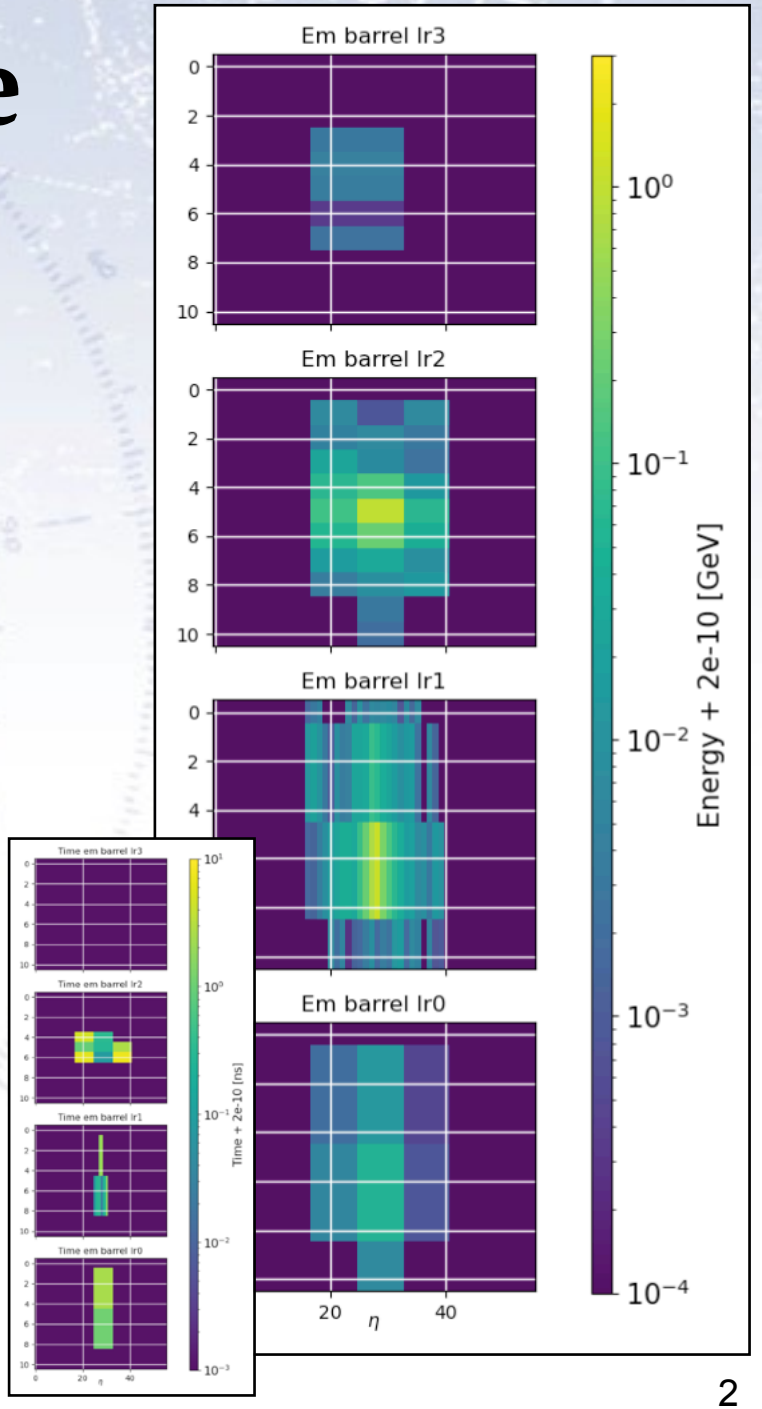
ML at Work: Electron Energy regression with CNN

Malte Algren*, Aske Rosted, and **Troels C. Petersen**
Niels Bohr Institute, Copenhagen (*now at Univ. of Geneva)

Outline

Outline of talk:

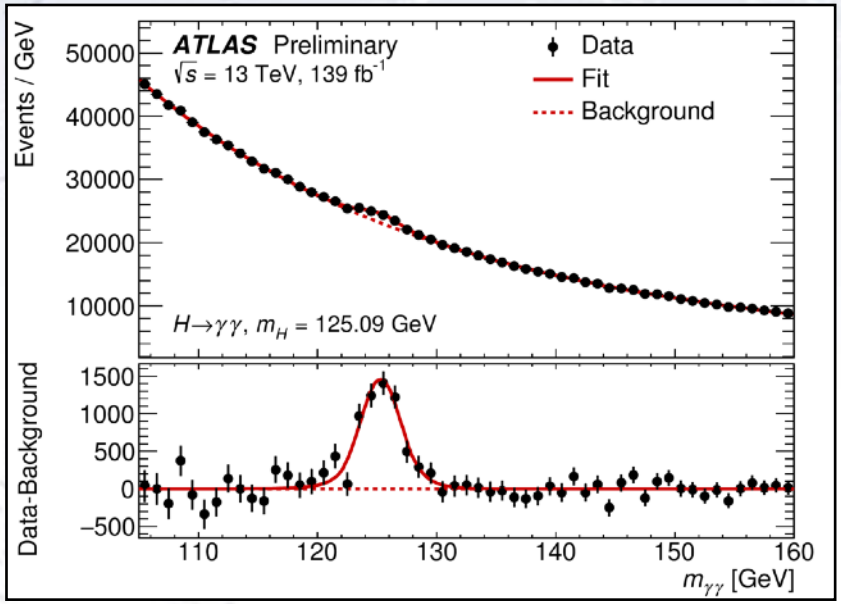
- Motivation
- Context
- Training a CNN for energy reconstruction:
  - The data
  - The selections
  - The input variables
  - The network architecture
  - Feature wIse Linear Modulation (FiLM)
- Results in MC
- Results in data (v1)
- Training in data and “simultaneous training”
  - Results in data (v2)
- Outlook



Motivation

Points of motivation:

- Improve H → ZZ* and **H → γγ** analyses
- Optimise searches for:
  - **HH → γγbb**
  - H → Zγ
  - H → γ*γ
- **Improve resilience to pile-up**
- Improve Z → ee reconstruction
- Utilise excellent data for testing:
  - CNN and GNN models
  - data+MC simultaneous training
  - e+γ simultaneous training
- Improve non-Higgs searches



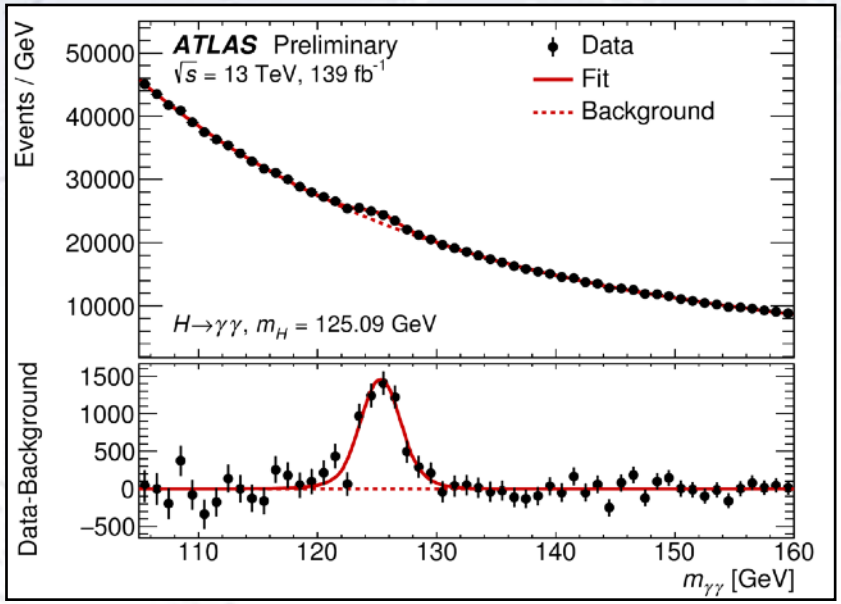
Goals of lecture:

- Give example of regression with CNN.
- Illustrate concept of attention and FiLM technique.
- Illustrate "target mismatch" and combined training.

Motivation

Points of motivation: **(You don't have to care - just know the list is long!)**

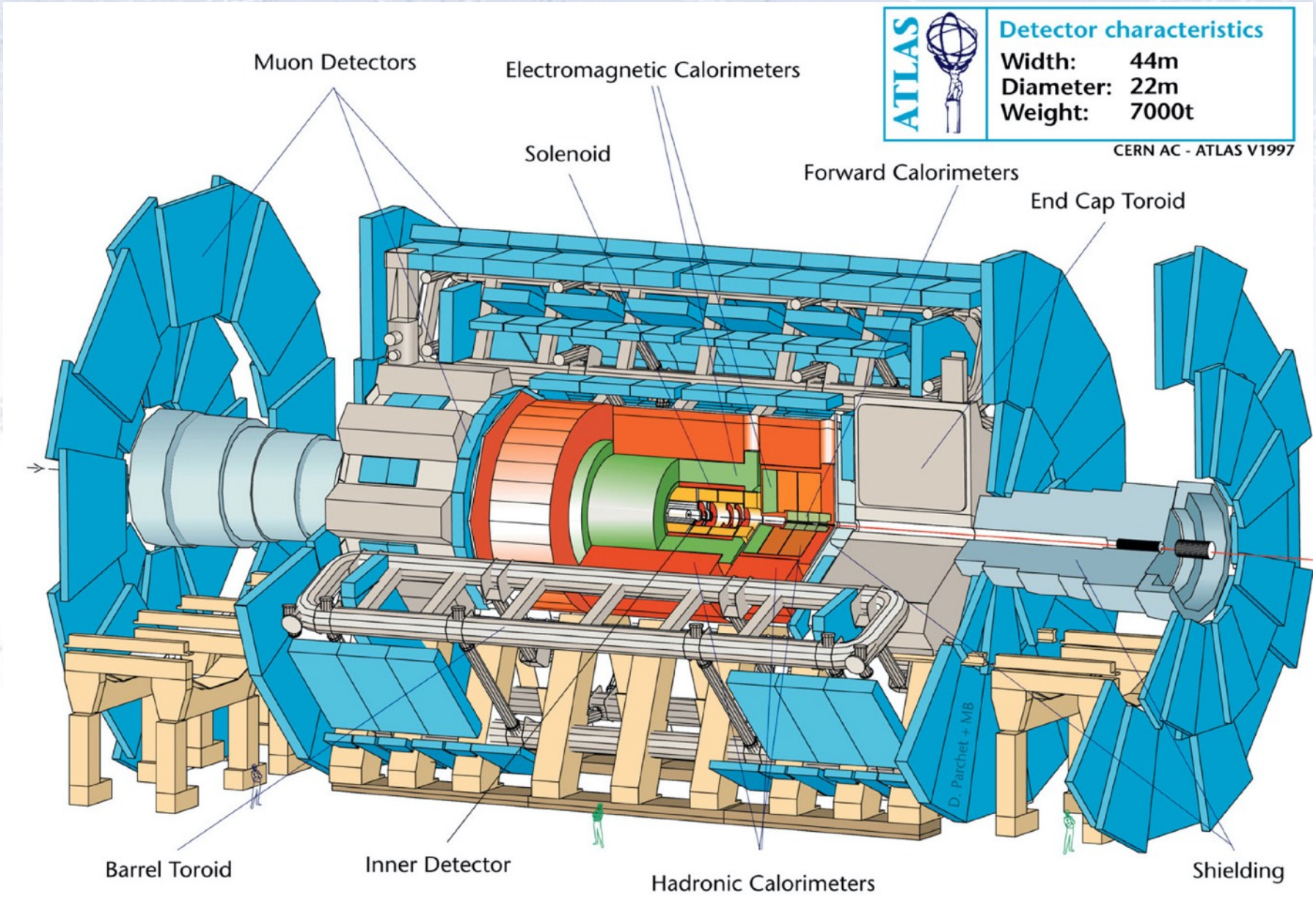
- Improve $H \to ZZ^*$ and $H \to \gamma\gamma$ analyses
- Optimise searches for:
  - $HH \to \gamma\gamma bb$
  - $H \to Z\gamma$
  - $H \to \gamma^*\gamma$
- **Improve resilience to pile-up**
- Improve $Z \to ee$ reconstruction
- Utilise excellent data for testing:
  - CNN and GNN models
  - data+MC simultaneous training
  - e+γ simultaneous training
- Improve non-Higgs searches



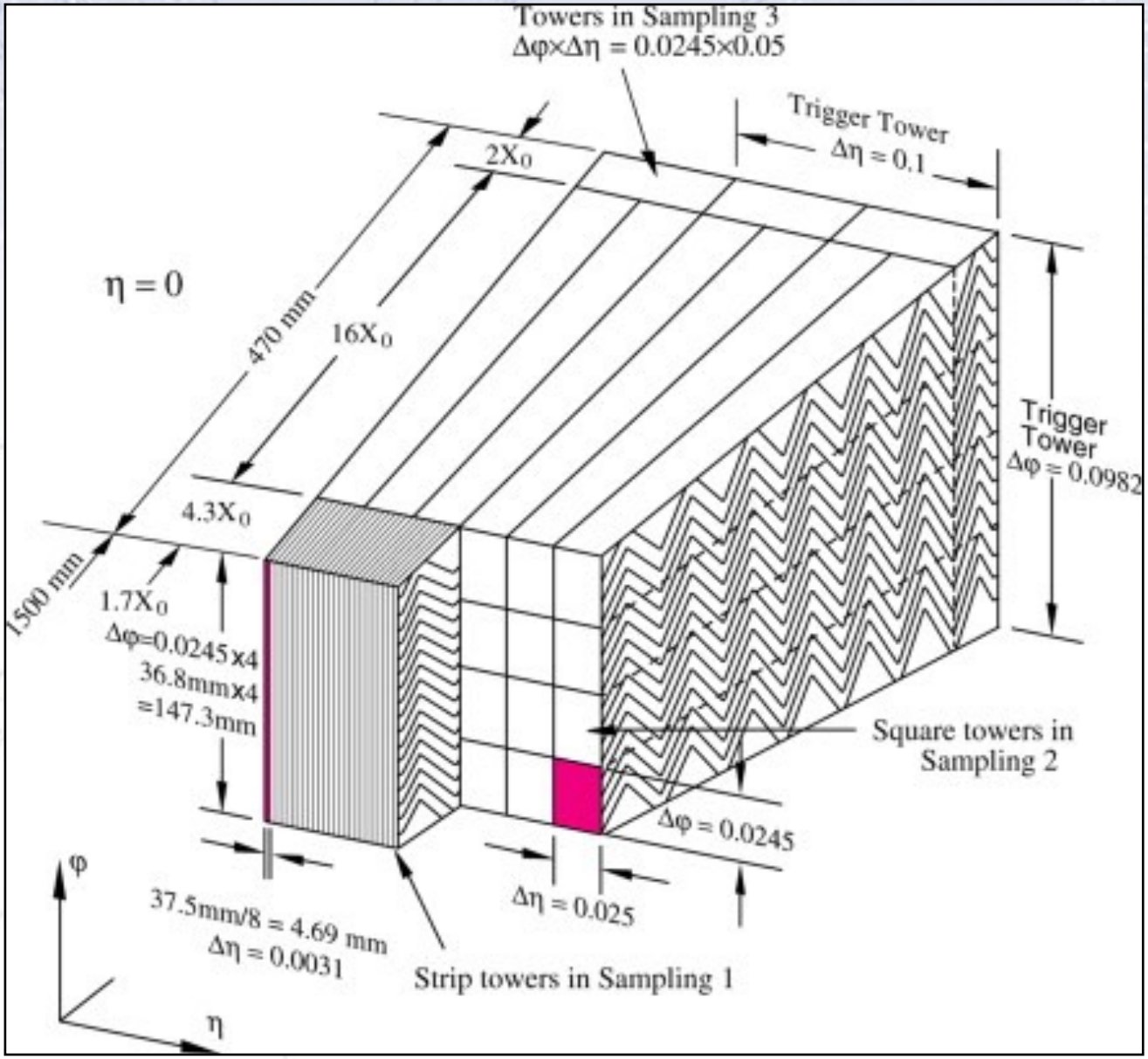
Goals of lecture:

