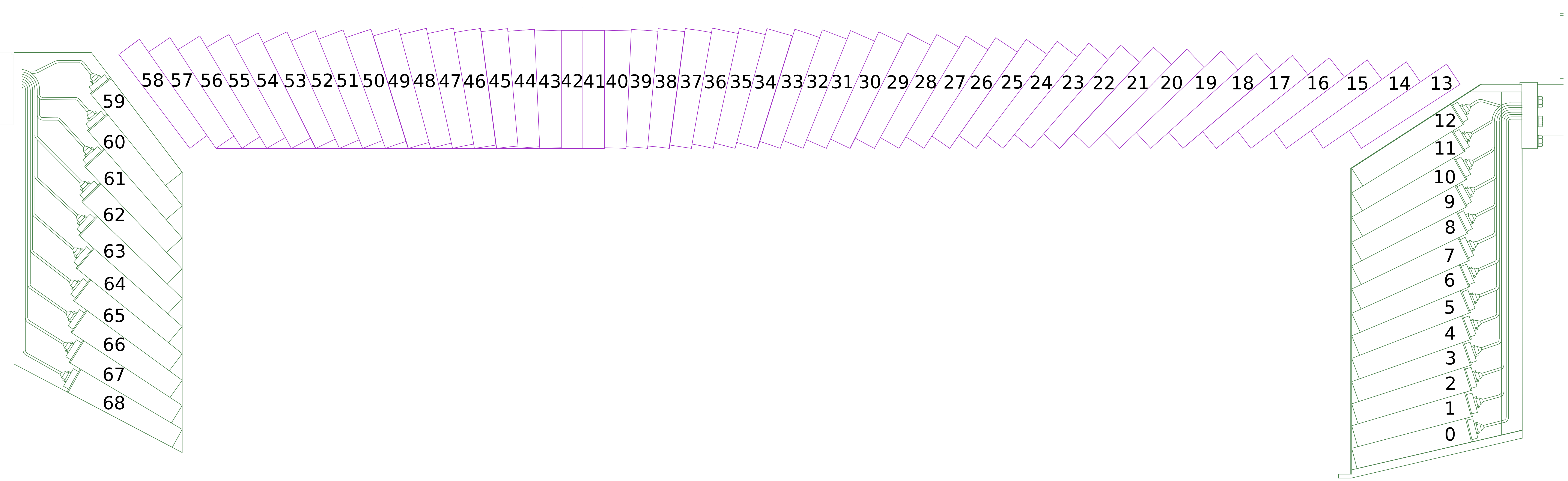
Electromagnetic calorimeter (ECL)





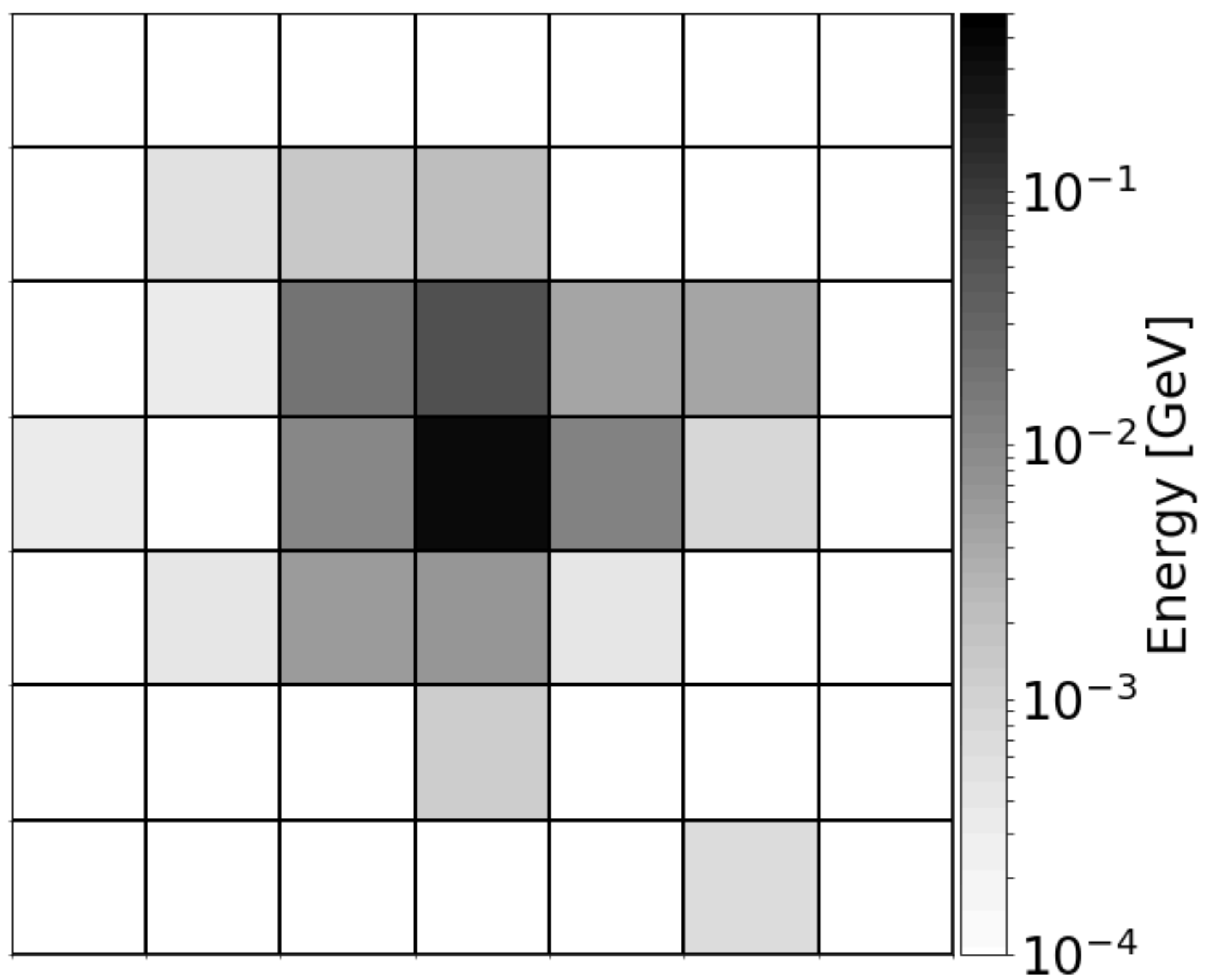
Challenges:

