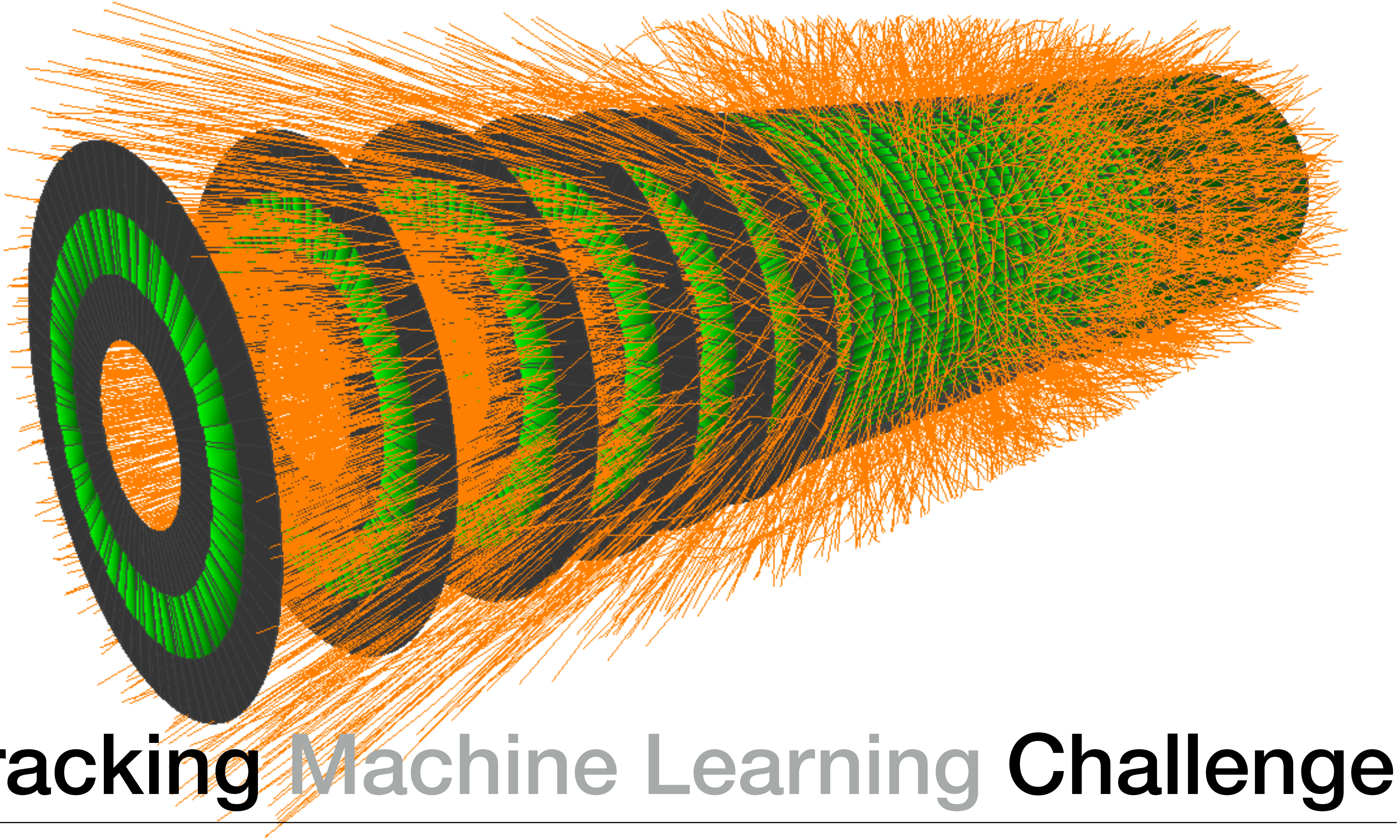
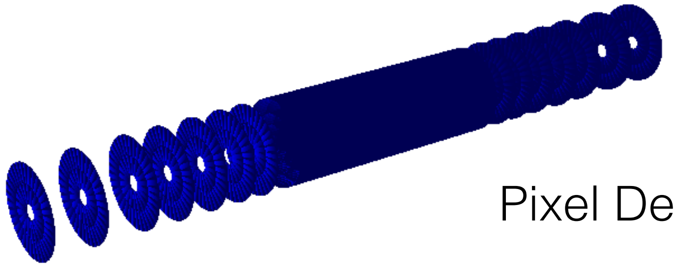


Tracking Machine Learning Challenge

towards a reference dataset of HEP

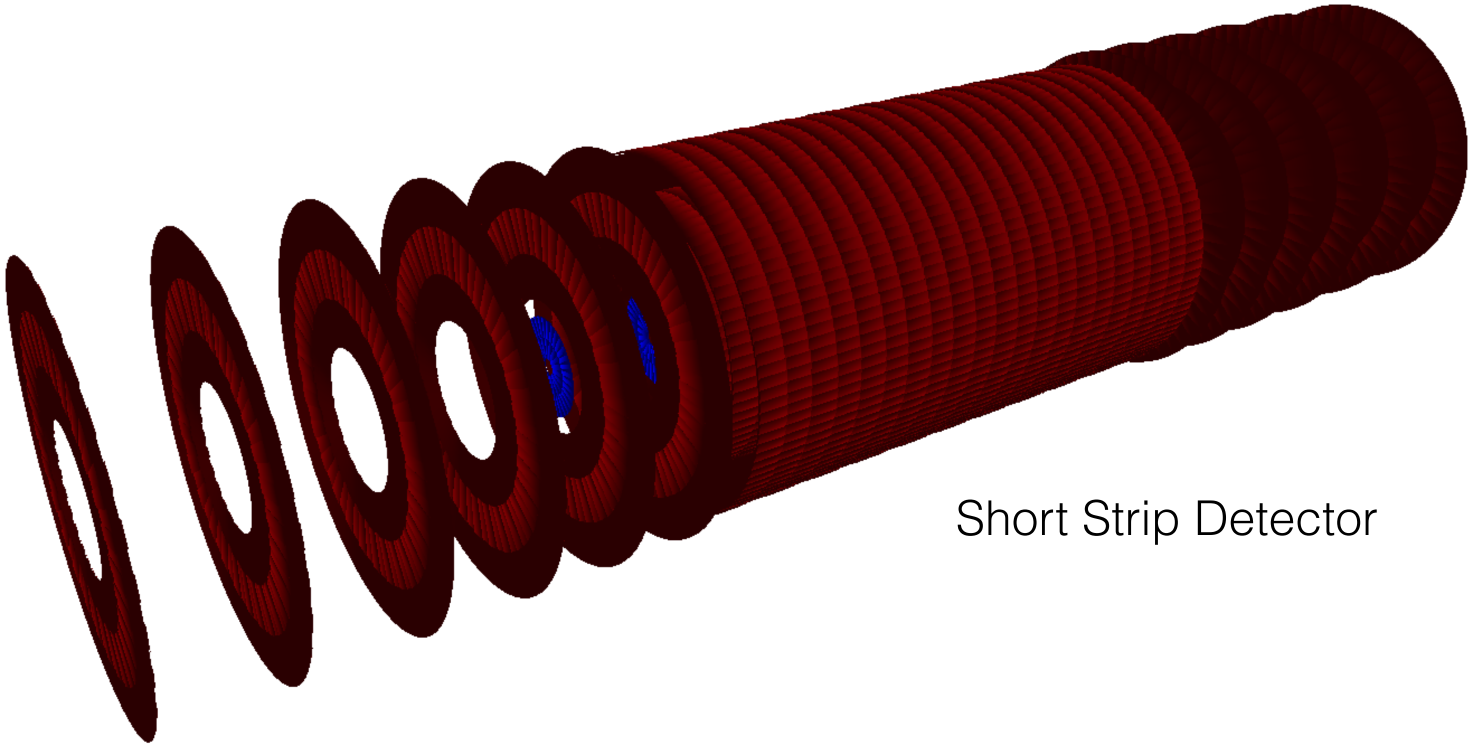
A. Salzburger (CERN) for the TrackML organisers
@SaltyBurger