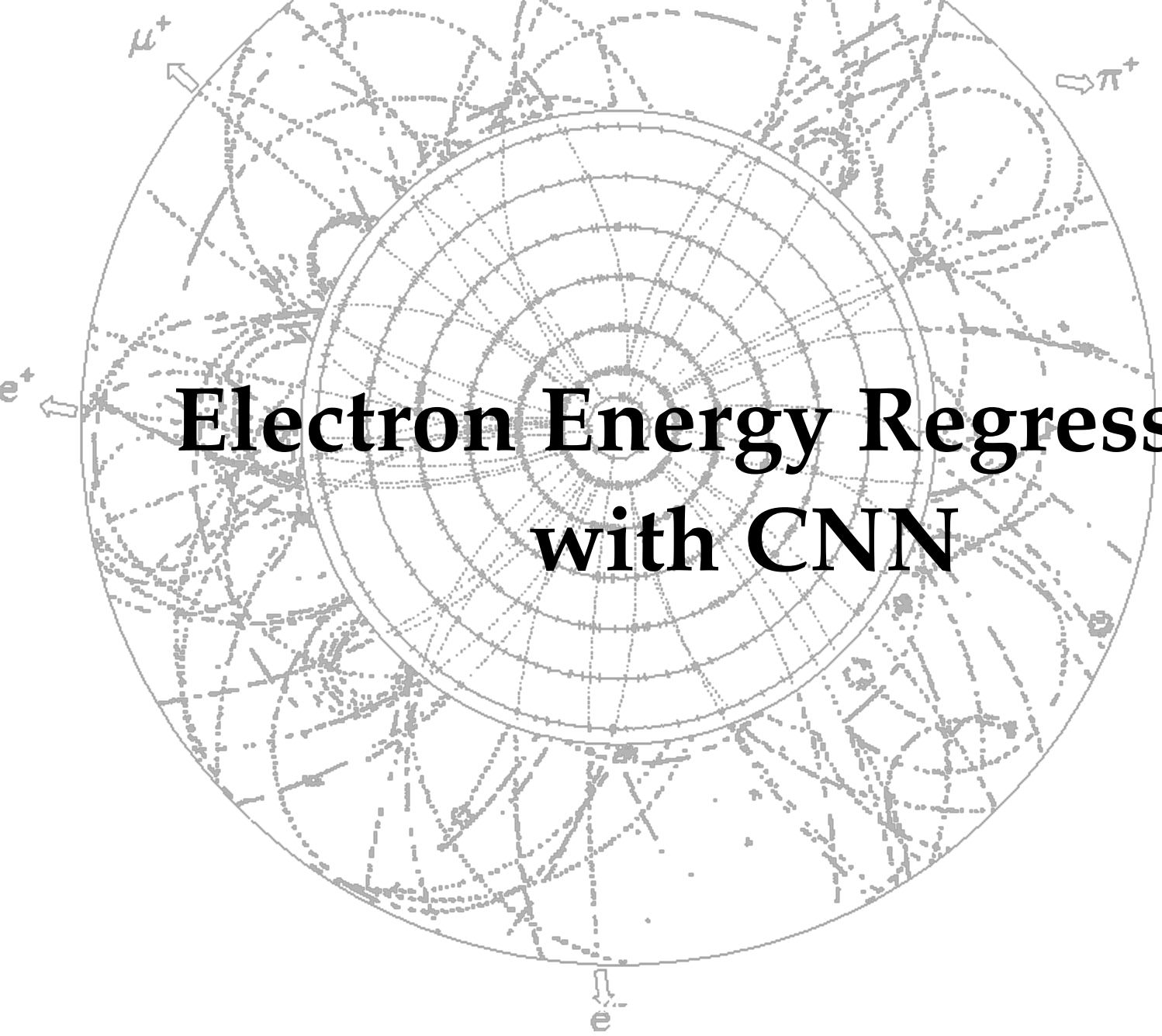
- Give example of regression with CNN.
- Illustrate concept of attention and FiLM technique.
- Illustrate "target mismatch" and combined training.

ATLAS Detector

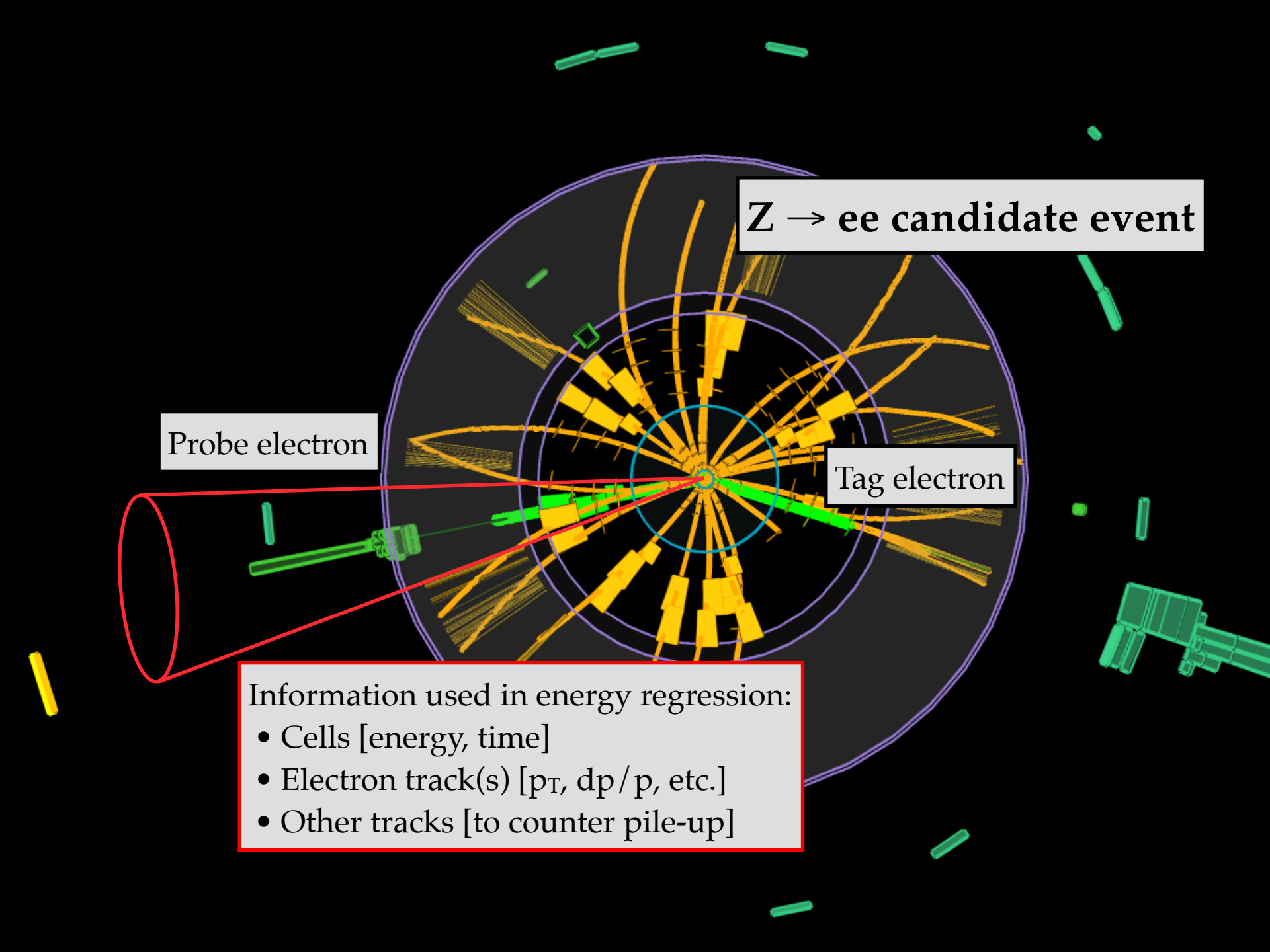


ATLAS Detector





Electron Energy Regression with CNN


Z → ee candidate event
Probe electron
Tag electron
Information used in energy regression:
• Cells [energy, time]
• Electron track(s) [p_T, dp/p, etc.]
• Other tracks [to counter pile-up]

The input variables

The variables are both scalar and cell based.
The scalars can be seen in table on the right.

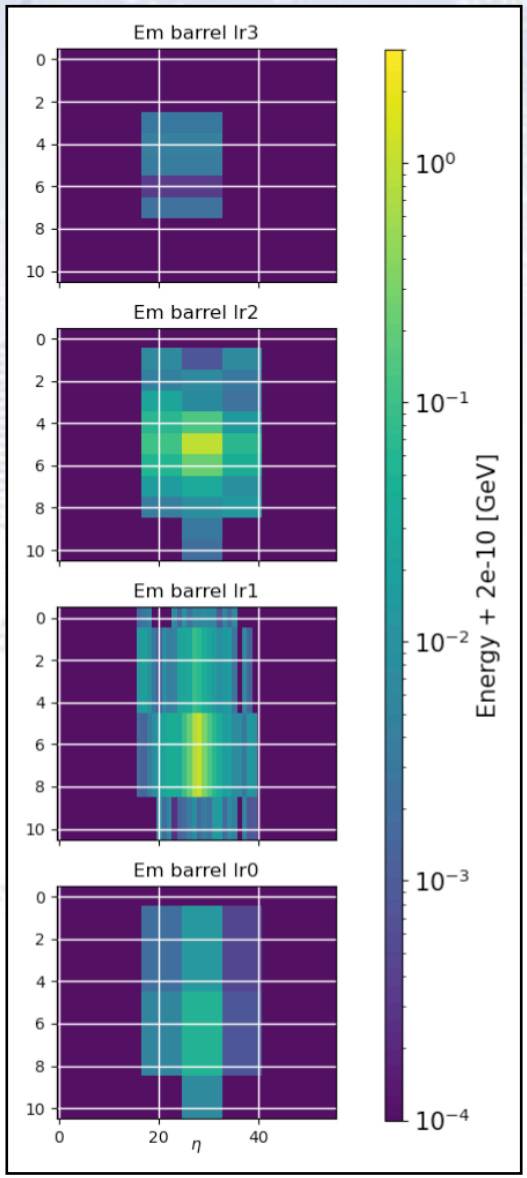
Type	Name	Description
Energy	E_{acc}	Energy deposit in layer 1-3 of ECAL.
	η_{index}	η cell index of cluster of layer 2.
	$f0_{cluster}$	Ratio of energy between layer 0 and E_{acc} in $\lvert\eta\rvert <$ 1.8 (end of layer 0).
	$R12$	Ratio of energy between layer 1 and 2 in the ECAL.
	p_t^{track}	p_T estimated from tracking for the particle (only e).
	E_{TG3}	Ratio between the energy in the crack scintillators and E_{acc} within $1.4 < \lvert\eta\rvert < 1.6$.
	$E_{tile-gap}$	Sum of the energy deposited in the tile-gap.
Geometric	η	Pseudorapidity of the particle.
	$\Delta\phi_2^{rescaled}$	Difference between ϕ, as extrapolated by tracking, use for ECAL momentum estimation and ϕ of the ECAL cluster.
	η_{ModCalo}	Relative η position w.r.t. the cell edge of layer 2 in the ECAL*.
	$\Delta\eta_2$	Difference between η, as extrapolated by tracking, use for ECAL momentum estimation and η of the ECAL cluster (only e).
	$poscs_2$	Relative position of η within cell in layer 2 in ECAL. $2(\eta_{cluster} - \eta_{maxEcell})/0.025 - 1$, $\eta_{cluster}$ is η of the barycenter of the cluster and $\eta_{maxEcell}$ is η of the most energetic cell of the cluster.
	$\Delta\phi_{TH3}$	Relative position in ϕ in a cell. $\text{mod}(2\pi + \phi, \pi/32) - \pi/32$.
Misc.	$\langle\mu\rangle$	Average proton-proton interaction per bunch crossing.
	n_{tracks}	# of tracks assigned (only e).
	$n_{vertexReco}$	Number of reconstructed vertices.

The input variables

The variables are both scalar and cell based. The scalars can be seen in table on the right.

We consider the cell energies in the LAr calorimeter as pixels in four images. The cells contain two (used) types of information:

- Energy (primary variable)
- Time of cell energy



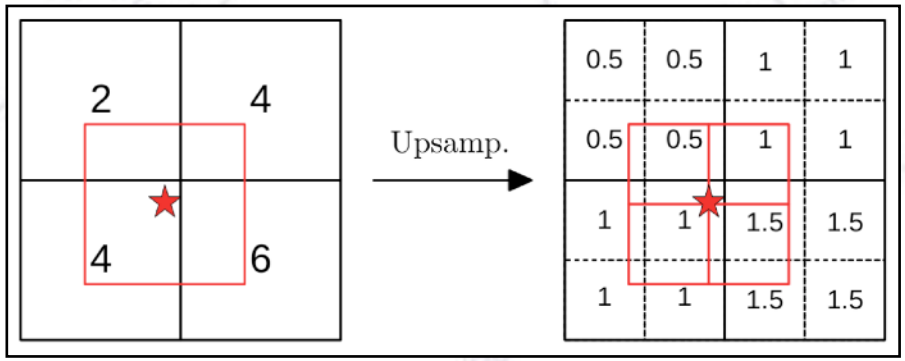
The input variables

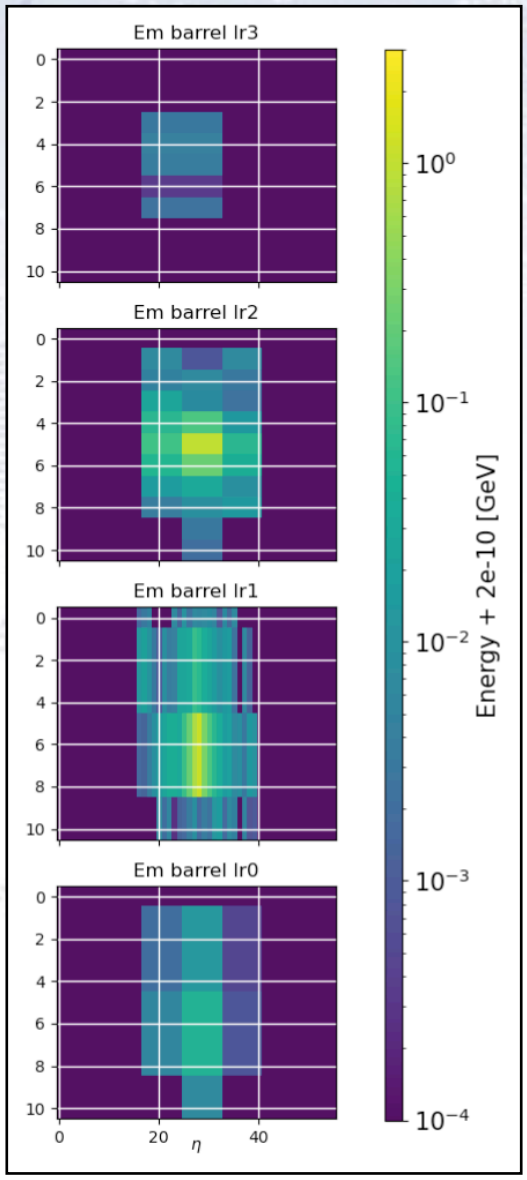
The variables are both scalar and cell based. The scalars can be seen in table on the right.

