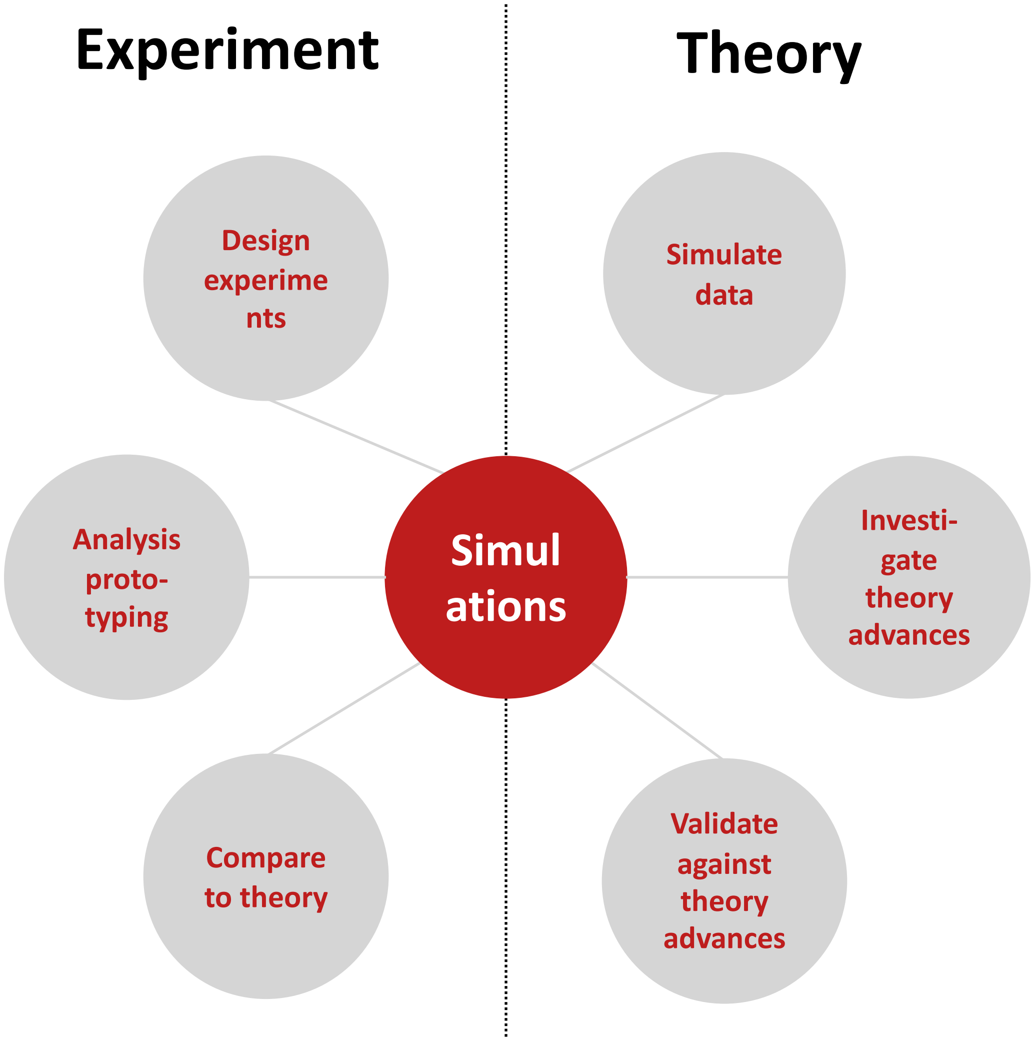
Event Generation and Simulation

Simulations in Experiment and Theory



The role of A.I. in simulations

Lesson learned High-precision QCD measurements require high-precision simulations

Statistical accuracy for precise hypothesis testing
- up to trillion of simulated events required (HL-LHC)
- often computationally intensive, in particular calorimeter simulations