dataset Pixel Detector



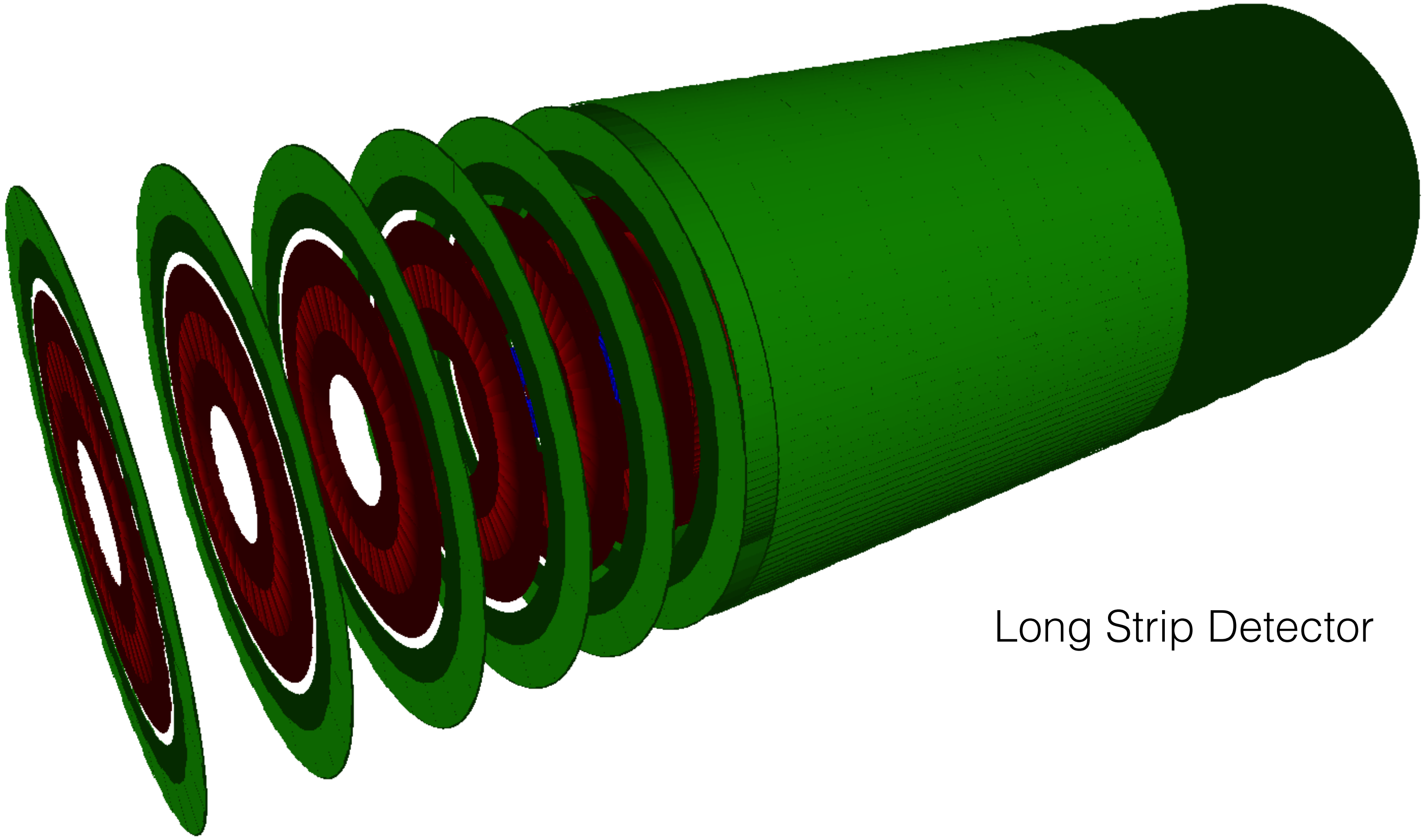
Pixel Detector

dataset Short Strip Detector

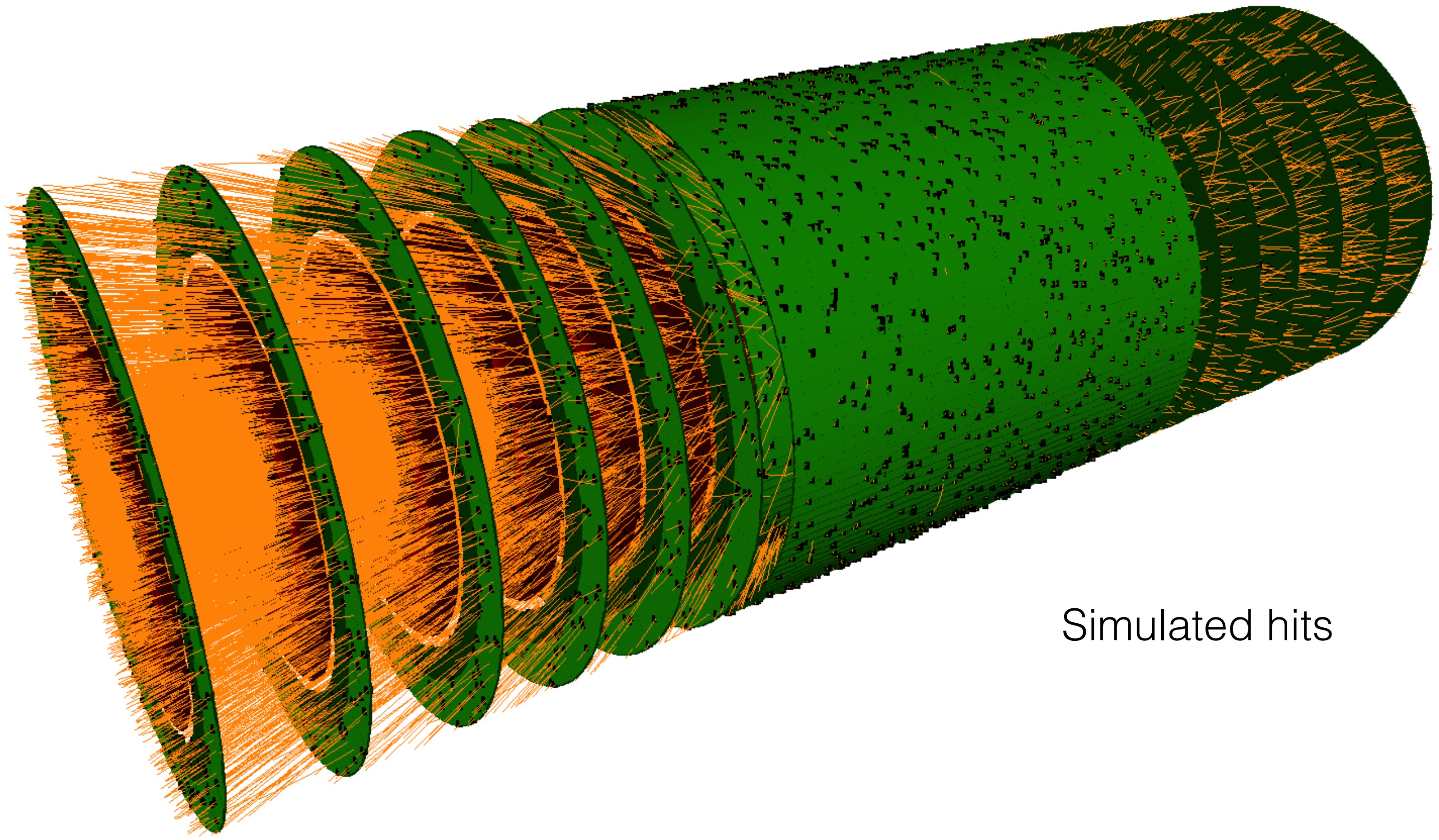


Short Strip Detector

dataset Long Strip Detector



TrackML dataset Simulation setup

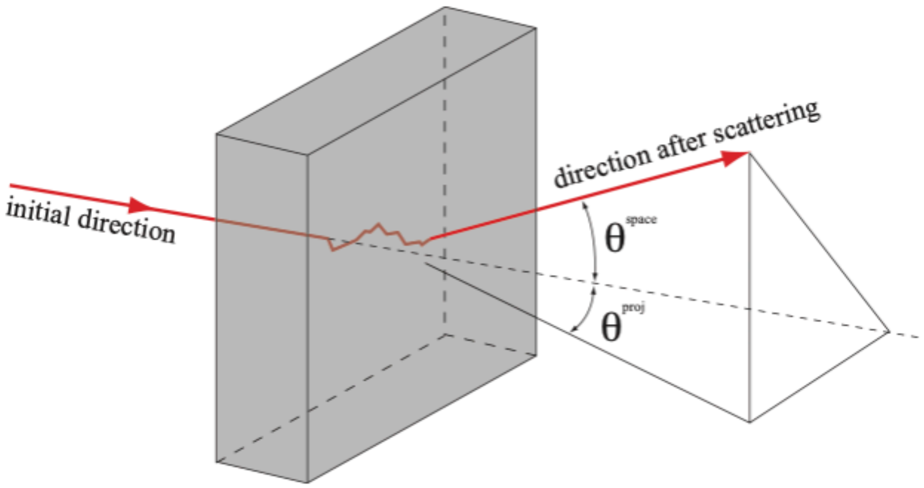


Simulated hits

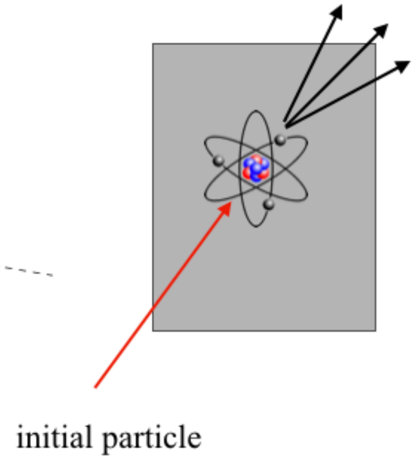
TrackML dataset Simulation setup

Main interactions of a particle with detector material

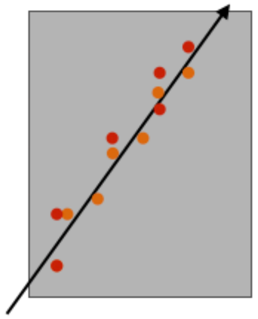
- scattering



- destructive (hadronic) interaction

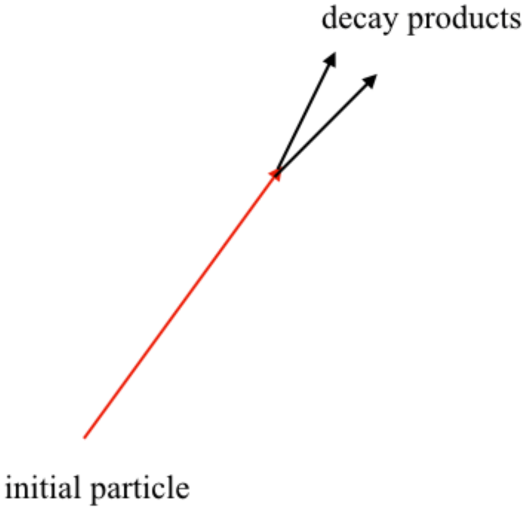


- energy loss



Simulated

- particle decay

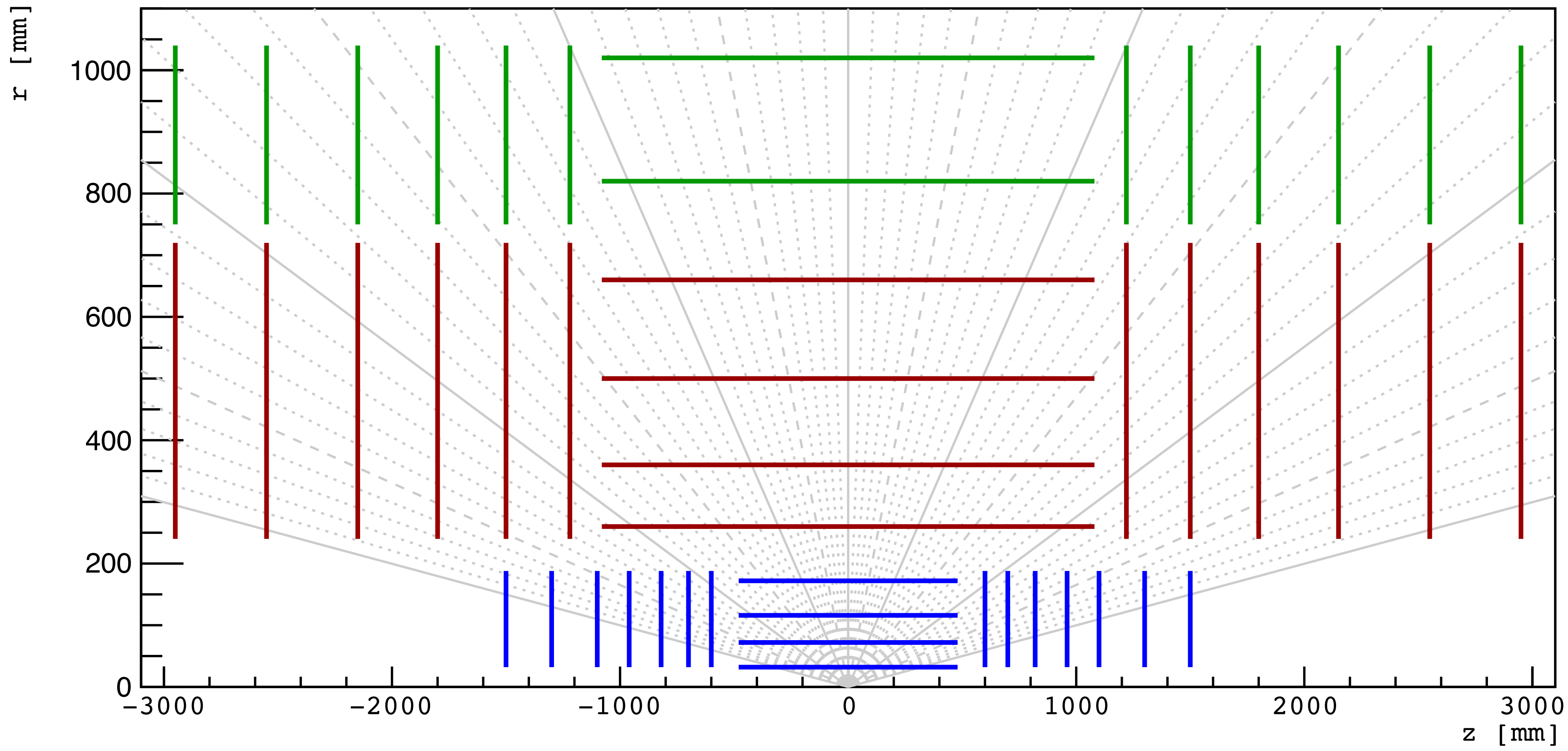


Not simulated

The **detector**

Defined a Phase-2 like detector

- full silicon detector with realistic resolution, material budget, magnetic field
- composed as Pixel, short strip, long strip
- restricted to size of tracking volume to $|\eta| < 3$

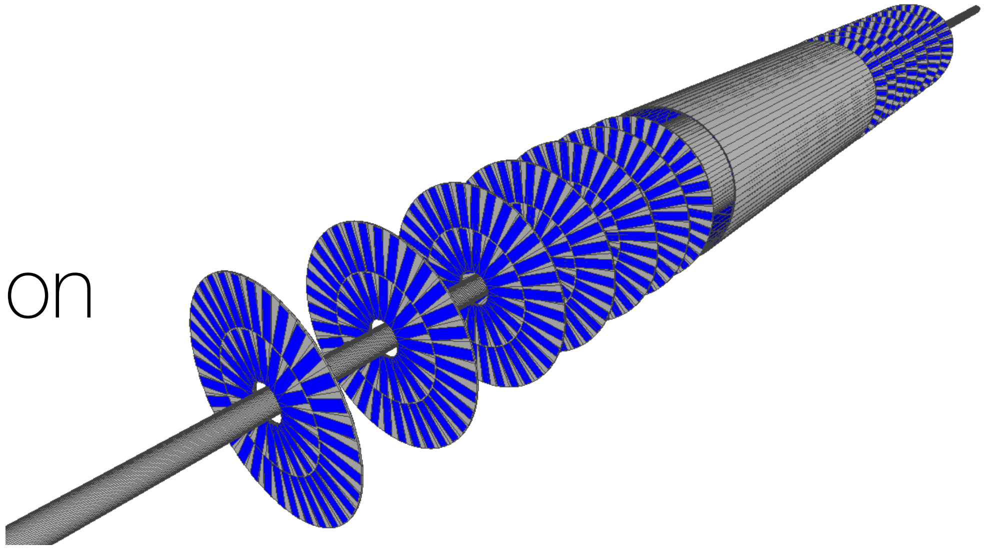


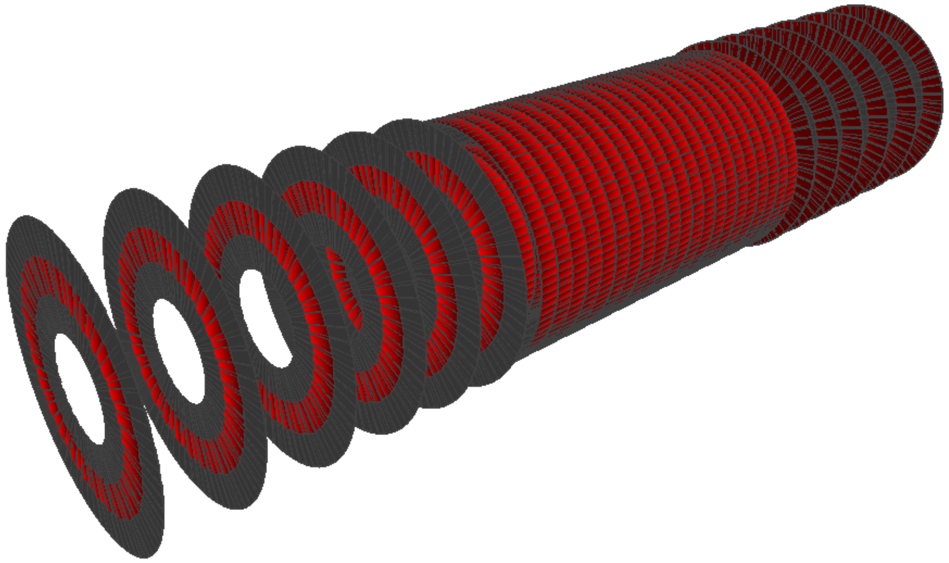
plot & image
(left) X0 distribution of the trackML detector
(right) longitudinal view of the trackML detector

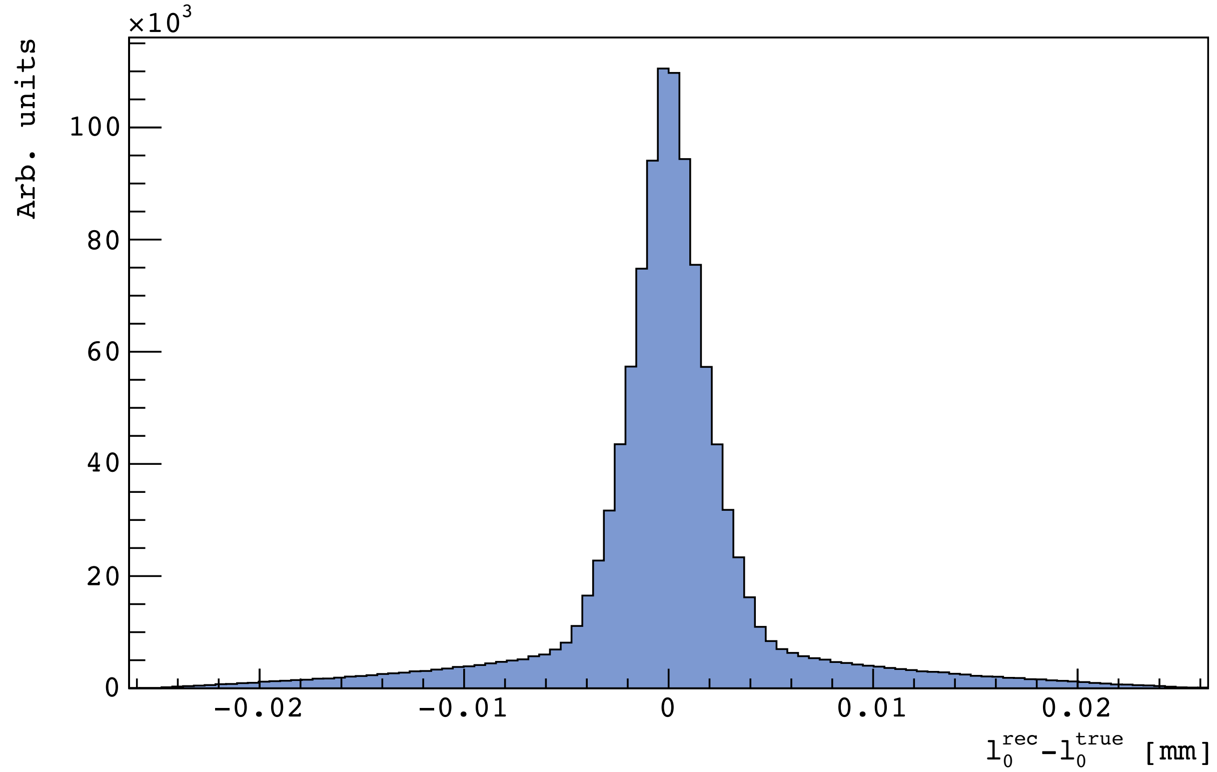
The **detector**

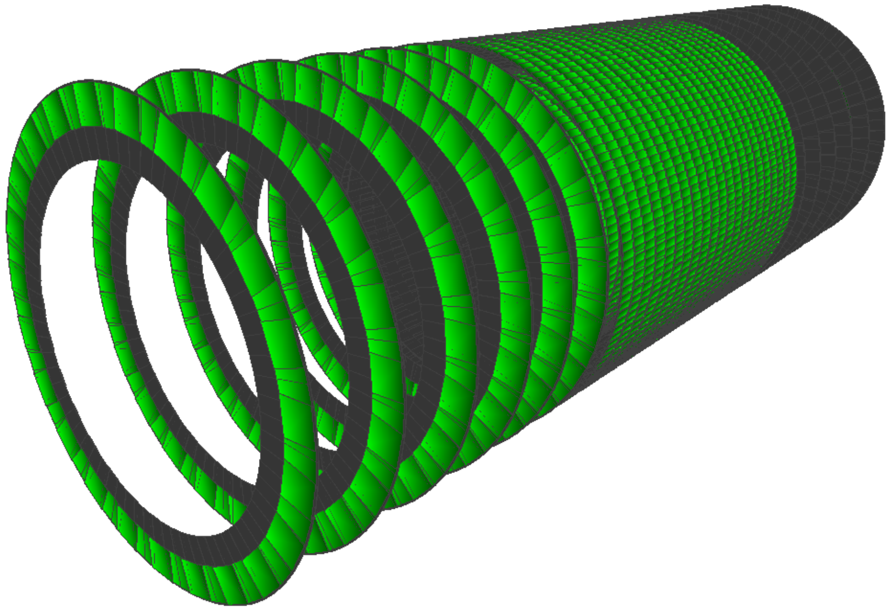
Dataset is simulation with ACTS fast simulation

- includes **multiple scattering, energy loss and hadronic interactions**
- includes **inefficiencies** and **noise/low momentum** particle hits
- includes pseudo-realistic **clustering model** (and hence resolutions)









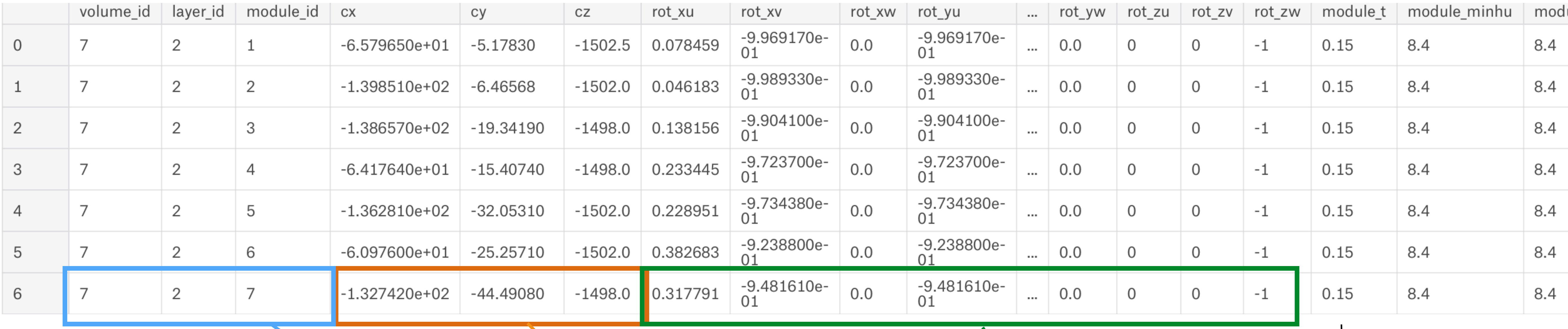
plot & images

(left) estimated pixel resolution distribution

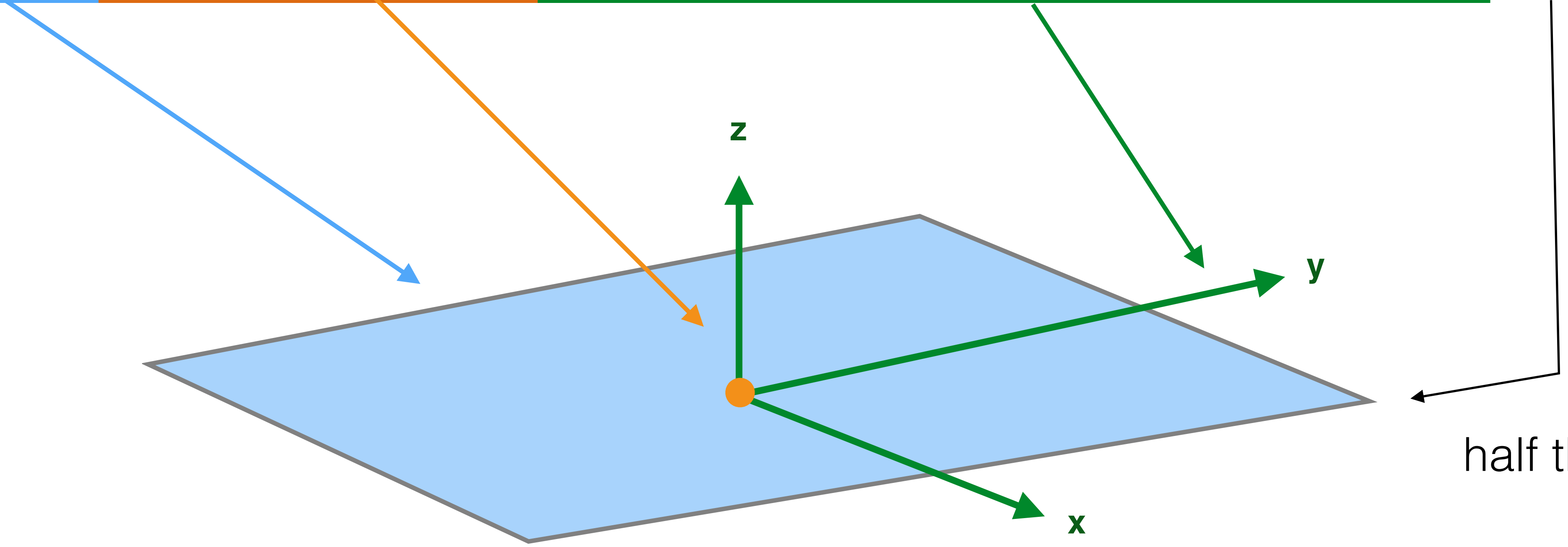
(right) 3D view of pixel, short strip and long strip detector

