



UNIVERSITÀ DEGLI STUDI DI MILANO

PIETRO DANIELE

TESI DI LAUREA TRIENNALE

ANNO ACCADEMICO 2019/2020

A machine learning approach to the electrons and photons classification with the ATLAS detector at the LHC

Relatore: Prof. Leonardo Carlo Carminati

Corelatori: Dott. Ruggero Turra, Dott. Davide Mungo

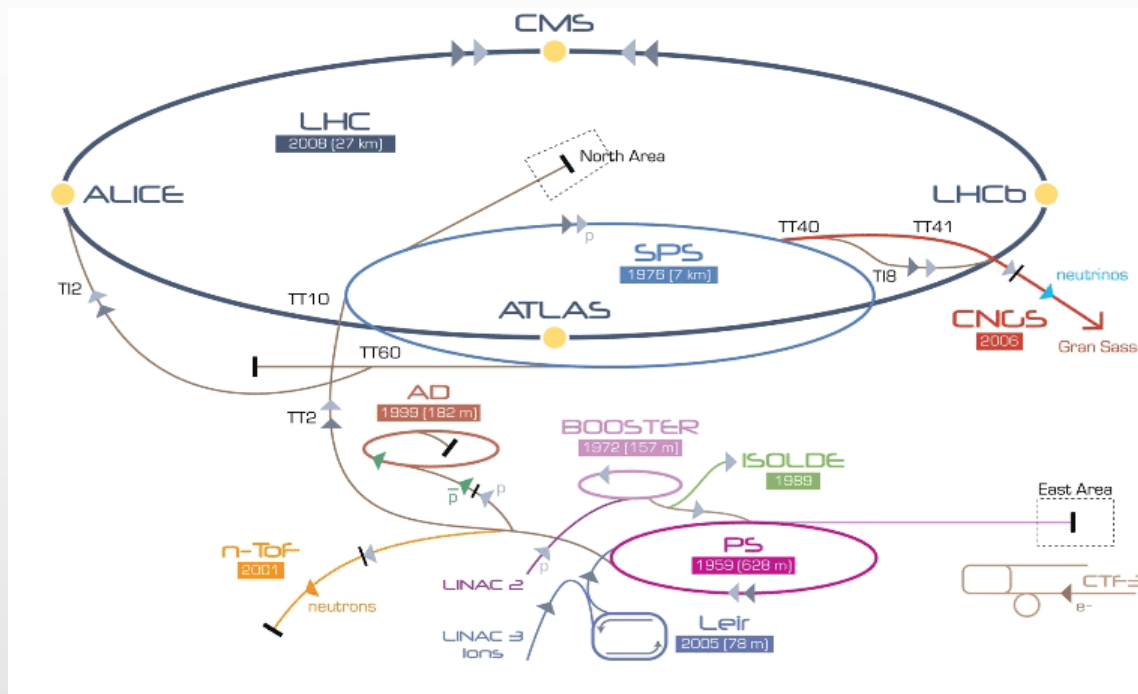
LHC

Large Hadron Collider



LHC

- LHC è un acceleratore di protoni e ioni pesanti installato nel tunnel di LEP lungo 27 km.
- Collisioni protone-protone a 13 TeV.
- Esperimenti ATLAS, ALICE, CMS, LHCb installati nei punti di intersezione

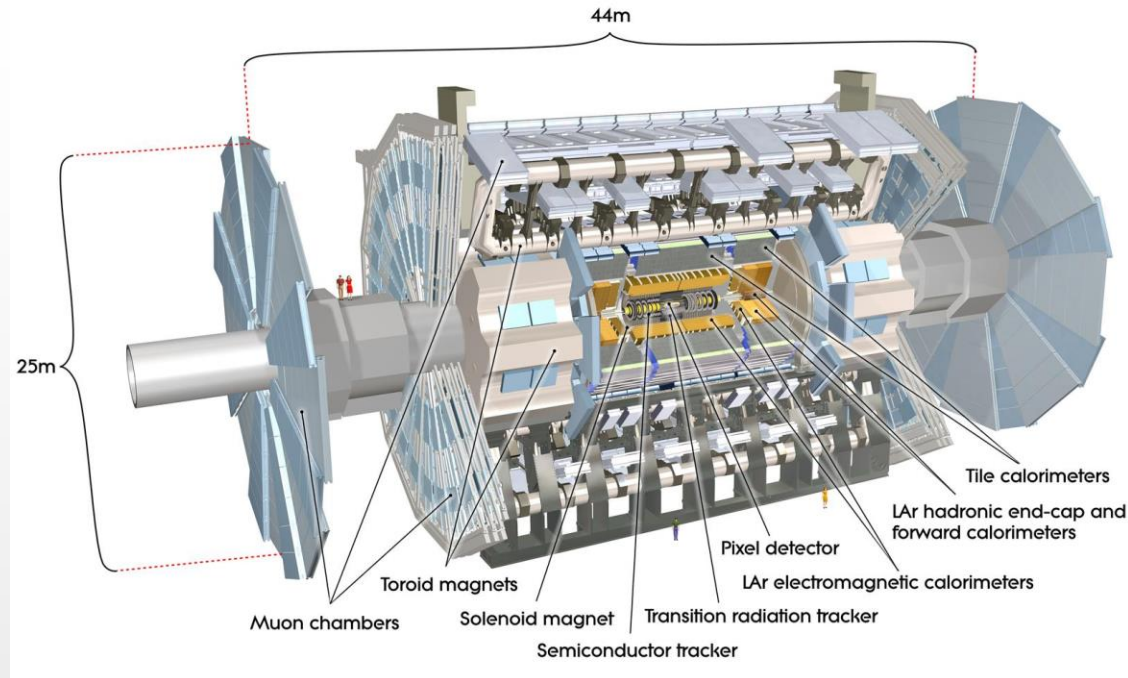


ATLAS



ATLAS: struttura

- ATLAS è uno dei quattro esperimenti posti lungo il tunnel di LHC. Esso composto da vari layer:
 - Tracciatore (ID);
 - Calorimetri;
 - Spettrometro a muoni.
- Campo magnetico:
 - Solenooidale;
 - Toroidale.
- Sistema di trigger.



