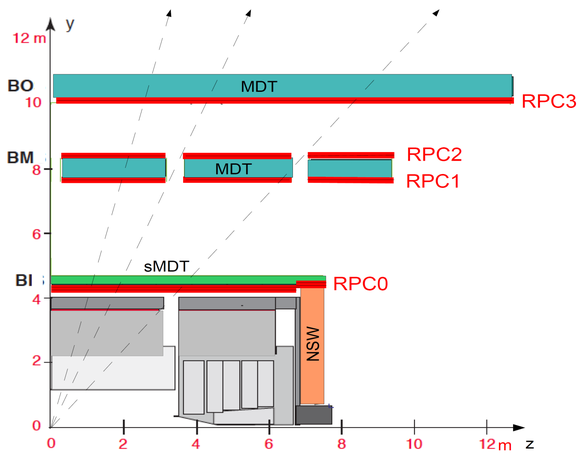
Trigger dei muoni L0 con reti neurali veloci

Martino Errico, Davide Fiacco, Stefano Giagu, Giuliano Gustavino, Valerio Ippolito, Graziella Russo
per la collaborazione ATLAS

Muoni in ATLAS Fase-II

Spettrometro per muoni del barrel
- Ricostruzione muoni con misure nel piano di curvatura (η, r) e in quello trasverso (ϕ, r)
- Diviso in 2 lati, 16 settori per lato
- RPC (Resistive Plate Chamber):
  - risoluzione spazio-temporale di 1 cm x 1 ns
  - 3 stazioni da 2 RPC, su piani cilindrici con asse sul fascio
  - per Fase-II (HL-LHC), nuova stazione di RPC (RPC0):
    - 3 strati di RPC, per un totale di 9 strati e 4 stazioni
    - compensa la diminuzione di efficienza degli RPC
    - permette schema di coincidenze con accettanza maggiore



Sezione dello spettrometro (ATLAS-TDR-026)

Trigger L0 (Livello 0) per Fase-II
- Implementato su FPGA (Field-Programmable Gate Array)
- 1 FPGA (XCVU13P) per settore, di cui metà utilizzata per la logica di settore (SL)
- Risorse per istanza di trigger:
  - 432k Look-Up Table (LUT)
  - 864k Flip-Flop (FF)
  - 23.6 Mb Block RAM (BRAM)

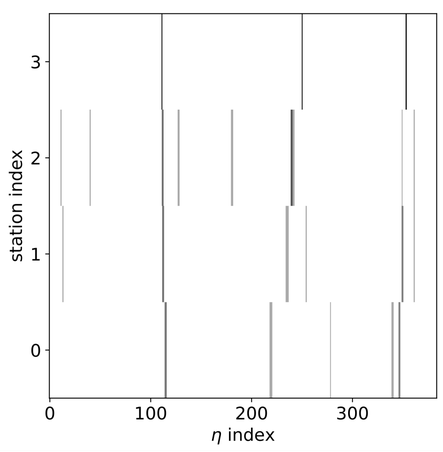
✦ Occupazione attuale:

Risorse	Occupazione
LUT	29%
FF	15%
BRAM	2%

Trigger con CNN

Logica di settore
- Soglie di selezione su p_T da 4 a 80 GeV
- Multipli candidati per semisettore
- Output per candidato:
  - soglia più elevata soddisfatta
  - p_T, η, q e ϕ
  - schema di coincidenza soddisfatto
  - z per stazione di RPC

Logica con rete convoluzionale
- Motivazioni:
  - bassa occupazione delle FPGA
  - capacità di identificare eventi meno comuni
- Input: immagini 4x384
  - un pixel acceso per hit sugli RPC
  - 1 super-strato per stazione (da 9 strati a 4)
  - 384 bin di η in (0, 1.05)
  - fino a 3 muoni simulati per evento, con p_T in (0, 30) GeV
  - rumore casuale simulato
  - densità media di immagini a 3 muoni: 3%
- ✦ Output della rete:
  - numero di candidati per semisettore in [0, 3]
  - ricostruzione di fino a 2 candidati: p_T, η e q

