Artificial Intelligence at Fermilab

Kevin Pedro
Associate Scientist
Scientific Computing Division / Particle Physics Division
Fermilab
August 3, 2021

Questions

What is AI?

When?
Now!

Why do we need AI?

How do we use AI?

Who?
A very long list…

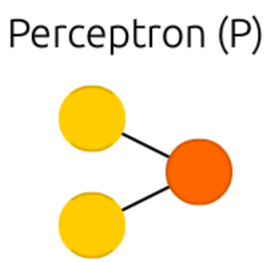
Where do we employ AI?

What is Artificial Intelligence?

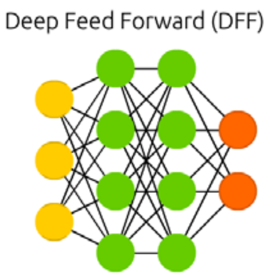
> "AI is whatever hasn't been done yet."
> – Douglas Hofstadter

- Today: *machine learning* (ML), which is *function approximation*:
  - ➢ map inputs to outputs, $\vec{x} \mapsto \vec{y}$
  - o $\vec{y} = F(\vec{x})$ unknown, probably not analytic
    $\rightarrow$ try to find approximation $\vec{y} \approx F'(\vec{x}; \vec{w})$ by optimizing *weights* $\vec{w}$
- *Deep learning* uses networks w/ many *layers* to derive *features* from inputs
  - o More "neurons" $\rightarrow$ more multiplications, weights (thousands–millions)

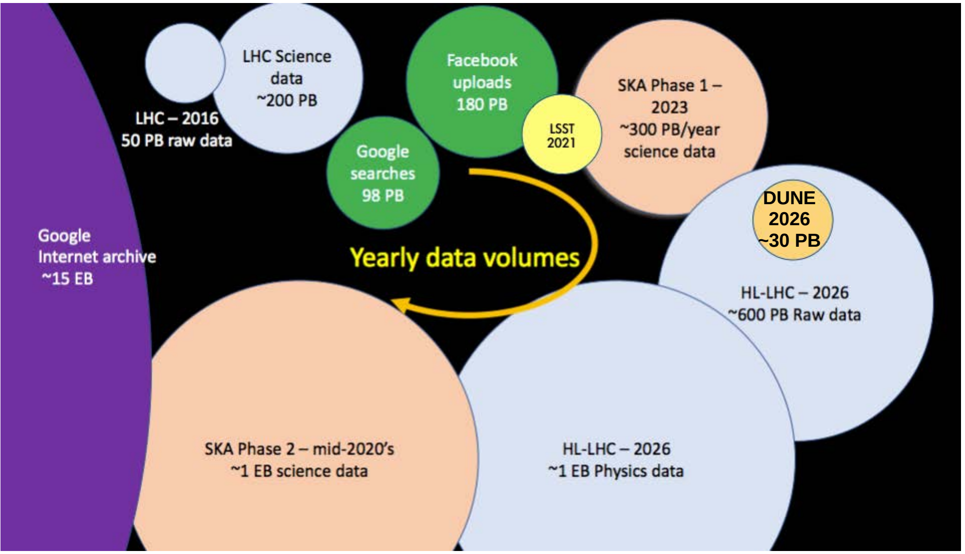
1. Training: *optimizing* weights to improve function approximation
2. Inference: applying optimized function to new data to make *predictions*



The Neural Network Zoo



Challenges…

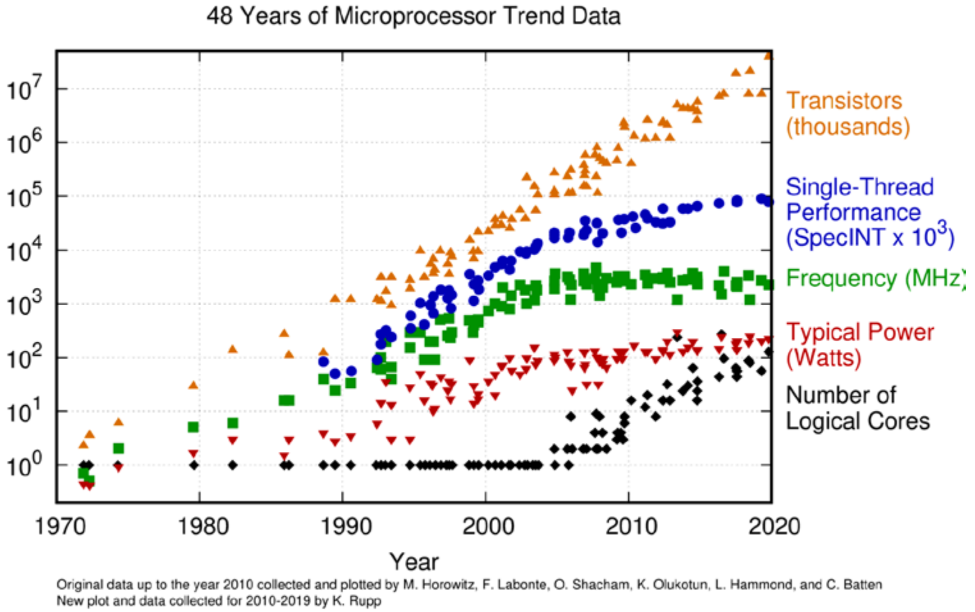


- HL-LHC, DUNE, LSST, SKA will produce up to *exabytes* of data *per year*

➢ *More than order of magnitude* above current dataset sizes

- Moore's Law continues
    - o But without Dennard scaling

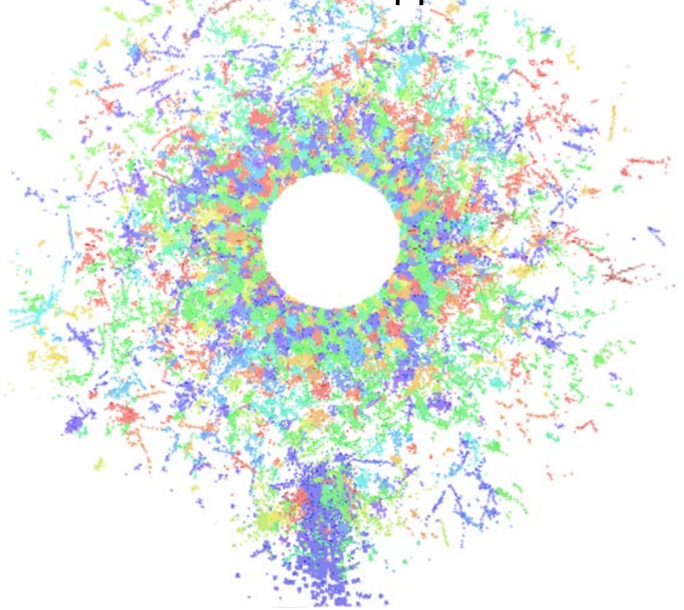
➢ Single-thread performance *can't keep up* with next-gen experiments

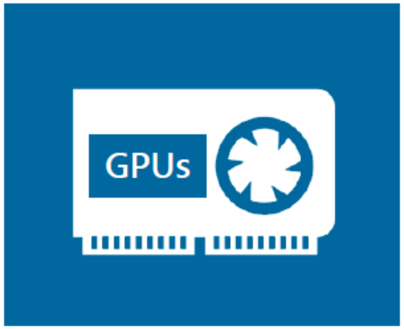


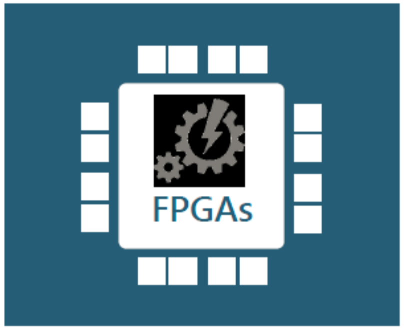
…provide Opportunities

- Not just more data: more *complex* data
- New discoveries rely on *precision measurements*
- Need to:
  - ➢ Extract more information to separate very small signals from very large backgrounds
  - ➢ Operate instruments at cutting edge performance for next-gen experiments to succeed