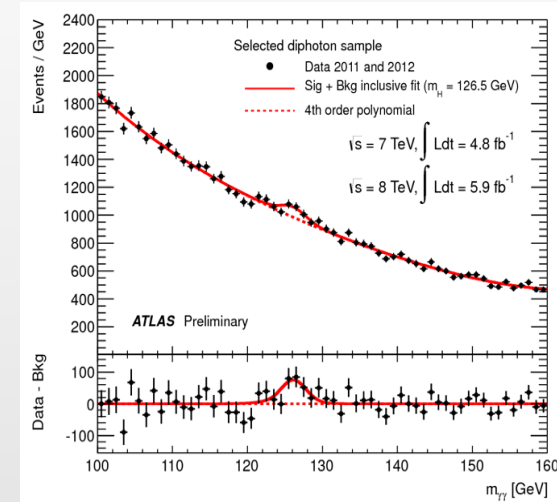
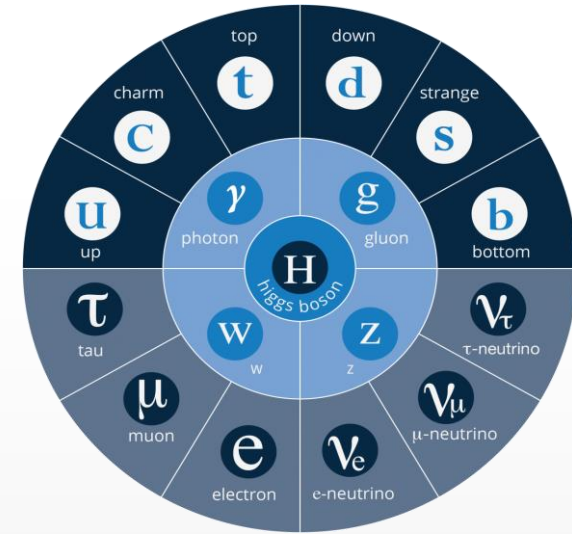
ATLAS: fisica

ATLAS esplora molti campi della fisica con lo scopo di studiare i costituenti fondamentali della materia e le loro interazioni:

- Standard Model (SM) con le produzioni previste di W, Z, top, Higgs;
- ricerca di nuova fisica \Rightarrow Supersimmetria, Dark Matter ...

Una delle più importanti scoperte di ATLAS è l'esistenza del bosone di Higgs: il 4 luglio 2012 ATLAS e CMS annunciano la scoperta di una nuova particella:

- massa ~ 125 GeV
- un bosone consistente con il bosone di Higgs predetto dalla teoria per giustificare la massa delle particelle fondamentali

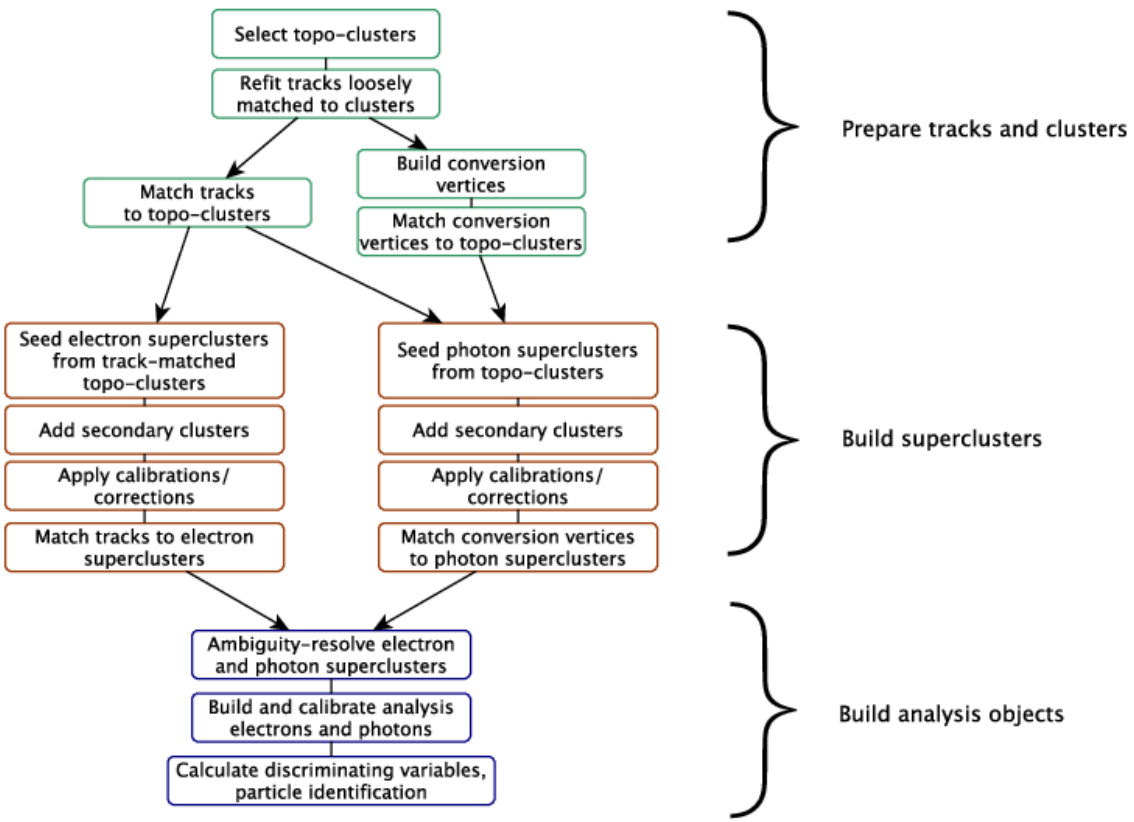


Ricostruzione elettroni e fotoni



e/γ : Algoritmo di ricostruzione

L'algoritmo di ricostruzione è molto complesso e segue gli step mostrati in figura.