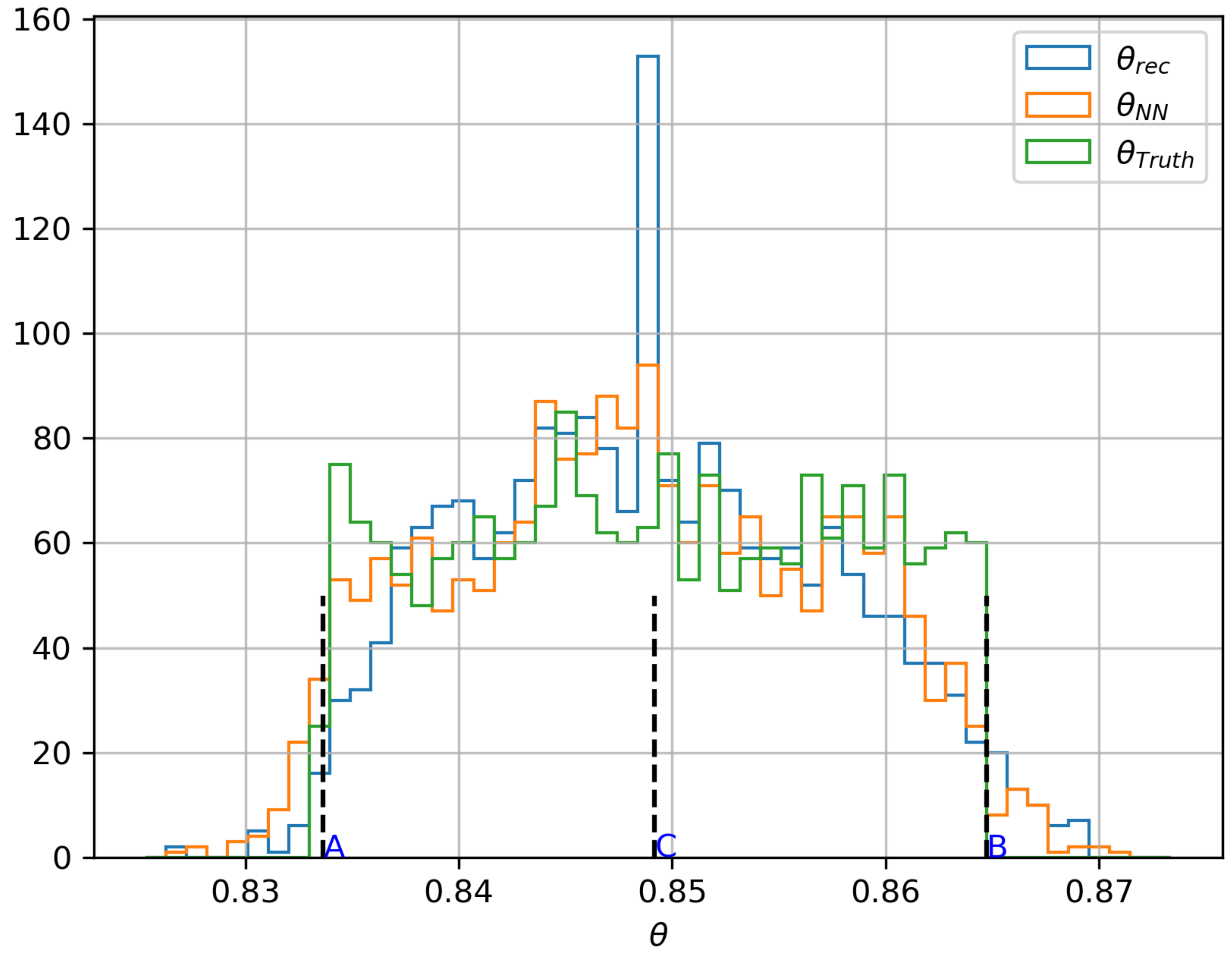
- strong position and time dependent backgrounds
- no translational invariance in endcaps
- Different crystal shapes in endcaps
- crystal staggering in barrel