Common alternatives
- fast simulations with computationally efficient approximations, e.g., parameterizations or look-up tables
- **still** insufficient accuracy for high-precision measurements

Promising alternatives
- fast generative models, e.g., GANs or VAEs
- A.I. driven design, e.g., Bayesian optimization

Contributions to discussions

Yaohang Li (ODU) **ETHER**

Cristiano Fanelli (MIT) **AI-optimized Detector Design**

Nicholas Polys (VT) **Advances In Human-Centered AI**

Noëlie Cherrier (Saclay) **Event classification with ML at CLAS12**

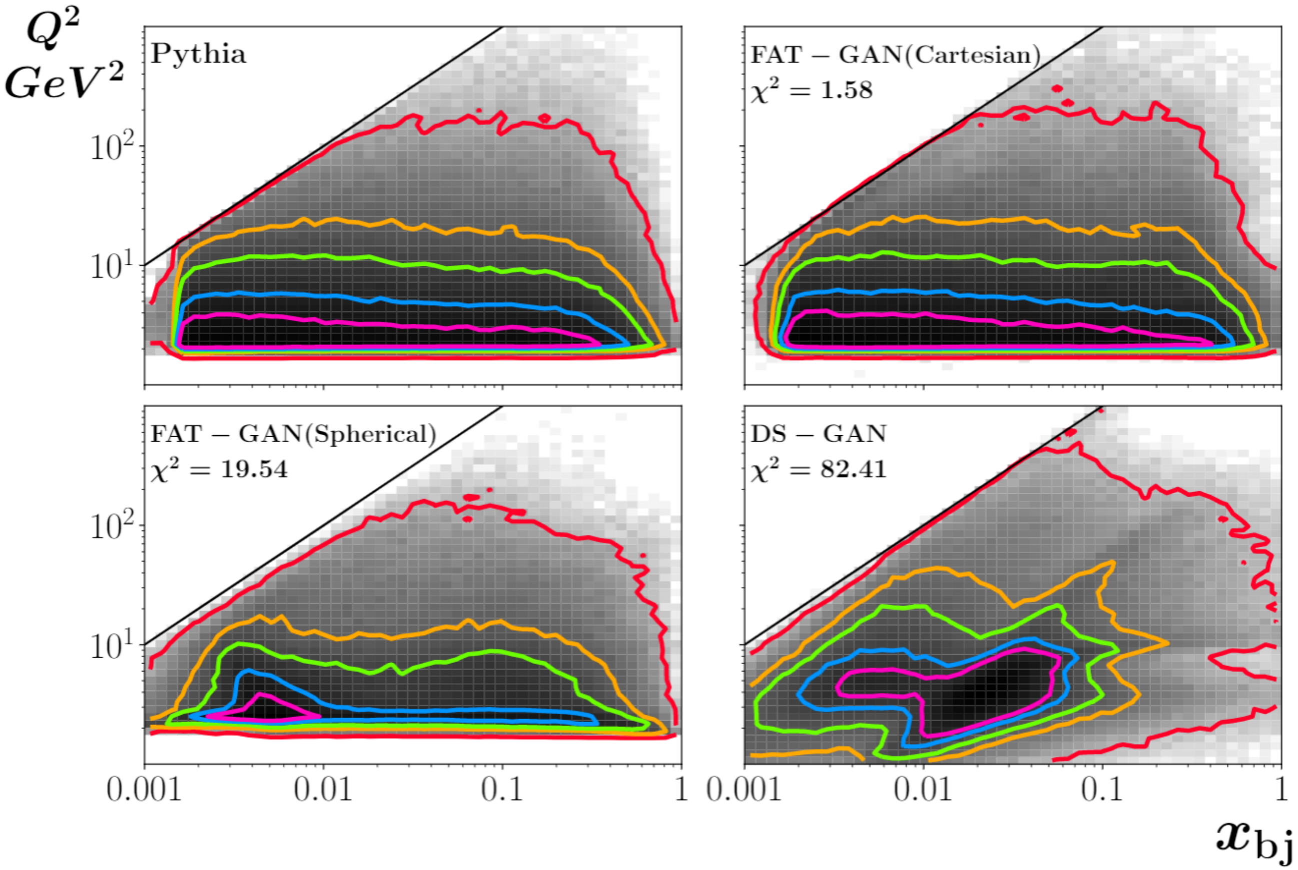
Veronique Ziegler (JLAB) **CLAS12 Tracking with ML**