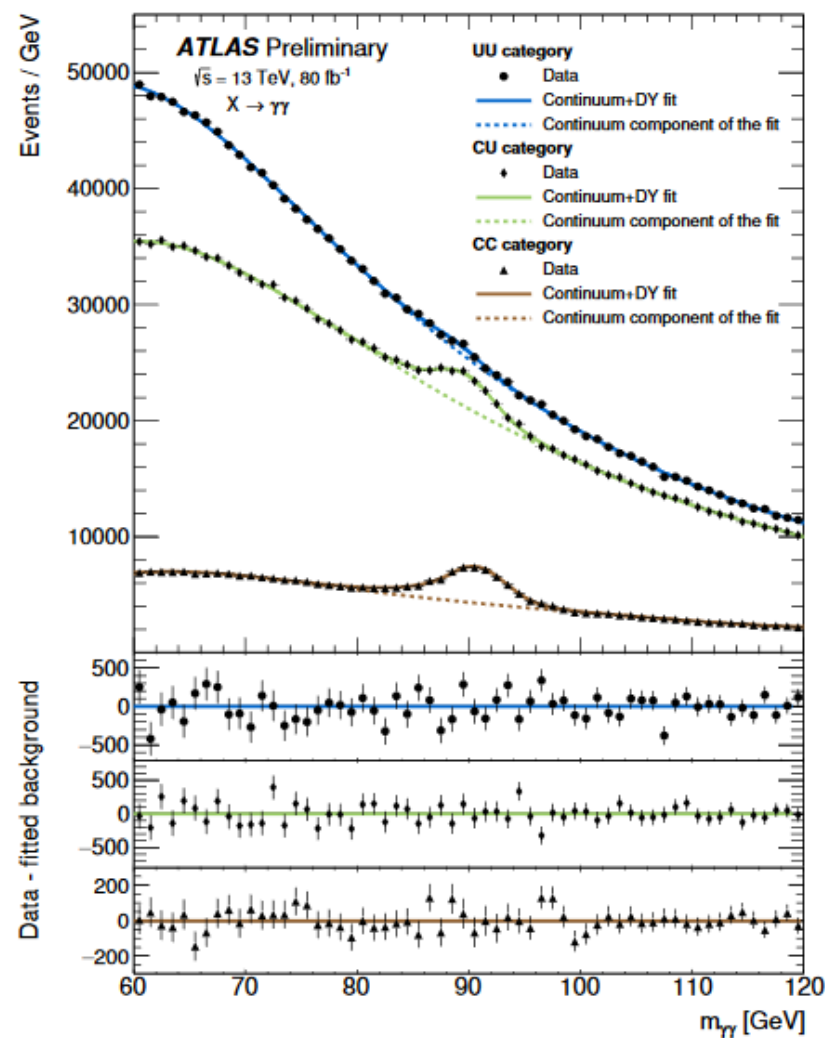
I e e γ sono definiti come:

- un elettrone è un oggetto costituito da un cluster e una traccia ad esso associata;
- un fotone può essere di due tipi:
 - fotone convertito: un cluster associato a un vertice di conversione;
 - fotone non convertito: un cluster non associato né a una traccia di elettroni né a un vertice di conversione.
- negli altri casi sono definiti come ambigui.



Ricostruzione e/γ :

- importante perché eventi con e/γ sono presenti in molte analisi fisiche;
- La ricostruzione può essere *ambigua* e possono essere presenti errori, pertanto una corretta classificazione è importante.
- Ricerca di bosoni di Higgs addizionali a $M_{\gamma\gamma} < 125$ GeV nel canale $\gamma\gamma$: nello spettro di due fotoni a 90 GeV è presente il picco da elettroni dello Z che sono stati (erroneamente) ricostruiti come fotoni.



e/γ : Ambiguity resolver

- Per ogni seed cluster si può produrre sia un elettrone sia un fotone, dunque ogni oggetto, dopo la costruzione dei supercluster, può essere ricostruito come e o γ
- Non è possibile salvare tutte le doppie ricostruzioni. Si usa così l'ambiguity resolver, un semplice algoritmo che:
 - se un particolare oggetto è facilmente identificabile con un e/γ allora solo la corrispondente ricostruzione è salvata;
 - altrimenti l'oggetto è chiamato *ambiguo* ed entrambe le ricostruzioni sono salvate per l'analisi.

In questa tesi sono utilizzati e messi a confronto due ambiguity resolver:

- lo standard ambiguity (old amb) resolver usato in Run 2;
- il nuovo ambiguity resolver (new amb) che verrà introdotto in Run 3, il quale salva più fotoni come ambigui, implementato modificando alcune parti di quello attuale.

Machine Learning e Gradient Boosted Trees



Machine Learning (ML)

Il Machine Learning (ML):

- un'applicazione dell'intelligenza artificiale;
- ha l'obiettivo di generare modelli matematici.

Supervised Learning è una categoria di ML:

- a partire da una variabile di input \vec{x} (che consiste in varie features) si vuole ottenere un output y ;
 - output differenti (categorical, real-value) sono utilizzati per vari problemi (classificazione, regressione);
- si usa un dataset di training per creare il modello.

Durante il training, si cerca di minimizzare la *funzione obiettivo*:

$$\text{Obj} = L(\vec{\vartheta}) + \Omega(\vec{\vartheta})$$

- training loss $L(\vec{\vartheta})$ valuta quanto il modello è predittivo, differenza predizione \Rightarrow realtà;
- termine di regolarizzazione $\Omega(\vec{\vartheta})$ valuta la complessità del modello.

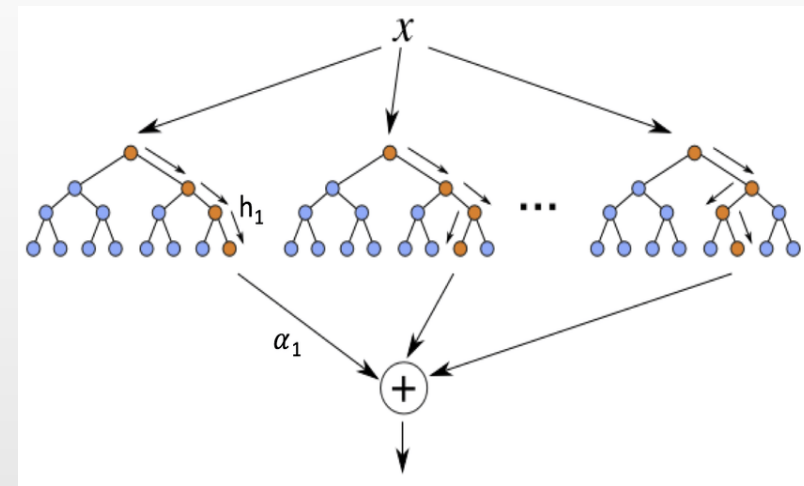
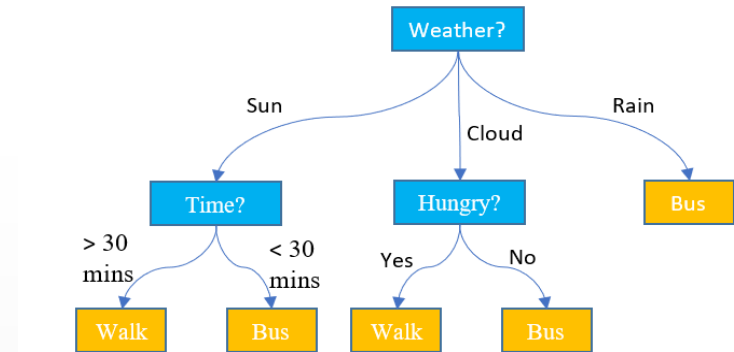
GRADIENT BOOSTED DECISION TREES (GBDT)

GBDT:

- è un algoritmo di ML;
- unità base il *decision tree*: una struttura gerarchica composta da nodi e linee dirette.