CMS HGCal simulation,
200 simultaneous pp collisions



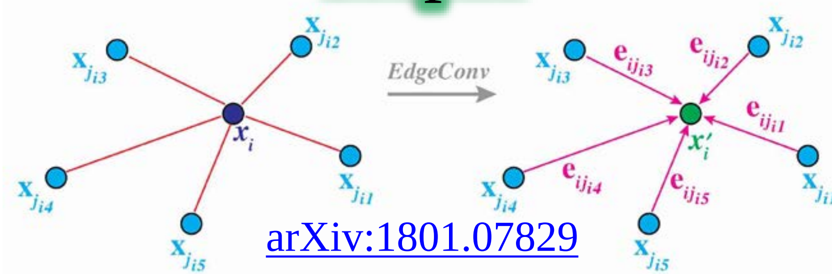




- Augment CPUs with new processors: GPUs, FPGAs, & more
- Deep learning is a natural fit for these devices
  - o Collaborate with industry and open-source communities

Themes in AI

Graphs



arXiv:1801.07829

Exploit *relationships* within data
(generalization of image recognition)

Heterogeneous Computing

Speed up AI algorithms and take advantage of new resources

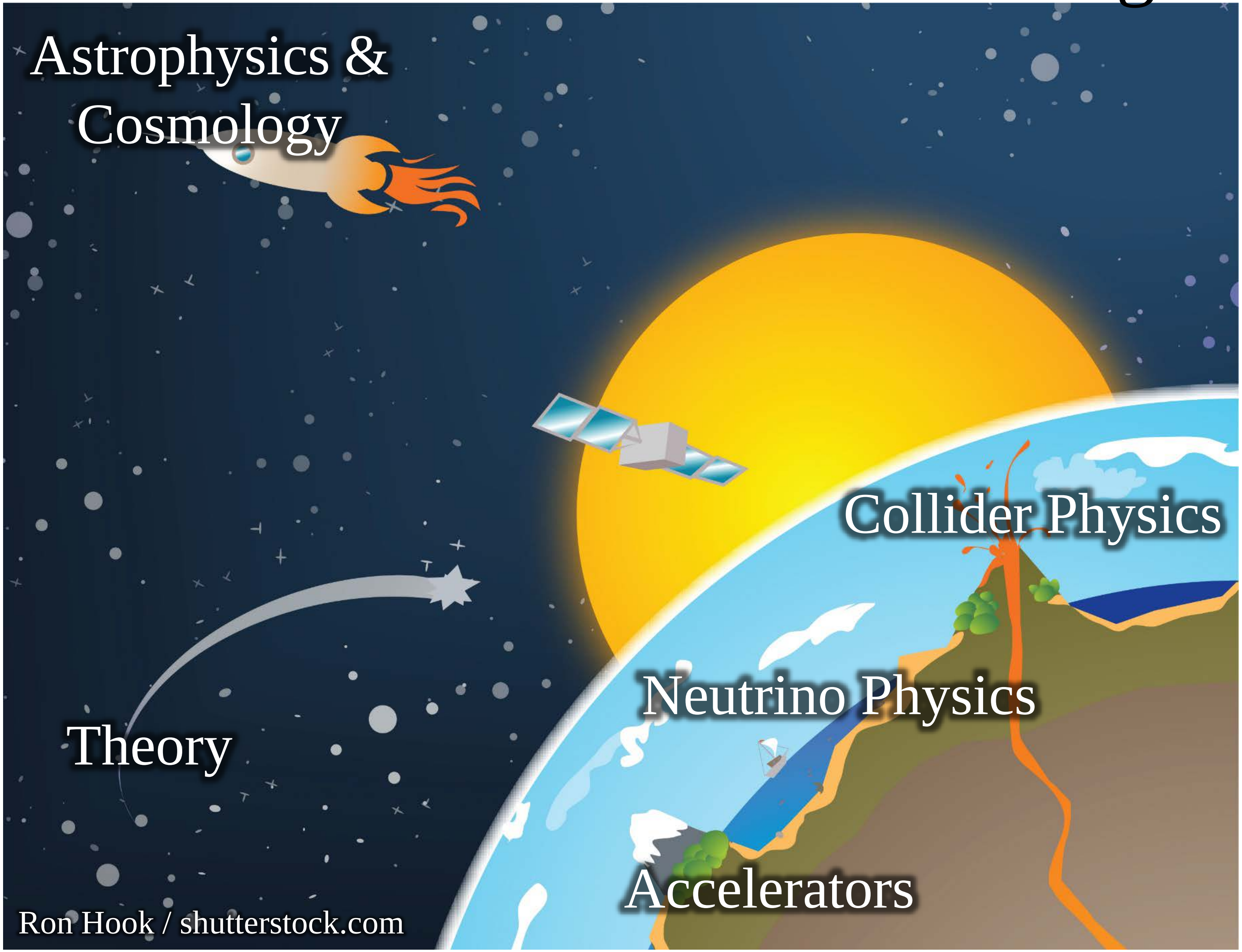
Real-Time

Deploy AI in operations, controls, sensors…

And more!

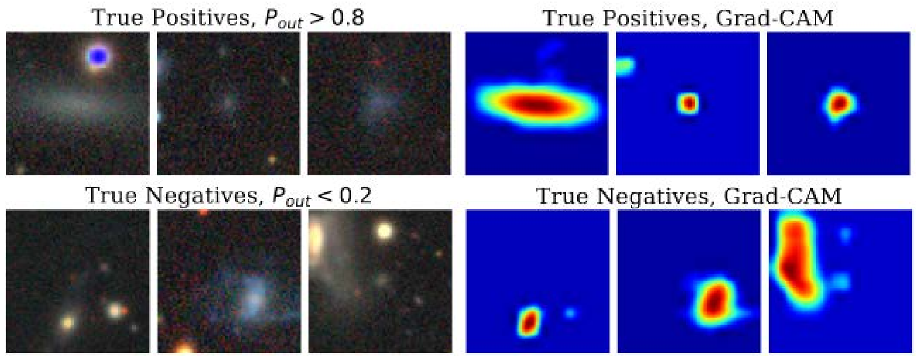
- Anomaly detection
- Invertible networks
- Robustness/uncertainty quantification & reduction

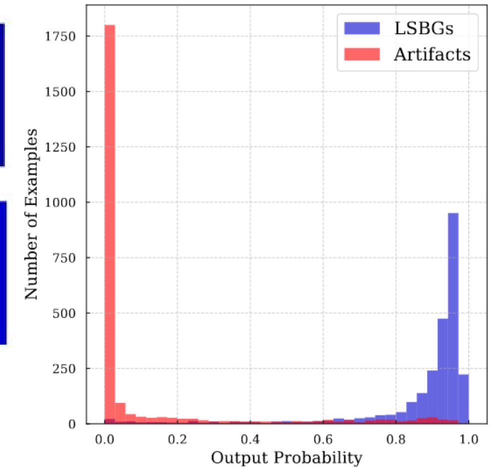
From Distant Galaxies to Miles Underground