We consider the cell energies in the LAr calorimeter as pixels in four images. The cells contain two (used) types of information:

- Energy (primary variable)
- Time of cell energy

In order to have the **same resolution** in each layer, we **upsample** the layers to the lowest common resolution (work by Lucas Erhke).





The input variables

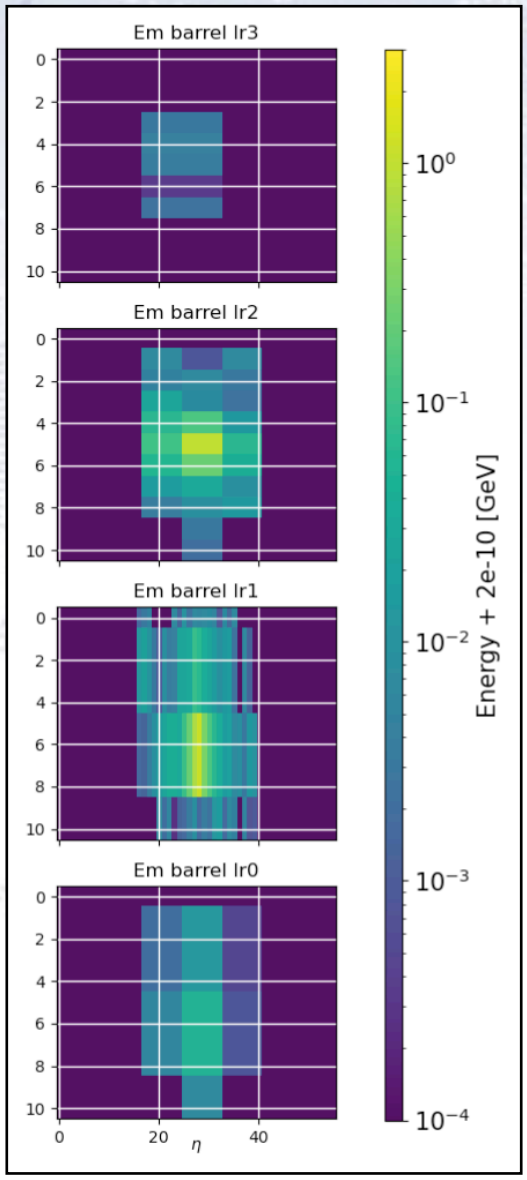
The variables are both scalar and cell based. The scalars can be seen in table on the right.

We consider the cell energies in the LAr calorimeter as pixels in four images. The cells contain two (used) types of information:

- Energy (primary variable)
- Time of cell energy

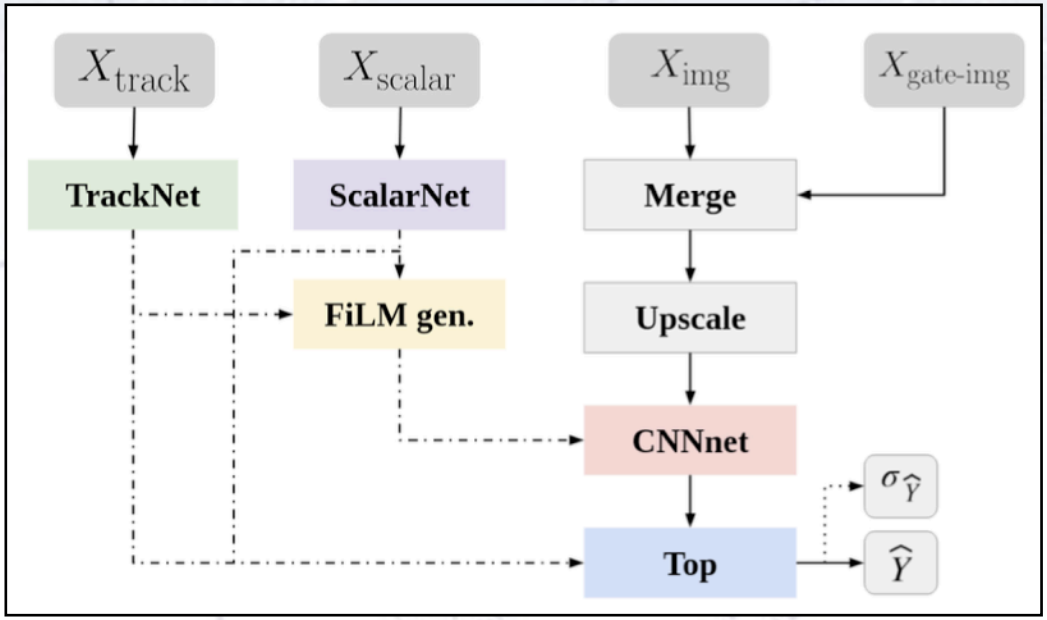
Finally, we consider the (up to) 10 nearest tracks in a "TrackNet" input:

Type	Name	Description
Energy	$p_{t,track}/q_{track}$	Transverse momentum of track divided by its charge q
Geometric	d_0/σ_{d0}	$d0$ is the signed transverse distance between the point of closest approach and the z-axis where σ_{d0} is its uncertainty
	ΔR	$\Delta R = \sqrt{(\phi_0 - \phi)^2 + (\eta_0 - \eta)^2}$
	$\text{vertex}_{\text{track}}$	Reconstructed vertex of the track
	z_0	Longitudinal distance between the point of closest approach and the z-axis.
	η_{track}	Reconstructed $\lvert\eta\rvert$ of tracks.
	ϕ_{track}	Reconstructed ϕ of tracks.
Misc.	n_{pixel}	Number of hits in the pixel detector
	n_{SCT}	Number of hits in the SCT
	n_{TRT}	Number of hits in the TRT



The network architecture

There are many ways to combine the input variables, and we have considered the following architectures, where the dashed lines are the considerations.



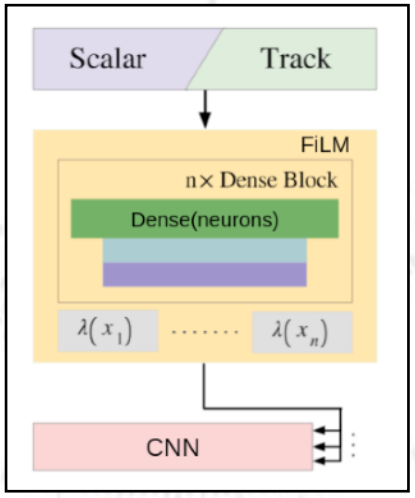
First, let us consider each part…

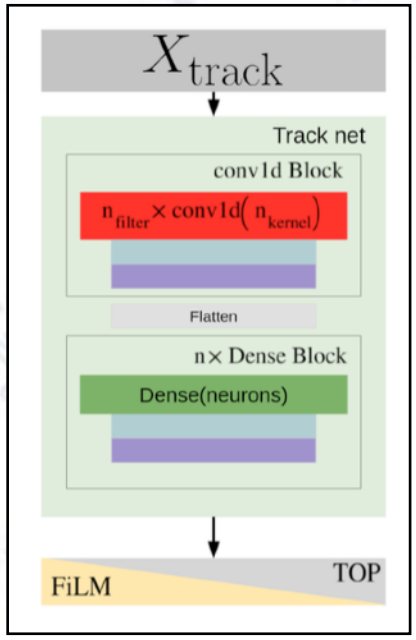
The network architecture

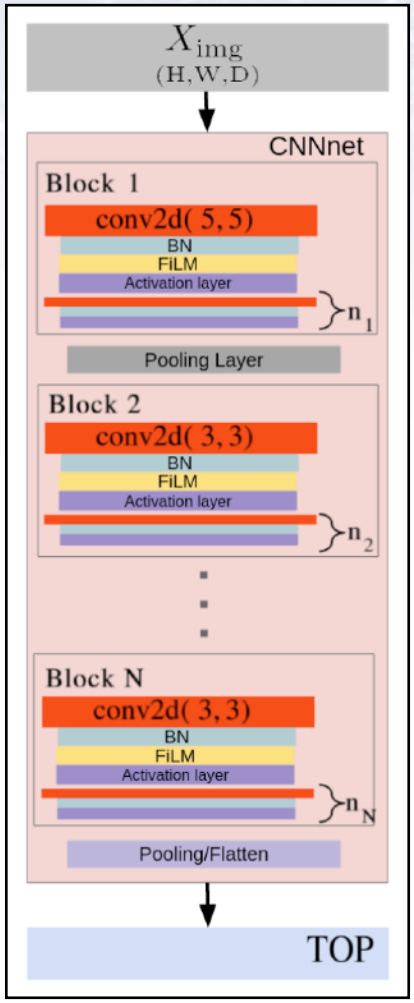
The CNN is the main estimator of the overall energy, and is relatively standard. The innovation lies in Feature wIse Linear Modulation (FiLM).

TrackNet is a pile-up-corrective input consisting of a combination of the reconstructed tracks close to the candidate. It can be input to both top layer or FiLM weights.

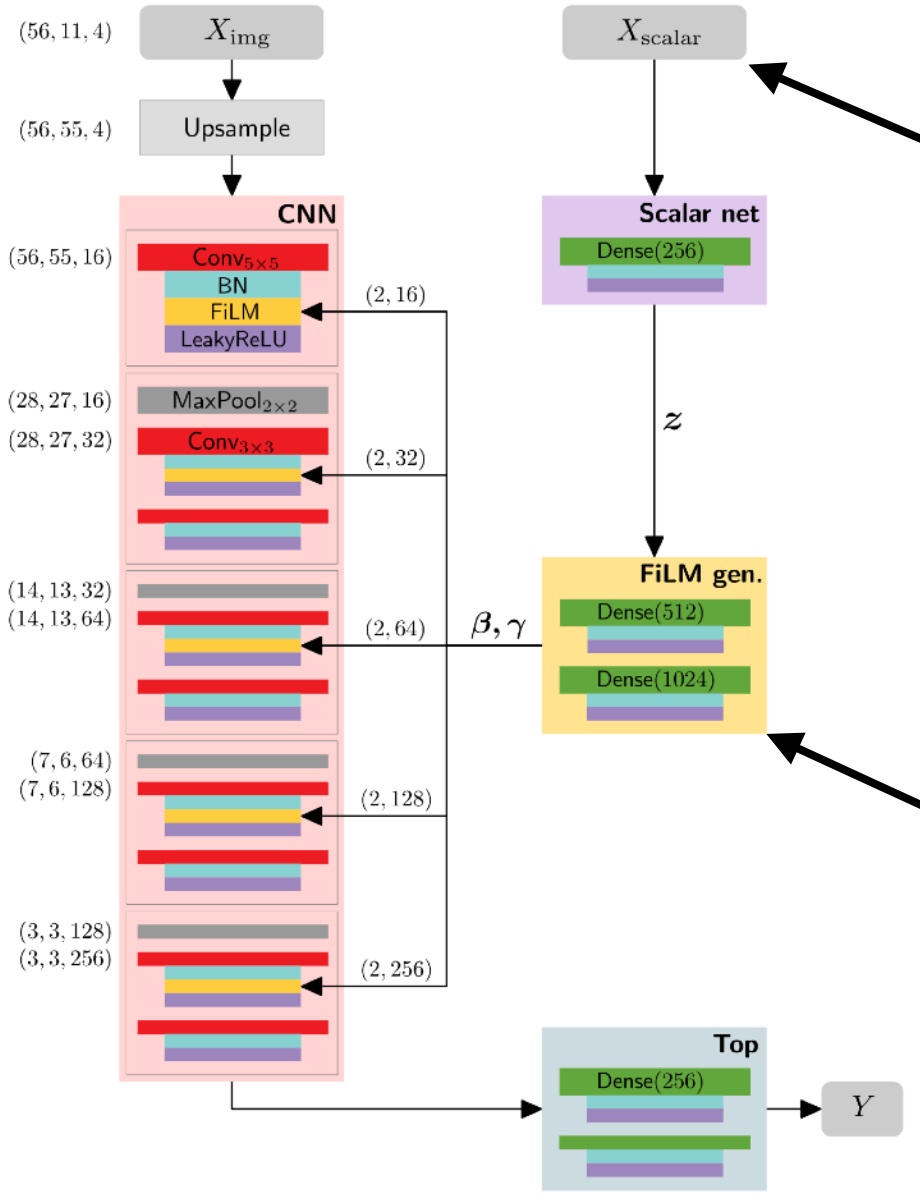
Finally, the scalars (and possibly tracks) are used in the FiLM.