Siccome un solo decision tree è una struttura troppo semplice per sistemi complessi si usano più alberi, *decision tree ensembles*:

- assume l'aggettivo *Gradient* quando è sottoposto a training;
- si sommano i risultati dei vari alberi;
- alberi aggiunti gradualmente, ottimizzando di volta in volta la funzione obiettivo \Rightarrow *Additive Training*.



Electron and photon classification



Classificazione con GBDT

- Il processo di ricostruzione:
 - È complicato e molto dispendioso a livello computazionale;
 - Si compie una sola volta;
- Il processo di analisi:
 - è più semplice;
 - può essere svolto più volte.
- Al momento tutte le particelle ambigue sono considerate o tutte elettroni o tutte fotoni (simple-cut) e prima di essere pre-classificate dagli ambiguity tools hanno entrambe le reco (*double reco*). In questa tesi si sviluppa un approccio alla classificazione basato su modelli BDT. Sono sviluppati tre modelli:
 1. modello *double reco*;
 2. modello *old amb*;
 3. modello *new amb*;



GBDT dataset

I campioni \Rightarrow eventi MC di singola particella (elettroni o fotoni).

È splittato in:

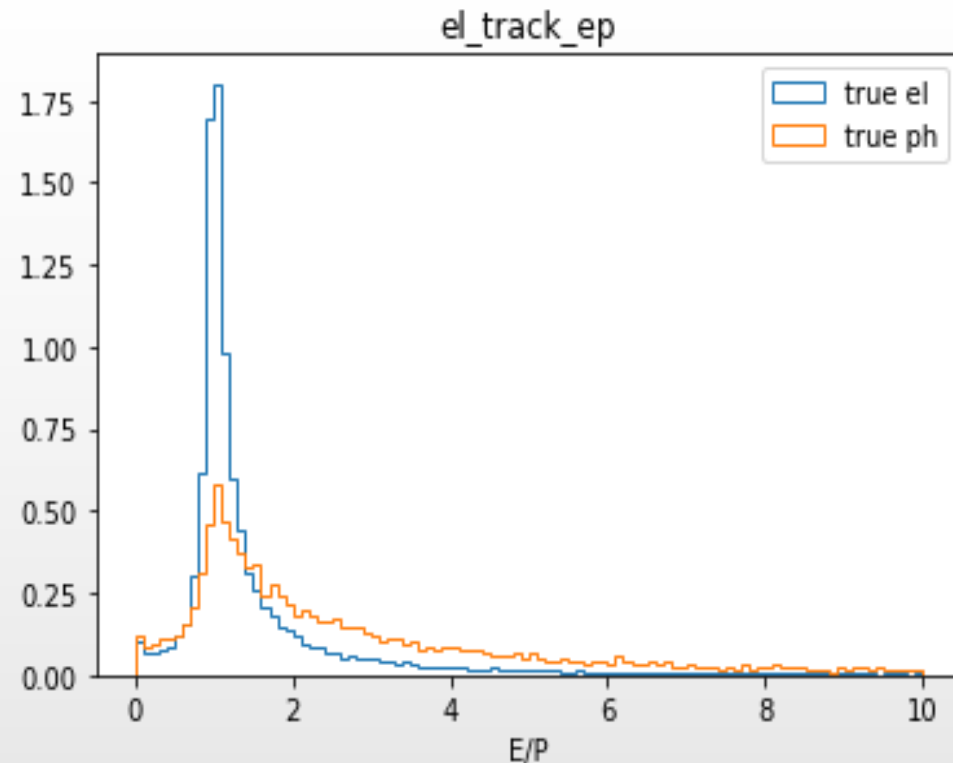
- 20% del set diventa test set;
- 80% del set diventa training set, splittato a sua volta:
 - 75% rimane training set;
 - 25% diventa validation set.



BDT Training: variabili di input

Ogni modello basa le sue scelte su variabili di input. Pertanto tanto più le variabili sono discriminanti tanto più il potere predittivo del modello è forte. Le variabili di input possono essere divise in tre categorie e sono:

- Variabili generali;
- Variabili elettrone, caratterizzano particelle ricostruite come e:
 - hit \Rightarrow elettroni hanno più hit nel primo layer del tracker;
 - qualità della traccia \Rightarrow gli elettroni hanno tracce migliori;
- Variabili fotone, caratterizzano particelle ricostruite come γ :
 - traccia proveniente da un fotone convertito è discriminante;
 - hit e p_T delle tracce associate al fotone conv.



BDT Training: variabili di input

Variabili generali

- cl_pt
- cl_eta

Variabili elettrone

- el_track_hasInnermostHits
- el_trkPixelHits
- el_trkSiHits
- el_trackz0
- el_track_ep

Variabili fotone

- ph_track1_hasInnermostHits
- ph_track2_hasInnermostHits
- ph_convtrk1nPixHits
- ph_convtrk1nSCTHits
- ph_convtrk2nPixHits
- ph_convtrk2nSCTHits
- ph_zconv
- ph_Rconv
- ph_pt1conv
- ph_pt2conv
- ph_ptconv
- pt1conv/ptconv