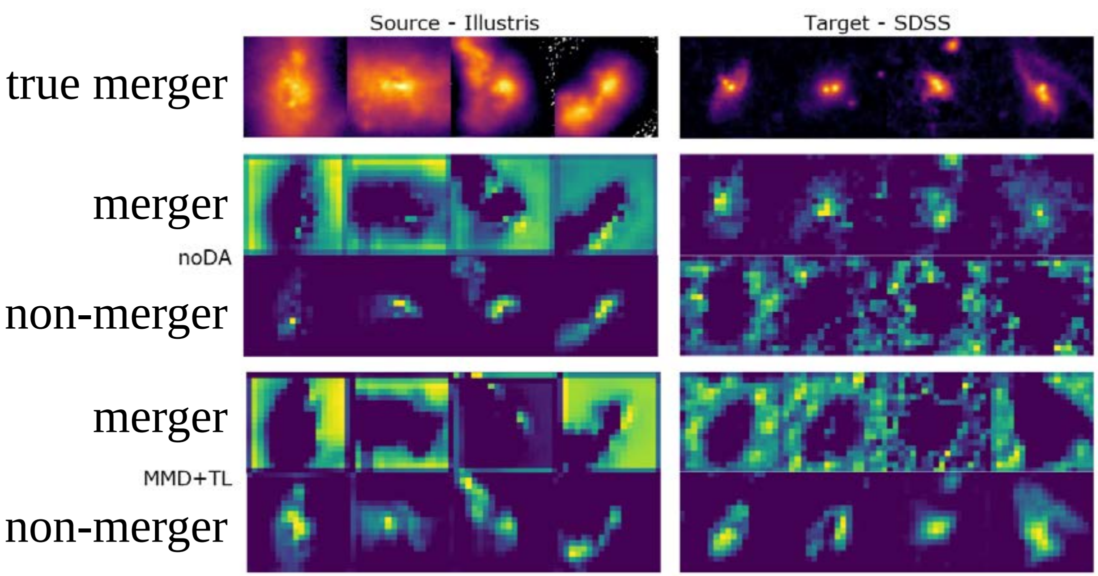


Astrophysics & Cosmology

DeepShadows: (arXiv:2011.12437)
- Convolutional NN to distinguish Low Surface Brightness Galaxies from artifacts in DES data
- 92% accuracy, vs. ~80% accuracy for simpler ML methods

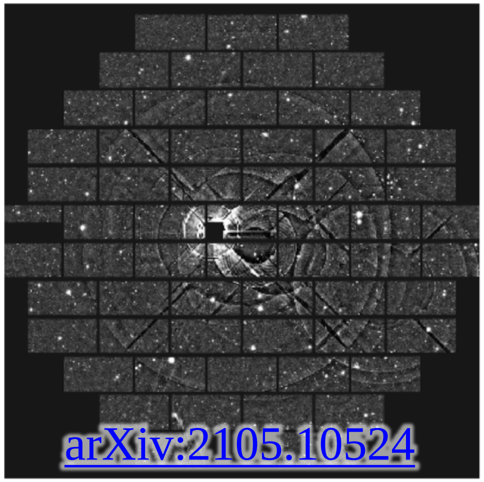






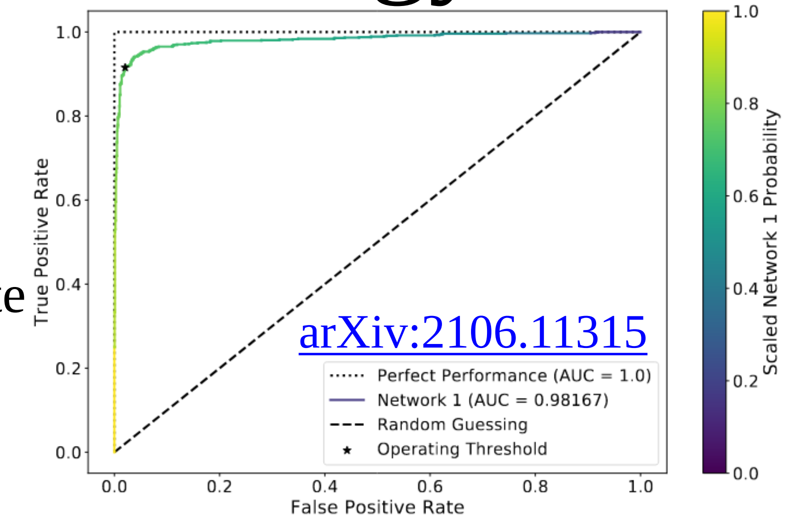
DeepMerge II: (arXiv:2103.01373)
- Goal: identify galaxy mergers
- Domain adaptation (bottom) to train CNN on simulation (left) and apply to data (right) with similar performance (vs. w/o domain adaptation, middle)

Astrophysics & Cosmology

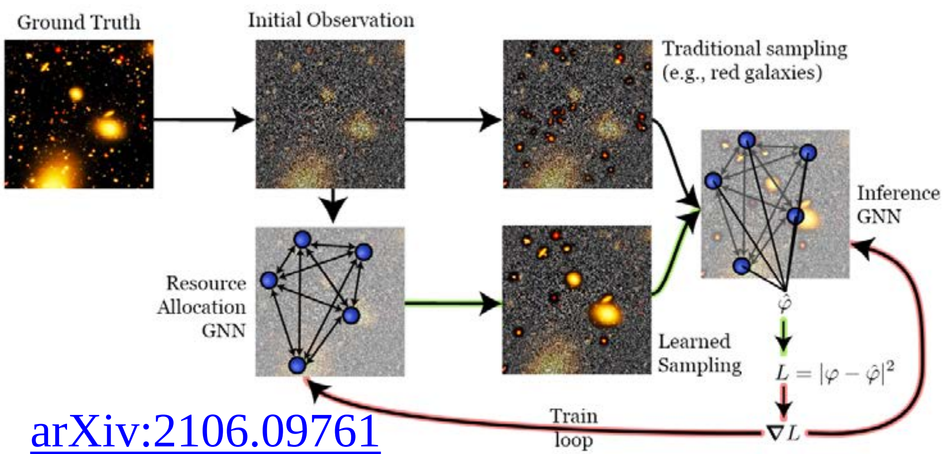


arXiv:2105.10524

- Left: CNN detects ~10× more DES images w/ ghosting/scattering vs. ray-tracing approach
- Right: CNN 95% accurate in removal of false detections in multi-messenger events



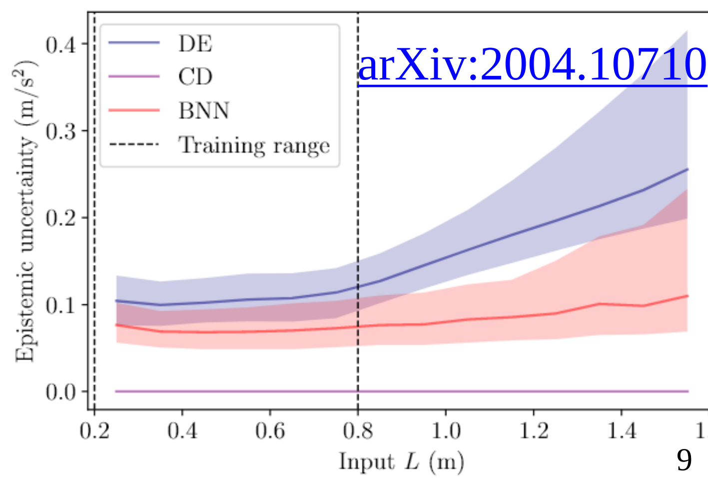
arXiv:2106.11315



arXiv:2106.09761

- Graph NN for unsupervised optimization of telescope time: pick best galaxies to observe
- Outperforms conventional strategies

- ML uncertainty quantification methods don’t reproduce analytic results; Deep Ensembles better than Concrete Dropout, Bayesian NNs