Feature wIse Linear Modulation



Type	Name	Description
Energy	E_{acc}	Energy deposit in layer 1-3 of ECAL.
	η_{index}	η cell index of cluster of layer 2.
	$f0_{cluster}$	Ratio of energy between layer 0 and E_{acc} in $\lvert\eta\rvert <$ 1.8 (end of layer 0).
	$R12$	Ratio of energy between layer 1 and 2 in the ECAL.
	p_t^{track}	p_T estimated from tracking for the particle (only e).
	E_{TC3}	Ratio between the energy in the crack scintillators and E_{acc} within $1.4 < \lvert\eta\rvert < 1.6$.
	$E_{tile-gap}$	Sum of the energy deposited in the tile-gap.
Geometric	η	Pseudorapidity of the particle.
	$\Delta\phi_2^{rescaled}$	Difference between ϕ, as extrapolated by tracking, use for ECAL momentum estimation and ϕ of the ECAL cluster.
	η_{ModCalo}	Relative η position w.r.t. the cell edge of layer 2 in the ECAL*.
	$\Delta\eta_2$	Difference between η, as extrapolated by tracking, use for ECAL momentum estimation and η of the ECAL cluster (only e).
	$poscs_2$	Relative position of η within cell in layer 2 in ECAL. $2(\eta_{cluster} - \eta_{maxEcell})/0.025 - 1$, $\eta_{cluster}$ is η of the barycenter of the cluster and $\eta_{maxEcell}$ is η of the most energetic cell of the cluster.
	$\Delta\phi_{TH3}$	Relative position in ϕ in a cell. $\text{mod}(2\pi + \phi, \pi/32) - \pi/32$.
Misc.	$\langle\mu\rangle$	Average proton-proton interaction per bunch crossing.
	n_{tracks}	# of tracks assigned (only e).
	$n_{vertexReco}$	Number of reconstructed vertices.

$$\text{FiLM}(x_i) = \gamma_i(\boldsymbol{z})x_i + \beta_i(\boldsymbol{z})$$

(FiLM: Feature wIse Linear Modulation of the CNN output layers based on the scalar input variables.)

Before the convolutions are pooled, they are weighted (linearly) by the "context". In this way, the best filters in the given case are given the most weight.

The network architecture

Testing all the different combinations yields the optimal architecture.

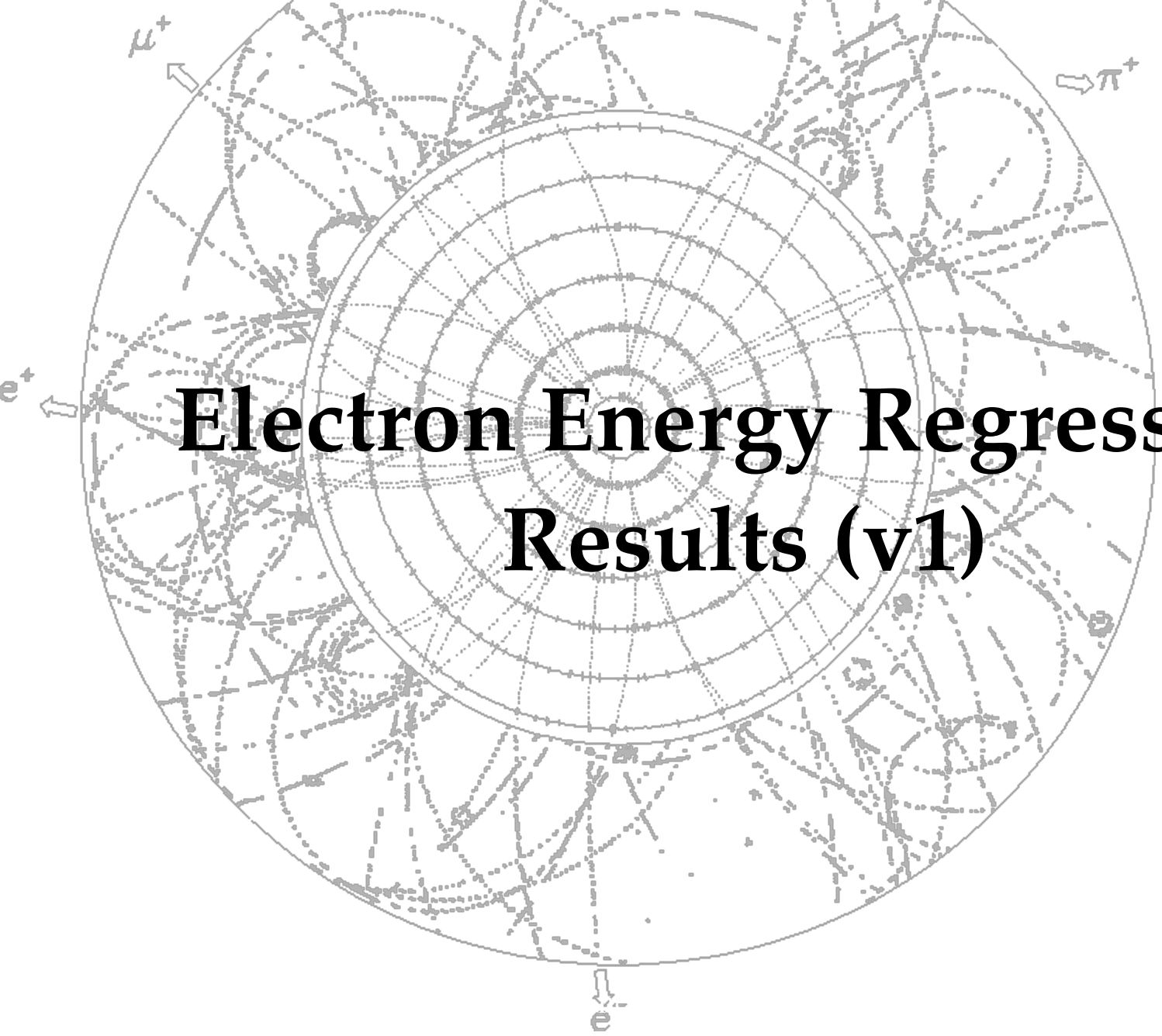
We evaluate the performance in the same way as previously done, namely the effective InterQuantile Range (eIQR) of the Relative Error (RE).

$$eIQR = \frac{P_{75}(RE) - P_{25}}{1.349}(RE), \qquad RE = \frac{E_{calib}}{E_{truth}},$$

	reIQR75	reIQR95
Basic	-0.121	-0.025
FiLM: scalar	0.229	0.257
FiLM: scalar - top: scalar	0.229	0.252
FiLM: scalar - top: scalar track	0.223	0.251
FiLM: scalar - top: track	**0.226**	**0.264**
FiLM: scalar track	0.228	0.265
FiLM: scalar track - top: scalar track	0.210	0.262
FiLM: track - top: scalar	-0.042	-0.067
FiLM: track - top: track	0.140	0.149
top: scalar	-0.154	-0.131
top: scalar track	0.213	0.233
top: track	0.136	0.164

Best Architecture

Hyperparameter	Parameter
`TrackNet`	
Units	(128, 64, 32,16)
Normalization	Batch
Kernel size & filters	5
Connected to	[Top]
`ScalarNet`	
Units	(256)
Normalization	Batch
Connected to	[FiLM]
`FiLM gen.`	
Units	(512, 1024)
Normalization	Batch
`CNNnet`	
Down-sampling	MaxPool
Globalpooling	MaxPool
Number of blocks	3
Depth of blocks	4
`Top`	
Units	(512, 512, 1)
Output activation	ReLU



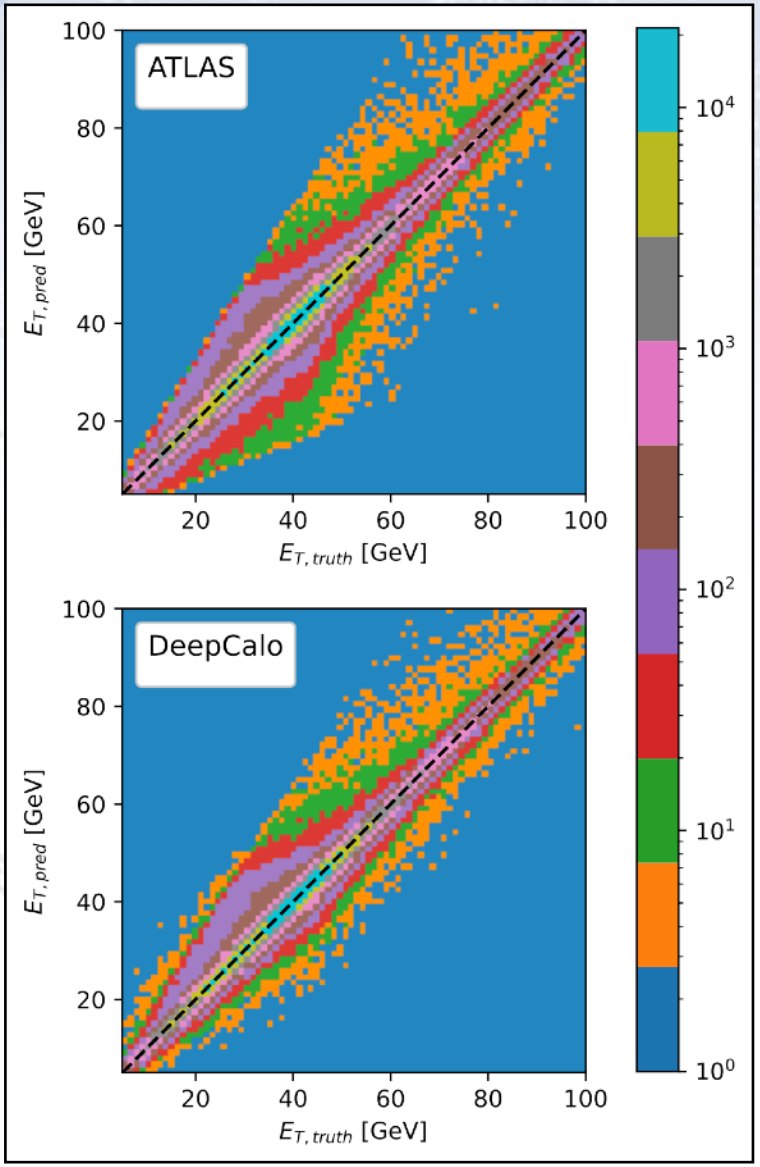
Electron Energy Regression Results (v1)

The results in 2D - MC

The E_T distribution for truth (x-axis) and reconstruction (y-axis) can be compared for the current ATLAS and the DeepCalo algorithms.

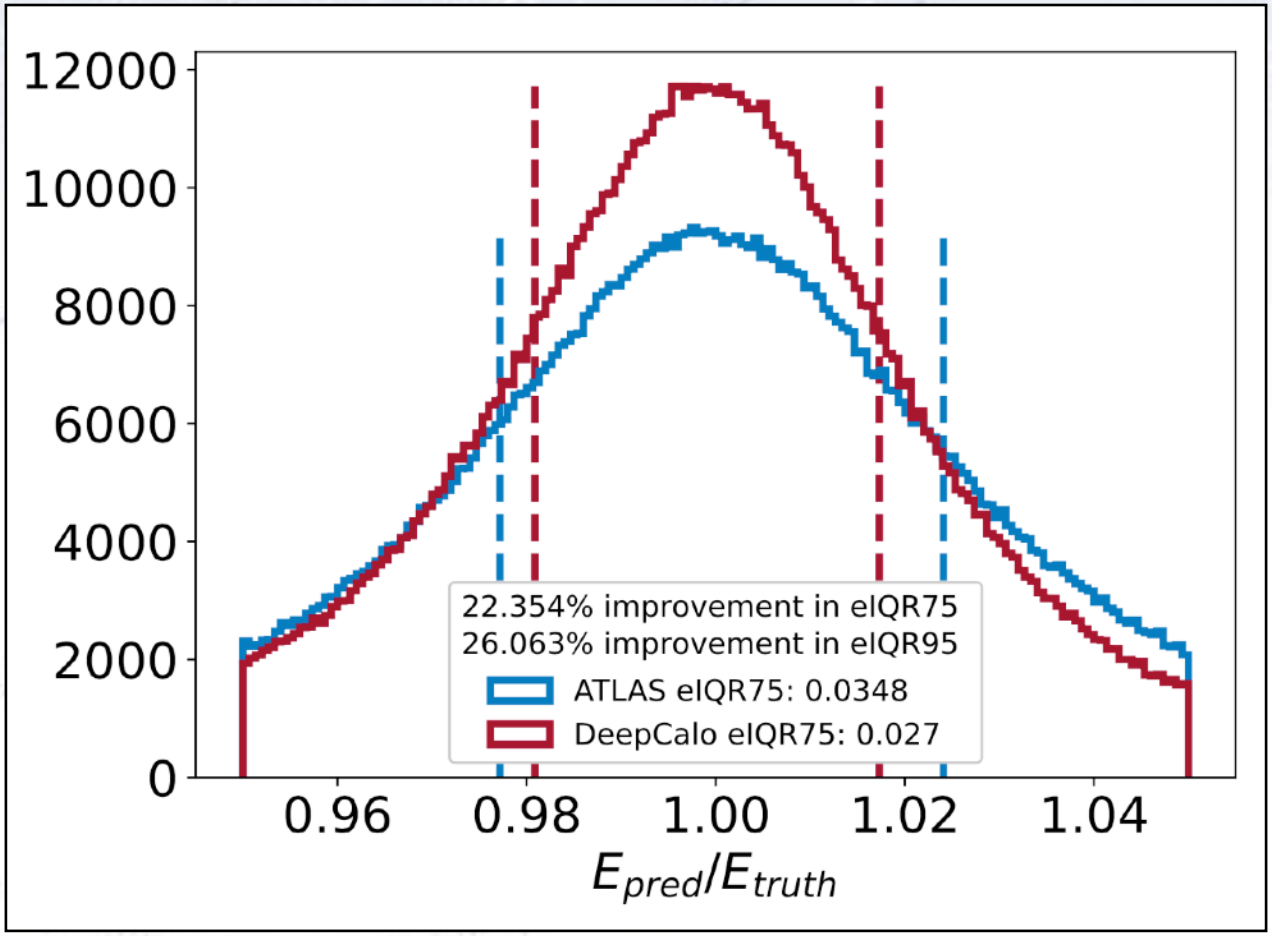
As the figure shows, both algorithms do well, and improve with energy.

As the statistics is largest around 40 GeV, this is where the comparison is most detailed, and here DeepCalo visibly has a significantly reduced lower edge.
Thus, the DeepCalo more rarely undershoots the energy.



The results in 1D - MC

Integrating the previous plot into 1D considering the RE distribution, we see a general sharpening. The improvement in relative eIQR (reIQR) is about 22%.