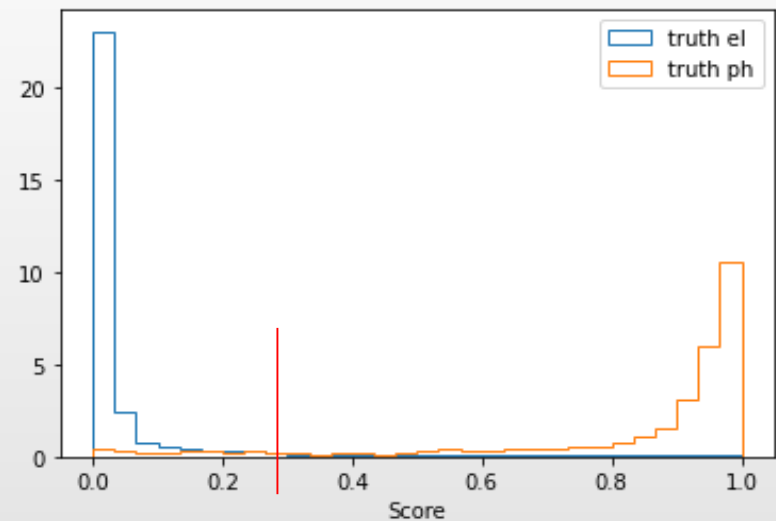
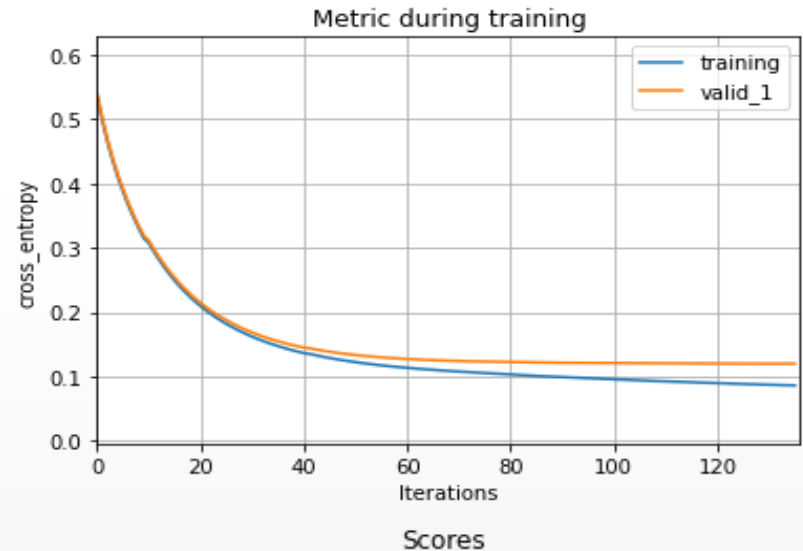
BDT Training and Scores

Finito il training, si può analizzare la qualità e la struttura del modello guardando:

- comportamento training loss.

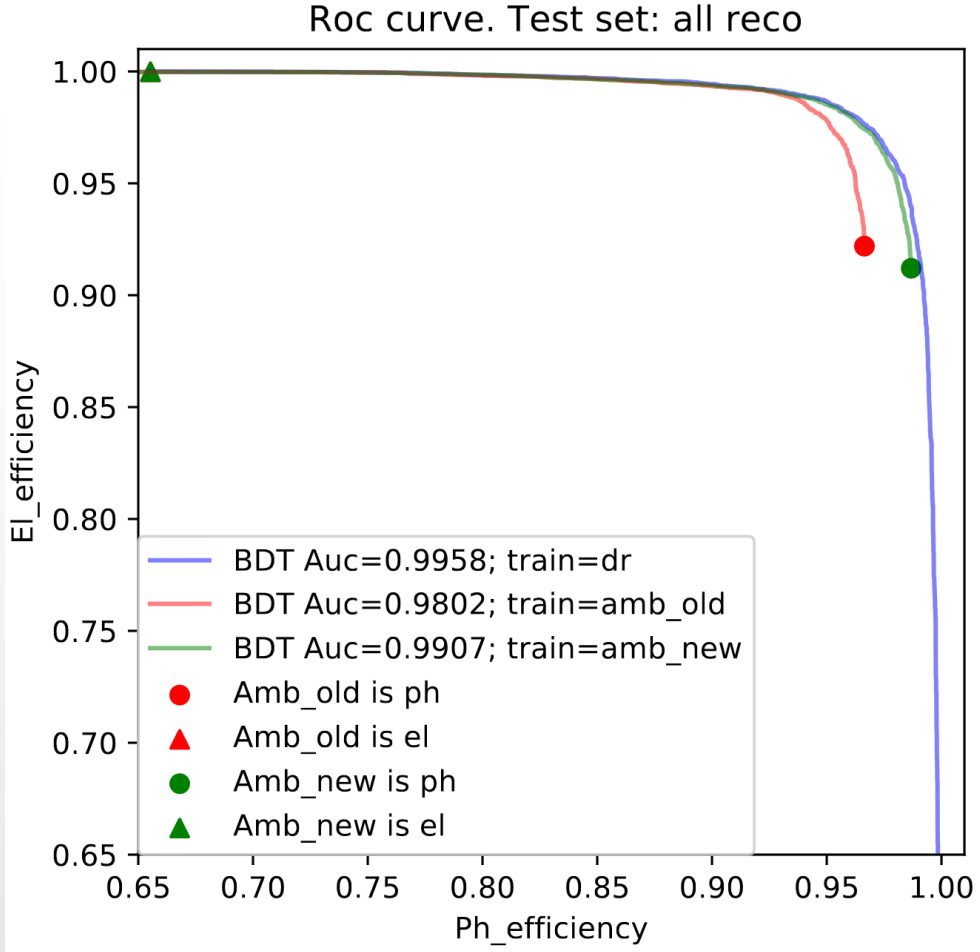
I modelli applicati al test set forniscono uno score compreso tra 0 e 1 come output.

- Classificazione con soglie
- Efficienze elettroni e fotoni
- ROC curve



Risultati

- All reco set: set composto da singole e doppie ricostruzioni;
- ROC curve: Receiver Operating Characteristic curve;
- AUC: Area Under Curve.



Conclusioni

Sono stati implementati tre BDT per la classificazione degli ambigui che testate sul set *all reco* ottengono ottimi risultati:

- modello *double reco* \Rightarrow $AUC \simeq 0.9958$
- modello *old amb* \Rightarrow $AUC \simeq 0.9802$
- modello *new amb* \Rightarrow $AUC \simeq 0.9907$

L'utilizzo del *nuovo* ambiguity resolver permette di ottenere:

- una più alta efficienza di classificazione;
 - un aumento massimo del 7% sull'eff. degli elettroni;
 - un aumento massimo del 3% sull'eff. dei fotoni
- Possibilità di selezionare il working point ideal a livello di analisi.

Si è così dimostrato che un approccio con BDT alla classificazione e/γ permette di ottenere ottimi risultati, tanto che il modello basato sul nuovo ambiguity resolver raggiunge il limite teorico. Dunque i modelli *old/new amb* possono essere utilizzati a livello di analisi.



LHC: collisioni protone-protone

Il numero di eventi N per secondo prodotti dalle collisioni p-p è proporzionale a:

- Luminosità $L = 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$
- Sezione d'urto $\sigma = 80 \text{ mb}$

Perciò $N = \sigma \cdot L \simeq 10^9 \text{ s}^{-1}$.