Abdullah Farhat (ODU) **ML to reconstruct DIS kinematics**

Google document for discussions

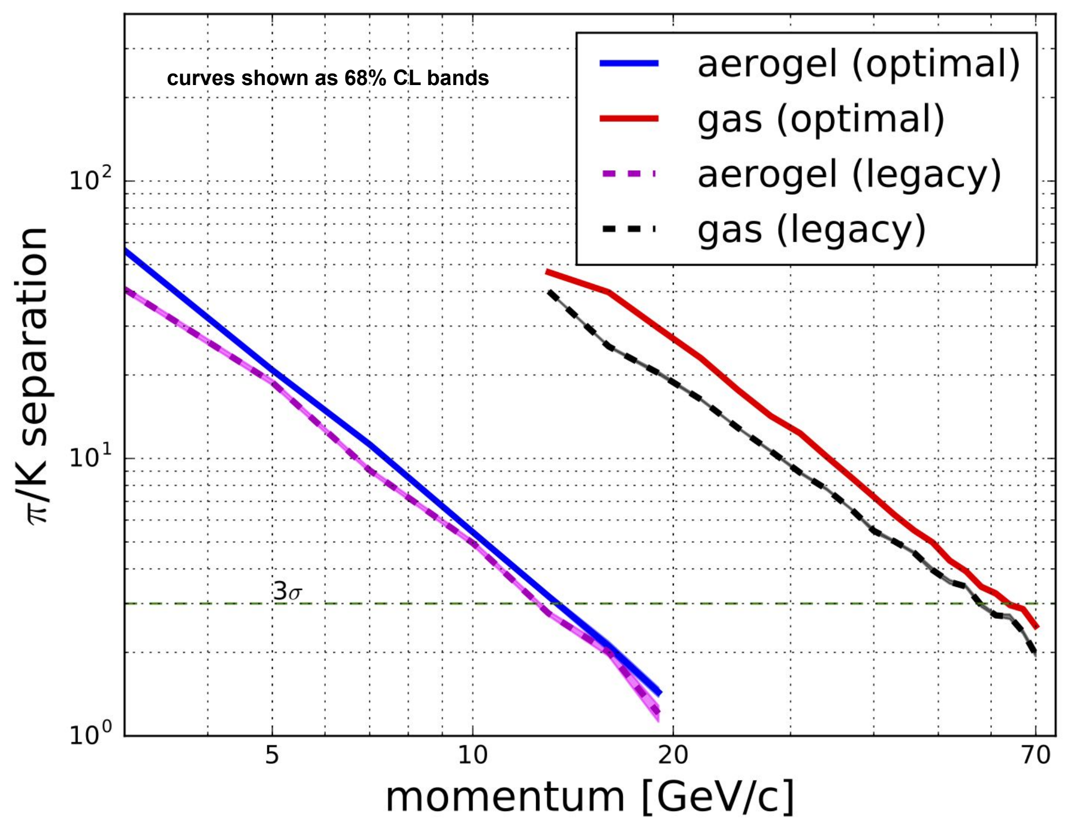
Empirically Trained Hadronic Event Regenerator (ETHER) Yaohang Li (ODU)

- LDRD project at Jefferson Lab: theorists interpolate across many different experiments, in a way that they could never do by stitching all the experiments together
- currently: study GAN as a repository of the behavior of the theory as expressed in Pythia (later real data)
- working well for single beam energy and inclusive single electrons / single electron and pion
- varying beam energy facing difficulty (variational GAN based event generators)