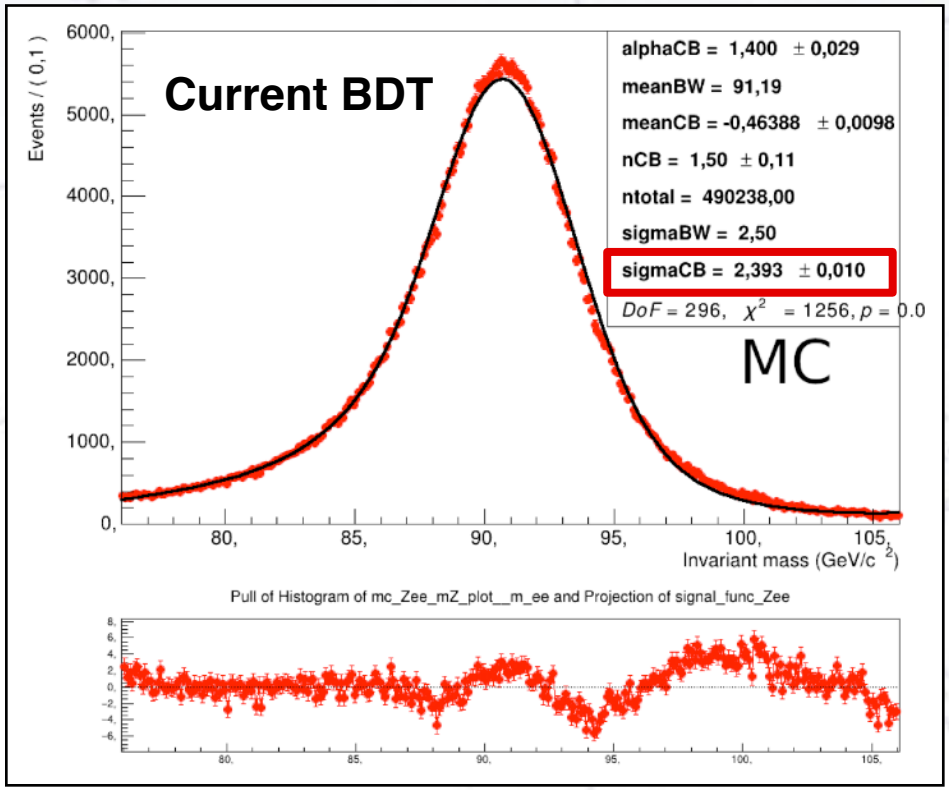


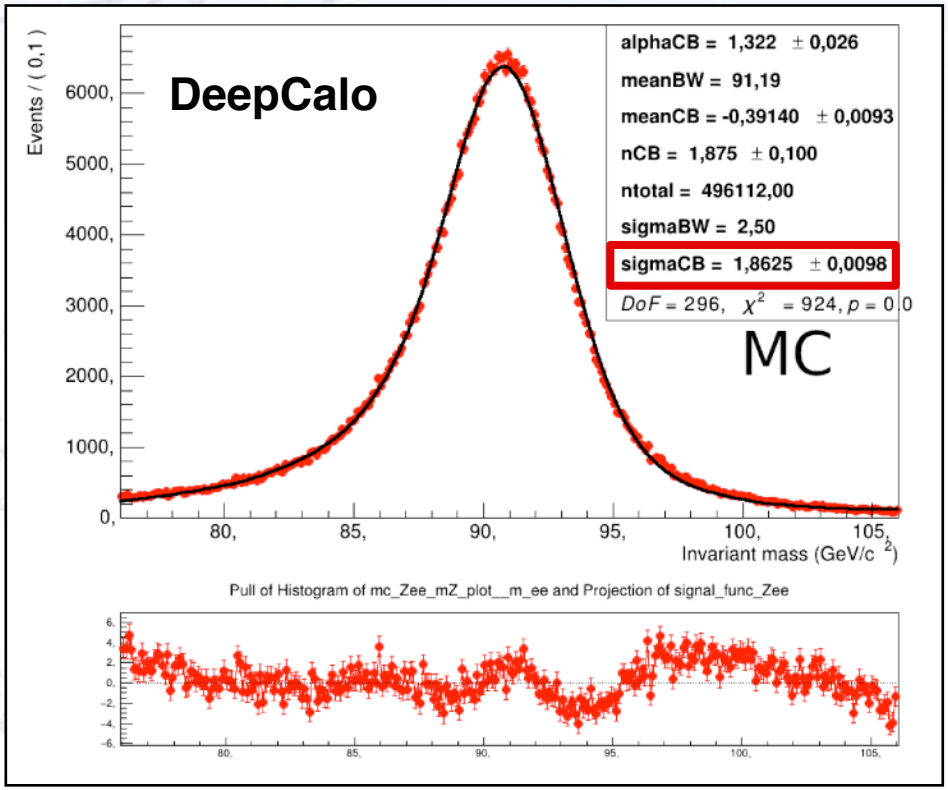
Naively, we would of course love to see a similar number in data!

Result in Zee - MC

On the Zee peak, we evaluate the improvement by fitting with a BW⊗CB fit, considering the CB width (sigmaCB) as the performance parameter. We get:

$$\langle 1 - \frac{\sigma_{CB}^{DeepCalo}}{\sigma_{CB}^{ATLAS}} \rangle = 1 - \frac{1.8310 \pm 0.006}{2.393 \pm 0.01} = 23.5 \pm 0.4\%$$

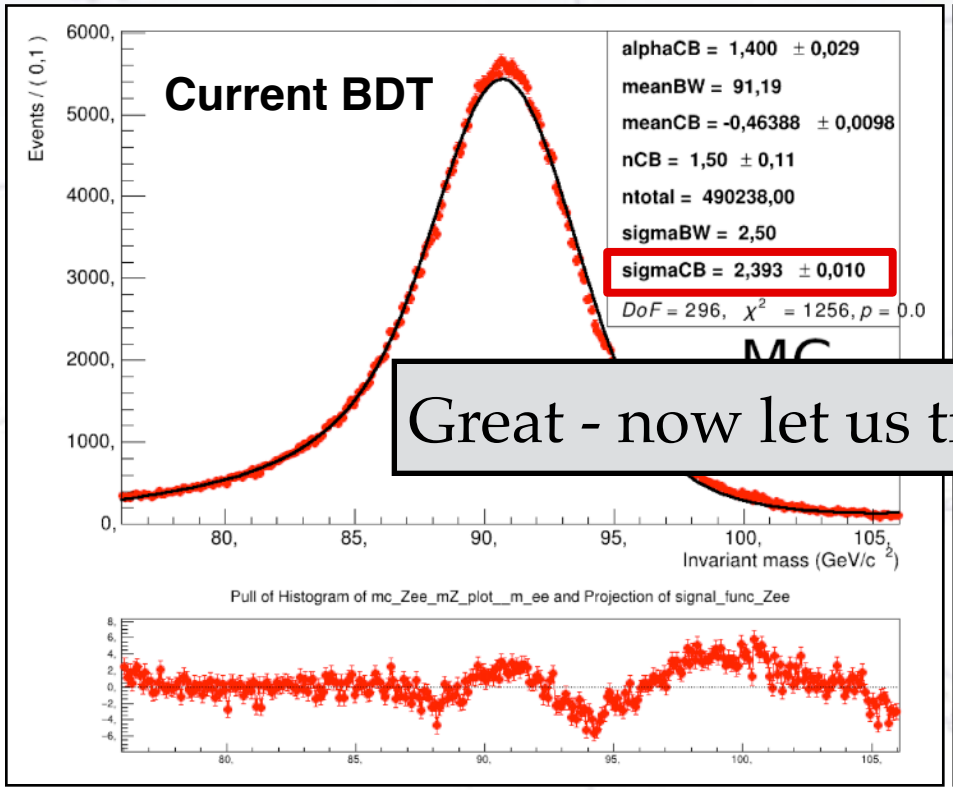


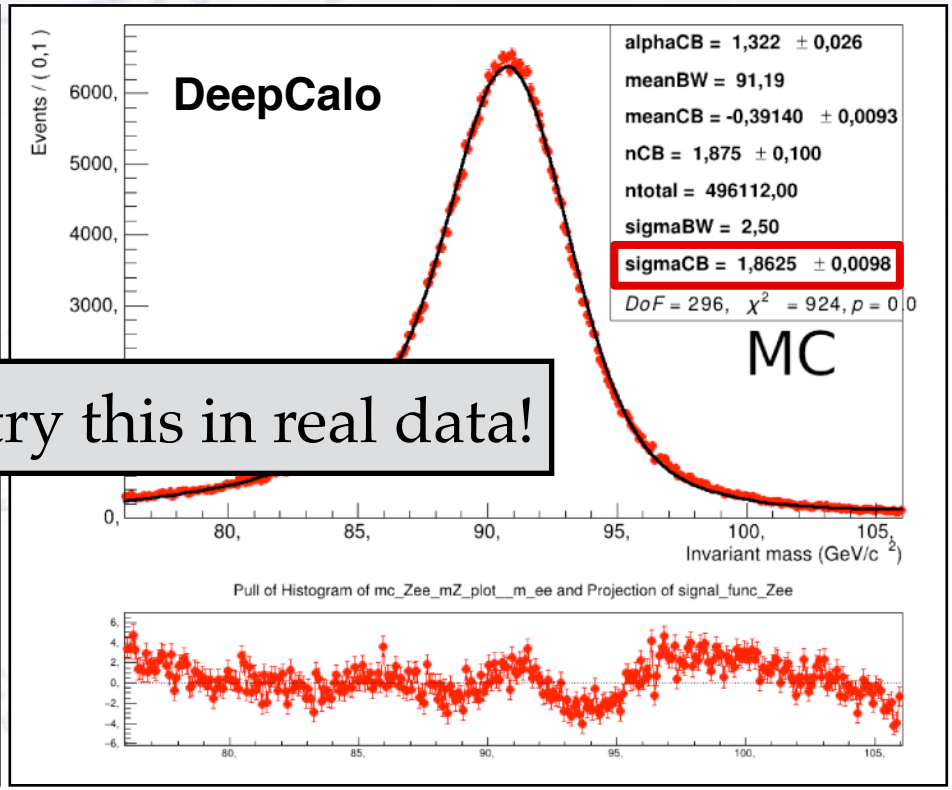


Result in Zee - MC

On the Zee peak, we evaluate the improvement by fitting with a BW⊗CB fit, considering the CB width (sigmaCB) as the performance parameter. We get:

$$\langle 1 - \frac{\sigma_{CB}^{DeepCalo}}{\sigma_{CB}^{ATLAS}} \rangle = 1 - \frac{1.8310 \pm 0.006}{2.393 \pm 0.01} = 23.5 \pm 0.4\%$$





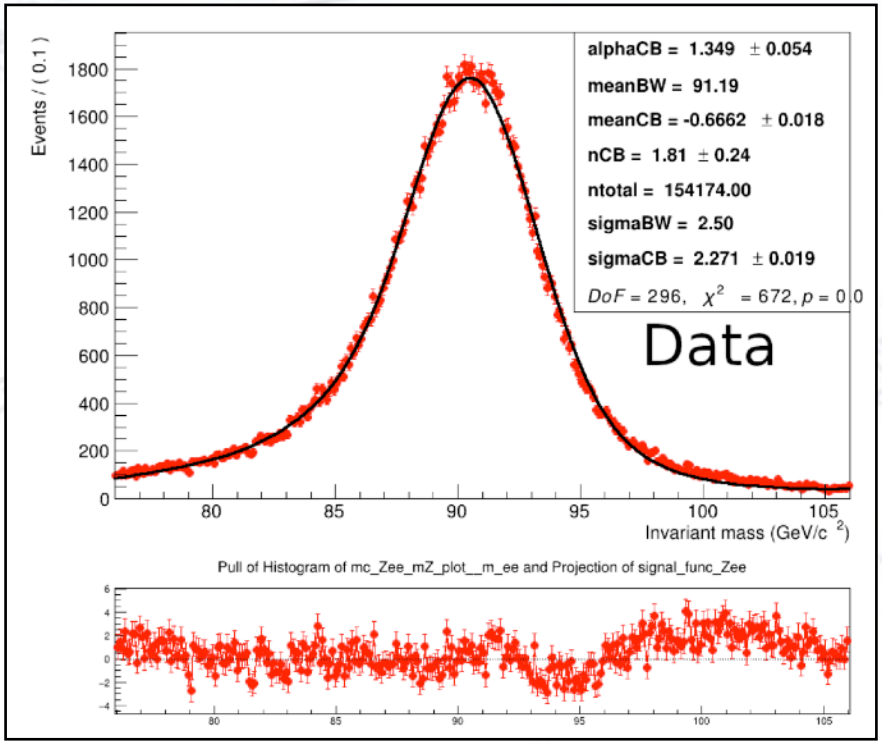
Great - now let us try this in real data!

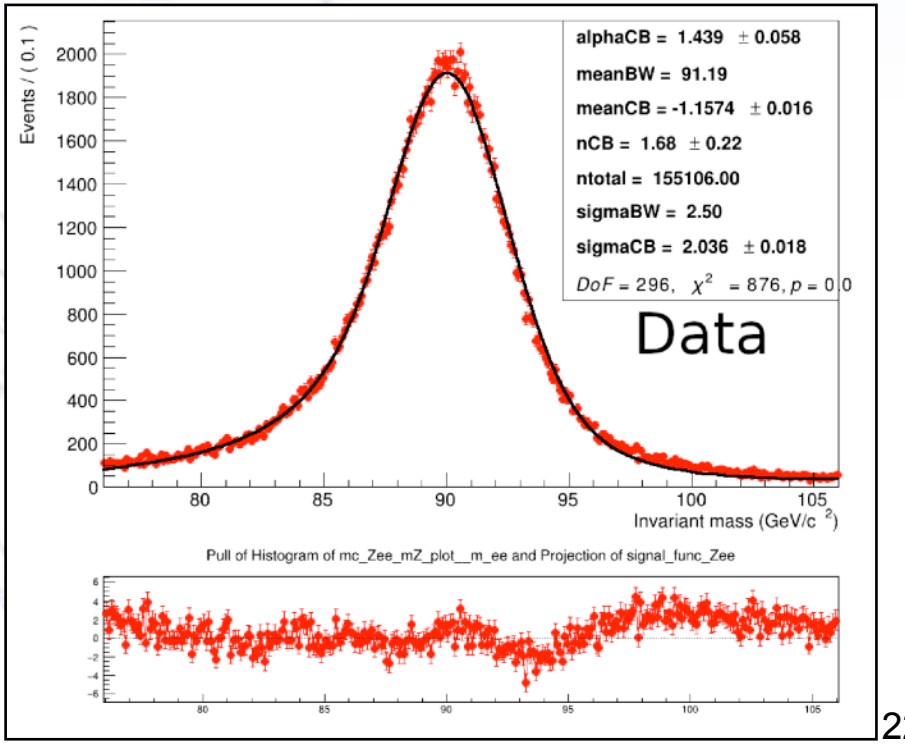
Results on Zee - data (v1)