Ci sono due tipi collisioni p-p:

- Soft: collisioni a grande distanza tra due protoni incidenti con momento trasferito ridotto. Particelle prodotte con alto momento longitudinale e basso momento trasversale p_T
- Hard: interazioni a piccola distanza tra quark (o gluoni) con la produzioni di particelle con alto p_T .

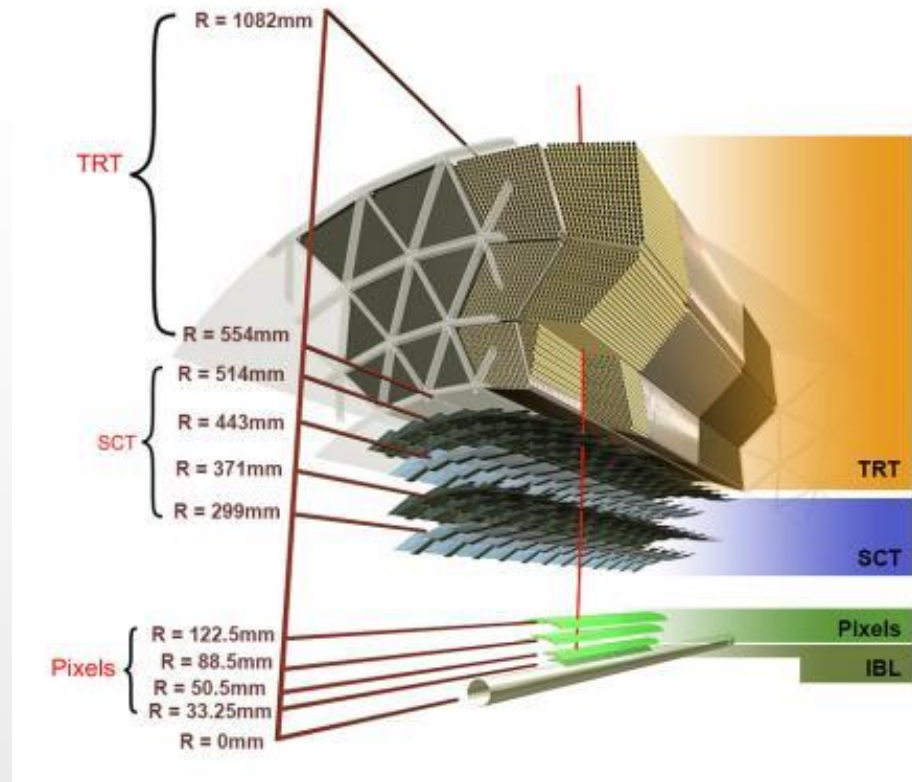
ATLAS: Inner Detector

ID è il layer più interno ed è costruito per fornire:

- un riconoscimento ermetico e robusto del pattern
- un'eccellente risoluzioni del momento
- una misura del primo e secondo vertice per le particelle cariche

Composto da 3 sub-detector:

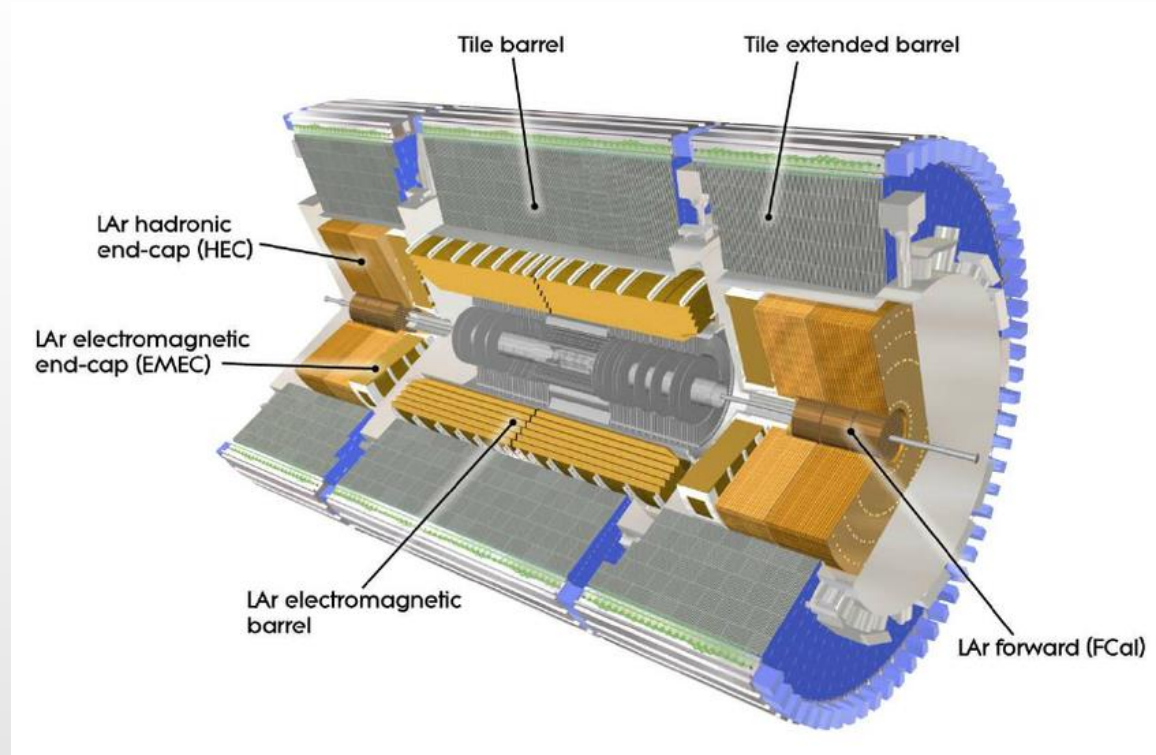
- Pixel Detector
- Semi Conductor Tracker (SCT)
- Transition Radiation Tracker (TRT)