Jefferson Lab

- automated, highly parallelized, self consistent framework for detector design
- specific application for the dual-RICH of the future EIC has been shown
- statistically significant improvement w.r.t. baseline design found
- tested with O(20) parameters, ways to deal with O(100) parameters, possible to add cost

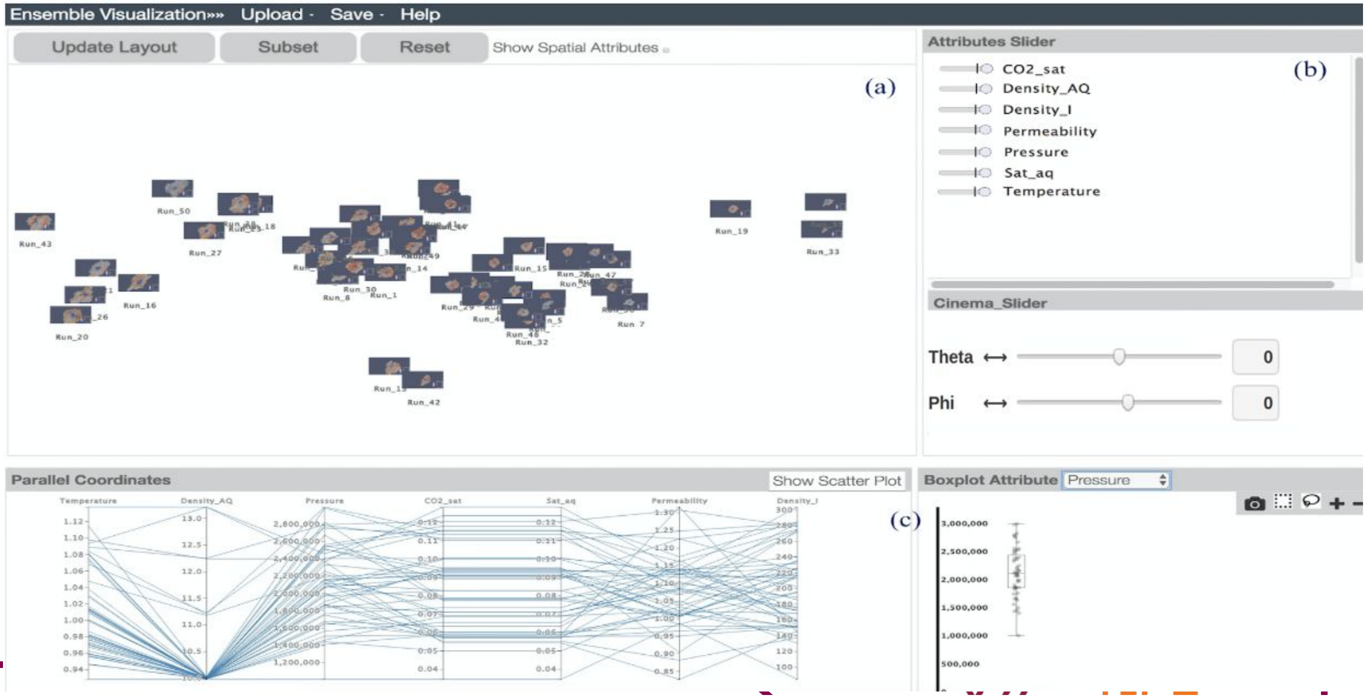


- Statistically significant Improvement in both parts.
- In particular in the gas region where the 5σ threshold shifted from 43 to 50 GeV/c and the 3σ one extended up to
- Notice that before this study we did not know “how well” the legacy design was performing.