ECL Photon position reconstruction

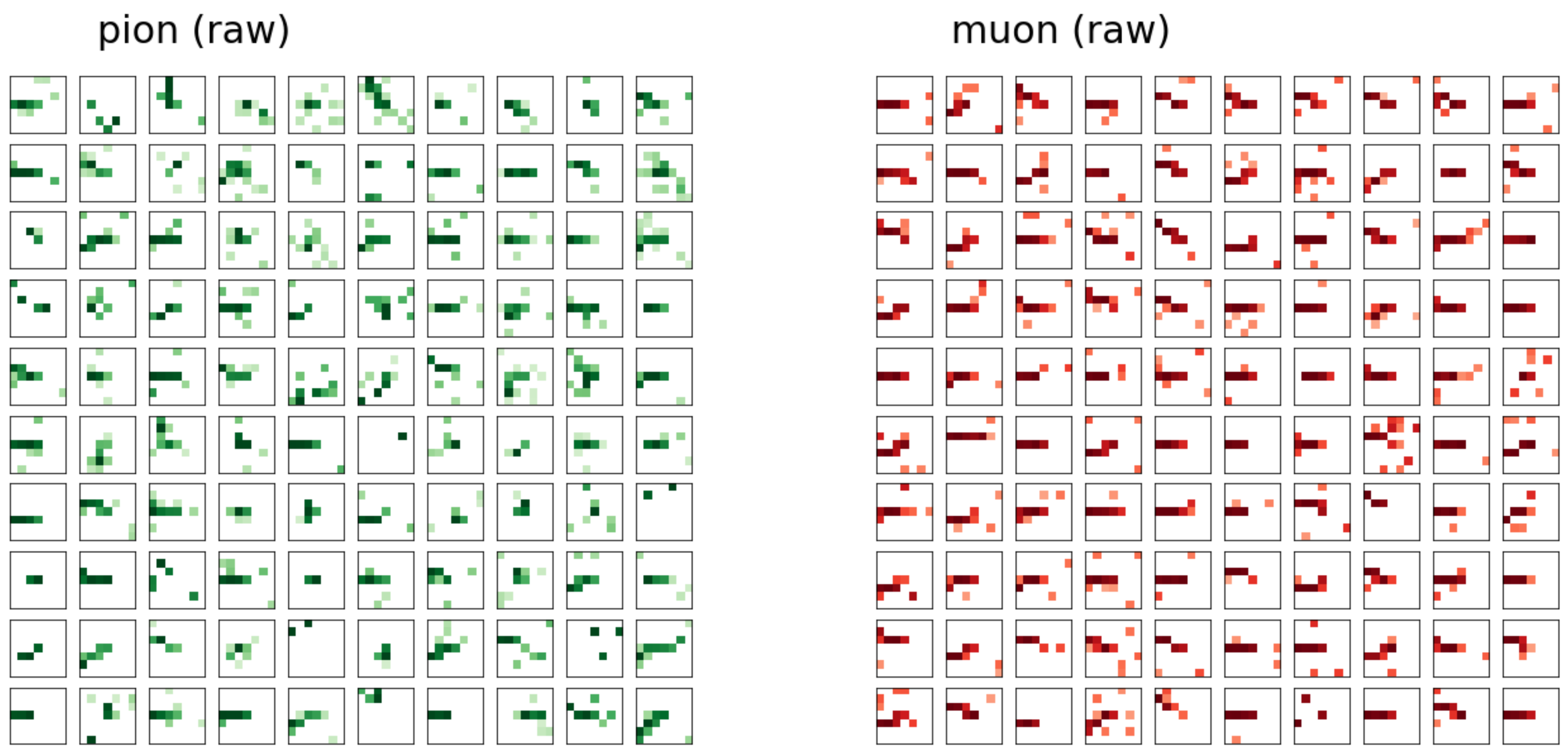
- Crystal calorimeter: most information contained in central crystal.
- Problem: **Very sparse information** leads to strong **bias towards towards central crystal** in non-ML approaches.
- Current ML approach uses “brute force” input 5×5×3 (energy, θ, Φ) and two targets θ_{Truth} and Φ_{Truth}. Barrel only. FC.
- Move to **generalized local position + bias** reconstruction next.