ATLAS: Calorimetri

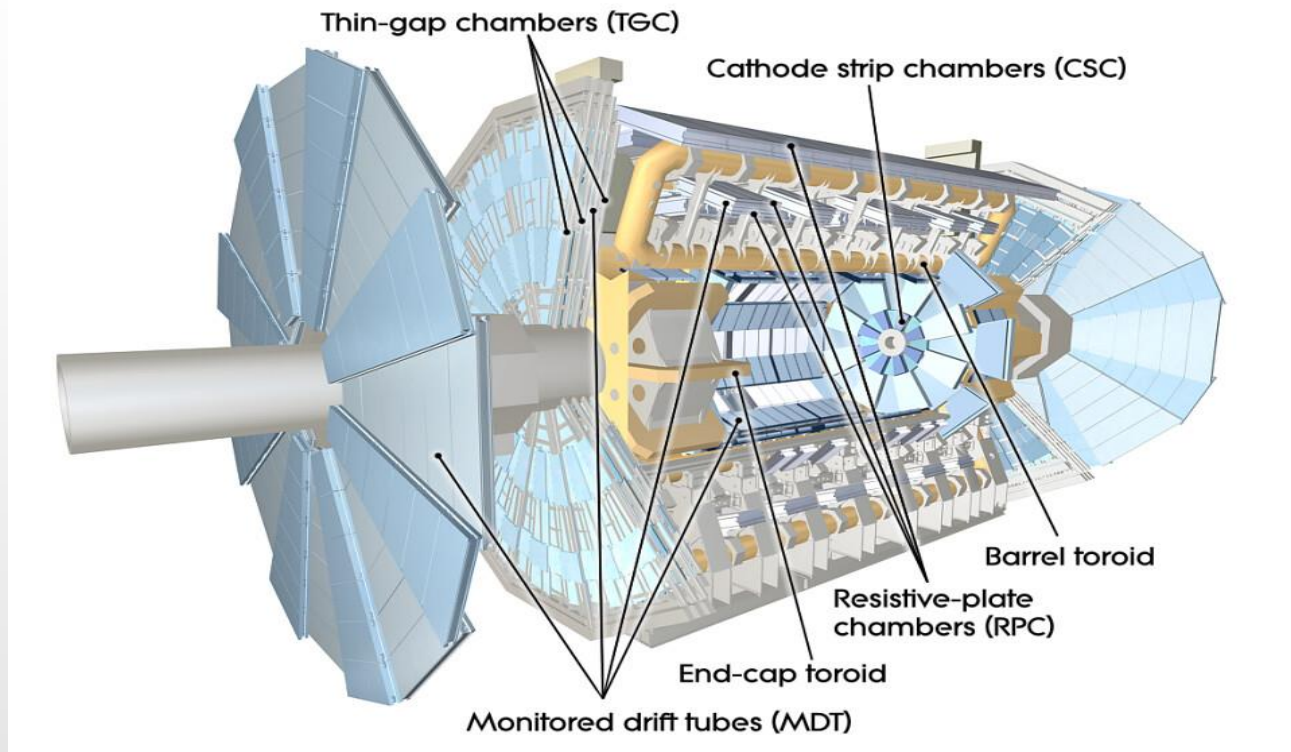
I calorimetri misurano l'energia persa dalle particelle che passano nel detector e in ATLAS ce ne sono di due tipi:

- Calorimetro elettromagnetico
- Calorimetro adronico



ATLAS: Spettrometro a muoni

Lo spettrometro a muoni è la parte più esterna di ATLAS ed è costruito con lo scopo di misurare il momento dei muoni, i quali superano i detector più interni. Questo viene fatto sfruttando un grande sistema magnetico toroidale che curva le traiettorie delle particelle cariche.

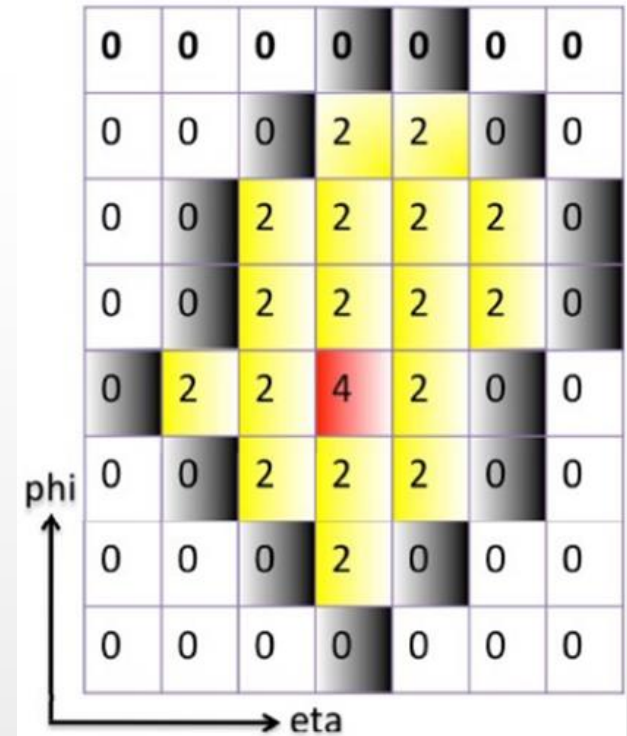


Topo-cluster

- Calorimetro diviso in celle, per ciascuna delle quali viene definito un valore di significanza

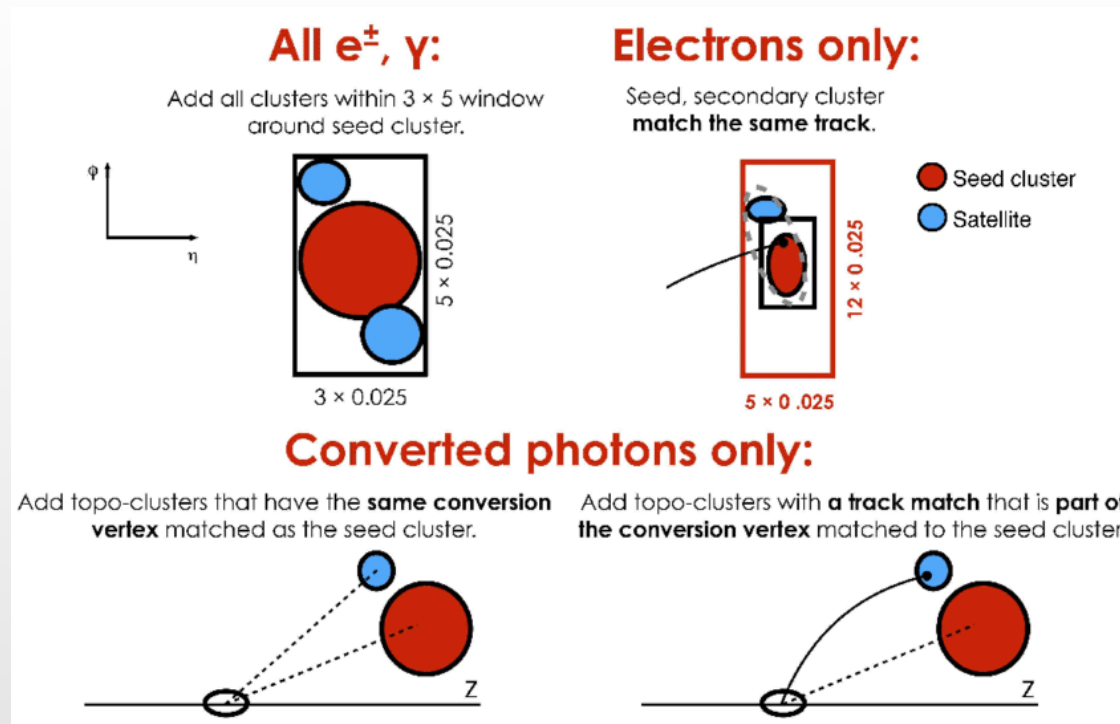
$$\zeta_{\text{cell}} = \frac{E_{\text{cell}}}{\sigma_{\text{cell}}}$$