arXiv:2004.10710

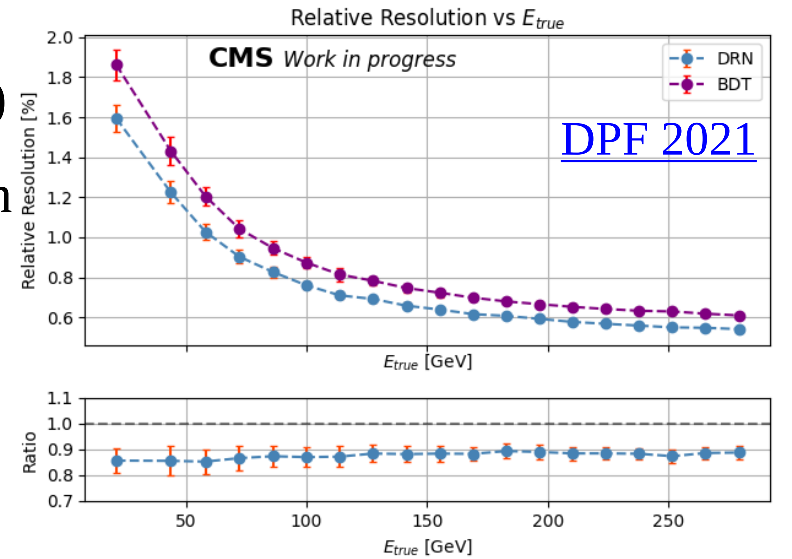
Collider Physics

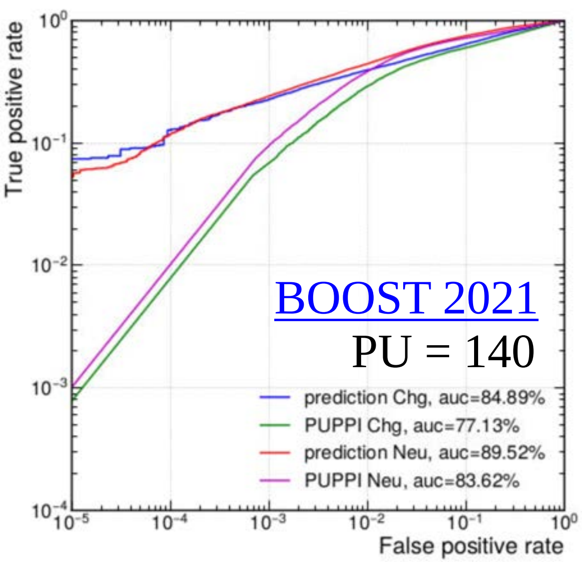
- Dynamic Reduction Network (arXiv:2003.08013)

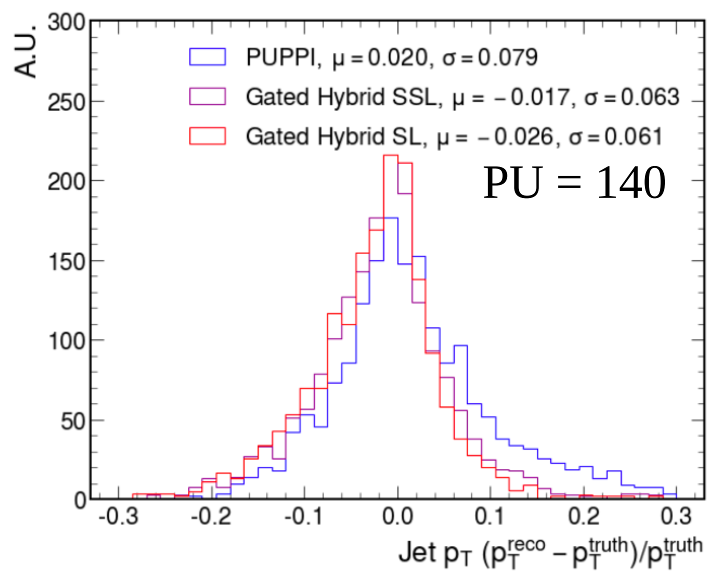
➢ Learn best graph of inputs & use it for regression

- Improve electron resolution by ***10%*** (vs. state of the art)

- Work in progress: apply to missing energy

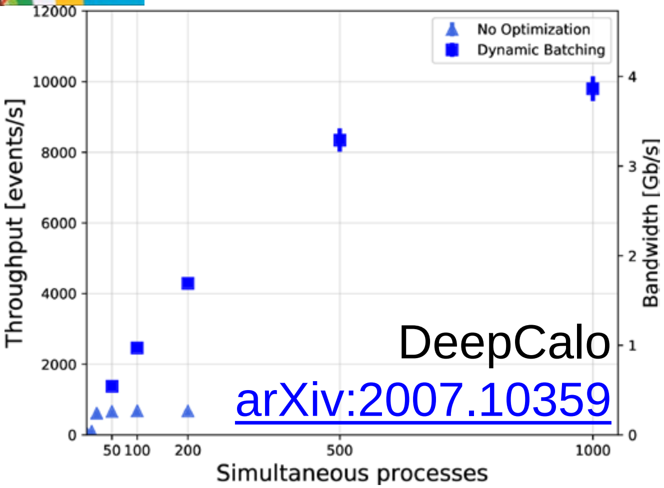






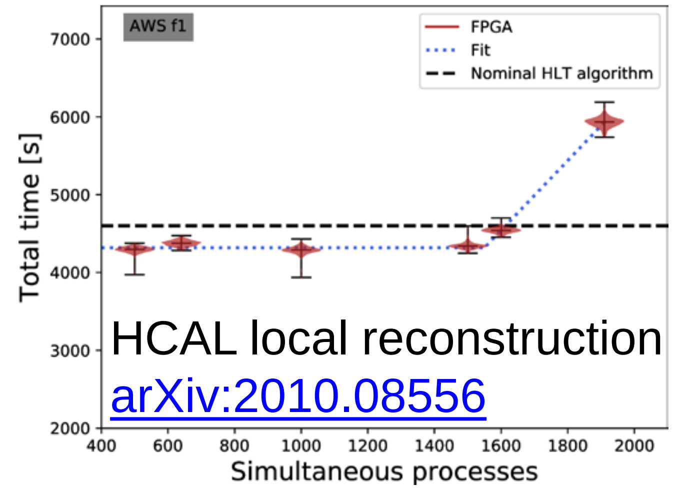
- Semi-supervised Graph NN to reject pileup: trained on charged particles → can use data!
- Significantly improves on classical algorithm

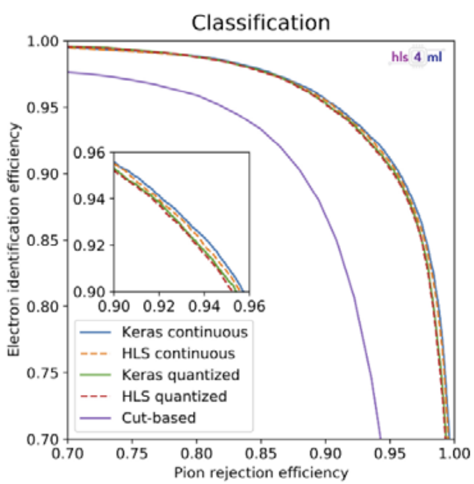
Collider Physics



Inference as a Service:

- GPU processes multiple events together: increase rate by 10× or more
- 1 FPGA can serve 1500 CPUs for an algorithm





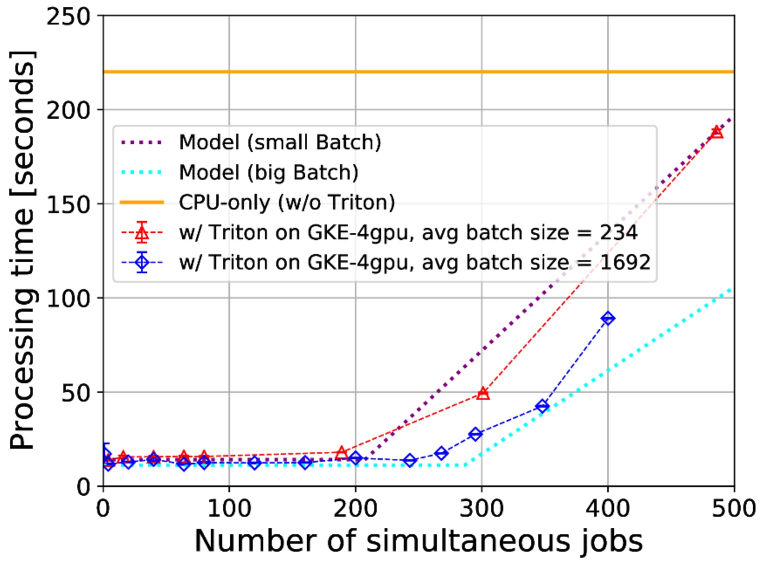
- Convert NNs to run on FPGAs (arXiv:2103.05579) for low-latency and low-power scenarios
- Simple NNs, CNNs (arXiv:2101.05108), GNNs (arXiv:2008.03601), & more!
  - Preserves GNN performance w/ ~1 μs execution time
- Quantization-aware pruning (arXiv:2102.11289) to improve computational efficiency
- Can also be used with ASICs