CF et al. AI-optimized detector design for the future EIC *arXiv:1911.05797* (2019)

Semantic Interaction

- Exploratory platform to find High-D relationships
- Offload the hard stuff to ML, but let the expert drive
- Manipulate CINEMA thumbnails (eg 3D projections)

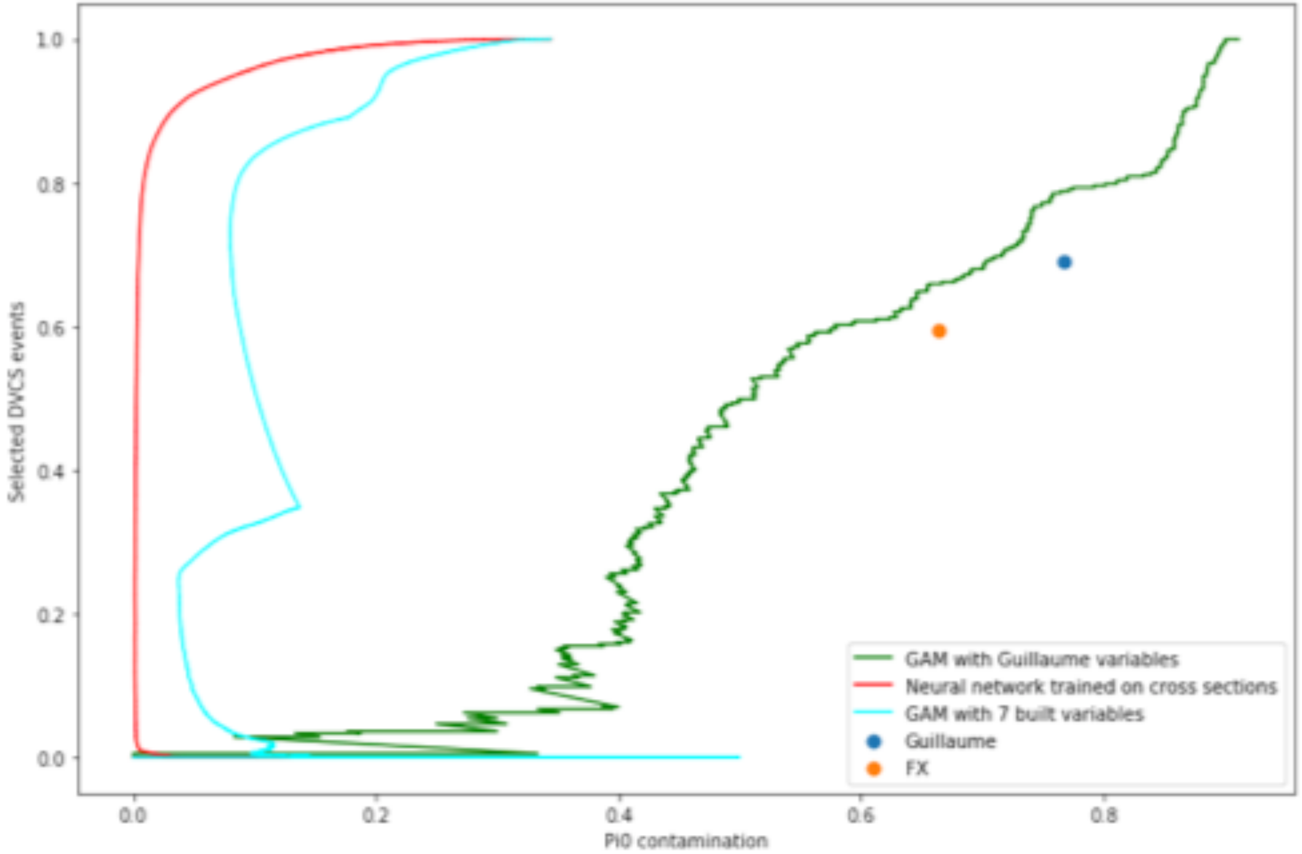


Visualization

- makes debugging models and code easier
- key component of discovery and communication
- better visualization tools can help build better models and analytic capabilities for A.I. / ML

