Lessons from HSF-India Scientific Computing Workshop

Lino Gerlach

HSF-India Scientific Computing Workshop

- One-Week-long workshop at IISER in Pune
- Designed for PhD- and Master students
- https://indico.cern.ch/event/1649033/timetable/
- 2 days of plenary lectures + 3 days of parallel “hackathons” (GPU or ML)
  - Hackathon projects shall be “real-world problems”
  - I chose: effects of radiation damage in SmartPixels

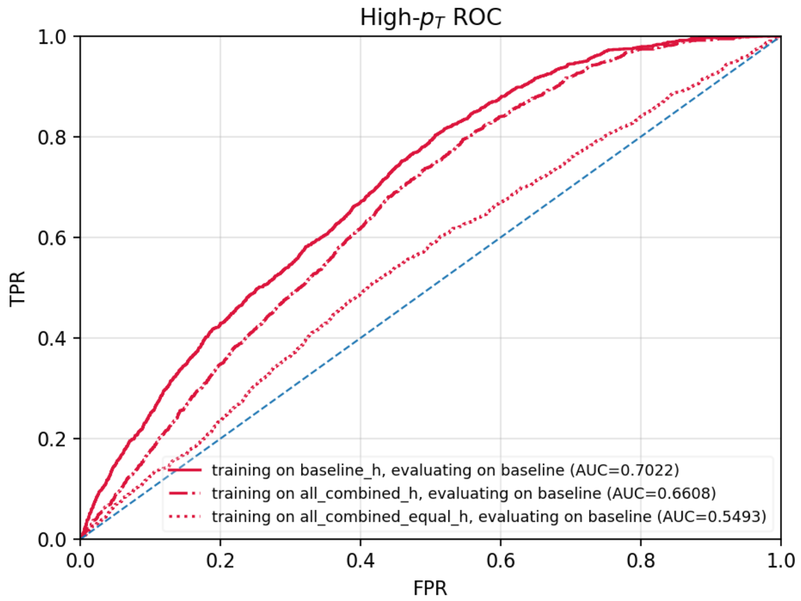
SmartPixels Hackathon Setup

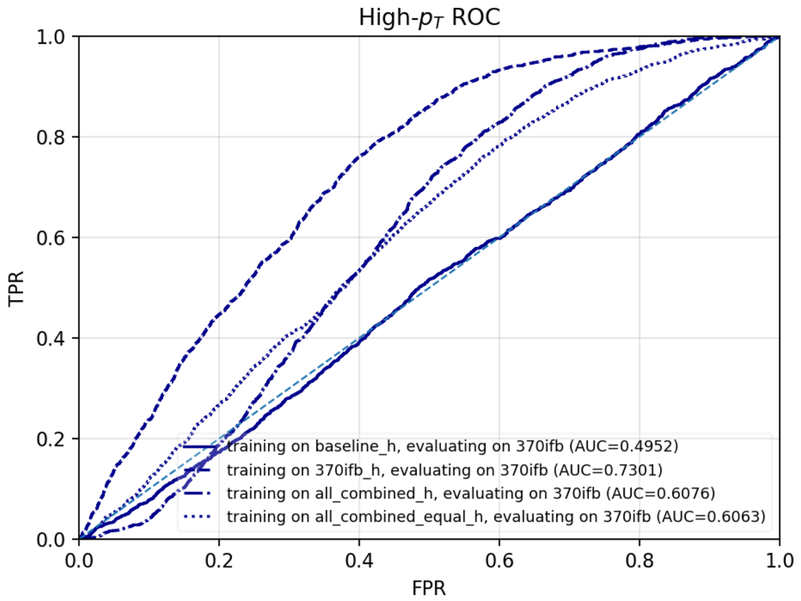
- Created smaller version of SmartPixels “co-design” dataset (zenodo 17180303)
  - Pick one sensor geometry, and 1/20th of events: ~40 GB -> ~2 GB
- Created simplified version of our SmartPixels GitHub repo (Mila & me)
  - https://github.com/ligerlac/SmartPixelsHackathon
  - Removed all LGN-related stuff, advanced training options, etc…

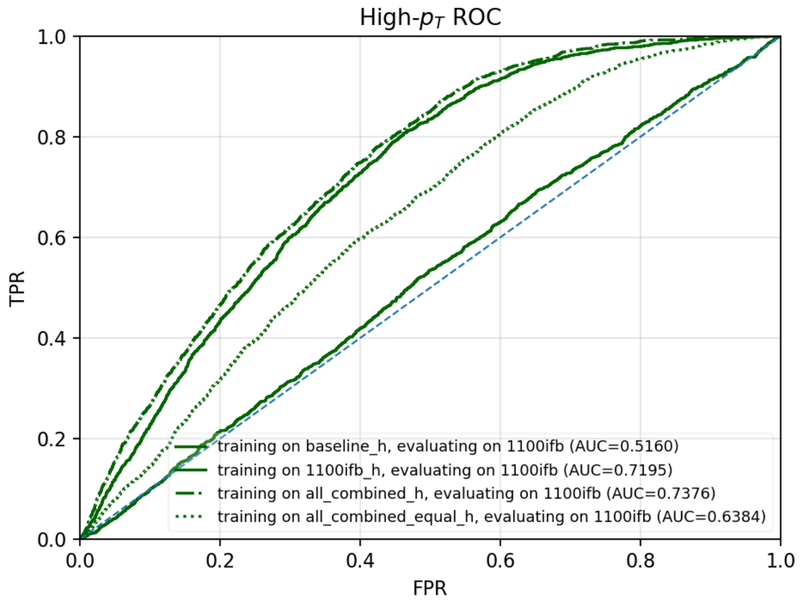
SmartPixels Hackathon Task

- We now have one dataset in 3 levels of irradiation: 0, 370, 1100 ifb
- Question: Can we design a model that performs well on all three?
  - Do not need to retrain / redeploy smart pixels chips throughout detector's lifetime
- Compare 4 approaches:
  - Train on baseline, apply to each (Should be worst)
  - Train on what is applied (this should be best, but unpractical)
  - Train on a mixed dataset
  - Train on a mixed dataset w/ irradiation level as additional feature (flag)

Results


High-p_T ROC
TPR
FPR
training on baseline_h, evaluating on baseline (AUC=0.7022)
training on all_combined_h, evaluating on baseline (AUC=0.6608)
training on all_combined_equal_h, evaluating on baseline (AUC=0.5493)


High-p_T ROC
TPR
FPR
training on baseline_h, evaluating on 370ifb (AUC=0.4952)
training on 370ifb_h, evaluating on 370ifb (AUC=0.7301)
training on all_combined_h, evaluating on 370ifb (AUC=0.6076)
training on all_combined_equal_h, evaluating on 370ifb (AUC=0.6063)


High-p_T ROC
TPR
FPR
training on baseline_h, evaluating on 1100ifb (AUC=0.5160)
training on 1100ifb_h, evaluating on 1100ifb (AUC=0.7195)
training on all_combined_h, evaluating on 1100ifb (AUC=0.7376)
training on all_combined_equal_h, evaluating on 1100ifb (AUC=0.6384)

Lessons Learned

- Creating a reduced dataset is not straightforward:
  - The different irradiation levels 0, 370, 1100 ifb do not consist of the exact same events (99% percent overlap, though)
  - pT profile differs across files (need to sample across a files carefully)
- Students independently found similar unexpected behavior as Mila (e.g. baseline model performing better on 1100 than on 370 ifb)
  - Might be a quirk of evaluating a 3-way classifier as a binary classifier w/ a different threshold
- Code base shall be simplified even further for such 3-day projects