The **detector**

Detector description is given as .csv file



	volume_id	layer_id	module_id	cx	cy	cz	rot_xu	rot_xv	rot_xw	rot_yu	...	rot_yw	rot_zu	rot_zv	rot_zw	module_t	module_minhu	modu
0	7	2	1	-6.579650e+01	-5.17830	-1502.5	0.078459	-9.969170e-01	0.0	-9.969170e-01	...	0.0	0	0	-1	0.15	8.4	8.4
1	7	2	2	-1.398510e+02	-6.46568	-1502.0	0.046183	-9.989330e-01	0.0	-9.989330e-01	...	0.0	0	0	-1	0.15	8.4	8.4
2	7	2	3	-1.386570e+02	-19.34190	-1498.0	0.138156	-9.904100e-01	0.0	-9.904100e-01	...	0.0	0	0	-1	0.15	8.4	8.4
3	7	2	4	-6.417640e+01	-15.40740	-1498.0	0.233445	-9.723700e-01	0.0	-9.723700e-01	...	0.0	0	0	-1	0.15	8.4	8.4
4	7	2	5	-1.362810e+02	-32.05310	-1502.0	0.228951	-9.734380e-01	0.0	-9.734380e-01	...	0.0	0	0	-1	0.15	8.4	8.4
5	7	2	6	-6.097600e+01	-25.25710	-1502.0	0.382683	-9.238800e-01	0.0	-9.238800e-01	...	0.0	0	0	-1	0.15	8.4	8.4
6	7	2	7	-1.327420e+02	-44.49080	-1498.0	0.317791	-9.481610e-01	0.0	-9.481610e-01	...	0.0	0	0	-1	0.15	8.4	8.4

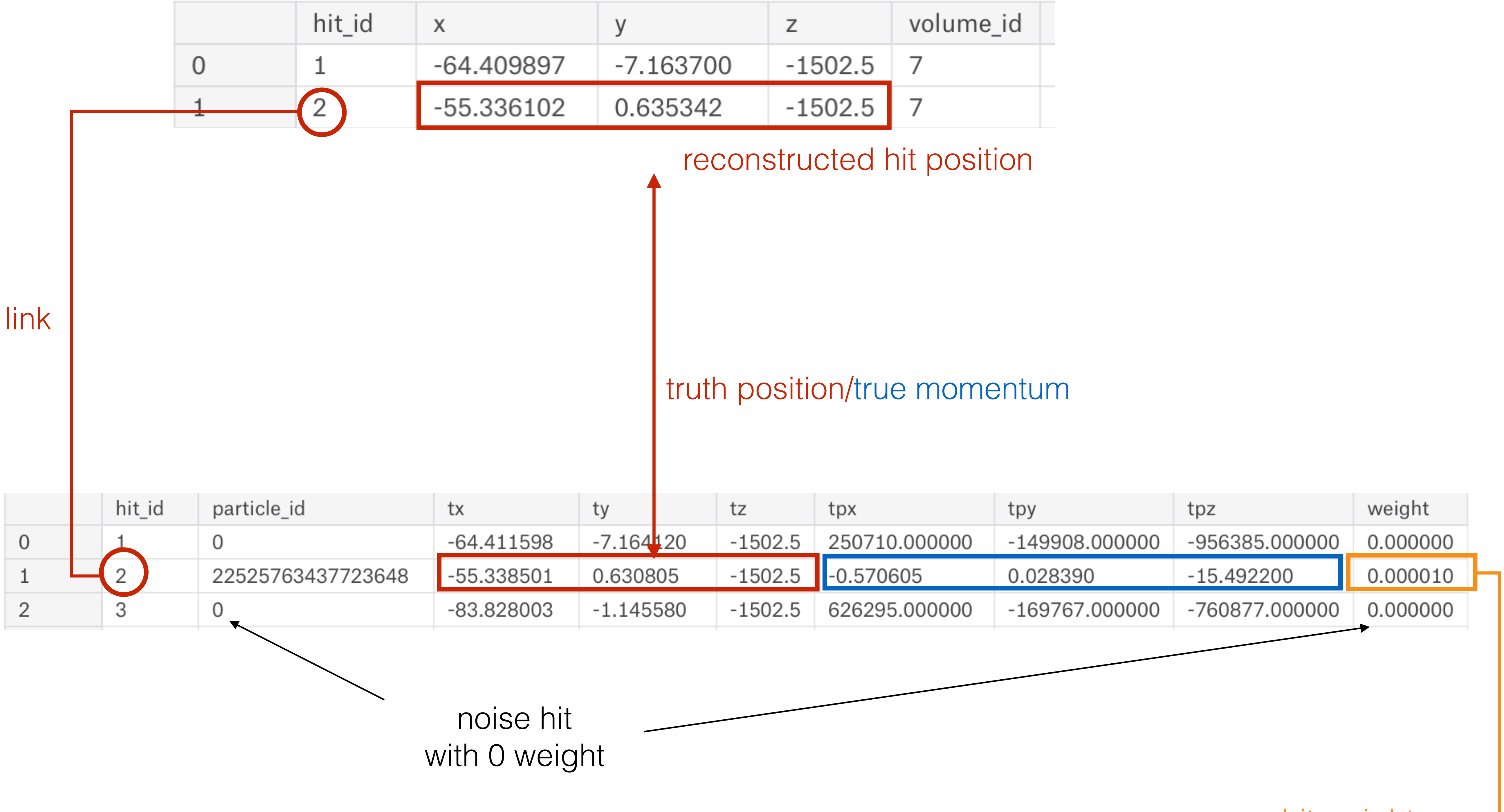


half thickness

plot & image
(top) csv file format for the detector
(bottom) module center and orientation

The training **dataset** - `eventXXXX-truth.csv`

hits:



	hit_id	x	y	z	volume_id
0	1	-64.409897	-7.163700	-1502.5	7
1	2	-55.336102	0.635342	-1502.5	7

	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	0	-64.411598	-7.164120	-1502.5	250710.000000	-149908.000000	-956385.000000	0.000000
1	2	22525763437723648	-55.338501	0.630805	-1502.5	-0.570605	0.028390	-15.492200	0.000010
2	3	0	-83.828003	-1.145580	-1502.5	626295.000000	-169767.000000	-760877.000000	0.000000

reconstructed hit position

link

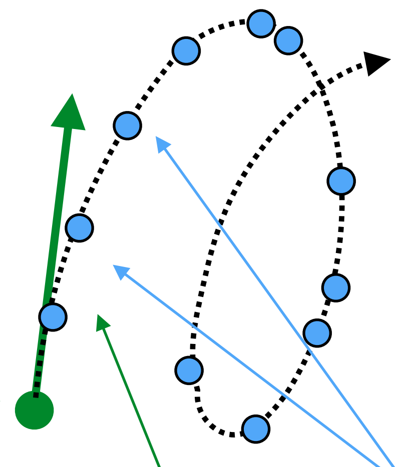
truth position/true momentum

noise hit
with 0 weight

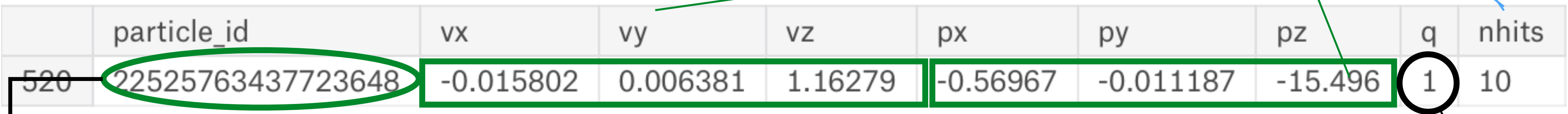
hit weight
for scoring (see later)

tables
(top) csv file format for the hit file
(bottom) csv file format for the truth file

The training **dataset** - `eventXXXX-particles.csv`

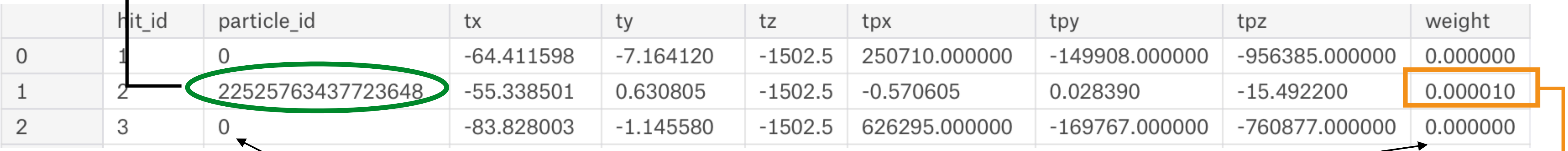


charge



	particle_id	vx	vy	vz	px	py	pz	q	nhits
520	22525763437723648	-0.015802	0.006381	1.16279	-0.56967	-0.011187	-15.496	1	10

link



	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	0	-64.411598	-7.164120	-1502.5	250710.000000	-149908.000000	-956385.000000	0.000000
1	2	22525763437723648	-55.338501	0.630805	-1502.5	-0.570605	0.028390	-15.492200	0.000010
2	3	0	-83.828003	-1.145580	-1502.5	626295.000000	-169767.000000	-760877.000000	0.000000

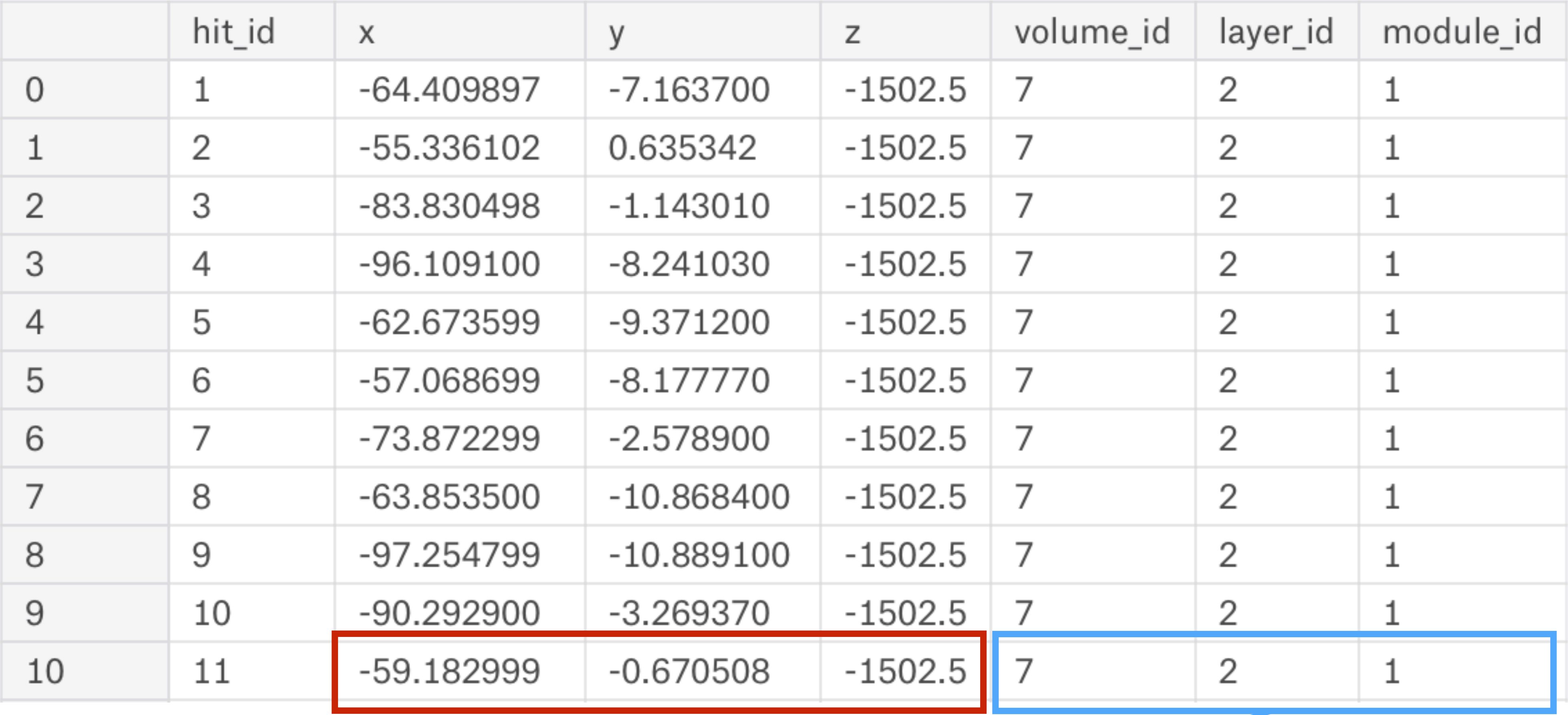
noise hit
with 0 weight

hit weight
for scoring (see later)

tables
(top) csv file format for the particle file
(bottom) csv file format for the truth file

The training **dataset** - `eventXXXX-hits.csv`

NeurIPS 2018 | A. Salzburger



	hit_id	x	y	z	volume_id	layer_id	module_id
0	1	-64.409897	-7.163700	-1502.5	7	2	1
1	2	-55.336102	0.635342	-1502.5	7	2	1
2	3	-83.830498	-1.143010	-1502.5	7	2	1
3	4	-96.109100	-8.241030	-1502.5	7	2	1
4	5	-62.673599	-9.371200	-1502.5	7	2	1
5	6	-57.068699	-8.177770	-1502.5	7	2	1
6	7	-73.872299	-2.578900	-1502.5	7	2	1
7	8	-63.853500	-10.868400	-1502.5	7	2	1
8	9	-97.254799	-10.889100	-1502.5	7	2	1
9	10	-90.292900	-3.269370	-1502.5	7	2	1
10	11	-59.182999	-0.670508	-1502.5	7	2	1

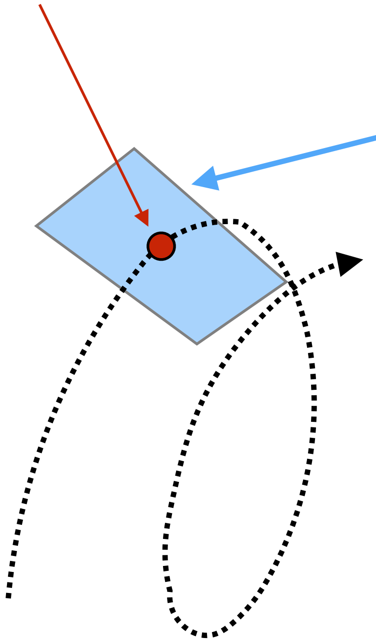


table & images

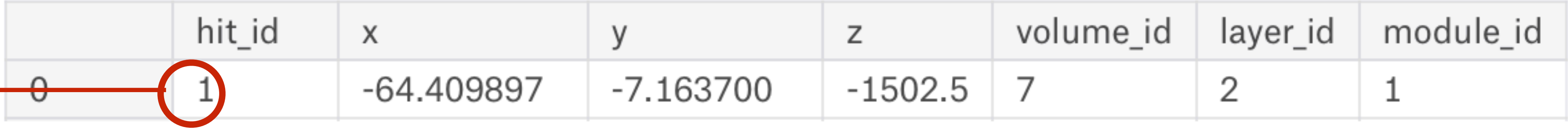
(top) csv file format for the hit file

(bottom) illustration of the hit information

The training **dataset** - `eventXXXX-cells.csv`

NeurIPS 2018 | A. Salzburger

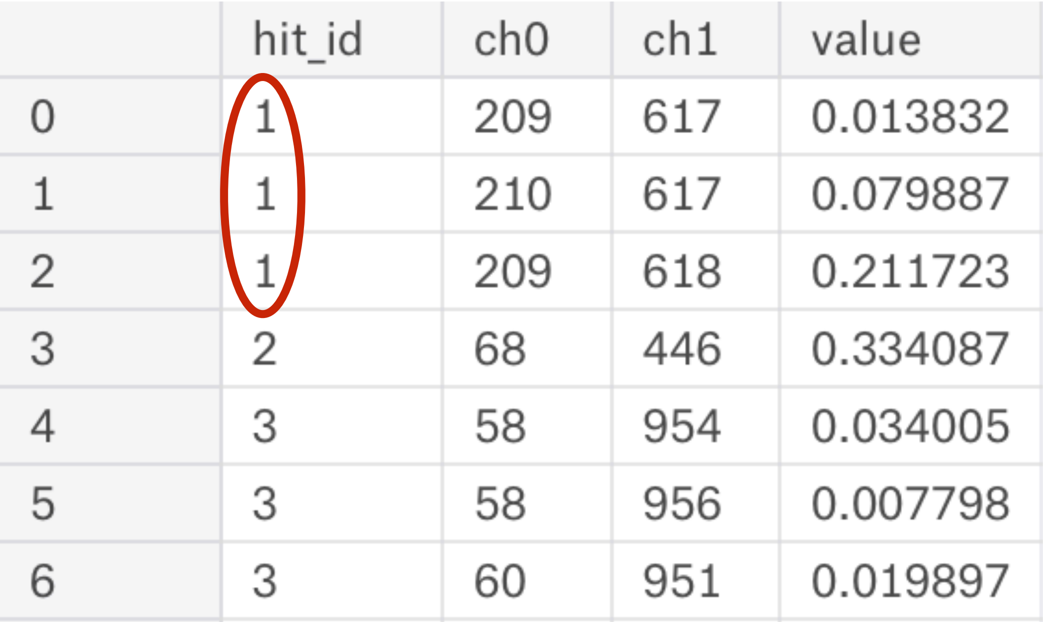
hits:



	hit_id	x	y	z	volume_id	layer_id	module_id
0	1	-64.409897	-7.163700	-1502.5	7	2	1

and cells:

link



	hit_id	ch0	ch1	value
0	1	209	617	0.013832
1	1	210	617	0.079887
2	1	209	618	0.211723
3	2	68	446	0.334087
4	3	58	954	0.034005
5	3	58	956	0.007798
6	3	60	951	0.019897