Event classification with ML at CLAS12

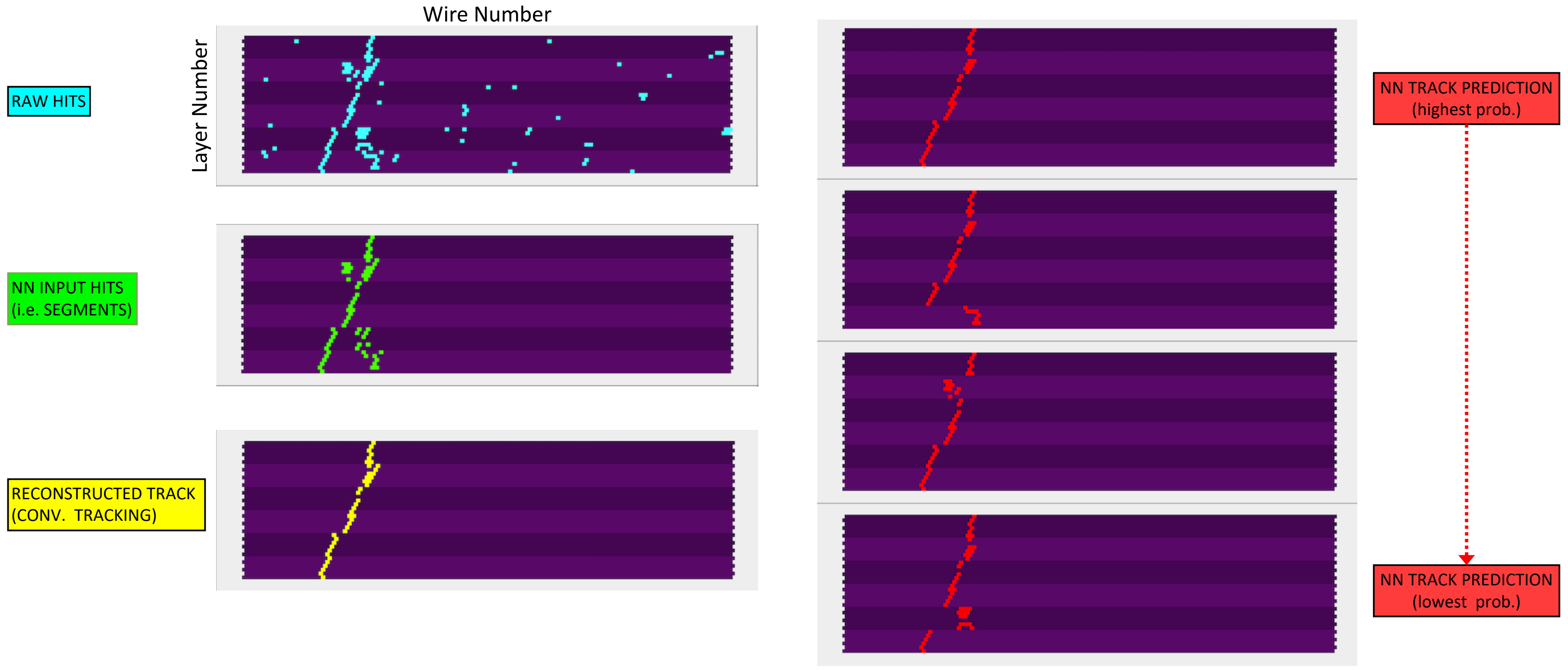
- build a selector for DVCS events
- uses feature construction to get new discriminative variables
- implementation in generalized additive models (GAMs)
- GAM makes better use of the correlations between the variables than other approaches, out-performs conventional approaches on efficiency and purity



Open questions

- fair comparisons of the different methods
- objectively assess interpretability
- how to apply to real data

- combinatorics in resolving ghost tracks, noise rejection takes considerable time:
  - AI-assisted tracking to speed it up
  - evaluate different NN approaches
- track reconstruction is ~5x faster using NN for segment finding
- NN tracking finds tracks missed with conventional tracking, in presence of high background. But also the reverse happens. Studying tracking efficiency is underway (currently ~99.5% accuracy).