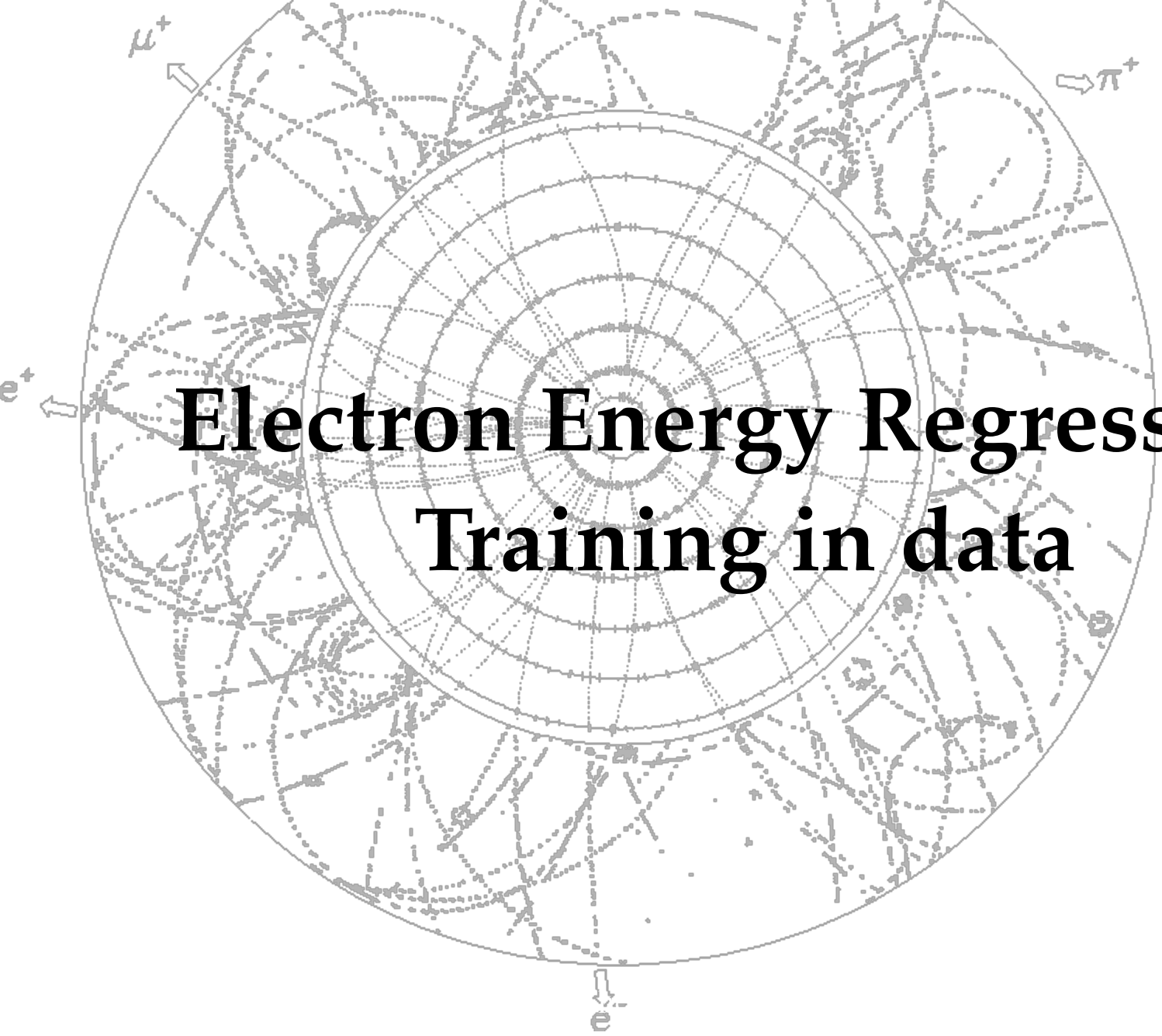
The result we get is a much more modest improvement:

$$\langle 1 - \frac{\sigma_{CB}^{DeepCalo}}{\sigma_{CB}^{ATLAS}} \rangle = 1 - \frac{2.058 \pm 0.010}{2.271 \pm 0.019} = 9.4 \pm 0.9\%.$$

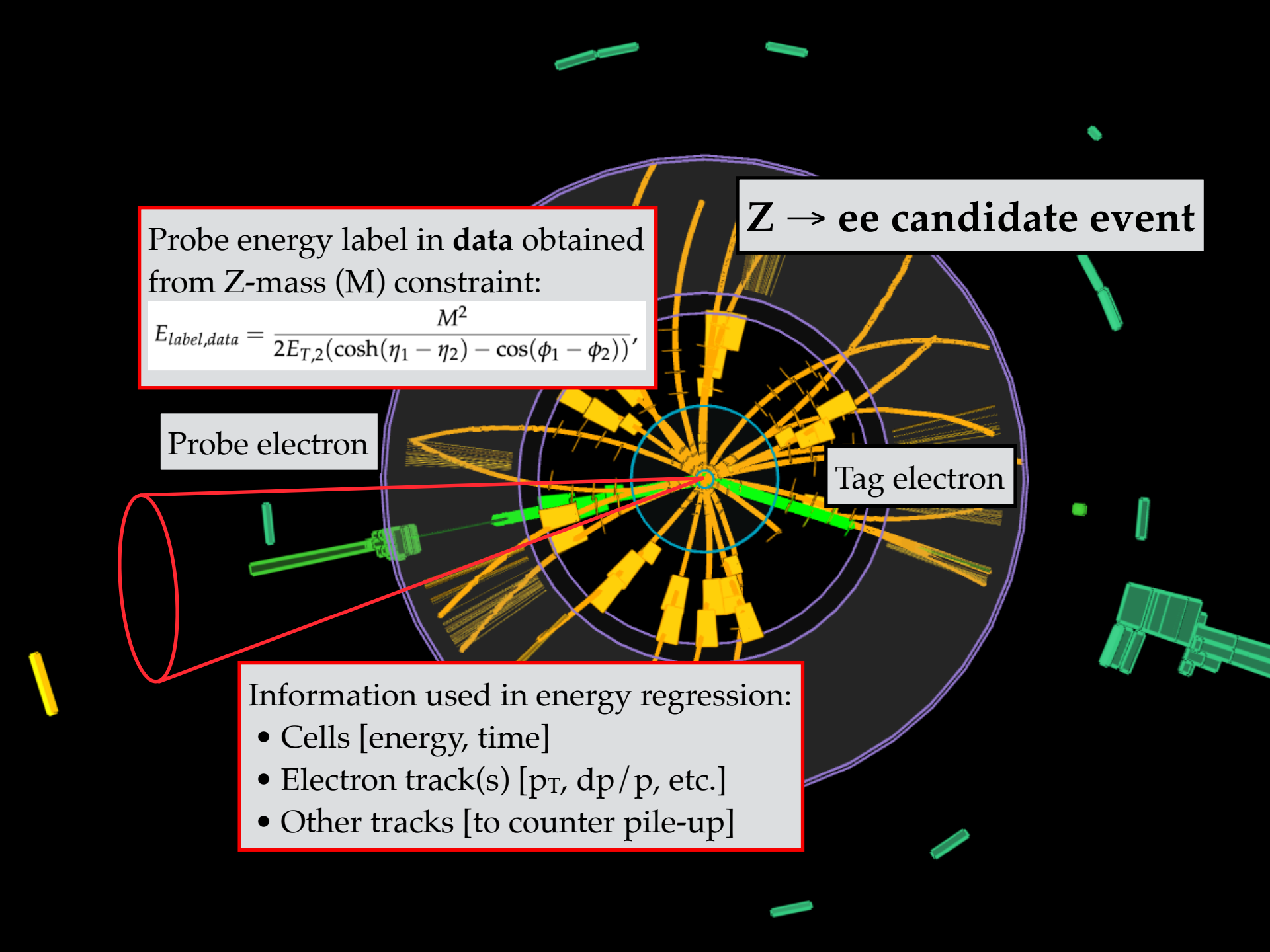
Though perhaps a little disappointing, this is not surprising, as we can not expect the MC to mimic data perfectly in the very large space considered. Also, models trained on Zee do not generalise well to all energies (EG, 6.8%).







Electron Energy Regression Training in data


Z → ee candidate event
Probe energy label in **data** obtained from Z-mass (M) constraint:
$E_{label,data} = \frac{M^2}{2E_{T,2}(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2))}$,
Probe electron
Tag electron
Information used in energy regression:
• Cells [energy, time]
• Electron track(s) [p_T, dp/p, etc.]
• Other tracks [to counter pile-up]

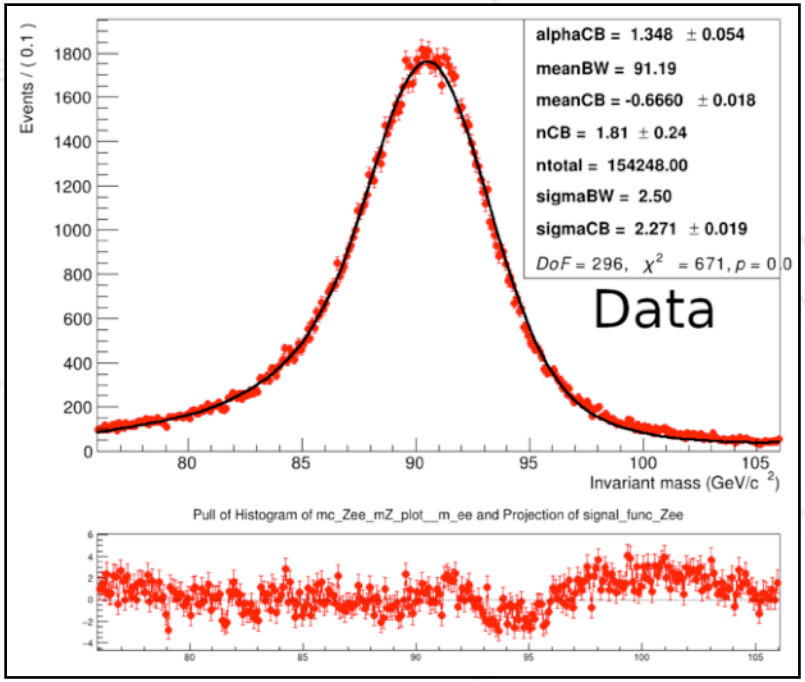
Training in data

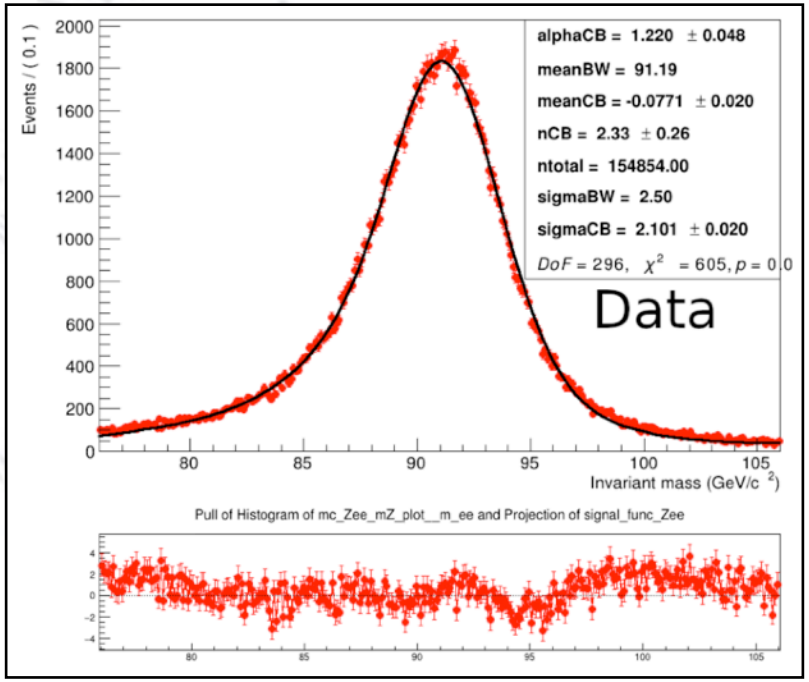
Using Zee events with invariant masses 86-97 GeV, one can get "approximate labels" in data, by assuming the true Z mass:

$$M^2 = 2p_{T,1}p_{T,2}(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2)), \quad p_T = E_T \updownarrow$$
$$E_{label,data} = \frac{M^2}{2E_{T,2}(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2))},$$
with $E_{T,2} = \text{Ecalib}^{(BDT)}$ and $M^2 = 91.19^2$

Using such labels, we train in data and get…

$$\langle 1 - \frac{\sigma_{CB}^{DeepCalo}}{\sigma_{CB}^{ATLAS}} \rangle = 5.9 \pm 0.9\%$$





Training in data

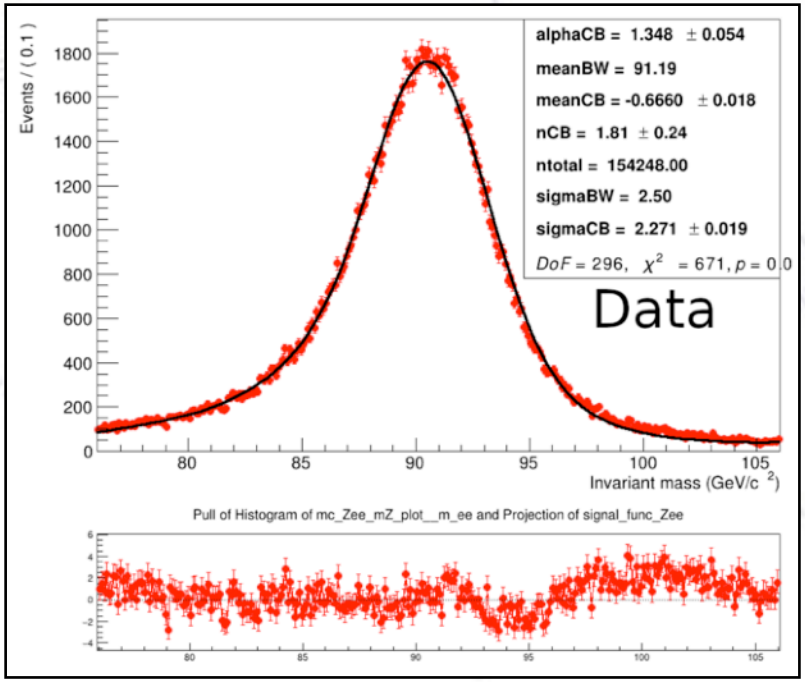
Using Zee events with invariant masses 86-97 GeV, one can get "approximate labels" in data, by assuming the true Z mass:

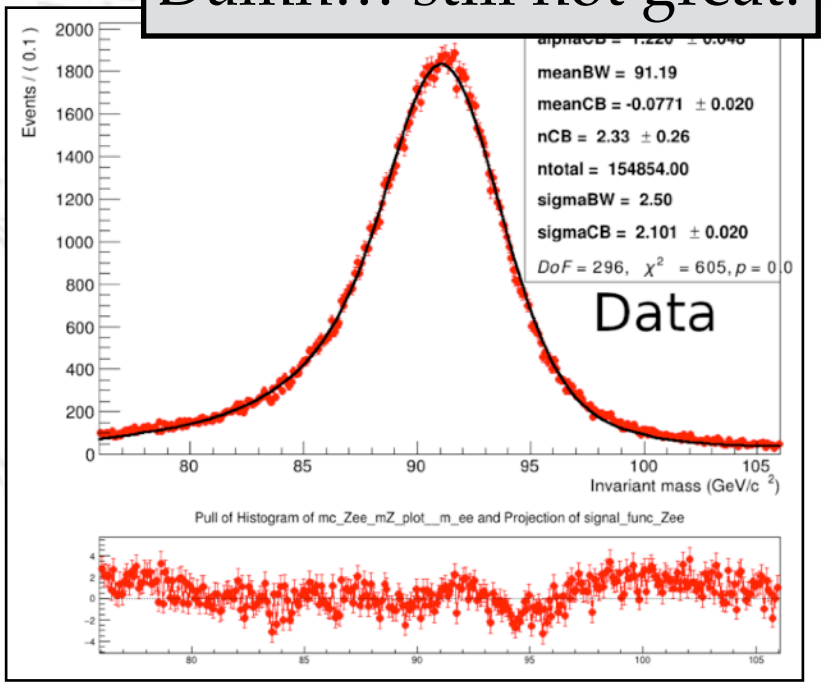
$$M^2 = 2p_{T,1}p_{T,2}(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2)), \quad p_T = E_T \updownarrow$$
$$E_{label,data} = \frac{M^2}{2E_{T,2}(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2))},$$
with $E_{T,2} = \text{Ecalib}^{(BDT)}$ and $M^2 = 91.19^2$