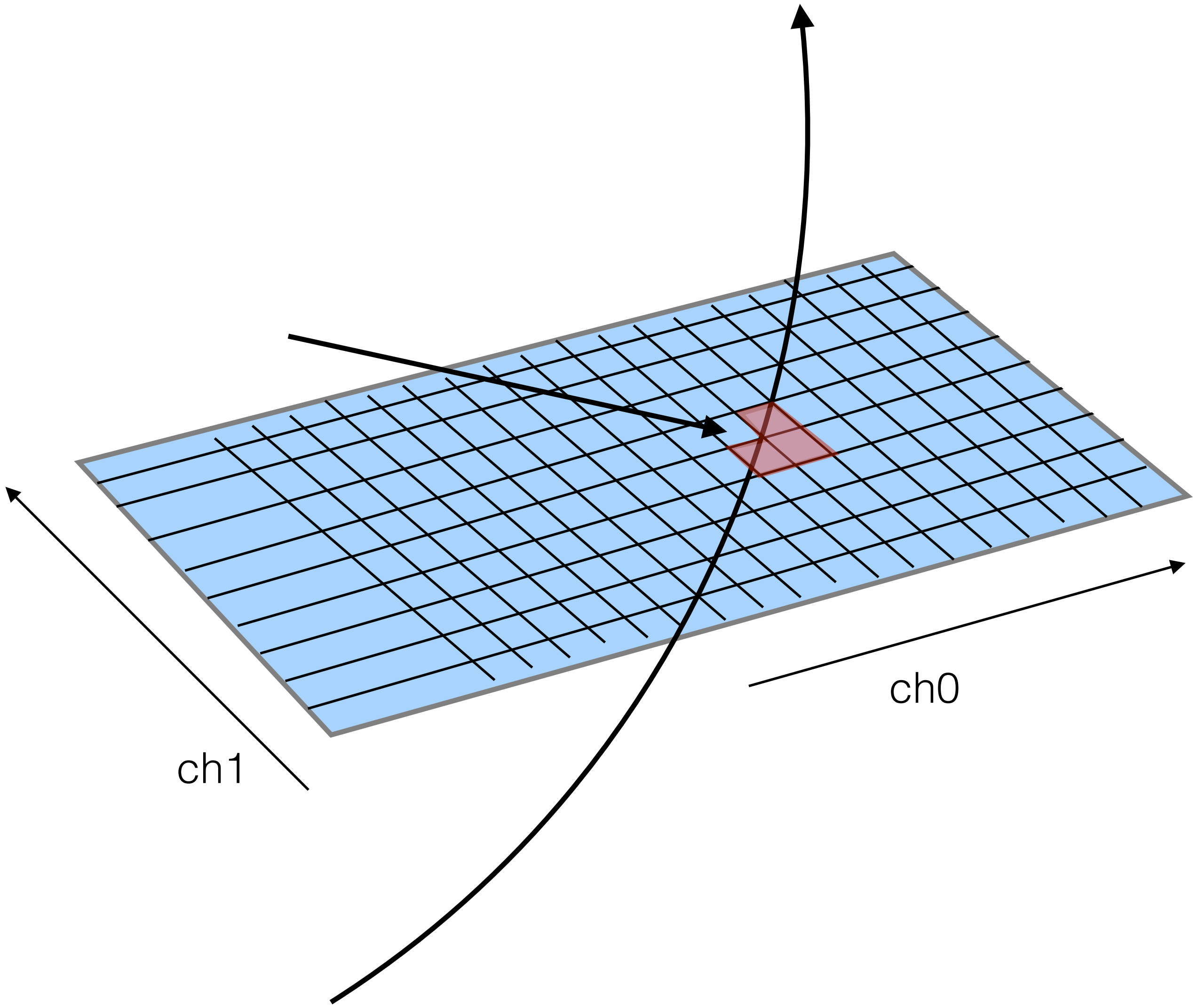
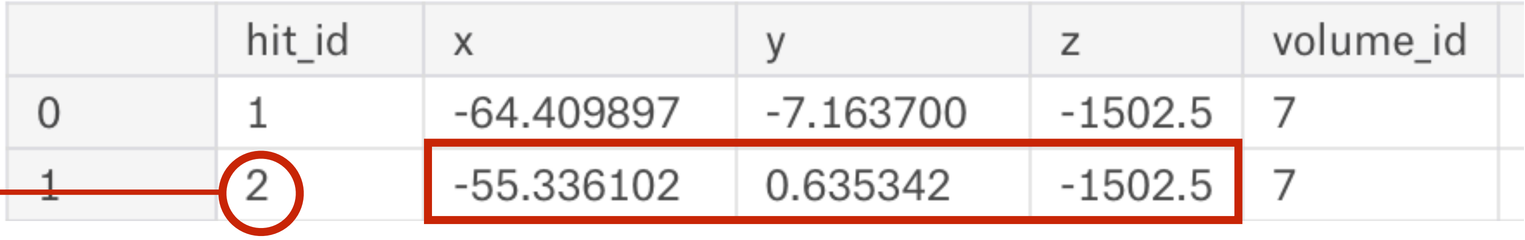


table & images
(top) csv file format for the hit file
(bottom left) csv file format of the cells information
(bottom right) cell information illustration

The training **dataset** - eventXXXX-truth.csv

hits:

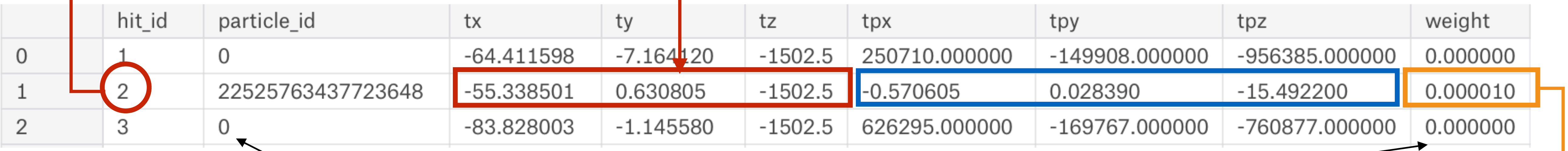


	hit_id	x	y	z	volume_id
0	1	-64.409897	-7.163700	-1502.5	7
1	2	-55.336102	0.635342	-1502.5	7

reconstructed hit position

link

truth position/true momentum



	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	0	-64.411598	-7.164120	-1502.5	250710.000000	-149908.000000	-956385.000000	0.000000
1	2	22525763437723648	-55.338501	0.630805	-1502.5	-0.570605	0.028390	-15.492200	0.000010
2	3	0	-83.828003	-1.145580	-1502.5	626295.000000	-169767.000000	-760877.000000	0.000000

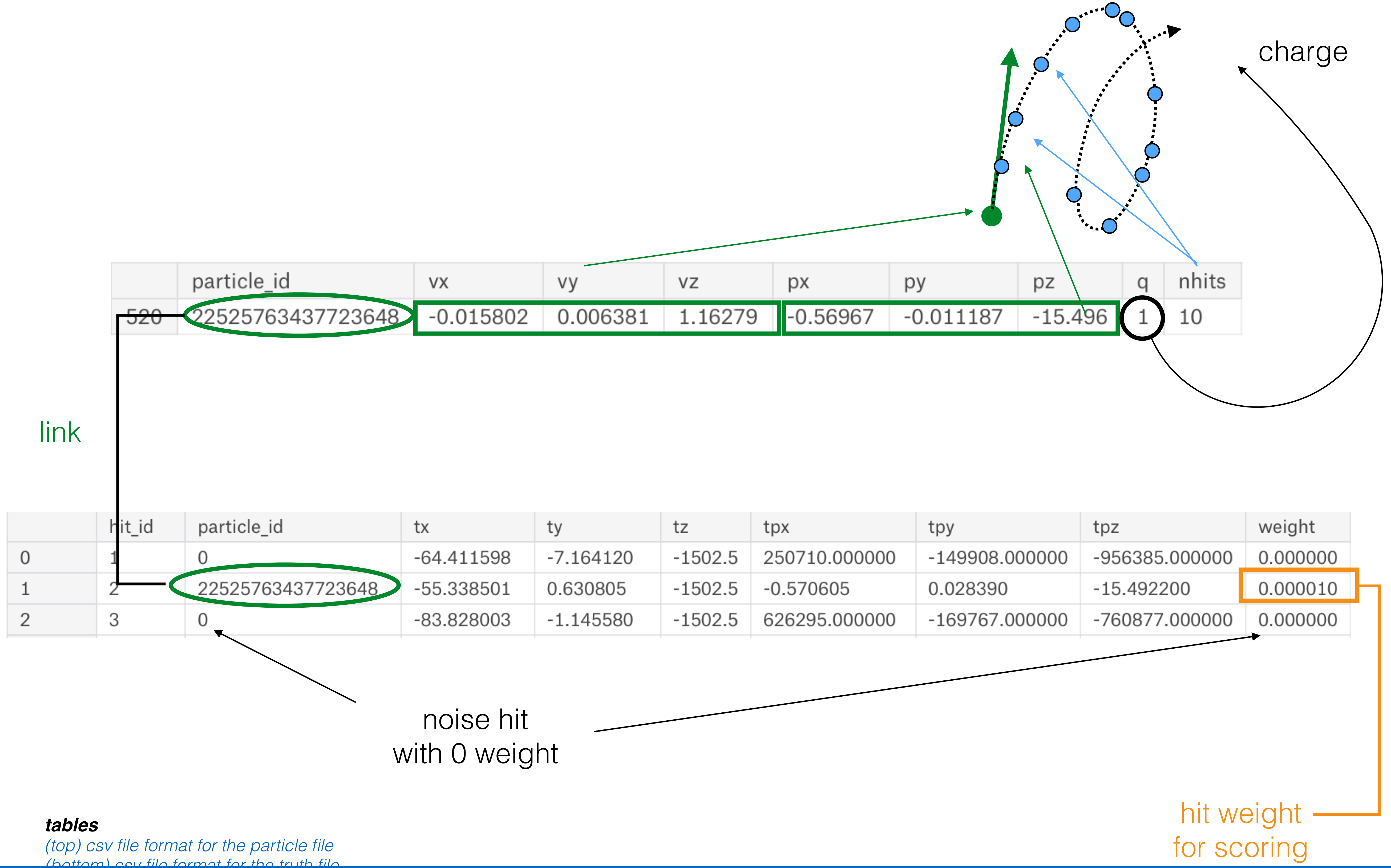
noise hit
with 0 weight

hit weight
for scoring (see later)

tables
(top) csv file format for the hit file
(bottom) csv file format for the truth file

The training dataset - `eventXXXX-particles.csv`

NeurIPS 2018 | A. Salzburger



	particle_id	vx	vy	vz	px	py	pz	q	nhits
520	22525763437723648	-0.015802	0.006381	1.16279	-0.56967	-0.011187	-15.496	1	10

	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	0	-64.411598	-7.164120	-1502.5	250710.000000	-149908.000000	-956385.000000	0.000000
1	2	22525763437723648	-55.338501	0.630805	-1502.5	-0.570605	0.028390	-15.492200	0.000010
2	3	0	-83.828003	-1.145580	-1502.5	626295.000000	-169767.000000	-760877.000000	0.000000

tables
(top) csv file format for the particle file
(bottom) csv file format for the truth file

TrackML dataset heavily in use

Hep.TrkX & Exa.TrkX project

- GNN full scale ML for high energy physics
- talk by Jean-Roch tomorrow

Hep.QPR project

- quantum annealing on D-wave
- talk by Jean-Roch today

Annoy & hashing

- Unsupervised learning with Spotify
- talk by Sabrina today

Various different other ML research

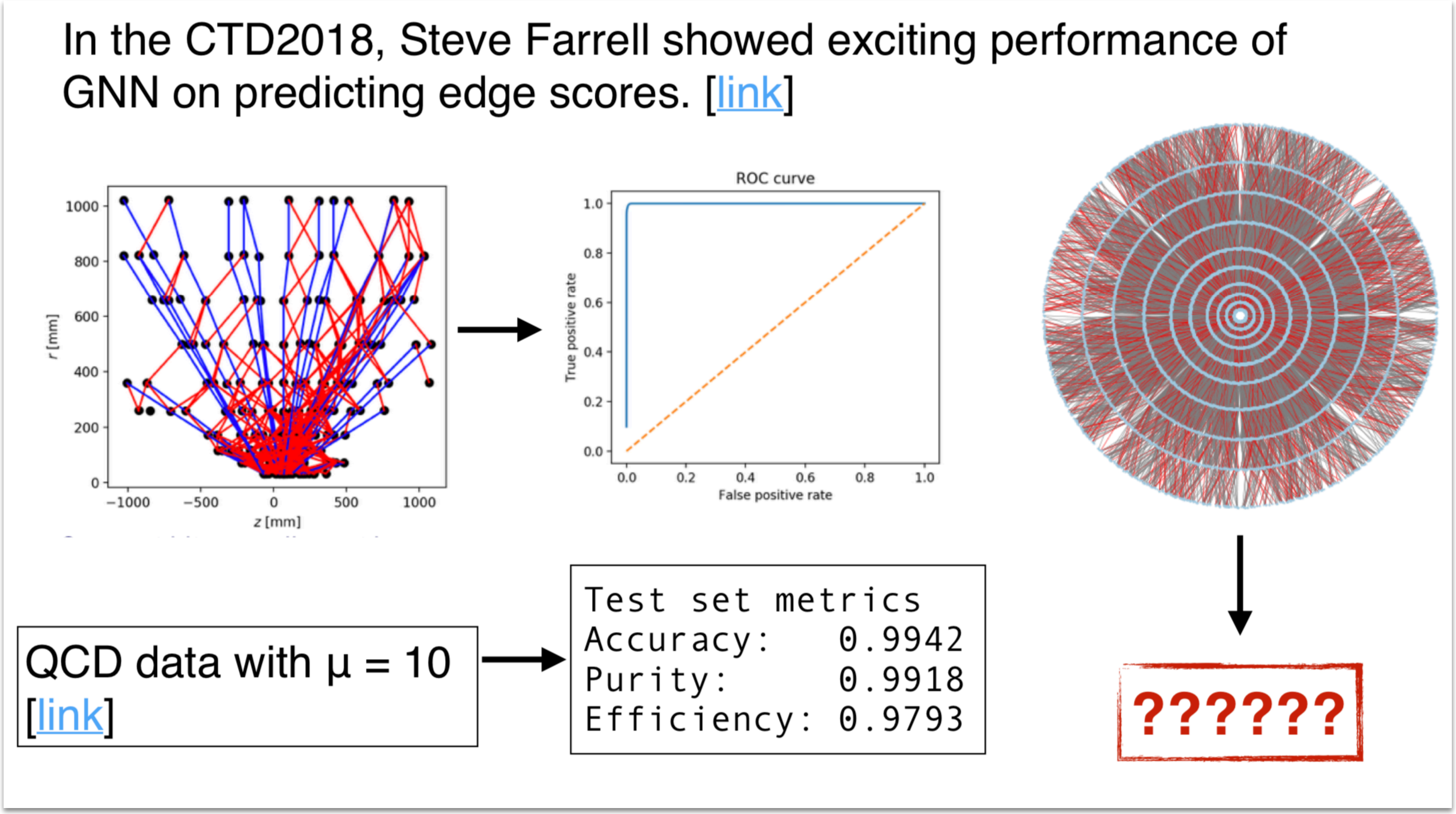
- Track seed classification using NNs

Track reconstruction algorithm templating

[See CTD 2019 contributions]

CTD Highlights Hep.TrkX & Exa.TrkX

Tracking ML challenge data



In the CTD2018, Steve Farrell showed exciting performance of GNN on predicting edge scores. [link]

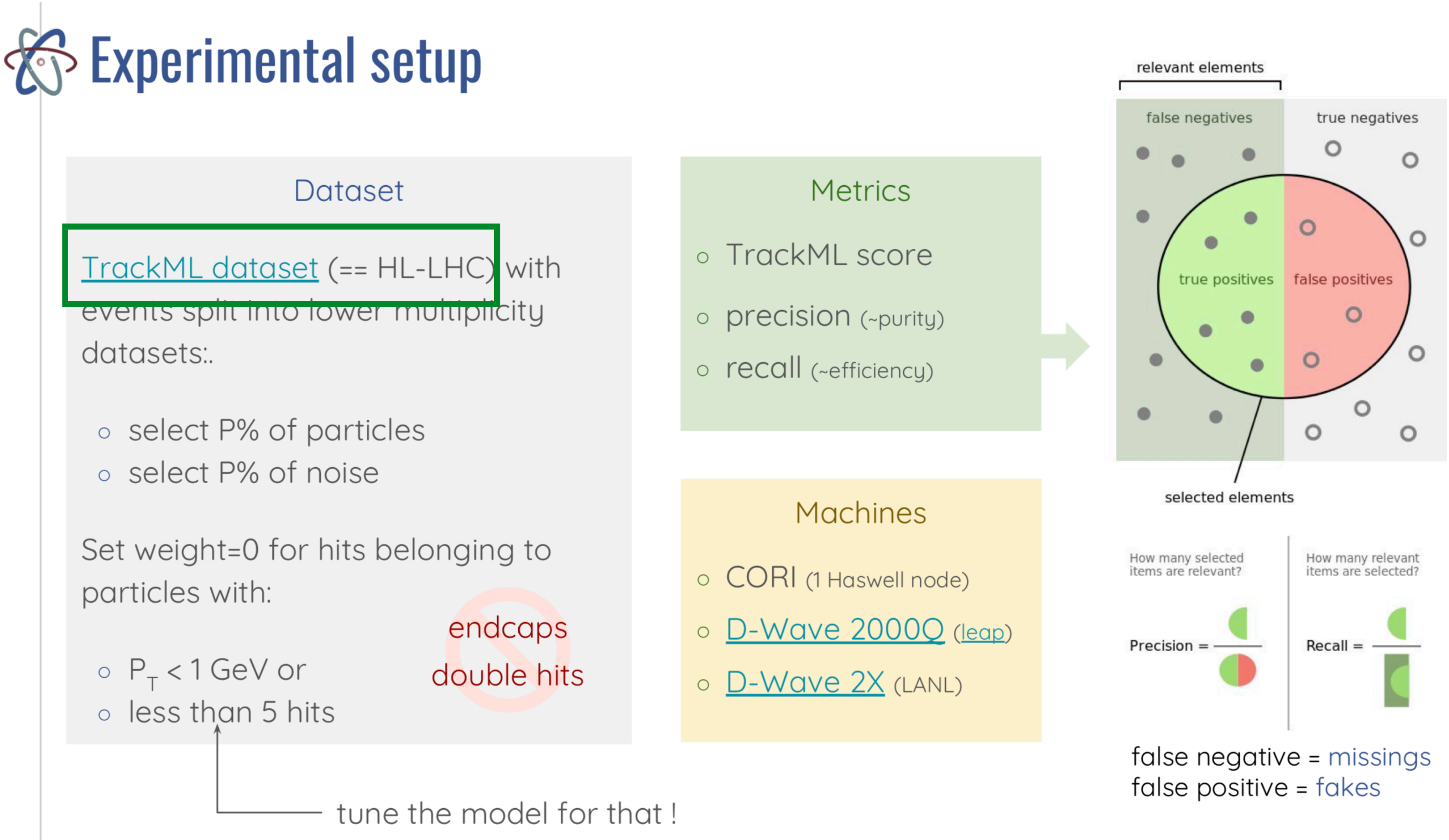
QCD data with μ = 10 [link]

```
Test set metrics
Accuracy:   0.9942
Purity:     0.9918
Efficiency: 0.9793
```

??????

talk by Jean-Roch

CTD Highlights Hep.QPR



Experimental setup

Dataset

TrackML dataset (== HL-LHC) with events split into lower multiplicity datasets:.