FAST MACHINE LEARNING LAB

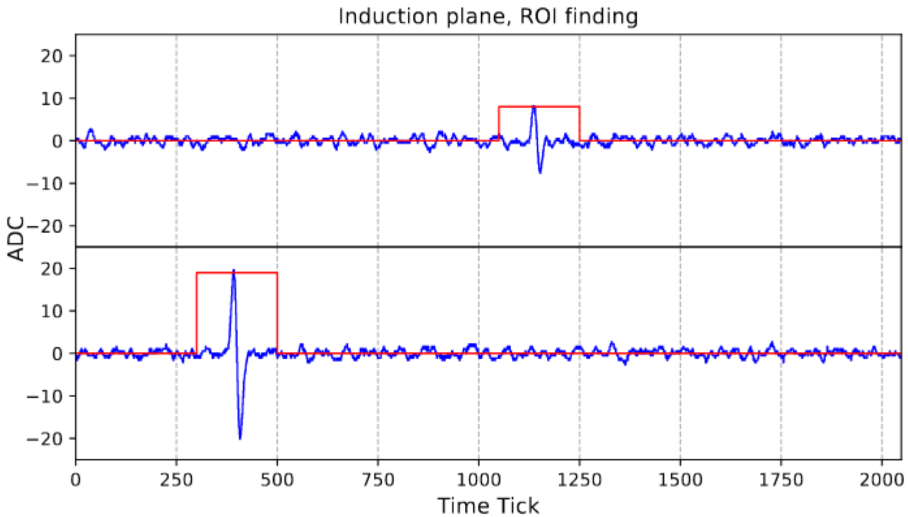
ASICs

Neutrino Physics

- ProtoDUNE data processing dominated by large CNN: 220/330 total seconds/event
- *GPU as a service*: CNN 17× faster, full workflow 2.7× faster (arXiv:2009.04509)

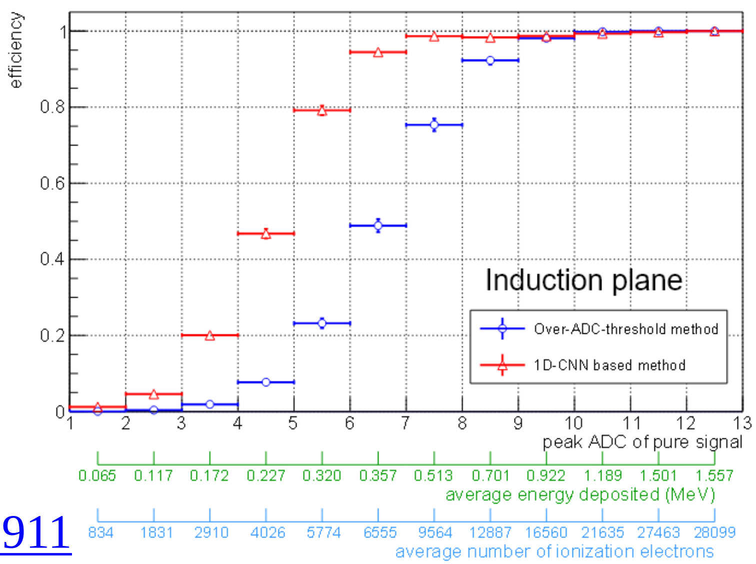


- 1D CNN can localize and extract low-energy signals in noisy LArTPC data
- Significantly more efficient than traditional approach



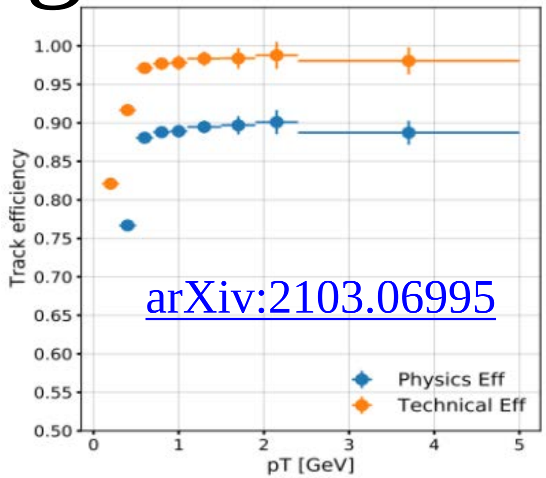
(a)

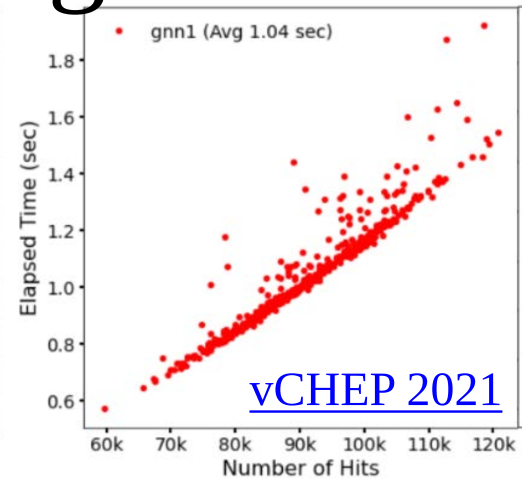
arXiv:2106.09911



Clustering & Tracking

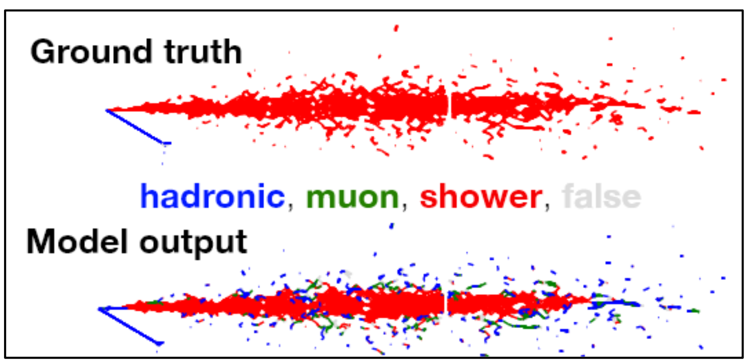
- Common set of tasks for collider & neutrino physics: combine low-level detector hits into *tracks* and *clusters*
- [Exa.TrkX](), [LDRD]():
  - ➢ Employ graph NNs to improve *accuracy* & *speed*
- Custom low-level operations contributed back to ML frameworks (TensorFlow, PyTorch)



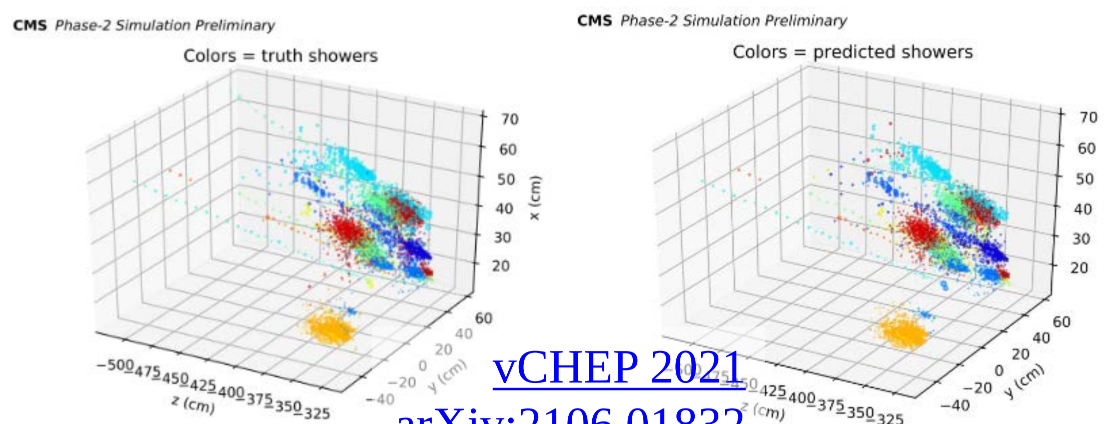


High efficiency & sub-quadratic inference time for LHC tracking

[vCHEP 2021]()