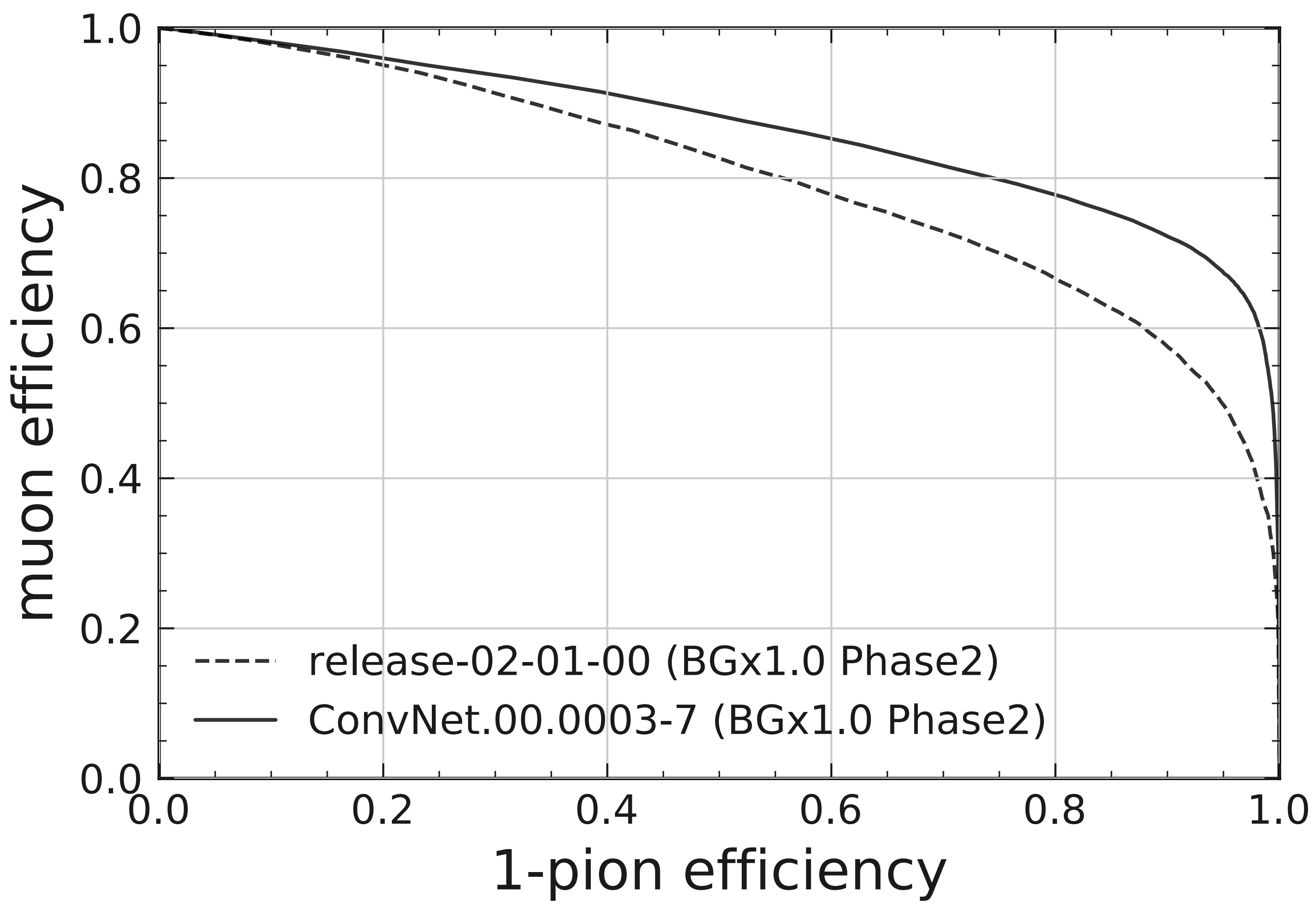




ECL Low p_t charged particle identification

- Low pt tracks will not reach the outer detector.
- **Seedless clustering** around extrapolated track impact point.
- **Preprocessing** to correct charge asymmetries and background fluctuations.
- **Image recognition** using convolutional networks.
- **Future**: Add non-image information in FC layers, use asymmetric images, use high dimensional image information (7×7×3..9) from digitized waveforms.