- select P% of particles
- select P% of noise

Set weight=0 for hits belonging to particles with:

- $P_T < 1$ GeV or
- less than 5 hits

endcaps
double hits

tune the model for that !

Metrics

- TrackML score
- precision (~purity)
- recall (~efficiency)

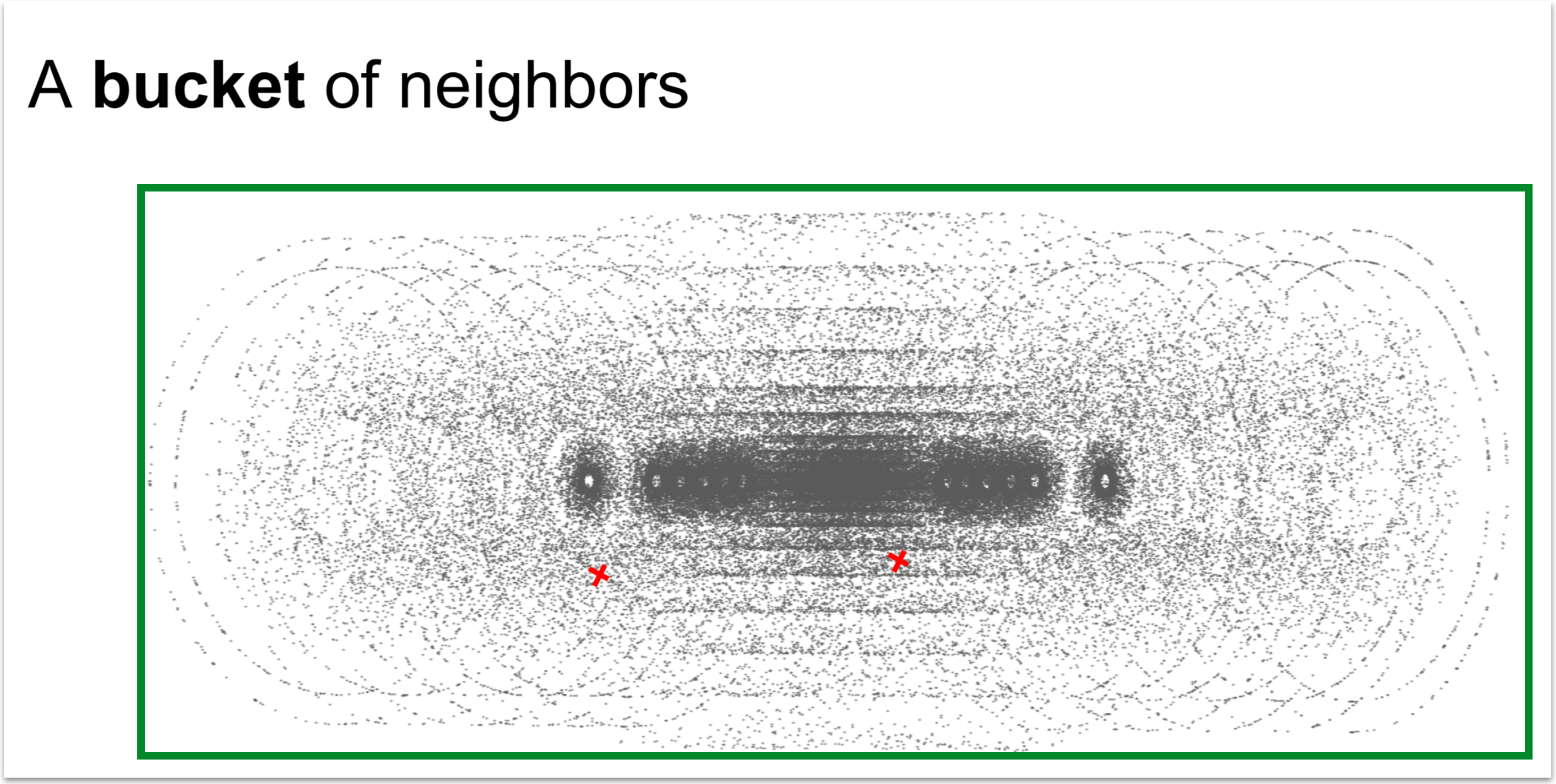
Machines

- CORI (1 Haswell node)
- D-Wave 2000Q (leap)
- D-Wave 2X (LANL)

false negative = missings
false positive = fakes

talk by Jean-Roch

CTD Highlights Annoy & Music

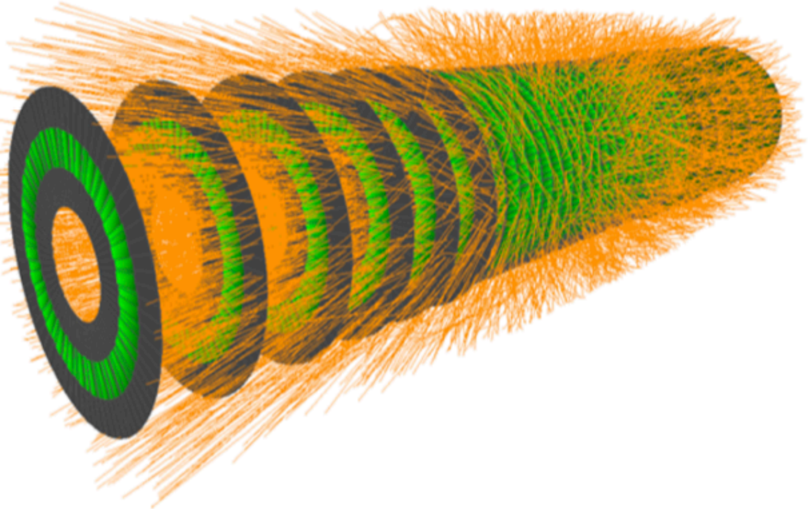


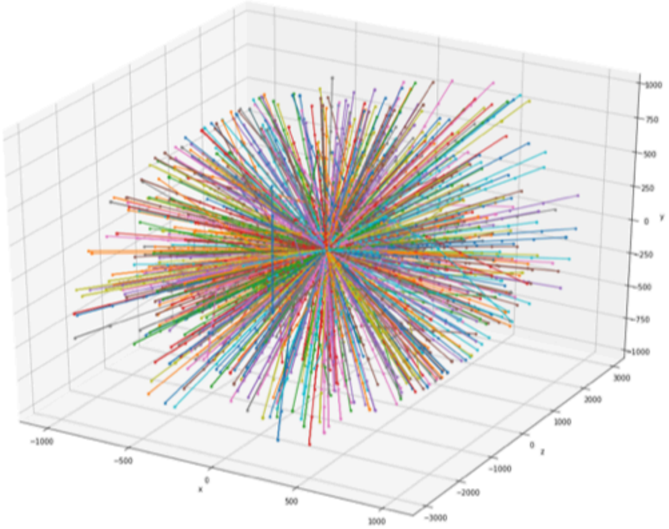
A **bucket** of neighbors

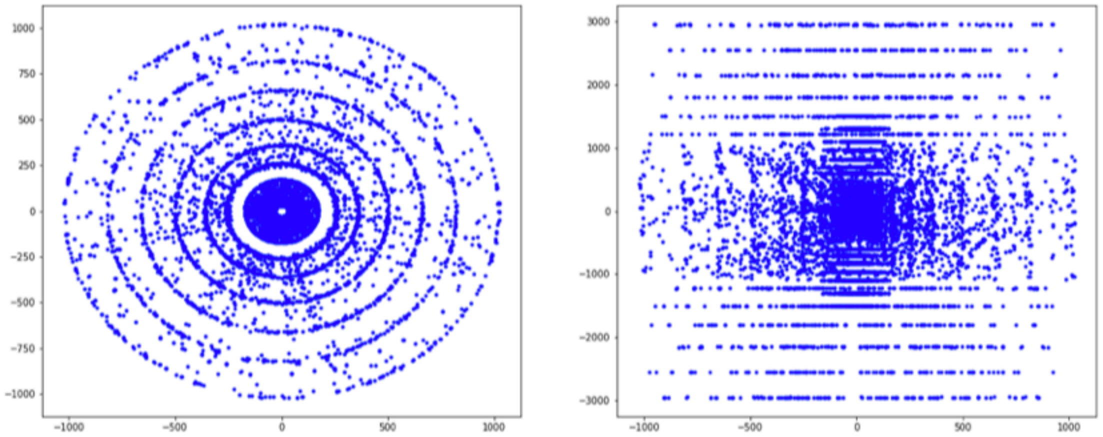
talk by Sabrina

CTD Highlights Seed Classification

- Use TrackML challenge dataset
- Simulated typical silicon HL-LHC detector
- Provides 3d hit coordinates and corresponding ground truth particles
- Generate "false" seeds by randomly interchanging hits







[talk by F. Dietrich]

TrackML dataset heavily in use

Hep.TrkX & Exa.TrkX project

- GNN full scale ML for high energy physics
- talk by Jean-Roch tomorrow

Hep.QPR project

- quantum annealing on D-wave
- talk by Jean-Roch today

Annoy & hashing

- Unsupervised learning with Spotify
- talk by Sabrina today

Various different other ML research

- Track seed classification using NNs

Track reconstruction algorithm templating

TML dataset/detector

- has several shortcoming:

not enough material
too much overlap
only fast simulation
no particle decay

CERN OpenData

CERN OPEN DATA PORTAL

[See CTD 2019 contributions]

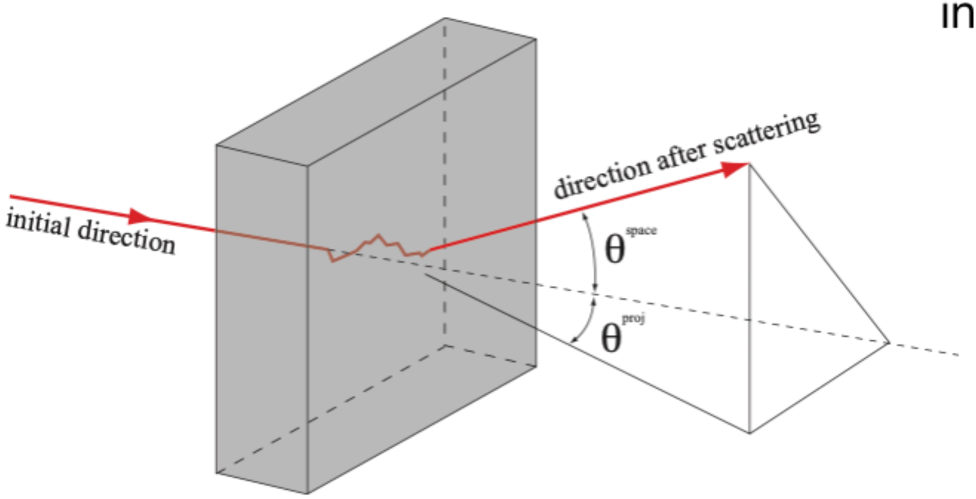
A realistic detector material

Simulating a realistic detector

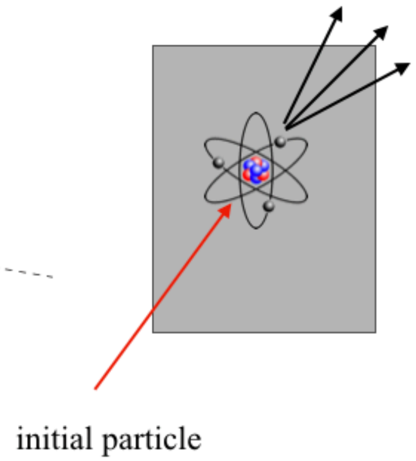
- excellent description of the material
- detailed modelling of the detection process

Determines the amount of “process noise”