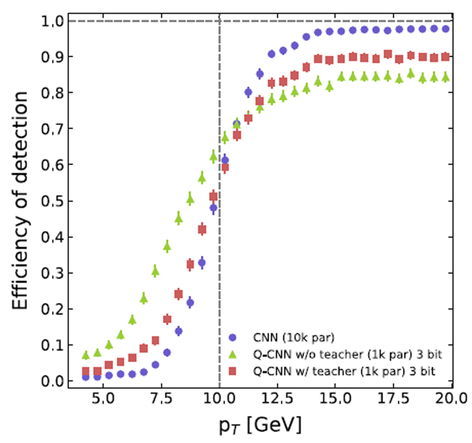


Esempio di immagine di input con 3 muoni e rumore casuale

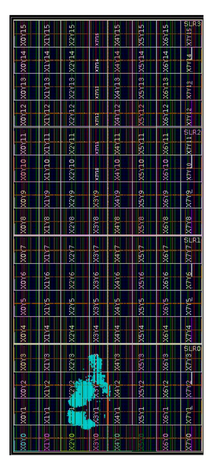
Implementazione CNN

Prima implementazione
- Modelli allenati e testati su eventi a singola traccia
- Modelli compressi per implementazione su FPGA:
  - knowledge distillation (modello student allenato con la guida di un modello teacher più grande)
  - quantizzazione (3 bit)
  - latenza richiesta: 400 ns (readout incluso)
- Metriche di qualità:
  - curva di turn-on di efficienza
  - frazione di eventi di fondo accettati
- Prestazioni desiderate:
  - plateau di efficienza ≥ 90%
  - efficienza sul fondo ≤ 0.2%

→ Risultati iniziali:
- ✔ tutti i requisiti di prestazione soddisfatti
- ✘ latenza raggiunta: 440 ns
- ✔ occupazione LUT, FF, BRAM < 1%



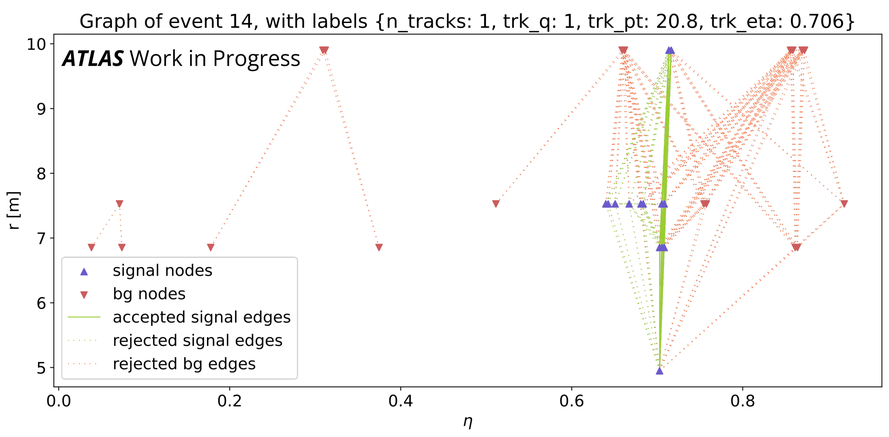
Confronto fra curve di turn-on del teacher (CNN) e dei due student quantizzati (Q-CNN), il secondo dei quali allenato con knowledge distillation (Eur. Phys. J. C 81, 969 (2021))



Occupazione FPGA

Trigger con GNN

Test rete a grafi come alternativa ad algoritmo con CNN
- Motivazioni test architettura GNN
  - gestione più efficiente di input non denso
  - migliore capacità di separare tracce
  - maggiore adattabilità a geometrie irregolari
  - maggiore flessibilità nell'aggiungere variabili ai dati
- ✦ Architettura non ancora definitiva:
  - modello a singola traccia
  - non quantizzato



Esempio di risultato della GNN usata per classificazione degli archi

Input:
- nodi: hit sugli RPC, come punti in (η, r)
- archi: connessioni fra nodi, realizzate se cinematicamente permesse ad un muone di $p_T \geq 3$ GeV con traiettoria circolare
- grafi statici (connettività costante)

✦ Output alternativi testati:
- classificazione archi
- ricostruzione impulso

Risultati preliminari

Estensione CNN a eventi multi-traccia
- Identificazione e ricostruzione fino a 2 muoni
- Prestazioni peggiorano al diminuire della distanza $|\Delta\eta|$ fra coppie di tracce in un evento