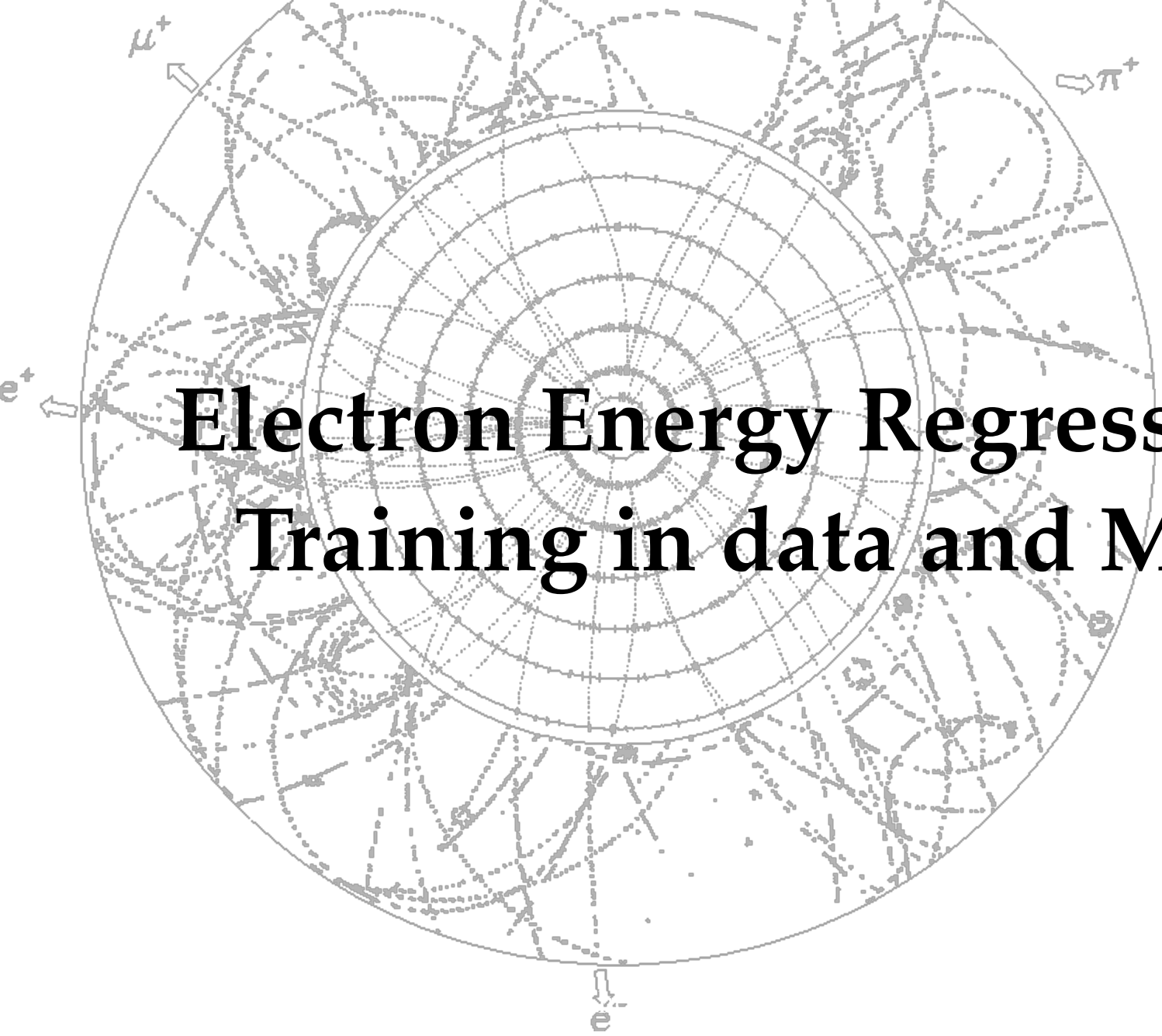
Using such labels, we train in data and get…

$$\langle 1 - \frac{\sigma_{CB}^{DeepCalo}}{\sigma_{CB}^{ATLAS}} \rangle = 5.9 \pm 0.9\%$$

Damn… still not great!







Electron Energy Regression Training in data and MC

Training in data and MC

Once we have labels in data, there is nothing keeping us from combining the loss functions of MC and data (they even have the same form), and thus training **simultaneously** in data and MC:

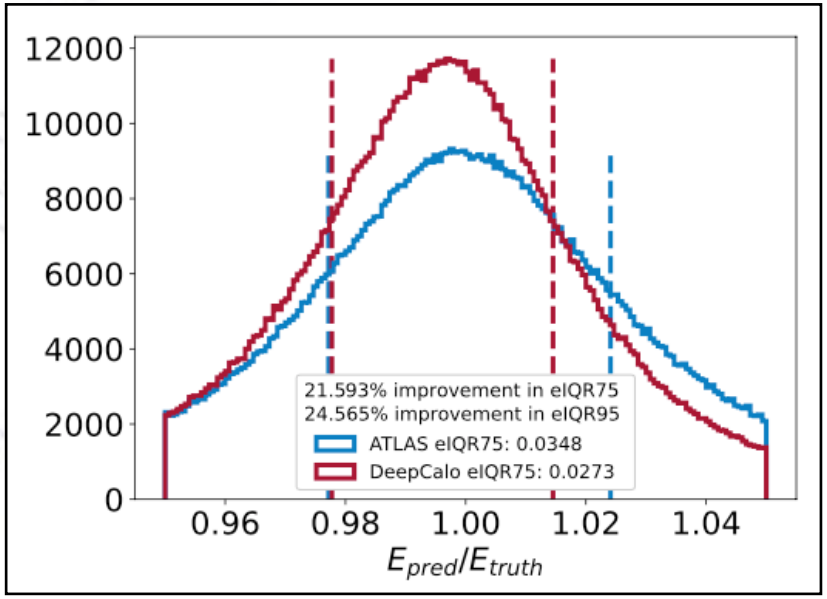
$$\mathcal{L}(y, \hat{y}) = \mathcal{L}(y_{(\text{Zee, MC})}, \hat{y}_{(\text{Zee, MC})}) + \mathcal{L}(y_{(\text{Zee, Data})}, \hat{y}_{(\text{Zee, Data})})$$

This allows the model to both use the "strength" of MC, but also learn the differences between MC and real data.

Doing this and trying out the result in MC first yields:

$$\langle reIQR_{75}^{DeepCalo} \rangle = 22.1 \pm 0.3\%$$

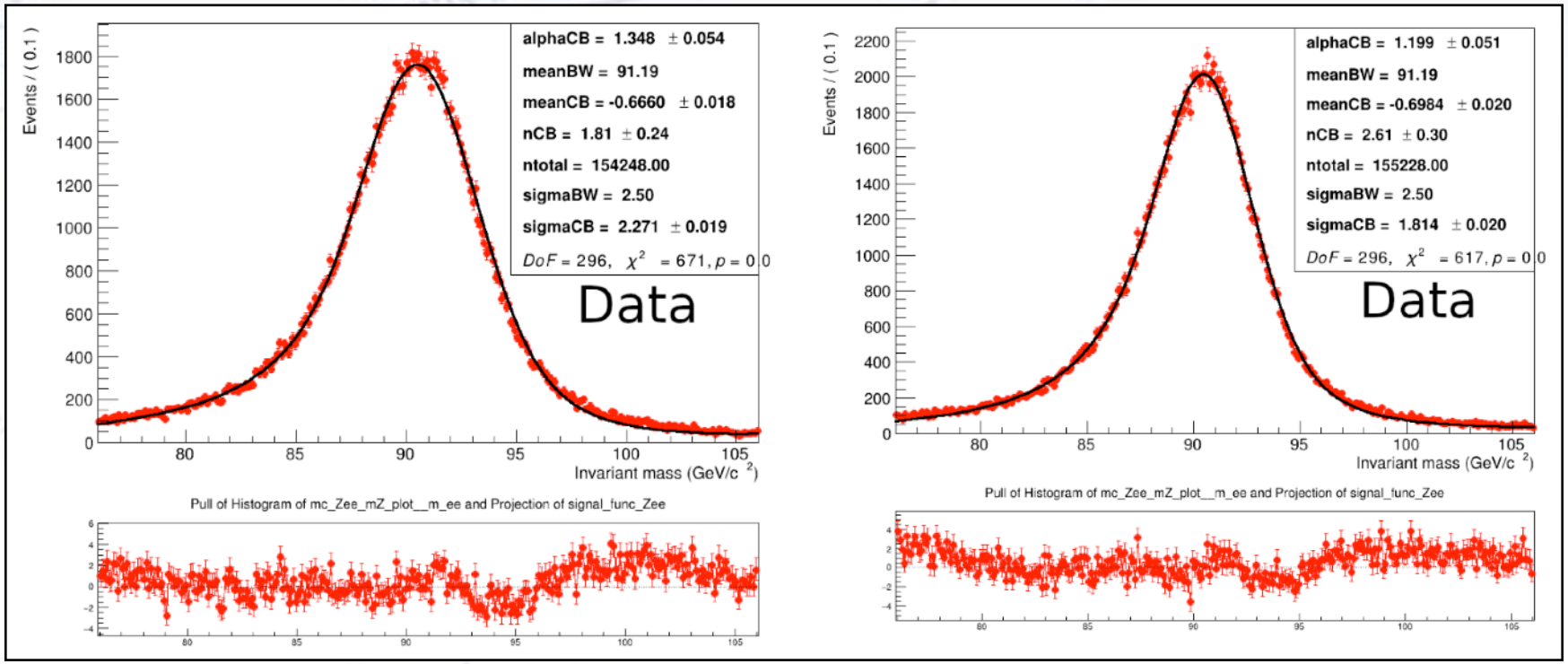
OK, so at least it doesn't ruin the model for MC. Now let us try data…



Result in data (v2)

The result in data is rather encouraging, and **greater than the sum of the improvements** from training separately in MC (9.4%) and data (5.9%).

$$\langle 1 - \frac{\sigma_{CB}^{DeepCalo}}{\sigma_{CB}^{ATLAS}} \rangle = 1 - \frac{1.86 \pm 0.010}{2.271 \pm 0.019} = 18.3 \pm 0.8\%,$$



Outlook

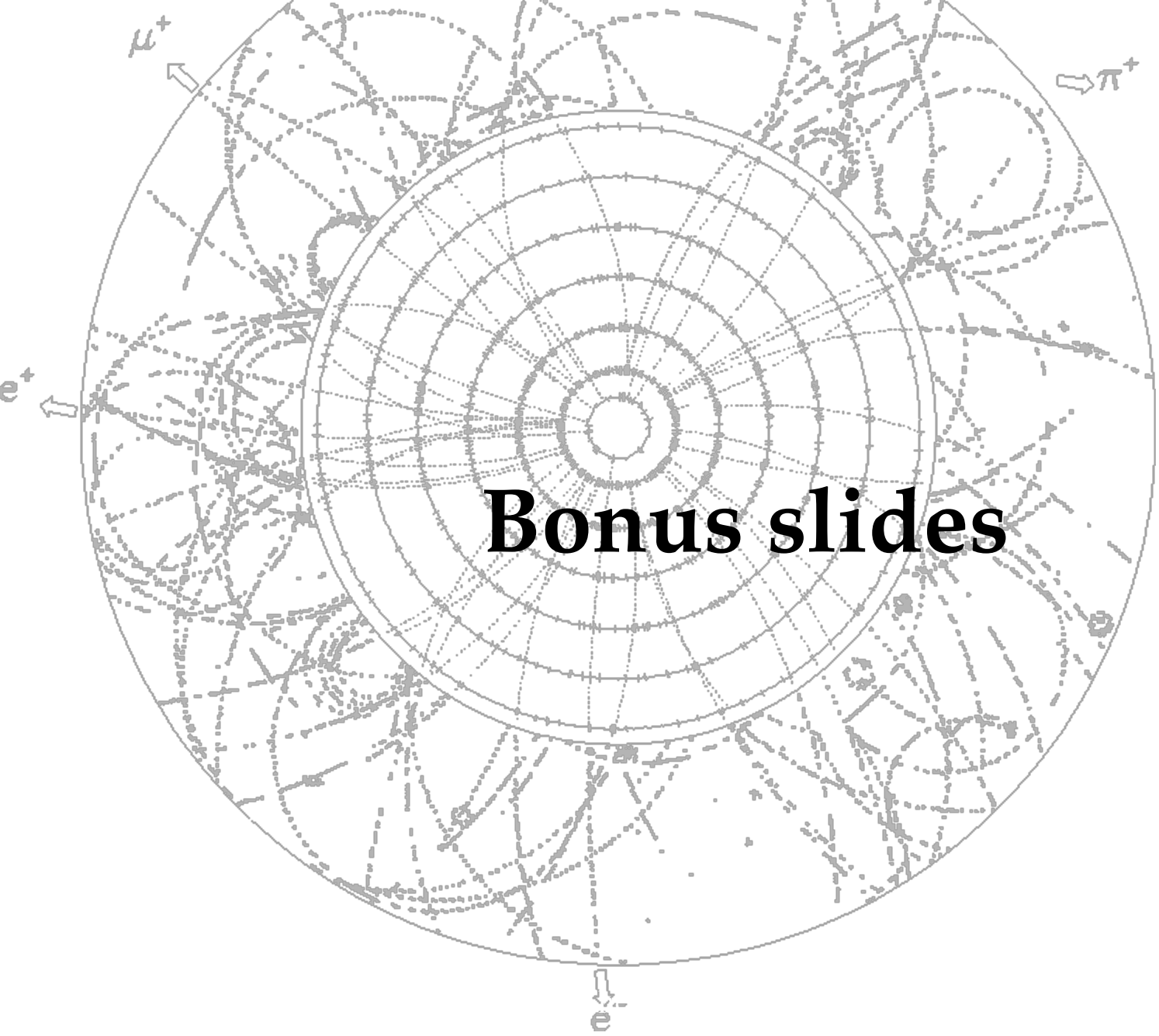
While this is still "only" an improvement in the electron energy regression, and only for lower energies (Zee range), the simultaneous training allows for extending the energy range, by including the Electron Gun MC.