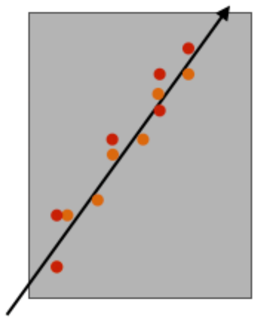
- scattering



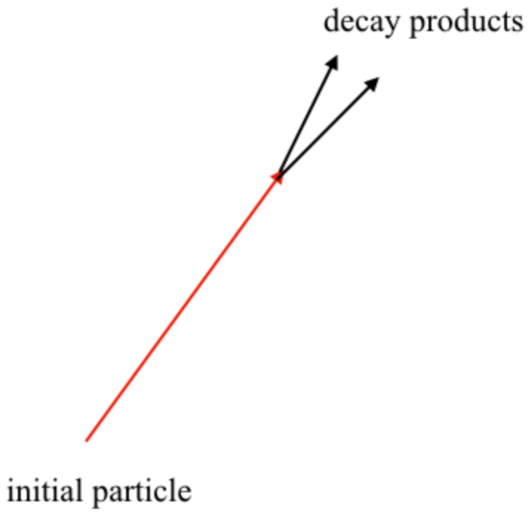
- destructive (hadronic) interaction



- energy loss



- particle decay

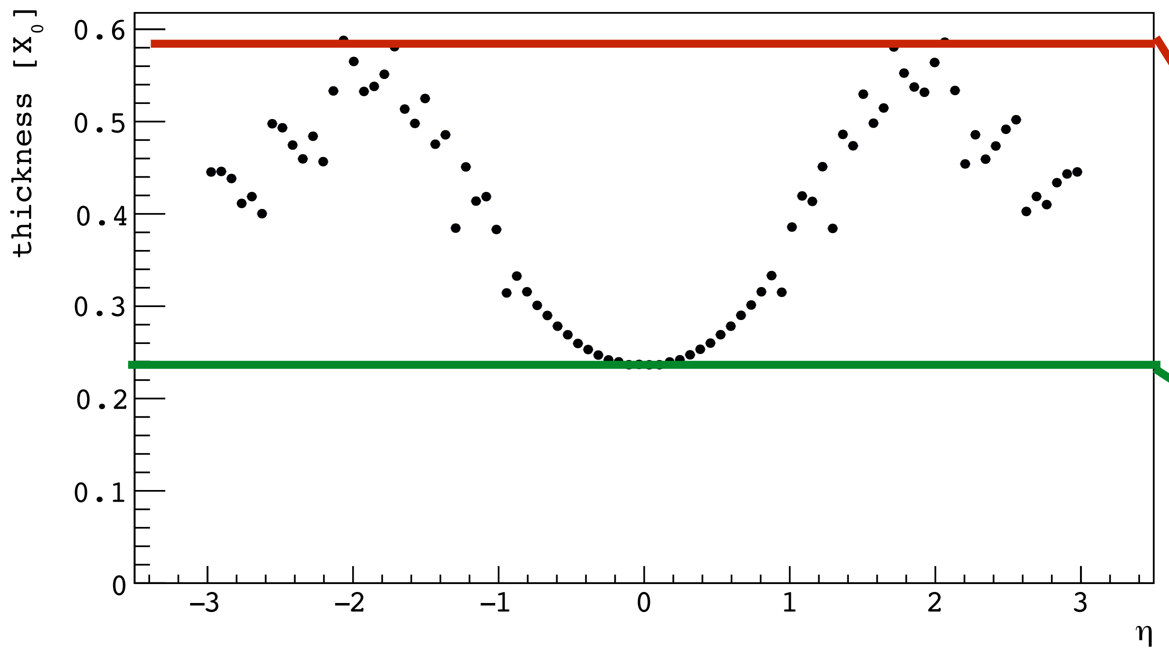


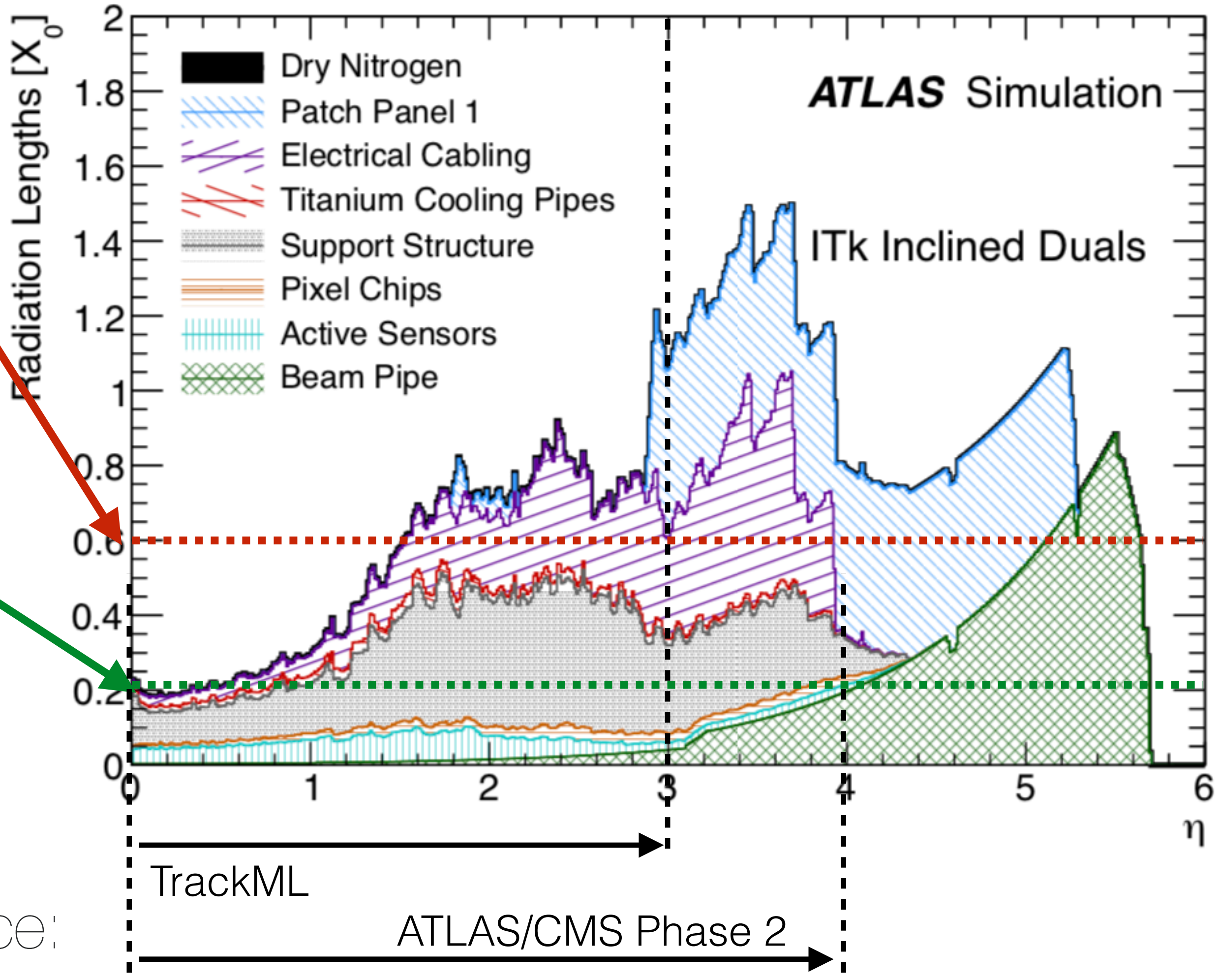
detector material & acceptance

Simulating a realistic detector

- excellent description of the material
- detailed modelling of the detection process

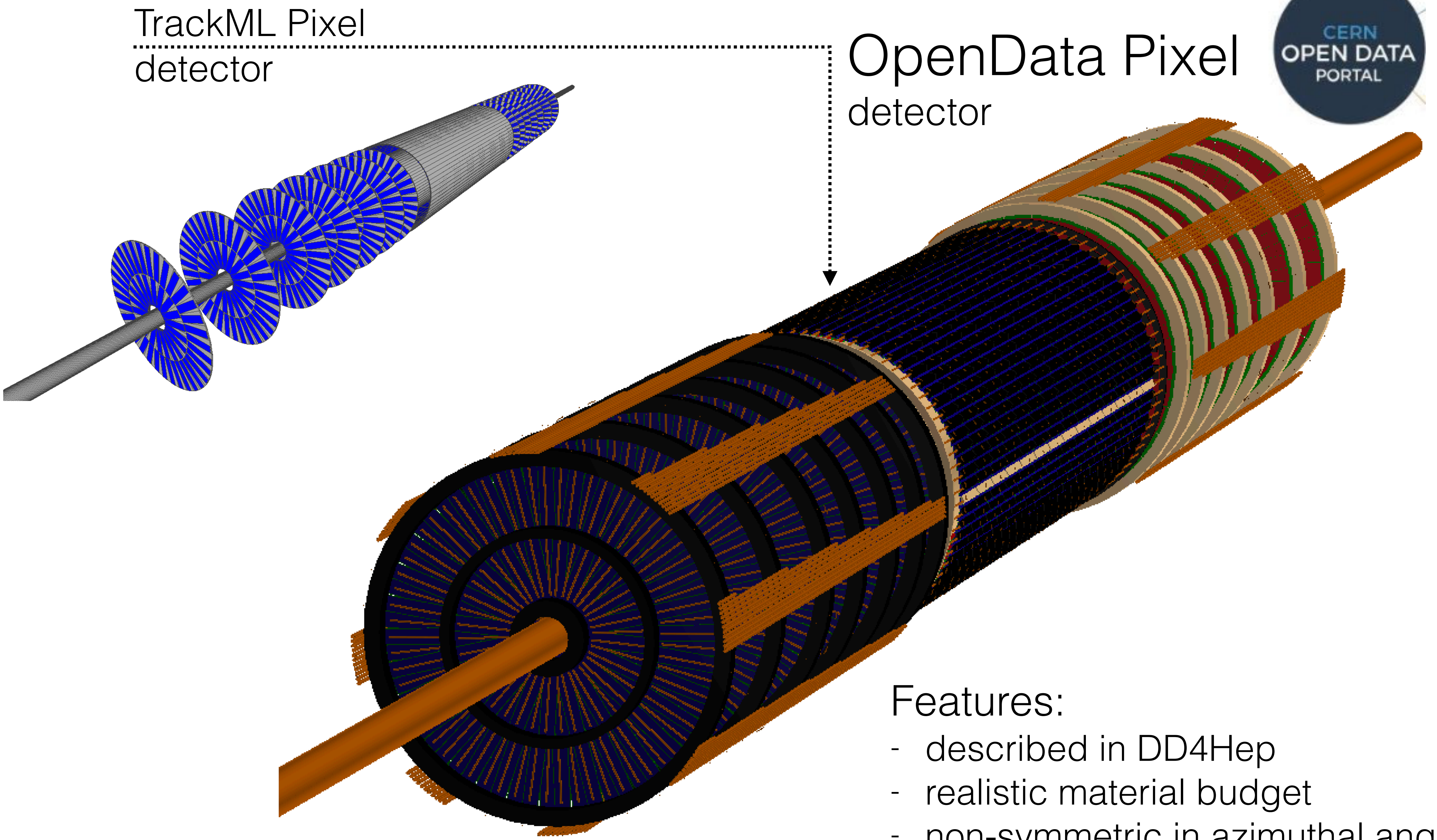
TrackML detector could use a bit of realism





Detector acceptance:

Sneak Preview



TrackML Pixel detector

OpenData Pixel detector

Features:

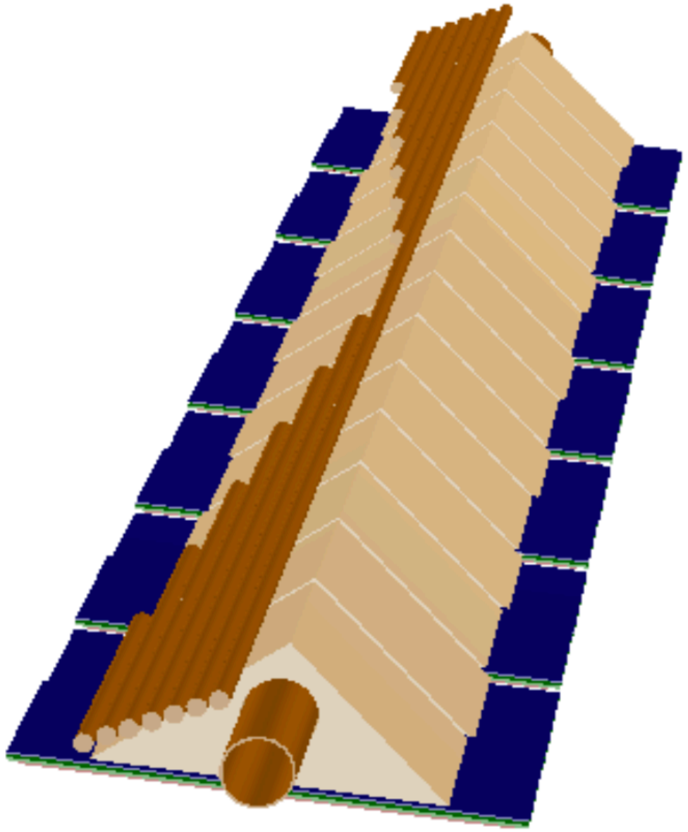
- described in DD4Hep
- realistic material budget
- non-symmetric in azimuthal angle
- full (G4) and fast (ACTS) simulation
- misalignment possibility

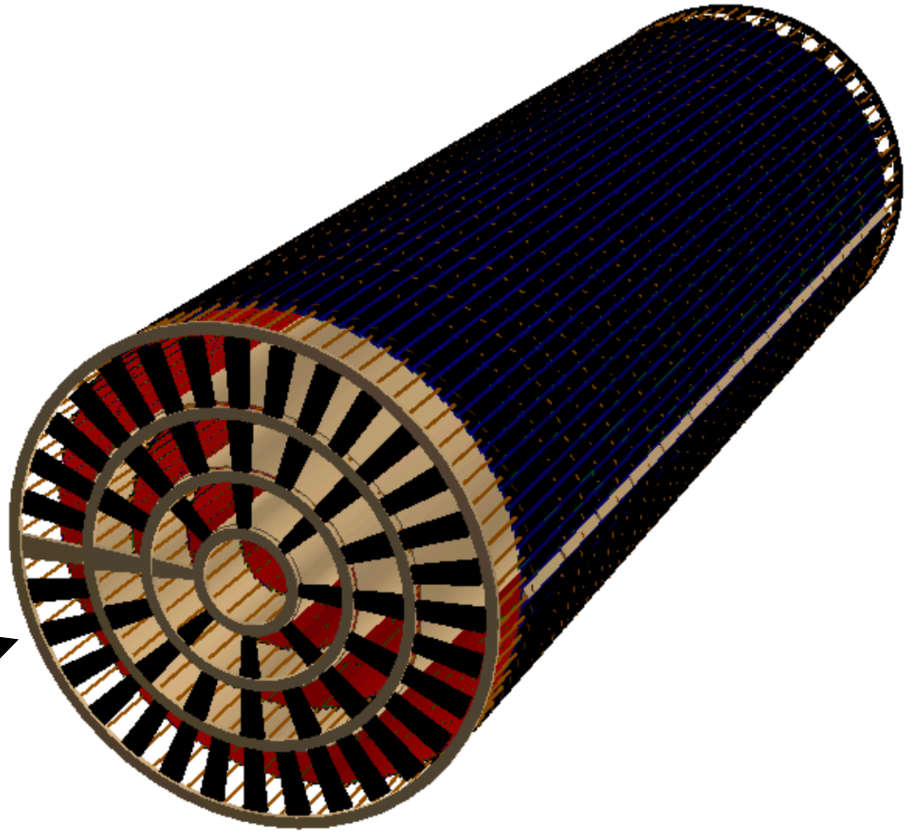
… to be released soon!

OpenData detector

Stave with multiplying services

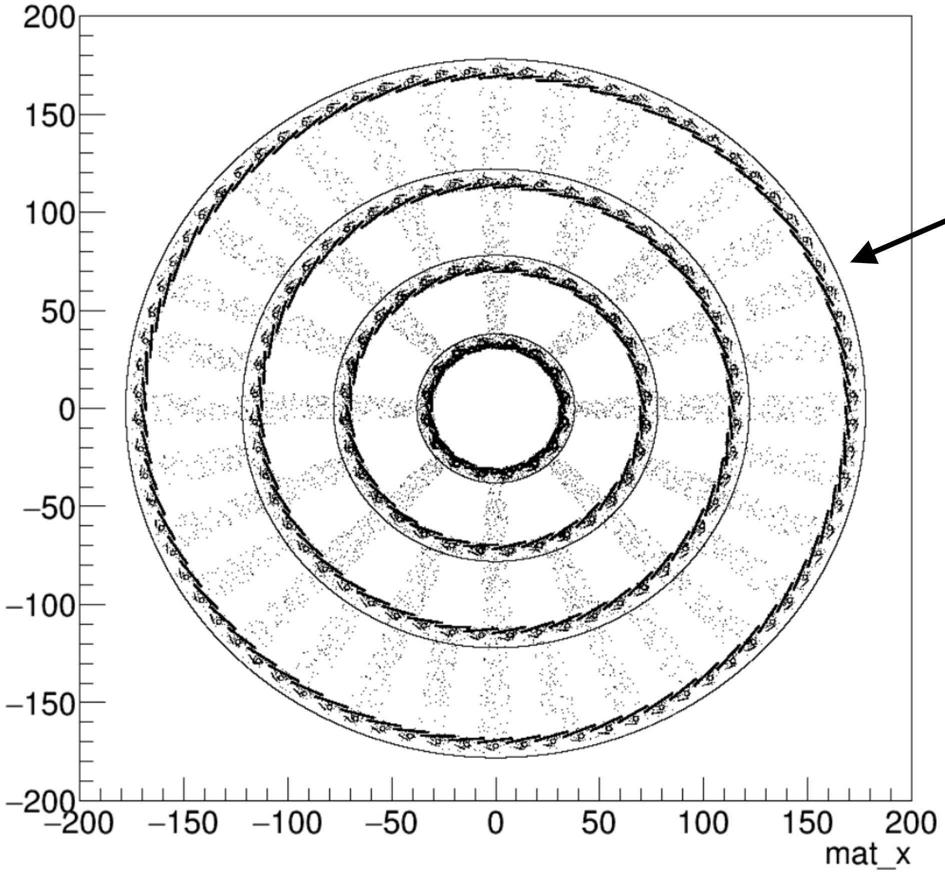
creates as more material a along the staves