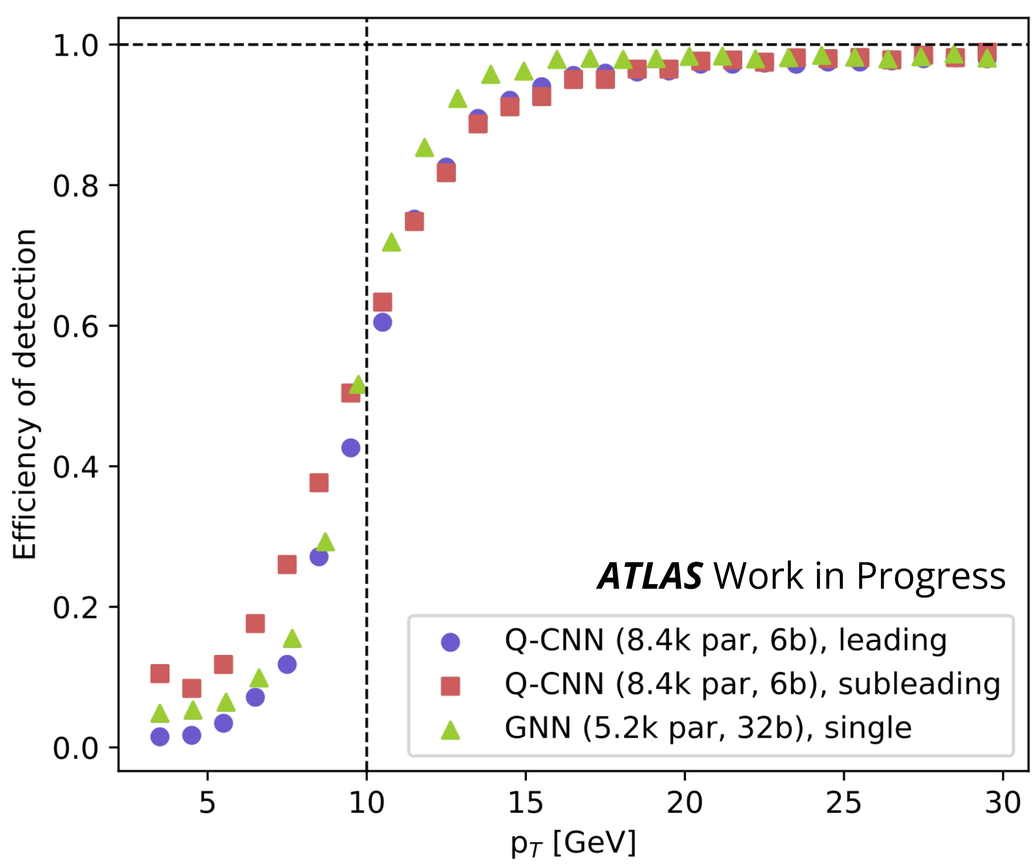
→ Prestazioni, misurate su campione di test con $|\Delta\eta| > 0.06$:
- ✔ efficienza > 90% da 14 GeV per entrambe le tracce
- ✘ efficienza minima 2% per traccia leading e 10% per traccia subleading
- ✔ efficienza complessiva sul fondo 0.17%

✘ Dimensioni modello maggiori che nel test iniziale (8.4k par a 6 bit vs 1k a 3 bit)
- Conclusione: anche su campione di test semplificato, scarsa reiezione delle tracce subleading sotto soglia e dimensioni eccessive del modello rendono architettura inadatta a caso multi-traccia

Algoritmo con GNN

→ Risultati modello a singola traccia:
- ✔ efficienza > 90% raggiunta prima della CNN
- ✔ efficienza minima 5%
- ✔ numero di parametri (5.2k) inferiore a CNN

→ Sviluppi futuri GNN:
- estensione a caso multi-traccia
- aggiunta degli altri output richiesti
- compressione e quantizzazione
- implementazione su FPGA



Confronto preliminare fra curve di turn-on degli student CNN multi-traccia quantizzati (6b) e della GNN

Incontri di Fisica delle Alte Energie 2025 - Cagliari - Aprile 2025