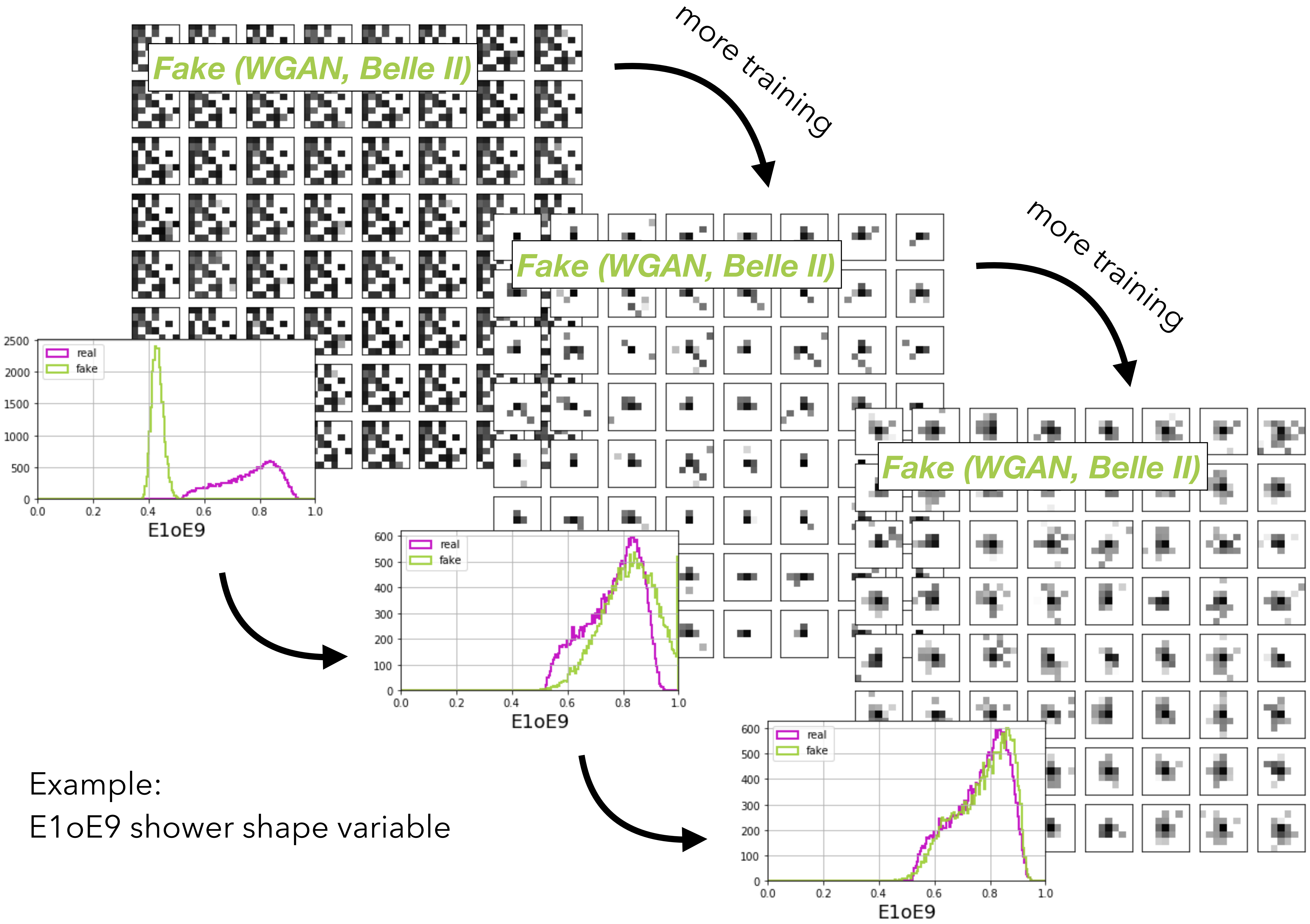




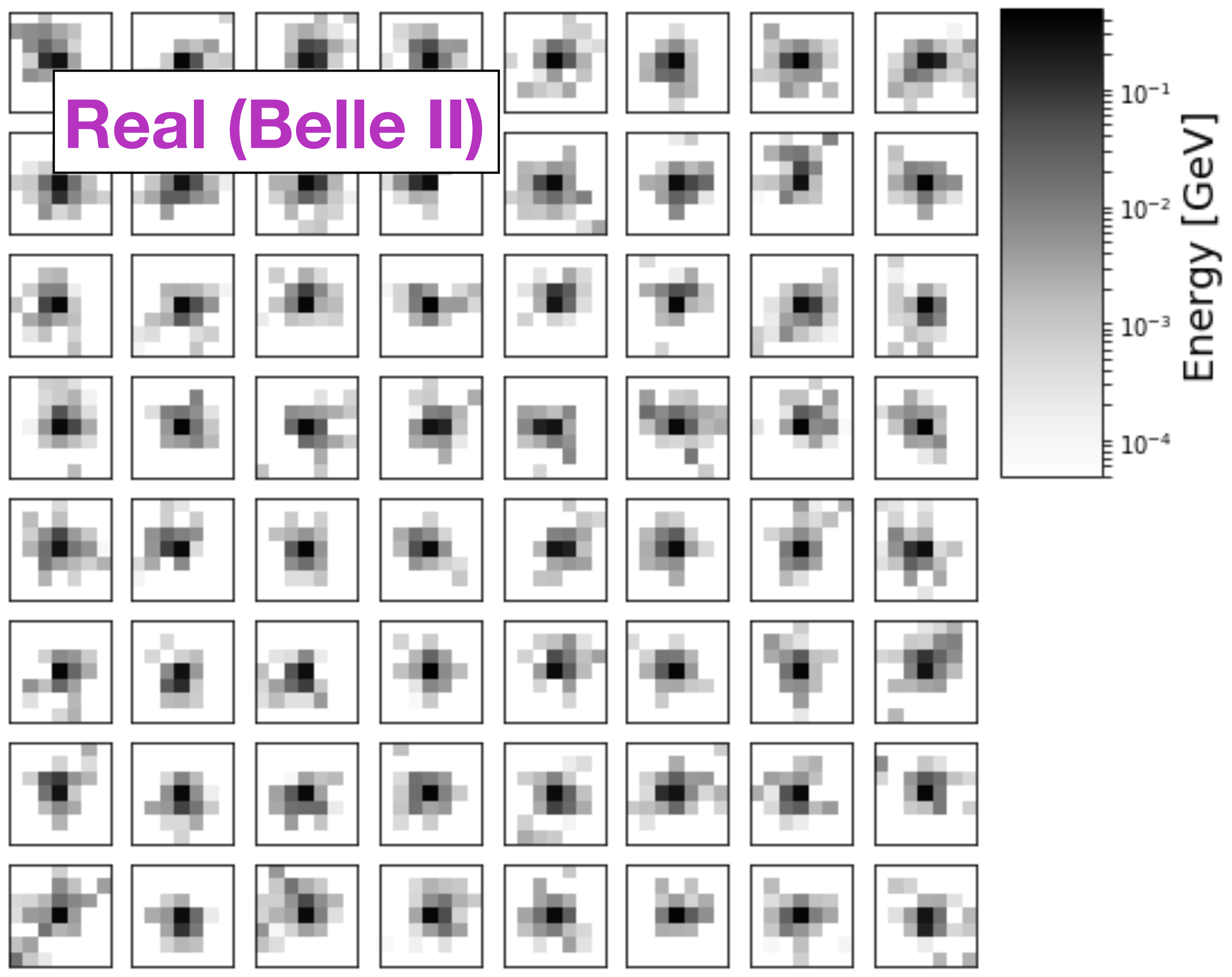
ECL cluster shape calibration

- High level user analysis is performed on reduced datasets with several expert-engineered **shower shape variables** per shower.
  - Used to separate photons and neutral hadrons.
  - Differences in data and simulation of shower shapes reduces experimental precision by introducing multiple ad-hoc corrections (one per shower shape).
  - Under study: Use Wasserstein refiner networks to calibrate shower images instead, before further analysis steps.