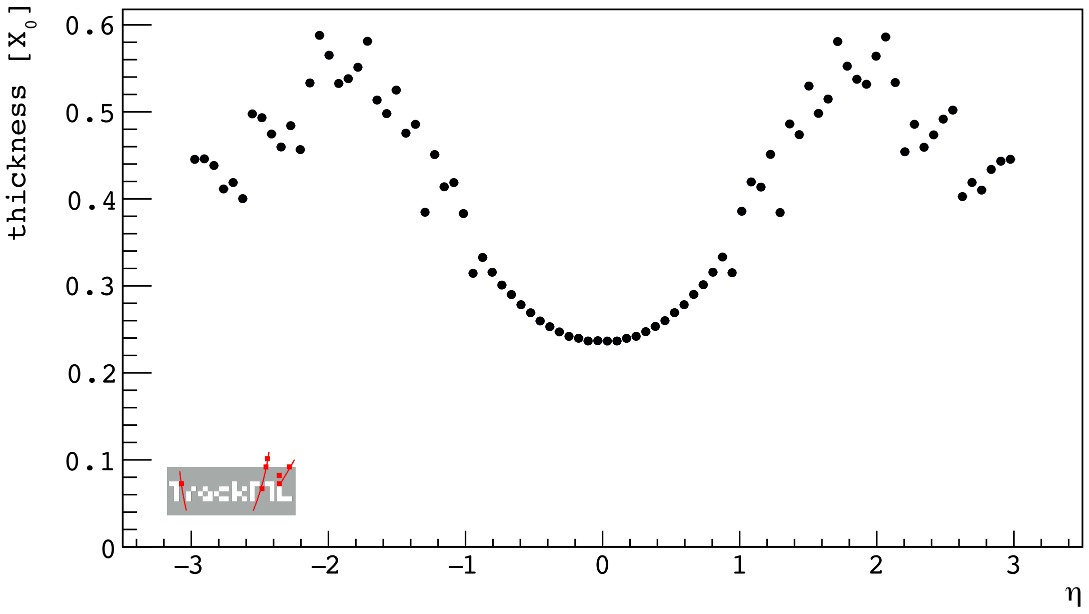
Packed into complicated structure

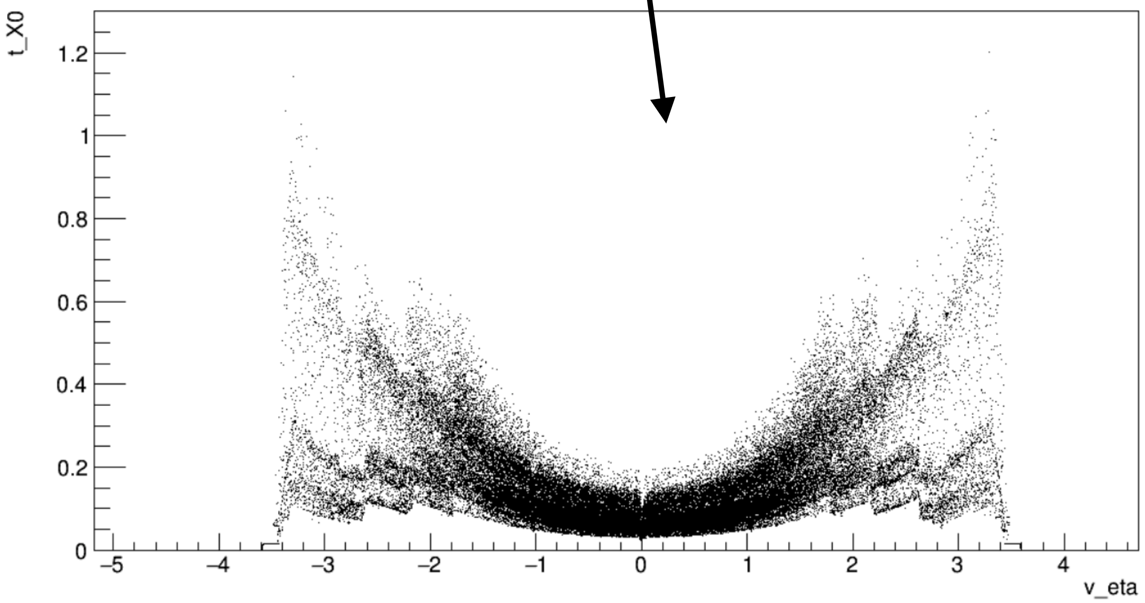
OpenData / Geant4 conversion



Geant4 hit map

Resulting material distribution





detector resolutions

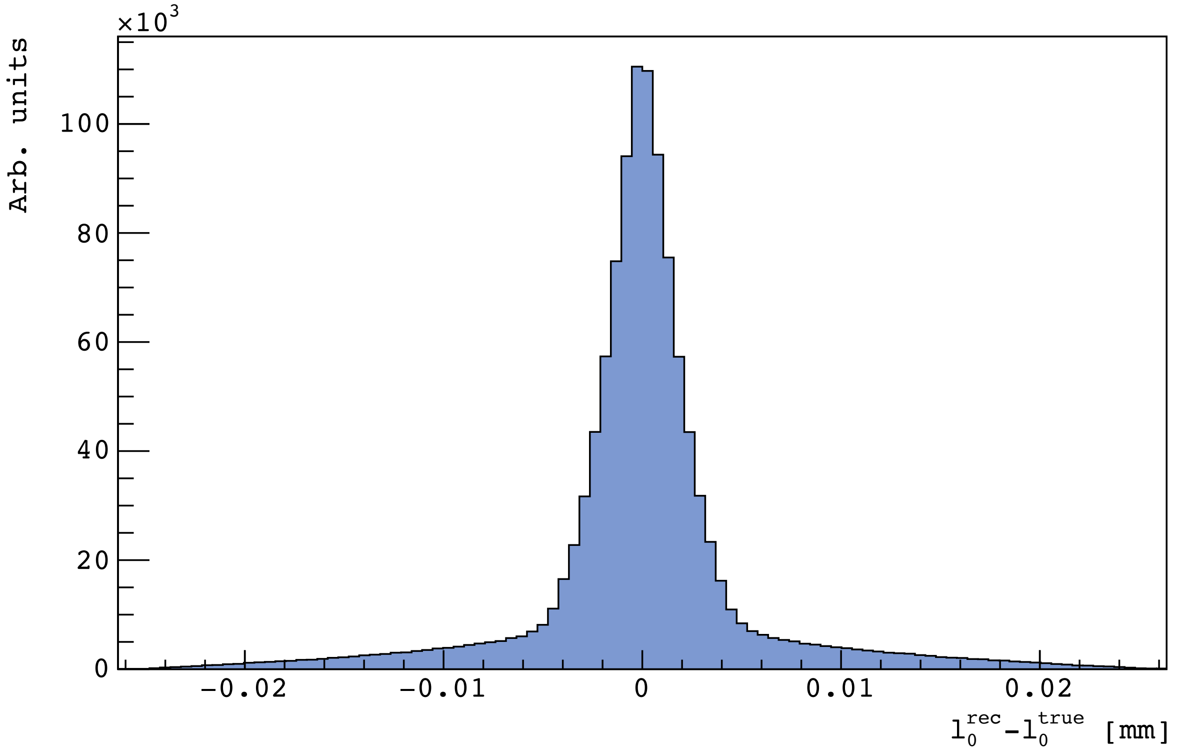
Simulating a realistic detector

- excellent description of the material
- detailed modelling of the detection process

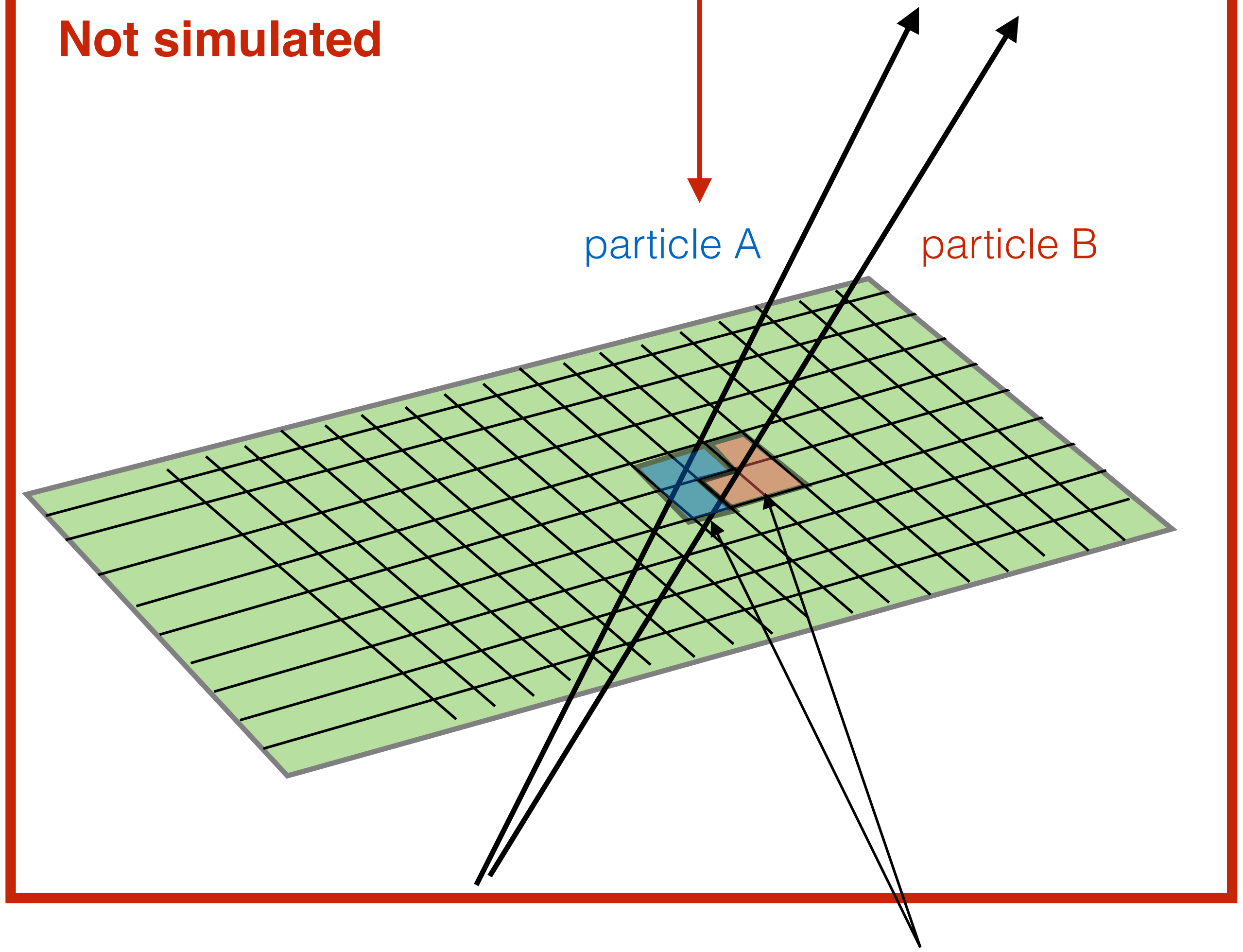
non-gaussian measurement resolution modelled by digitisation

Cluster merging not simulated so far

Simulated



Not simulated



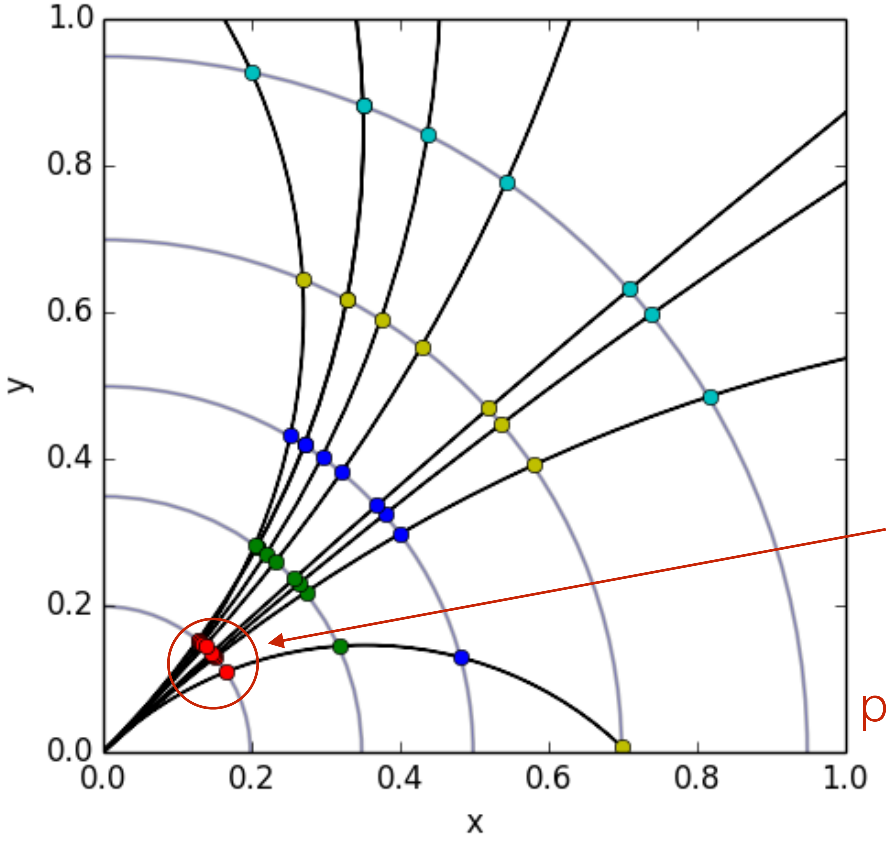
2 measurements simulated

detector resolutions

Simulating a realistic detector

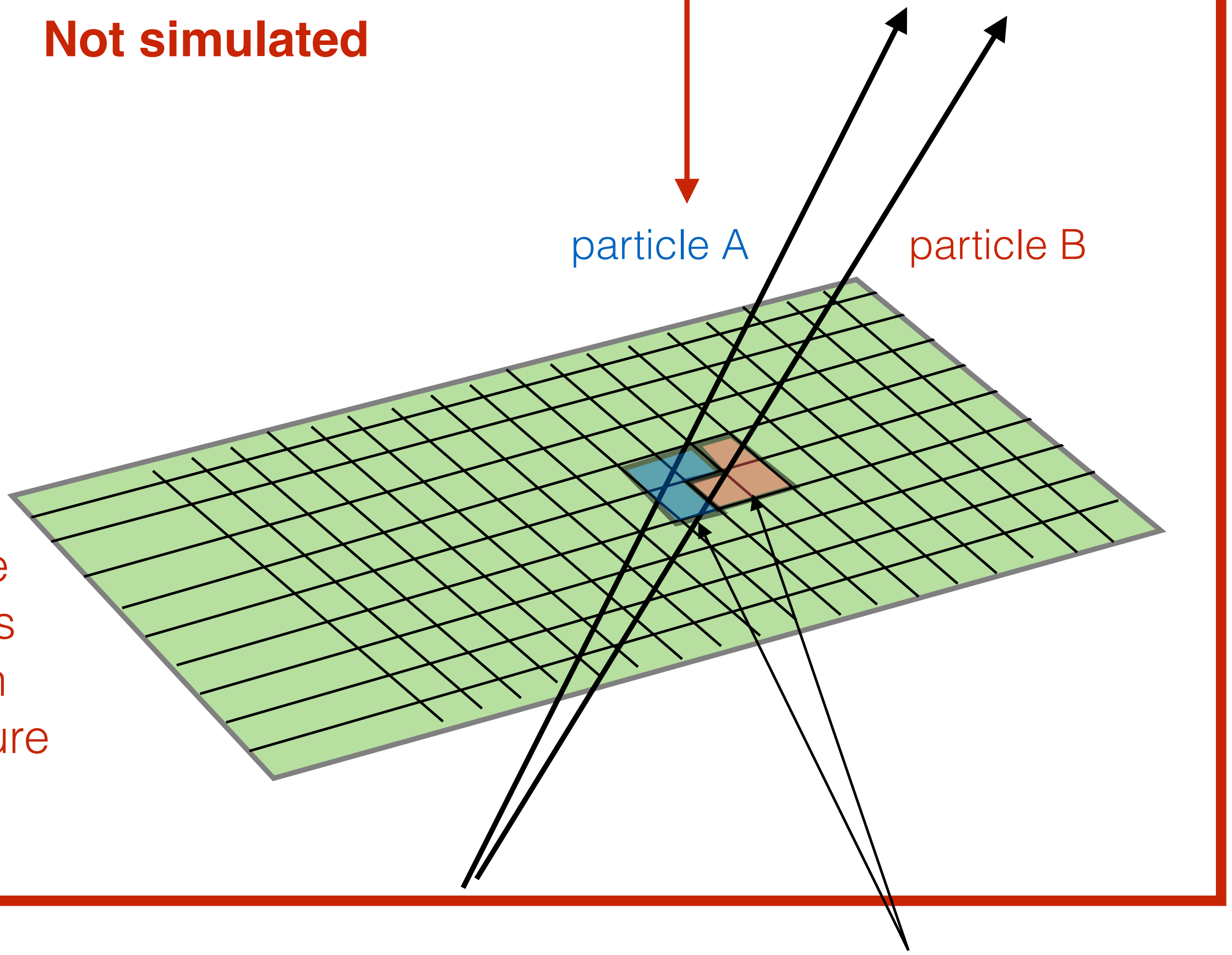
- excellent description of the material
- detailed modelling of the detection process

Cluster merging not simulated so far



Locally dense environment is Very common physics signature

Not simulated



2 measurements simulated

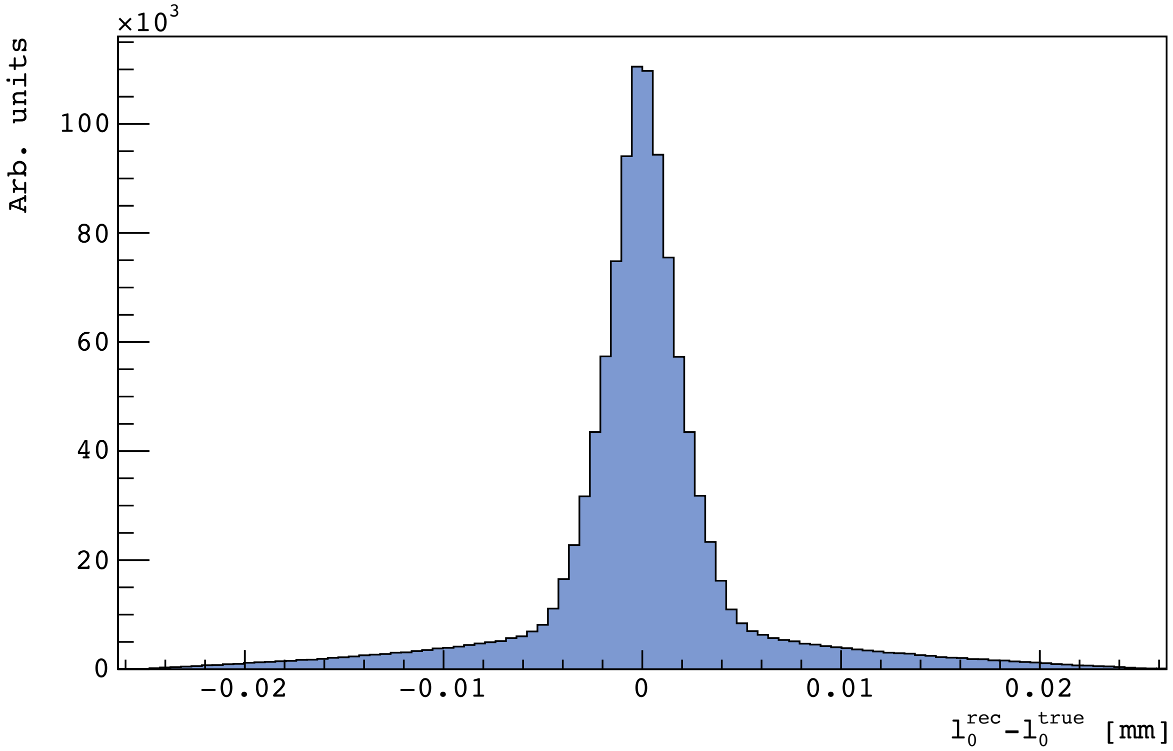
detector resolutions

Simulating a realistic detector

- excellent description of the material
- detailed modelling of the detection process