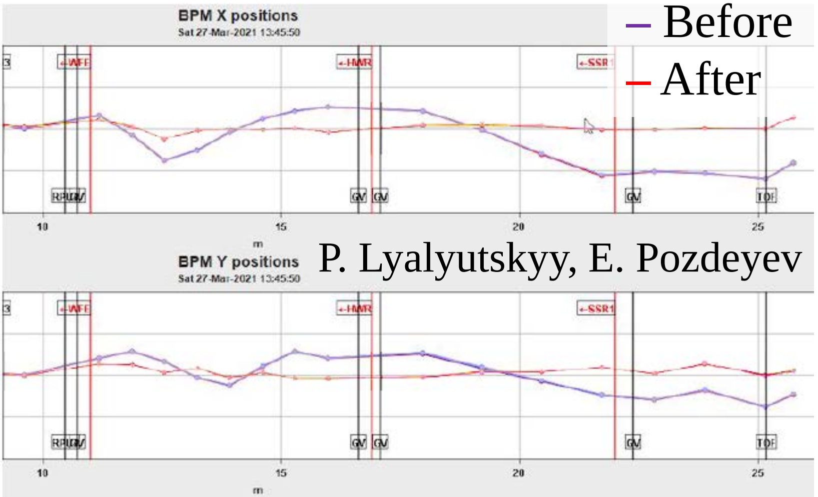
84% edge efficiency for LArTPC



[vCHEP 2021]()
[arXiv:2106.01832]()

Reconstruct multiple clusters in CMS high granularity calorimeter

Accelerators



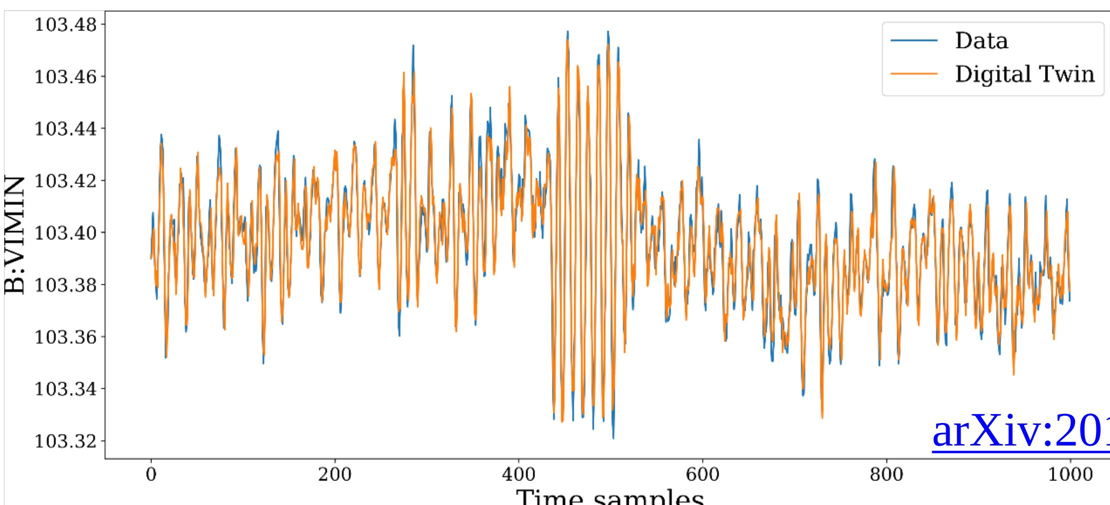
P. Lyalyutskyy, E. Pozdeyev

- Bayesian optimization for beam alignment at PIP2IT
- Converges faster than Simplex

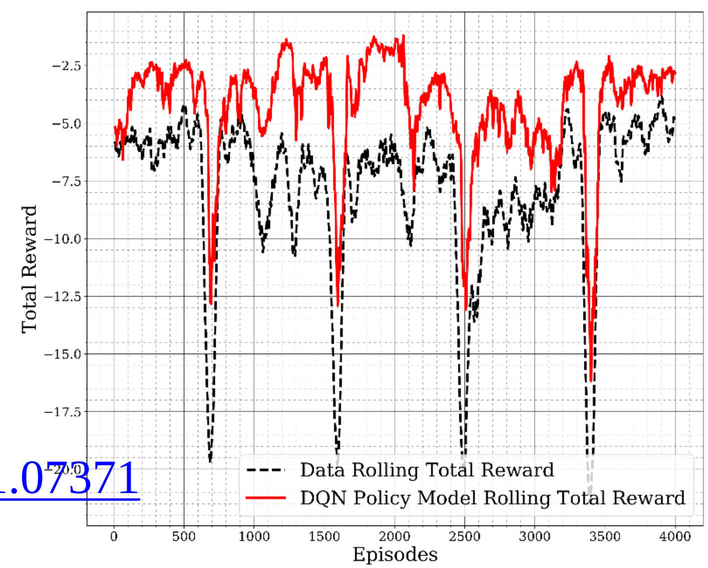
AI for Gradient Magnet Power Supply @ FNAL Booster:

- LSTM surrogate model reproduces system dynamics from data (bottom left)
- MLP agent performs ~2× better than existing regulation circuit (bottom right)

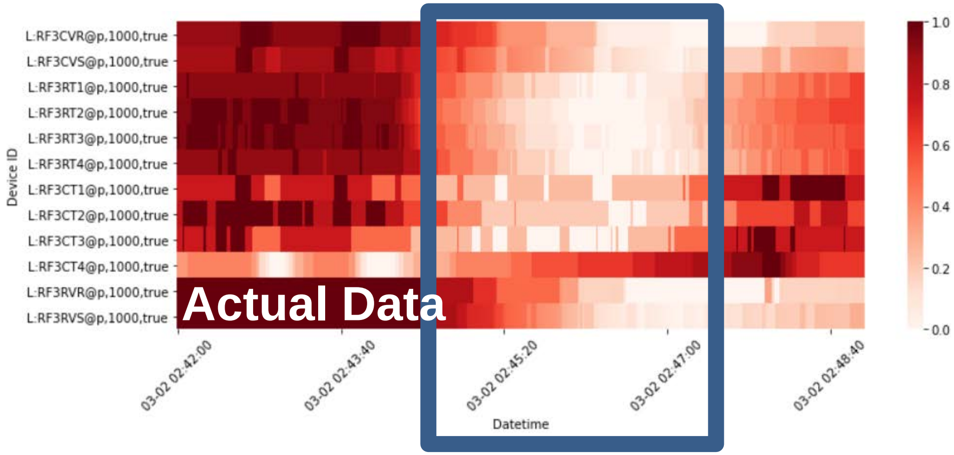
➢ Agent optimized for FPGA w/ hls4ml, inference at 15 Hz



