AI in the LHCb group

Jacco & Gerhard

1.1) Where is AI having a major impact internationally in production? With what computing resource requirements?

- Reco/selection: majority of data is triggered by online ML models (MLPs / lipschitz NNs), relatively small and fast inference, no huge computing resources (but online on GPUs, requires software infrastructure dev.)
- Particle ID: relatively simple classifiers, no real demands
- Flavour tagging*: relatively simple classifiers, no real demands

2.1) What is the involvement of Nikhef in ML R&D?

- MLOps / engineering within LHCb trigger (“RTA”)
- ‘To some extent’ on various topics within IRIS-HEP and CERN summer student programmes
- ‘To some extent’ on various student projects in Maastricht (e.g. anomaly detection)
- PINNs for track extrapolation across the magnet
- (variational quantum algorithms for velo track reco)

2.2) Who are the in-house experts? On what?

- No ‘real’ ML experts in-house
- To some extent: Roel Aaij, Maarten van Veghel, MLOps
- To some extent: Jacco de Vries, Xenofon Chiotopoulos (PhD), on model building
- (Gerhard?)
- At Maastricht DACS: collaboration with Kurt Driessens (prof. In AI, mostly on anomaly detection and transformers)
- At Maastricht/Hasselt: Jochen Schuetz, mathematician.

2.3) MSc/BSc projects using AI/ML? Involve CS students?

- Very positive experiences with MSc. CS students reported by multiple people. (Andrii, Maarten, Maastricht,Roel,UvA/VU)
- Involvement of many BSc students in Maastricht

2.4) Do you use AI-based tools?

- Co-pilot(s) sometimes used but sparsely
- ChatGPT to various degrees of dependency

3.1) Near-future deployment that brings strong physics perf?

- ‘Inclusive’ flavour tagging with DeepSets / transformer-based models
- Generative models for fast simulation: LHCb CALO (and RICH), even parameterized reco: GANs, normalizing flows
- PINNs to replace places with heavy computation like magnetic field lookups
- Online: small models with new sophisticated loss functions / FoMs / training, Replacing current heuristic / simplified steps in reconstruction with NN-based tools

- Online: Run Hlt2 fully on GPUs, plenty opportunity for (larger) NN’s
- (velo tracking with GNNs)
- (PV finding with CNN or GNN, joint effort with Atlas)
- Optimize control and data flow within trigger selections
- Data Quality monitoring (identify anomalies, classify - reinforcement learning?)

3.2) Major novel ML/AI applications in the pipeline?

- ‘Inclusive’ flavour tagging with DeepSets / transformer-based models
- Generative models for fast simulation: LHCb CALO (and RICH), even parameterized reco: GANs, normalizing flows
- PINNs to replace places with heavy computation like magnetic field extrapolation
- Anomaly detection, anomalous topologies (VAE, transformers?)
- Reconstruction of complex signatures like LLP’s, (semi-)emerging jets, new exotics with e.g. normalized autoencoders

- Full event interpretation (with sequential GNNs)

https://indico.cern.ch/event/1433541/attachments/2893486/5072673/ML4LHCb_CERNSeminar_JGP_100724.pdf

3.3) Expected comp. Resource requirements?

- use the LHCb GPUs in the trigger farm
- For genAI use in simulation, benefits GPUs in datacenters / WLCG tiers
- Current infrastructure (stoomboot) GPUs decent for our current applications
- “More GPUs I guess are always nice”

- (for quantum: GPU backends for simulation, or access to hardware)

4.1) near-future Nikhef involvement - ambition?

- Jet reconstruction / substructure at large pileup, learn from Atlas/Alice?
- Particle isolation tools (at high pileup) (???)
- Flavour tagging with transformer-based models
- Tracking (mag.field extrap., high pile-up combinatorics, “incomplete” tracks)
- (Global) Trigger Optimization
- Online deployment of AI/ML algorithms

- (quantum: R&D in track reco)

4.2) Clear leading institutes / consortia in these R&D?

- MIT is leading in several areas
- Coruna started exploring their flavour tagging ambitions
- Online / high-performance (throughput): not so much, ours is a niche case.
- Soon: Spain in collab. with its supercomputers and dedicated funds (Barcelona, Valencia, Santiago)

4.3) How do Nikhef efforts compare?

- Less dedicated focus on AI, though local infrastructure is excellent
- We are a bit behind other institutes, but expected to reach a reasonably similar level soon
- Online: leading on inference, less so on training / architecture design

- (quantum in HEP: everyone is new, we lead in LHCb)

4.4) Who are your partners (inter)nationally?

- SURF
- DACS in Maastricht
- Coruna* / ‘LHCb’

- (for quantum: QuTech and Eindhoven, also SURF, internationally CERN QTI)

4.5) What expertise, person-power and infra are you missing to compete?

- Person power (PhD’s / PD’s) to do the work
  - both exploratory (what works?) and production (finished product!)
- ML + Statistics expertise, with focus on the math

4.6) Expectations regarding AI-driven tools?

- Would be great to have LHCb-specific chatGPT (or other LLM) to help train newcomers (or old professors), train on internal notes / documentation
- Seems to have a lot of promise for speeding up coding, but also a danger of not understanding the functionality ‘under the hood’
  - “Reviewing code generated by chatGPT does not fill me with joy”
- Private paid chatGPT service: used extensively, very welcome for both coding and document review.

5.1) Disruptively big: what current problems of interest are currently unsolvable / paradigm shifting solutions?

- Pattern recognition for tracking in high-pileup environments (combinatorics)
- Lattice QCD advancements
- “Generic data-driven model-independent anomaly detection”
- Detector alignment / calibration
- Global re-optimisation of software e.g. (re)structuring a collection of trigger lines
- Optimal detector (hardware) design: maximize performance, (uniform) efficiency, minimize compute resources, systematic uncertainties, cost, ...
- Alternative architectures for fast inference of large networks in various places in the data pipeline (GPU vs. FPGA, neuromorphic) to scale well with extreme lumi / combinatorics
- (various quantum algorithms with speedup, quantum sensing?)

6.1) Other?

- Mass production development / deployment of novel small-ish networks should be also taken into account
  - Ecosystem must be ‘machine learning’ adaptable
  - Versioning, characterization, training, pipelines; all with proper provenance
- Structured training of people beyond superficial use of packages
  - Including understanding / quantifying the performance of ML algorithms, beyond looking at a trivial ‘figure of merit’ to be optimized (eg. consider systematic effects)