ECL cluster shape calibration

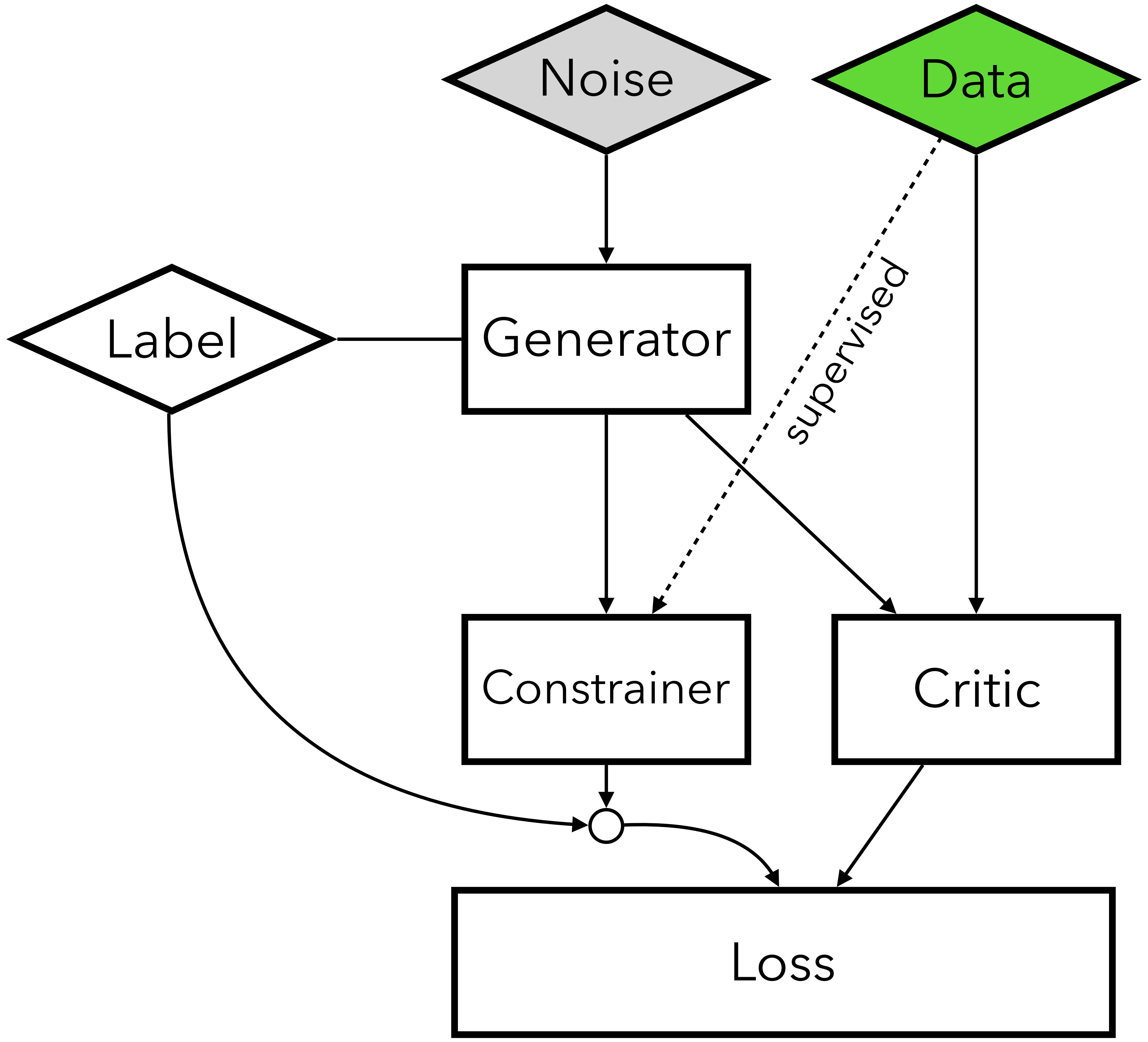


Semi-supervised learning: Wasserstein GAN learns to create ‘fake’ images that look like real Belle II images.