arXiv:2011.07371



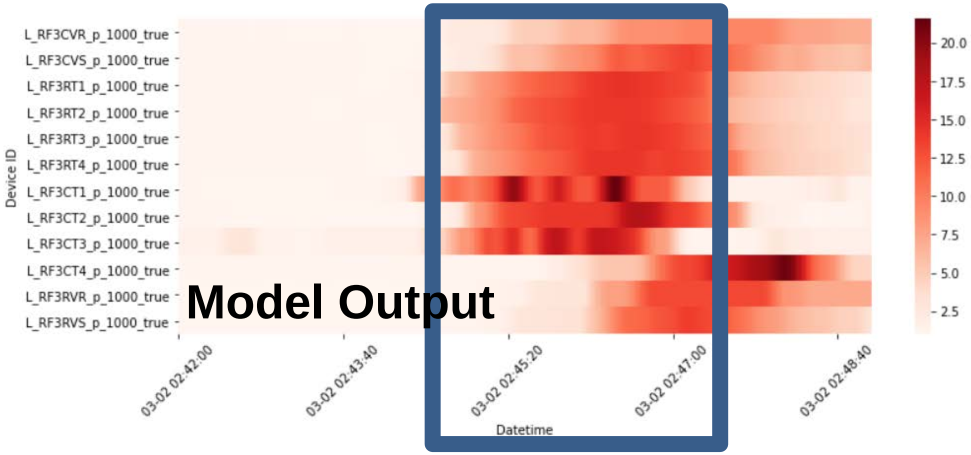
Accelerators

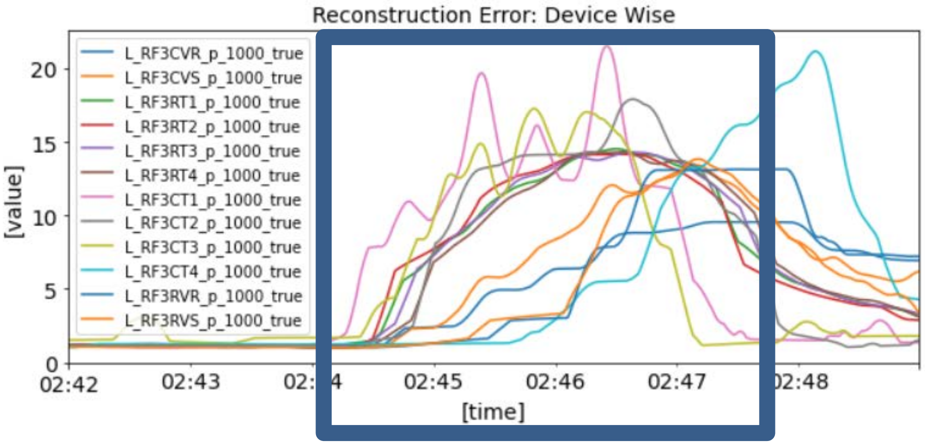


L-CAPE: Linac Conditional Anomaly Prediction of Emergence

- ~3000 unique device data streams
  - Frequencies: 66 ms, ~2–3 min

Accelerators

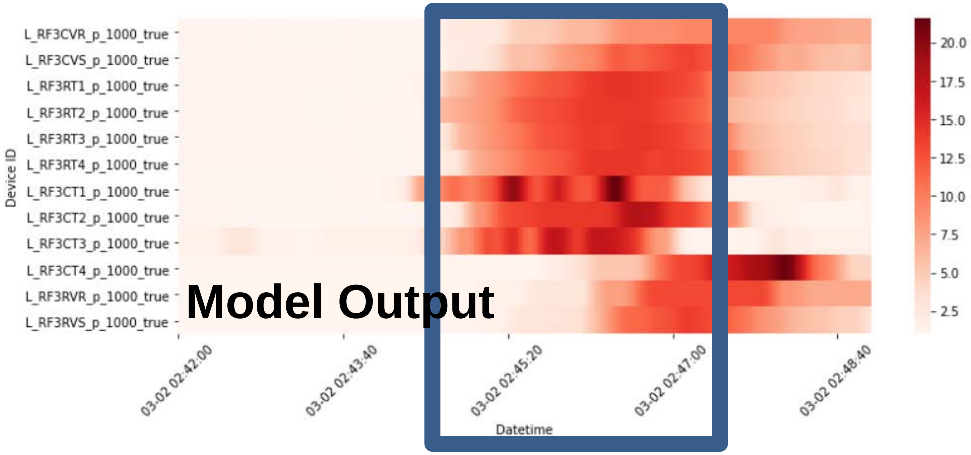




L-CAPE: Linac Conditional Anomaly Prediction of Emergence

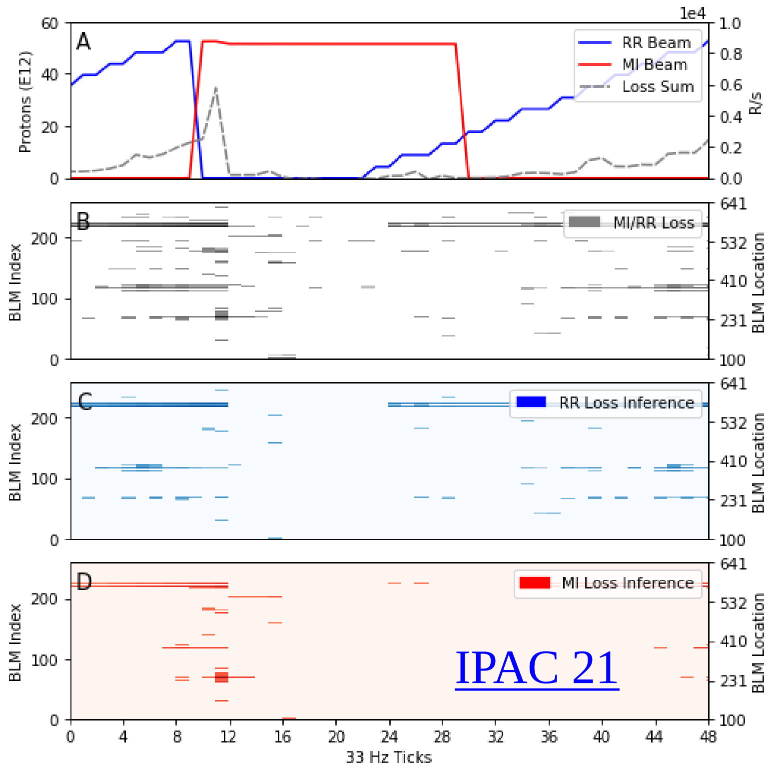
- ~3000 unique device data streams
  - o Frequencies: 66 ms, ~2–3 min
- LSTM autoencoder identifies outage precursors as anomalies

Accelerators



L-CAPE: Linac Conditional Anomaly Prediction of Emergence

- ~3000 unique device data streams
  - o Frequencies: 66 ms, ~2–3 min
- LSTM autoencoder identifies outage precursors as anomalies

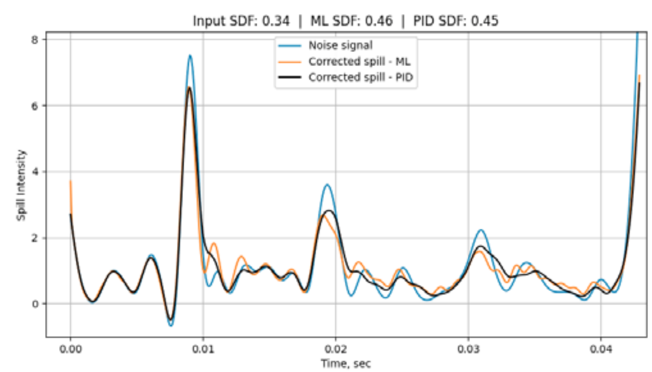


[IPAC 21]()

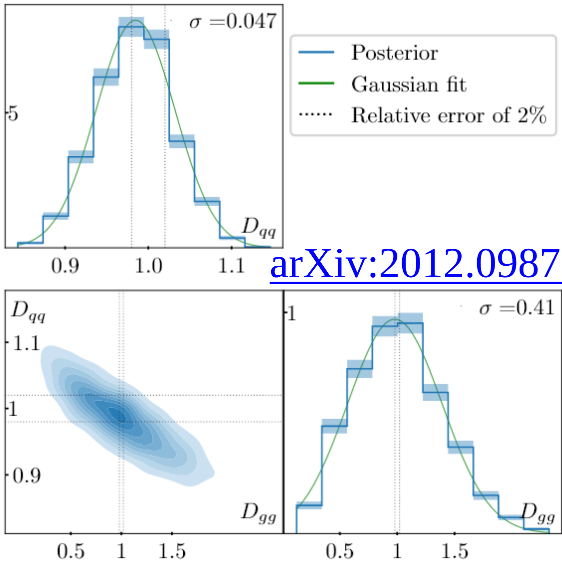
READS: Real-Time Edge AI for Distributed Systems ([arXiv:2103.03928]())

- MI/RR beam loss deblending (left)
- Mu2e slow spill regulation (right)