Furthermore, this training might be extended to include photons, as these behave much the same as electrons, and suffer the same sources of uncertainties and smearing.

For improving the $H \rightarrow \gamma\gamma$ resolution, one might use the following loss function and related training samples:

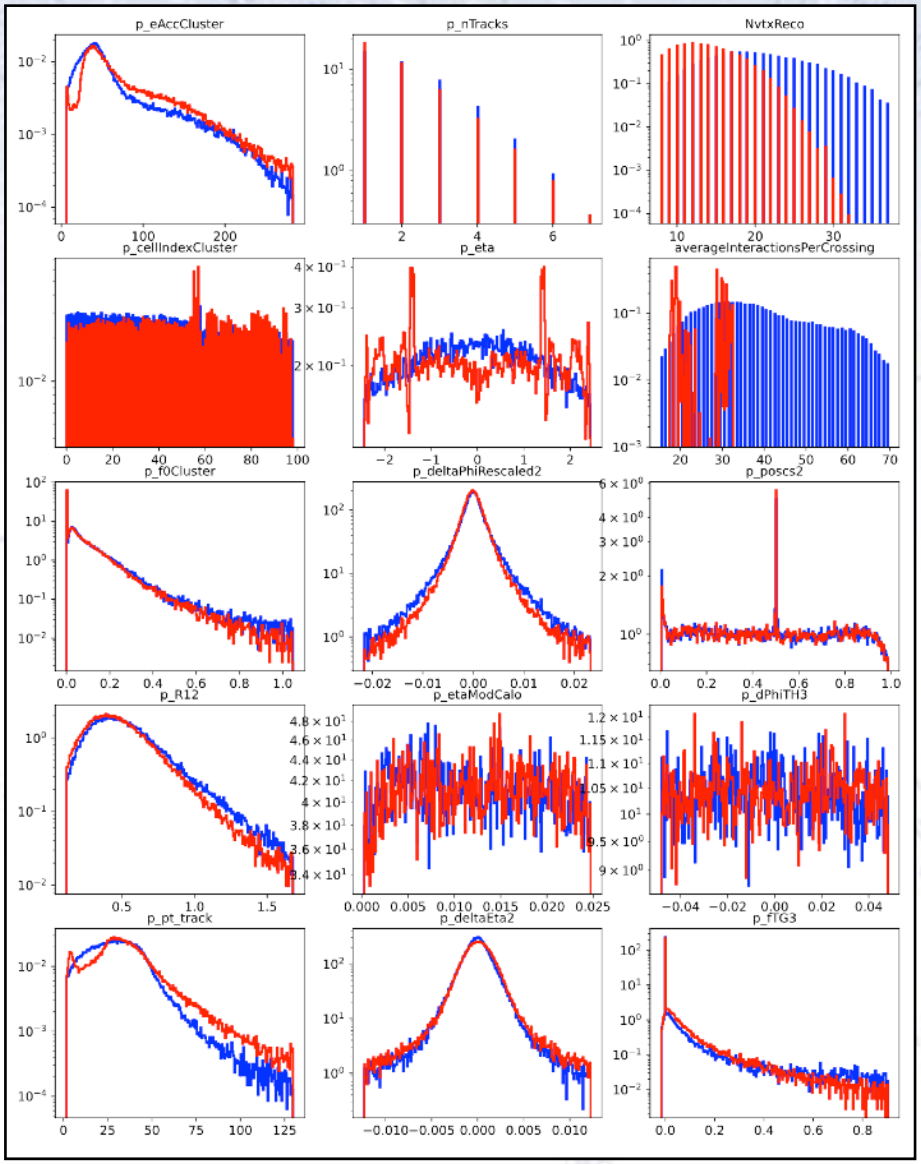
$$\begin{aligned}\mathcal{L}(y,\hat{y}) = \mathcal{L}(y_{(\text{Zee, MC})},\hat{y}_{(\text{Zee, MC})}) + \mathcal{L}(y_{(\text{Zee, Data})},\hat{y}_{(\text{Zee, Data})}) + \\ \mathcal{L}(y_{(Z\mu\mu\gamma,\text{ MC})},\hat{y}_{(Z\mu\mu\gamma,\text{ MC})}) + \mathcal{L}(y_{(Z\mu\mu\gamma,\text{ Data})},\hat{y}_{(Z\mu\mu\gamma,\text{ Data})}) + \\ \mathcal{L}(y_{(H\gamma\gamma,\text{ MC})},\hat{y}_{(H\gamma\gamma,\text{ MC})})\end{aligned}$$

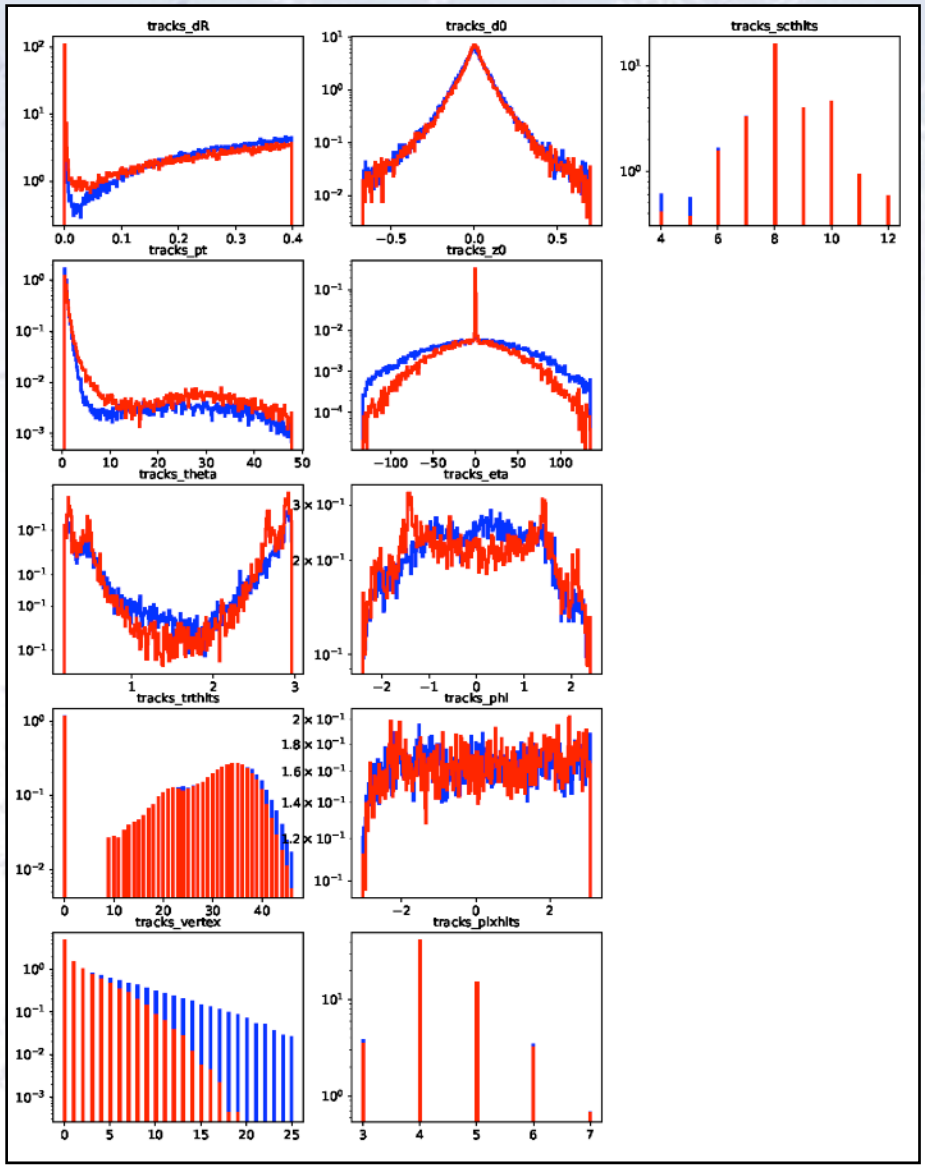
Meanwhile, we are trying to write this up somehow (but Malte is now a Ph.D. in Geneva).



Bonus slides

The input variables



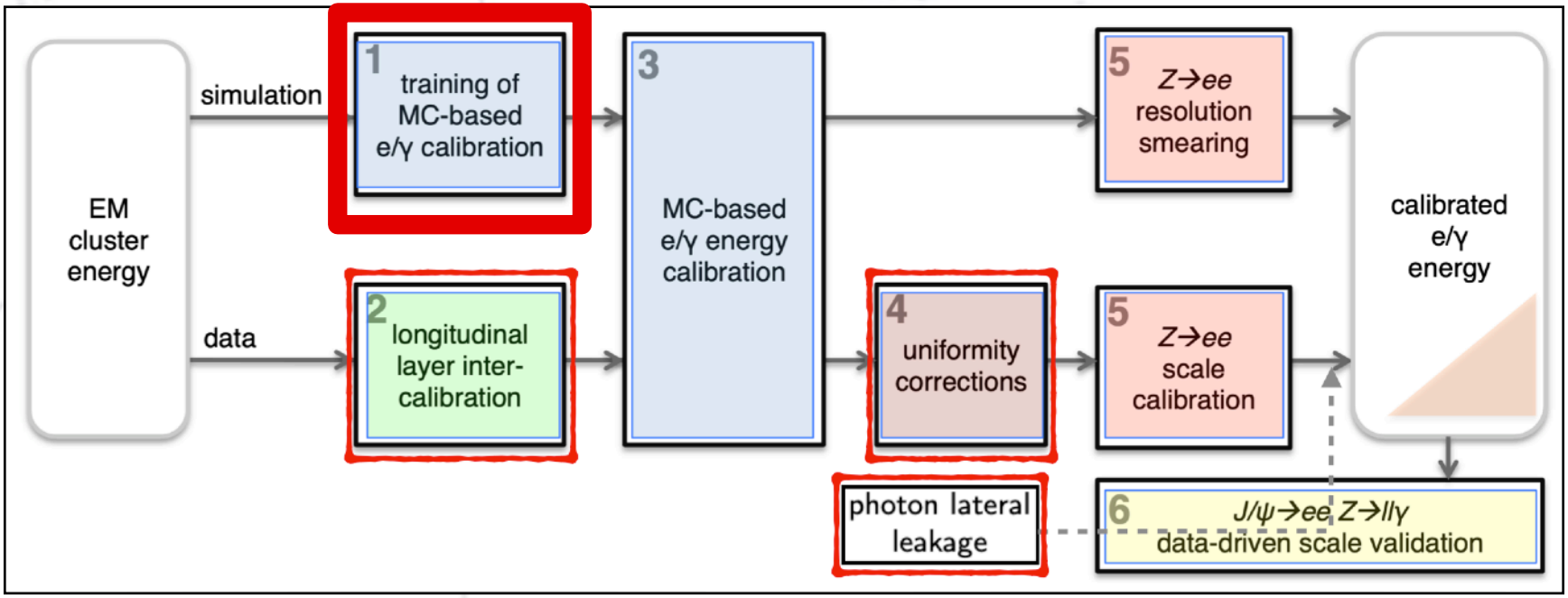


Context

We have for the most part worked **only on MC** (step 1 below), comparing our CNN approach to the "**ATLAS BDT**". Here we see significant improvements.

Lacking the remaining "hard work" of corrections and calibration to match data, our performance improvements **in data** have been decent but "mediocre".

While we have lately included data in training also, the following results will almost surely further improve with the subsequent calibration.



The data

We have used "millions" of mainly Zee decays and Electron Gun.
The data retained for **testing** is as follows:

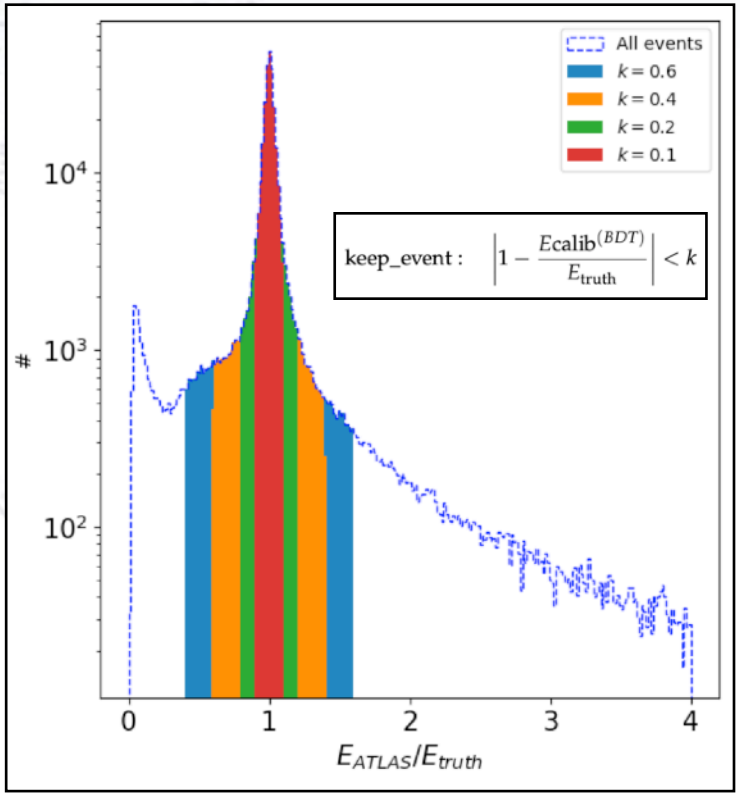
Channel	MC	Data
$Z \rightarrow ee$	1.000.000	450.000
$Z \rightarrow \mu\mu\gamma$	350.000	400.000
$H \rightarrow \gamma\gamma$	310.000	No data available
Electron Gun	1.100.000	No data available

The selection

We applied a general (loose) selection to the different channels in order to obtain large unbiased samples in both MC and data. In the following, we consider mainly the Zee channel.

For the MC, we furthermore required the truth energy to match the reconstructed energy (by ATLAS) to avoid mis-matches (k = 0.6).

$Z \to \mu\mu\gamma$		$Z \to ee$
$\mu\mu$	γ	ee
> 9.5GeV	> 9.5GeV	> 9.5GeV
Loose	Loose	Loose
$\sum Q = 0$	• $N_\gamma = 1$	$\sum Q = 0$
• $N_{\mu\mu} = 1$	Tight	• $N_{ee} = 1$
Trig	• $N_\gamma = 1$	**Event dropped**
• $N_{\mu\mu} = 1$	**Event dropped**	
$m_{\mu\mu} < 82$ GeV		
$m_{\mu\mu} > 20$ GeV		
• $N_{\mu\mu} = 1$		
Loose vs. Tight		
• $N_{\mu\mu} = 1$		
Event dropped		



Differential results - MC

Comparing on electron gun MC in a "known" ATLAS figure style, the improvement is isotropic in eta, and decreases slightly with energy.