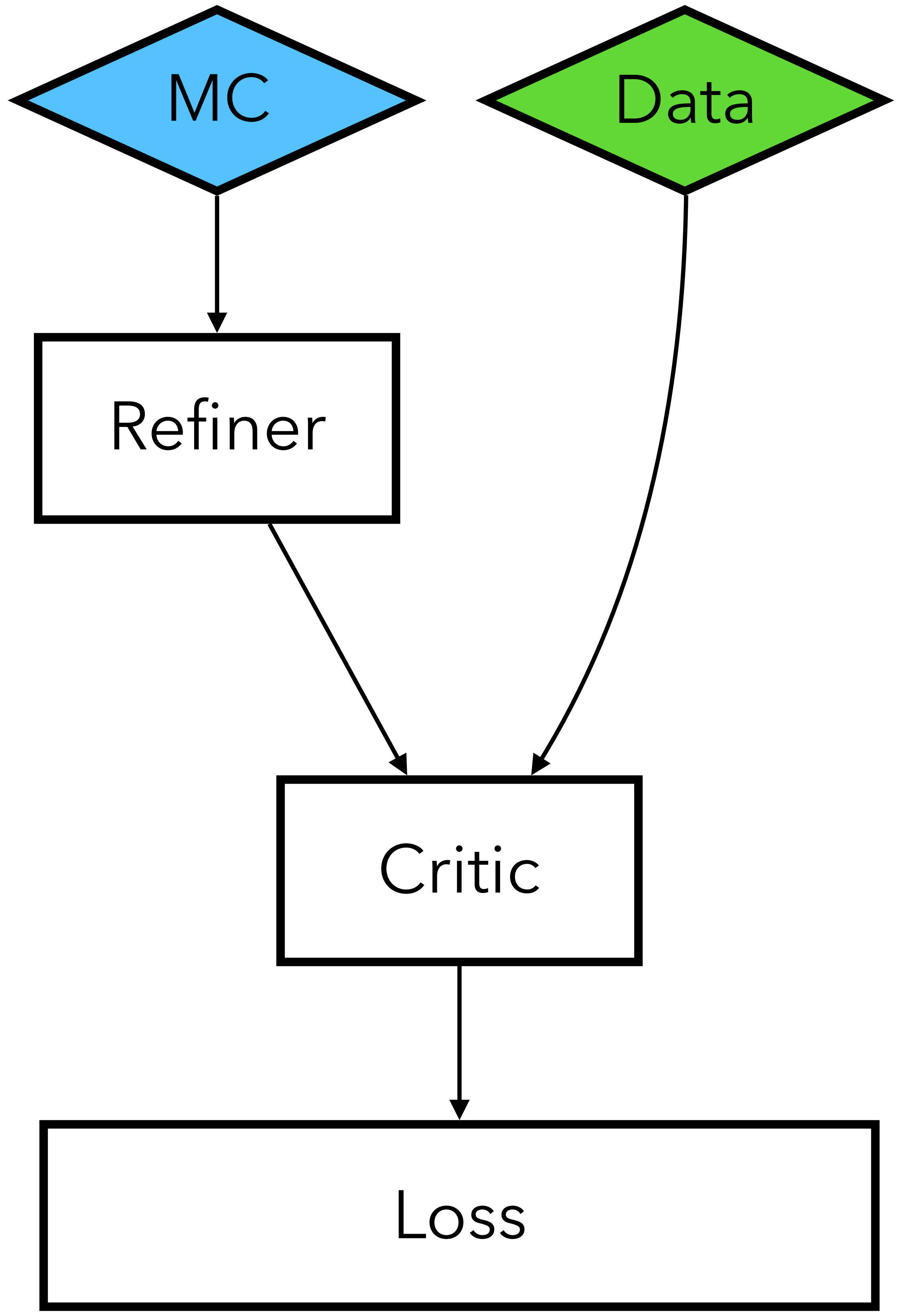
Example:
E1oE9 shower shape variable

ECL cluster shape calibration



Wasserstein Generative Adversarial Network: **WGAN**
(with supervised auxiliary constrainers: AC-WGAN)

Under study at DESY:
Improve existing MC simulations using data before further analysis steps.



Wasserstein Refiner Network

Erdmann et al (https://arxiv.org/pdf/1802.03325.pdf)

Challenges/random thoughts

- Different tasks (development, tuning, finalization, application) require different tools (Maxwell, HTCondor, local resources). Workflow needs to be understood and optimized.
- Ultimately we need to run ML where our data is → HTCondor.
- Belle II calorimeter image problems are not (few images) × (large image size) but (many images*) × (small image size). Sill learning to utilize full GPU benefits.
- Precision calibration using ML is not used in our field yet: Synergy at DESY?
- Best practices to address systematic uncertainties?

* about 3 million cluster images per second

Summary

- Not covered here: Most Belle II analyses use ML during high-level analysis.
- Strong trends towards fully python-based analyses at Belle II @ DESY.
- ML focus at Belle II @ DESY:
    - Full Event Interpretation
    - ECL reconstruction and particle identification.

Contact

DESY.

Deutsches Elektronen Synchrotron
www.desy.de

Torben Ferber
torben.ferber@desy.de