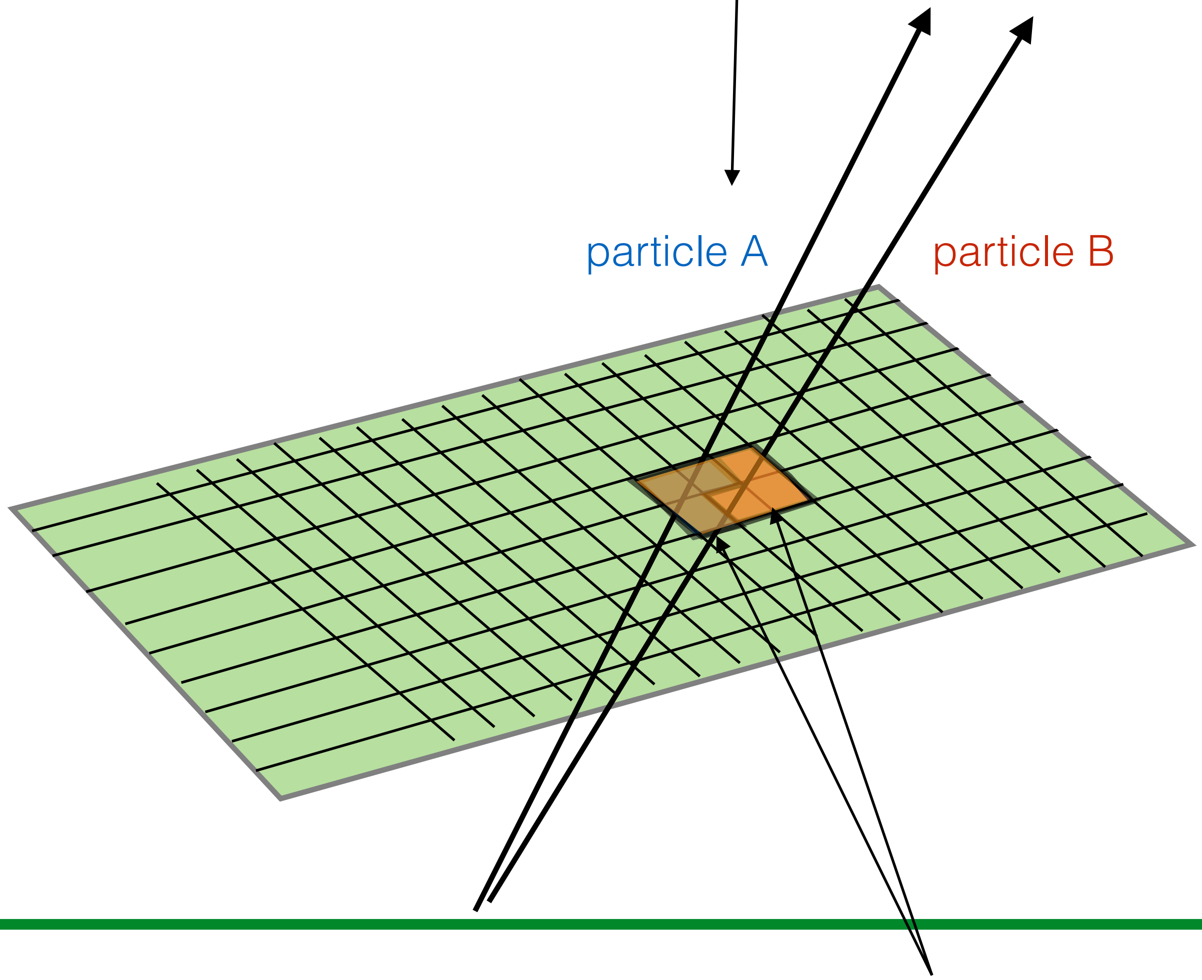
non-gaussian measurement resolution
modelled by digitisation

Simulated



no unique cluster labelling anymore

will have to change the score for this



2 measurements simulated

+ merged into 1 measurement

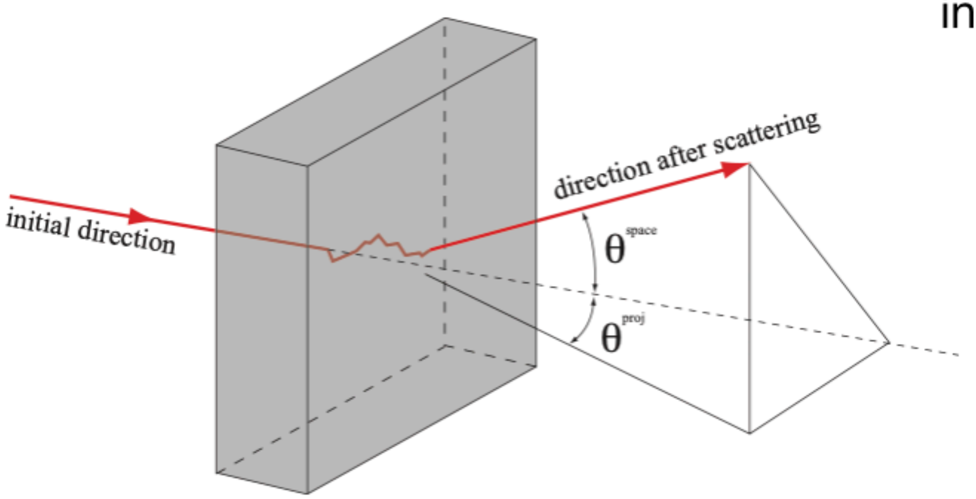
realistic problem physics

Simulating a realistic experiment

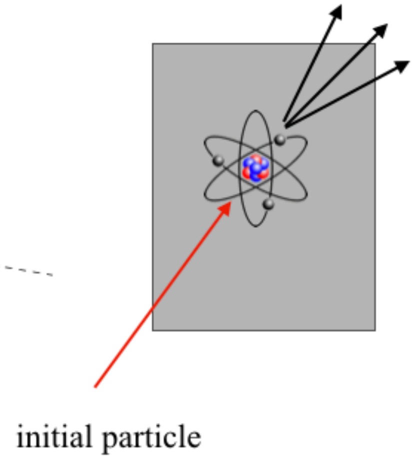
- excellent description of the material
- detailed modelling of the detection process
- detailed simulation of relevant models

simulation of particle decay for testing of large radius tracking

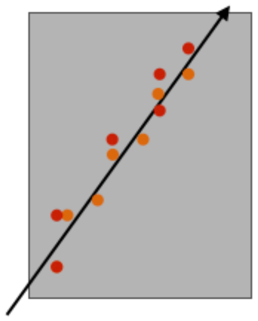
- scattering



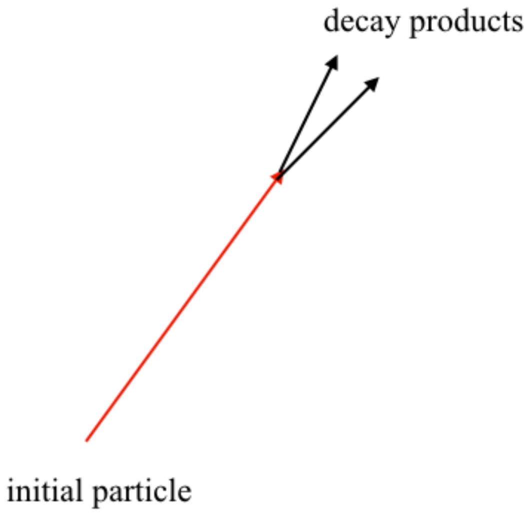
- destructive (hadronic) interaction



- energy loss



- particle decay



The dataset - physics

Pythia configured with:

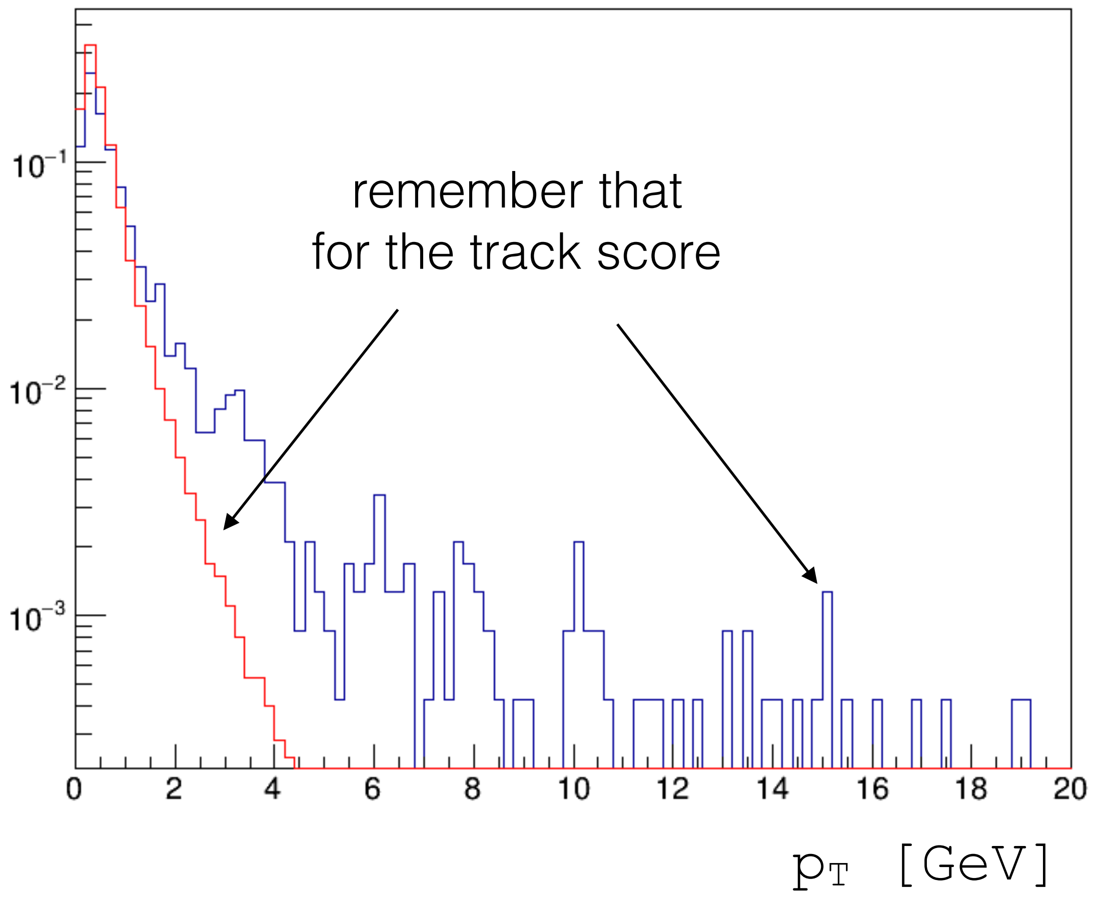
- HS: **“Top:gg2ttbar = on”**
- PU (@200): **“SoftQCD = on”**

Smeared beam spot

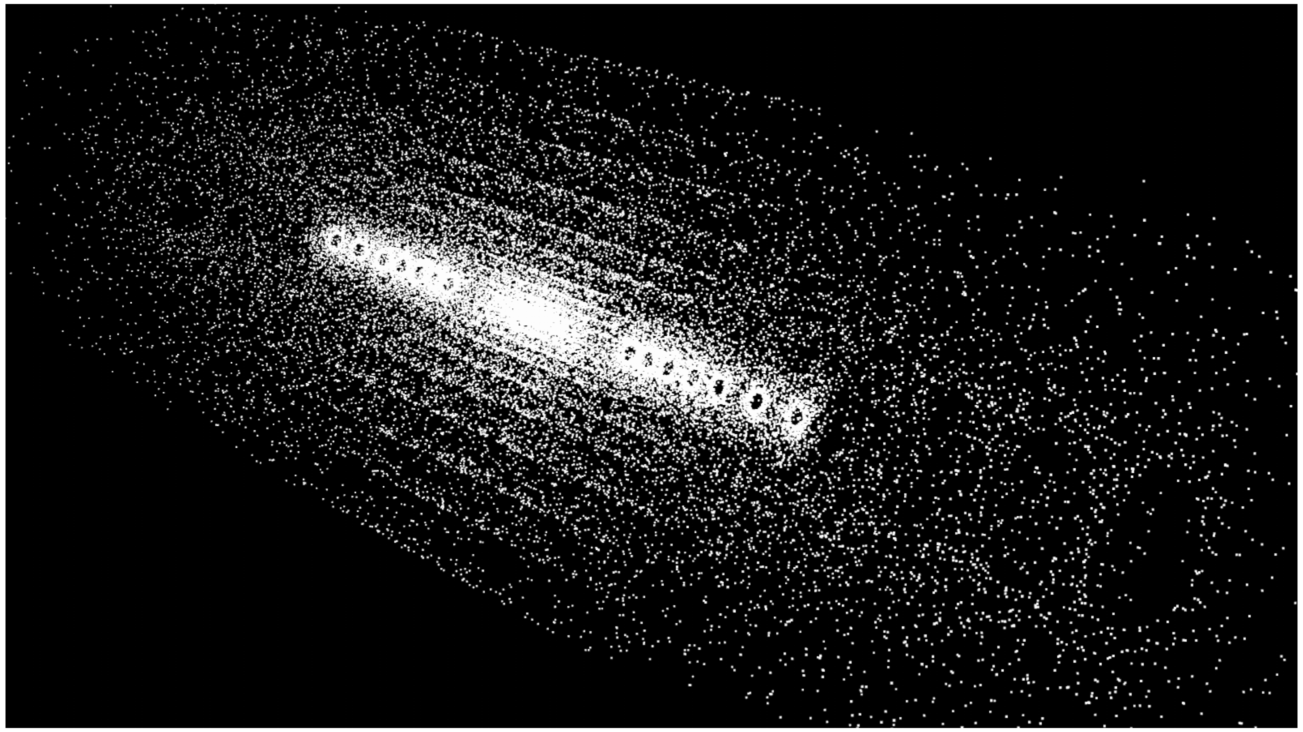
- σ_Z = 5.5 mm, σ_T = 15 µm

Charged particles are simulated

- p_T > 150 MeV



remember that
for the track score



large benchmark dataset (100s Gb)
to be released as CERN OpenData

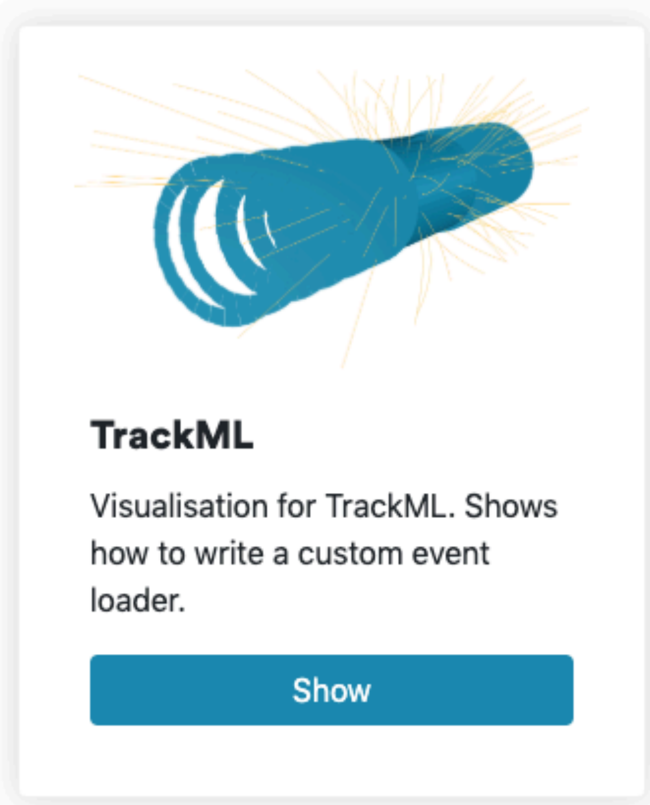
plot & image
(top) transverse momentum distribution for hard scatter and pileup event
(bottom) hits produced in one single event

CERN OPEN DATA PORTAL and PHOENIX

Application for visualizing High Energy Physics data.

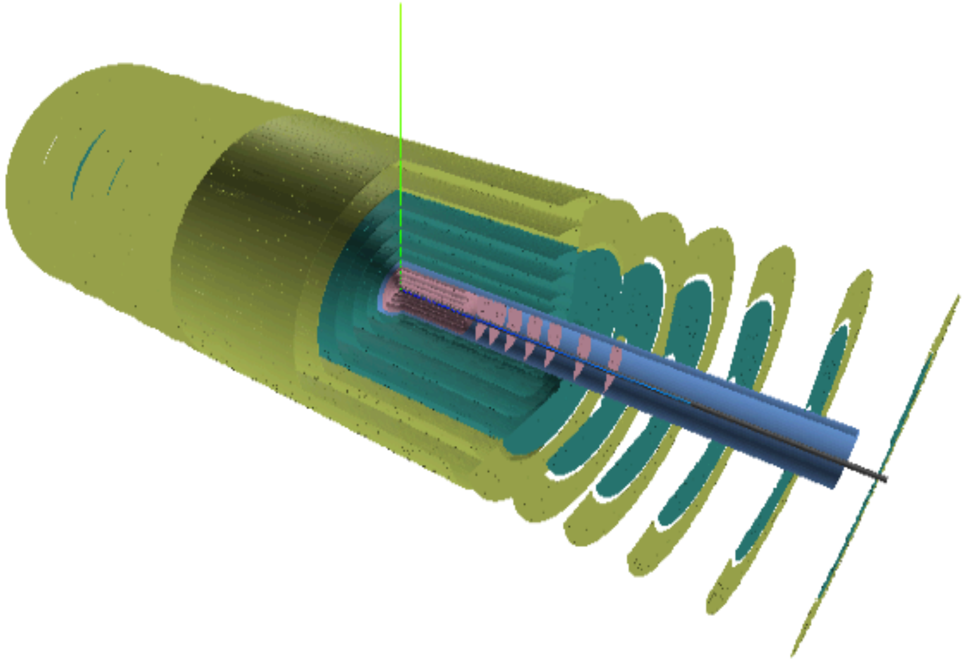
Common geometry/event display project

- part of the HSF (Hep Software Foundation) projects
- built-in support for TrackML/OpenData Detector



TrackML

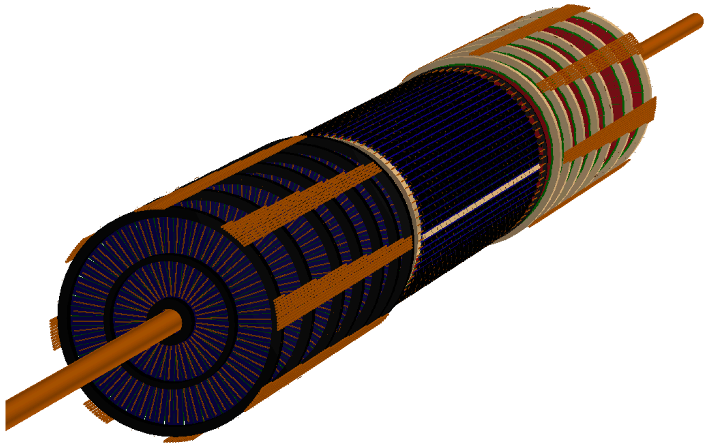
Visualisation for TrackML. Shows how to write a custom event loader.



Release timeline

Tentative release timeline for OpenData detector

Consolidation of detector & simulation



Dataset production:
Geant4 simulation
(small statistics validation sample)

ACTS-Fatras simulation
(large statistics sample)

Sep 2019

July/Aug 2019