- Algoritmo di ricostruzione 4-2-0:
 - Cella significanza $\zeta_{\text{cell}} \geq 4$ cluster seed;
 - Aggiunta 3D delle celle con significanza $\zeta_{\text{cell}} \geq 2$ al seed;
 - Unione cluster con vicini in comune;
 - Aggiunta di una corona delle celle più vicine con $\zeta_{\text{cell}} \geq 0$.

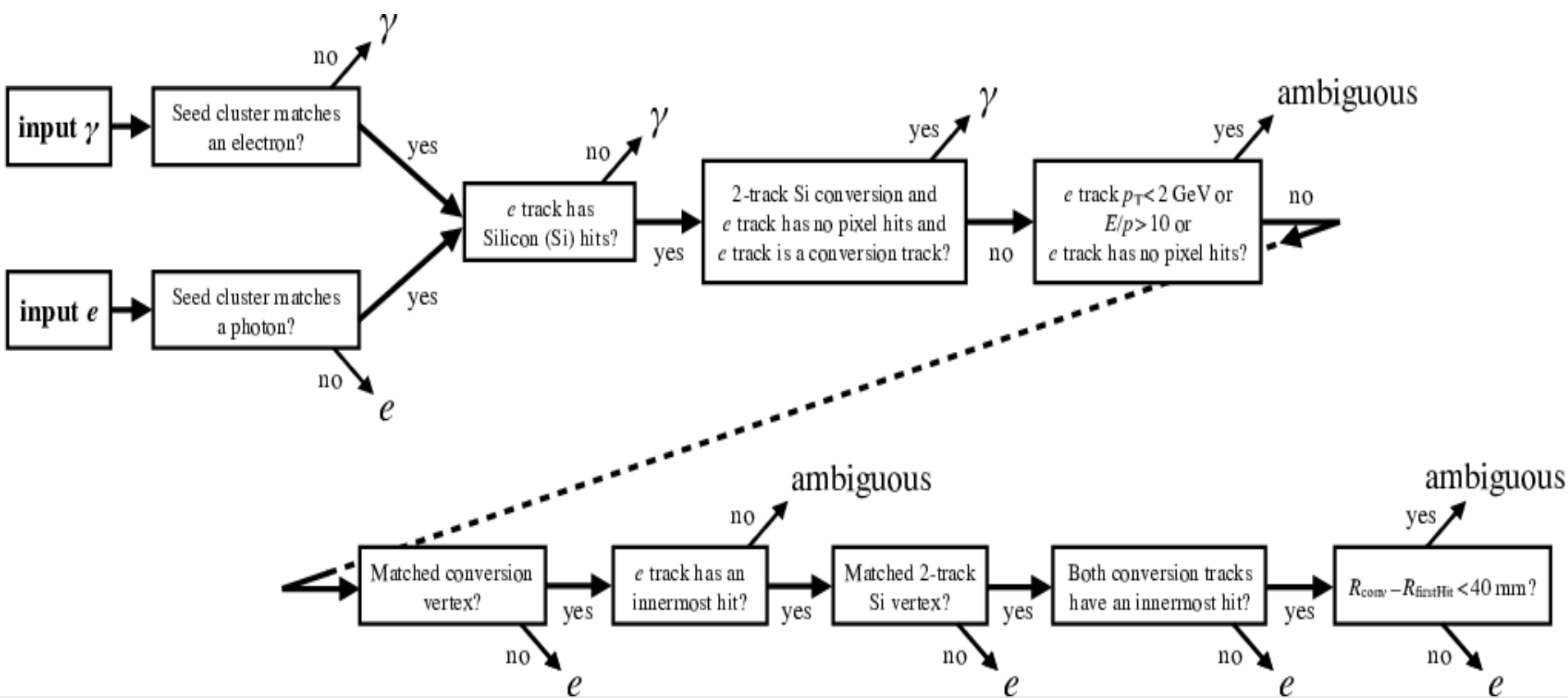


Superclusters

- Input: cluster associati a tracce o vertici di conversione
- Aggiunta di cluster satelliti (da Bremsstrahlung) o unione di clusters da formazione coppie



e/γ : Ambiguity resolver



BDT Training: ottimizzazione iperparametri

Gli iperparametri:

- parametri regolabili LGBM;
- regolano il processo di training;
- determinano le performance del modello.

Ottimizzazione degli iperparametri:

- ha l'obiettivo cercare nello spazio degli iperparametri la migliore configurazione sul training set;
- la ricerca, usata in questa tesi, è di tipo Bayesiano;
- gli iperparametri ottimizzati sono:
 - bagging;
 - numero di foglie;
 - feature fraction;
 - learning rate.

Machine Learning (ML): funzione obiettivo

Durante il training, per valutare quanto bene il modello descrive il training set si definisce la *funzione obiettivo*:

$$\text{Obj} = L(\vec{\vartheta}) + \Omega(\vec{\vartheta})$$

- training loss $L(\vec{\vartheta})$ che valuta quanto il modello è predittivo
- termine di regolarizzazione $\Omega(\vec{\vartheta})$ che valuta la complessità del modello
- $\vec{\vartheta}$ è il set dei parametri del modello

Un buon bilanciamento tra i due termini della funzione obiettivo porta alla creazione di un buon modello.

Alto $L(\vec{\vartheta})$ e basso $\Omega(\vec{\vartheta}) \Rightarrow$ Underfit

Basso $L(\vec{\vartheta})$ e alto $\Omega(\vec{\vartheta}) \Rightarrow$ Overfit