➢ Aim to deploy on FPGAs



Theory

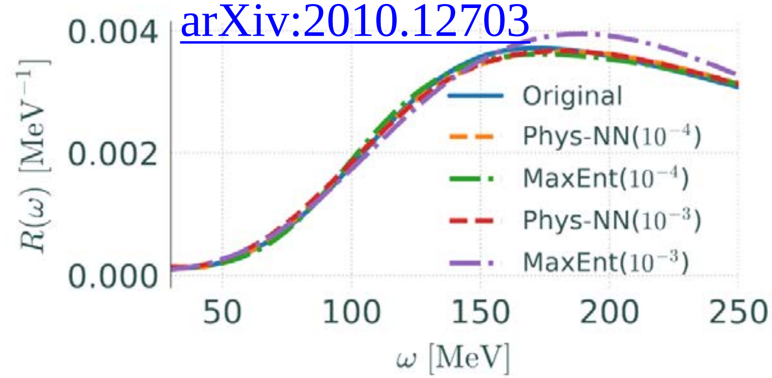


arXiv:2012.09873

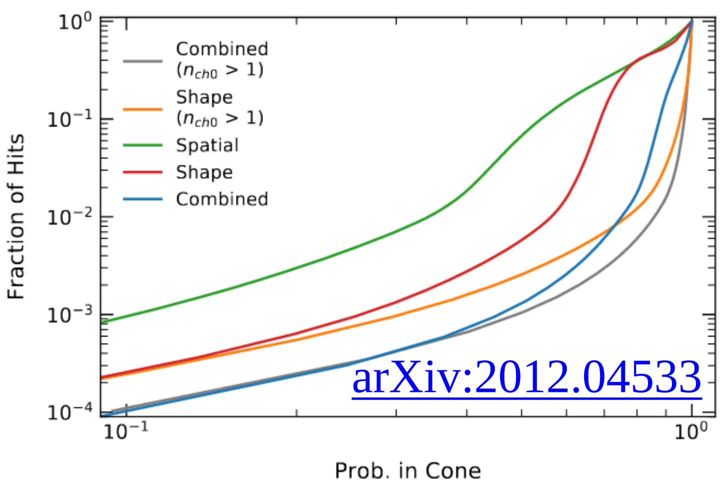
- Invertible NN enables LHC measurements of QCD splitting parameters w/ precision comparable to LEP

- Invertible NN outperforms classical method to reconstruct nuclear response functions

arXiv:2010.12703



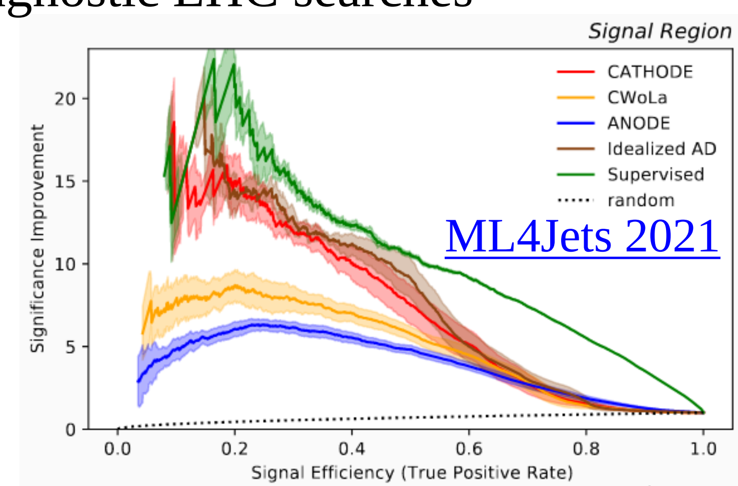
- NN combines cluster shape & spatial info for LHC tracking



arXiv:2012.04533

- Reduces fake combinatorical backgrounds while preserving efficiency

- CATHODE: combine unsupervised anomaly detection techniques for huge sensitivity increase in model-agnostic LHC searches



ML4Jets 2021

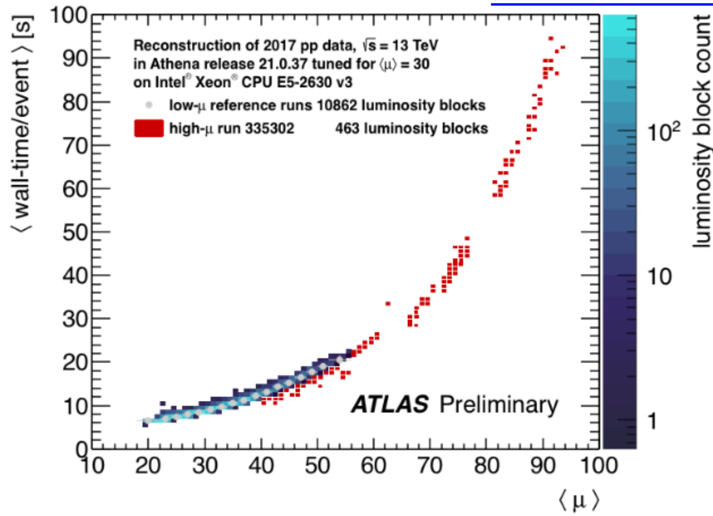
Final Thoughts

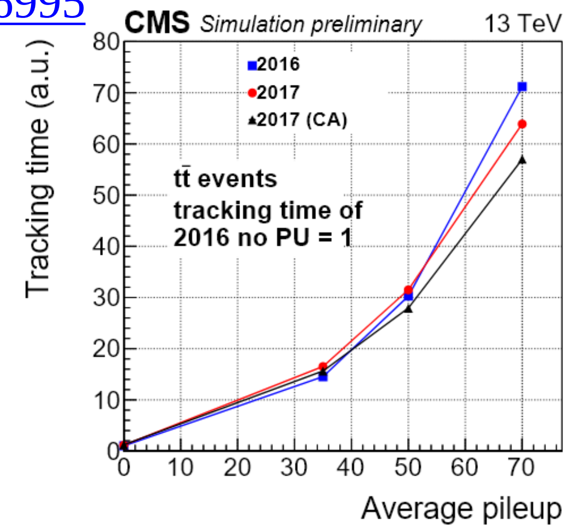
- AI at FNAL is reaching escape velocity
  - *Dozens* of projects ongoing
  - Not just strong, but *leading* results
  - Maturing and moving into production
- In addition, many new efforts starting!
  - Next year's talk will have a *whole different* set of results
- ➢ AI can solve our big data & computing challenges, but it is *not* egalitarian
  - Past decades: classical algorithms, one CPU is ~as good as another
  - Today: better devices (GPUs, FPGAs, etc.) lead to better results
    - Both hardware and support *cost more*
- AI can be used for good or evil: be wary of bias and misuse
  - FNAL scientists, engineers, technicians, users have a responsibility to promote *scientific* and *humanitarian progress*

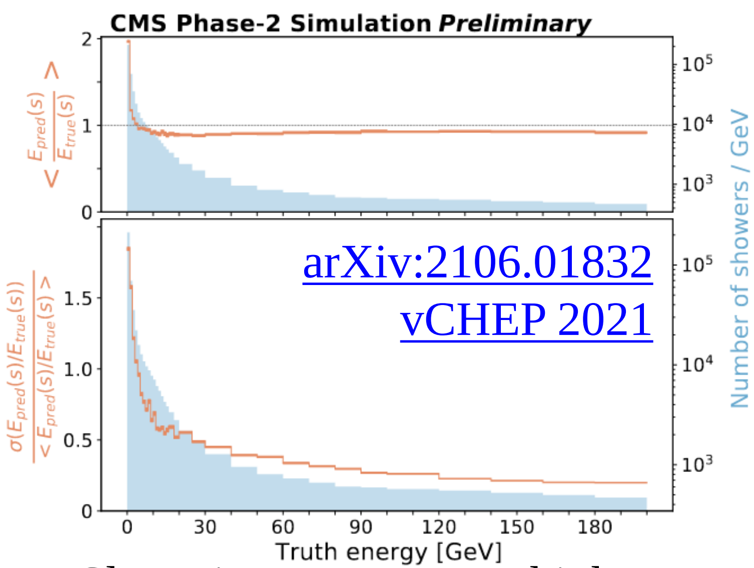
Backup

Clustering & Tracking

[arXiv:2103.06995](https://arxiv.org/abs/2103.06995)







[arXiv:2106.01832](https://arxiv.org/abs/2106.01832)

[vCHEP 2021]()

Clustering reconstructs high granularity calorimeter energy