identify which track segments are consistent with being on-track and use these NN-found segments for seeding recons

ML to reconstruct DIS kinematics

Abdullah Farhat (ODU, EIC^2)

Method	Required Measurements	Strengths	Limitations
Electron	$E_{e'}, \theta_{e'}$	Precise	Sensitive to QED radiation
Jacques-Blondel	$\delta_{\mathcal{H}}, P_{T,\mathcal{H}}$	Resistant to QED radiation	Needs precise energy measurements
Double-Angle	$\theta_{e'}, \gamma_{\mathcal{H}}$	Does not need precise energy measurements	Poor resolution at low x, low Q^2

Bin	Events	Q^2 (GeV^2)	x	x RMSE	Q^2 RMSE
1	114606	80 – 160	0.0024 – 0.010	NN: 0.0040 EL: 0.0029 JB: 0.0042 **DA: 0.0012**	NN: 22.705 **EL: 14.810** JB: 204.39 DA: 20.753
2	65501	160 – 320	0.0024 – 0.010	NN: 0.0049 EL: 0.0014 JB: 0.0053 **DA: 0.0013**	NN: 35.068 **EL: 29.609** JB: 405.88 DA: 36.397
3	74382	320 – 640	0.01 – 0.05	**NN: 0.0053** EL: 0.0226 JB: 0.0086 DA: 0.0063	**NN: 60.198** EL: 64.426 JB: 311.52 DA: 82.069
4	47055	640 – 1280	0.01 – 0.05	**NN: 0.0046** EL: 0.0061 JB: 0.0103 DA: 0.0047	**NN: 96.406** EL: 105.55 JB: 792.58 DA: 151.91
5	60684	1280 – 2560	0.025 – 0.150	**NN: 0.0102** EL: 0.0262 JB: 0.0194 DA: 0.0154	**NN: 195.70** EL: 216.84 JB: 1012.1 DA: 283.20
6	46242	2560 – 5120	0.05 – 0.25	**NN: 0.0154** EL: 0.0333 JB: 0.0303 DA: 0.0249	**NN: 410.11** EL: 435.00 JB: 1694.9 DA: 509.29
7	47380	5120 – 10240	0.06 – 0.40	**NN: 0.0197** EL: 0.0358 JB: 0.0452 DA: 0.0327	**NN: 712.45** EL: 745.37 JB: 3368.6 DA: 831.62
8	28507	10240 – 20480	0.10 – 0.6	**NN: 0.0288** EL: 0.0454 JB: 0.0791 DA: 0.0433	**NN: 1553.4** EL: 1660.8 JB: 7096.9 DA: 1796.4

- reconstruct kinematic variables x and Q^2 at collider via ML
- using ZEUS MC at HERA

- still working on the low kinematic range, outperforms conventional methods elsewhere
- did consider dividing the network into several for the different regions of the detector, for now decided to work with a single network covering the full detector

Summary: Multidisciplinary approach

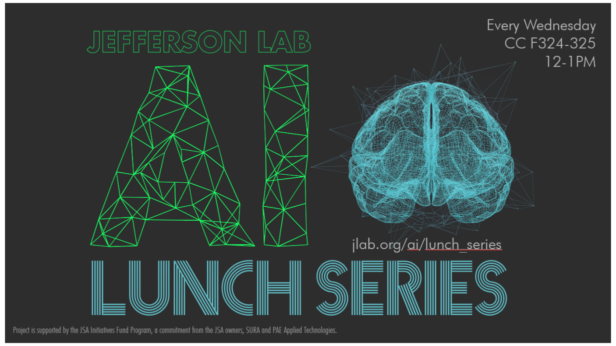
- interplay between **Mathematics**, **Computer Science**, and **NP**
  - computer scientists need problems to solve
  - NP problems give insights into research in computer science and mathematics
  - great opportunity for education

- **related to** in NP (and HEP) need closer connection between experiment and theory

- **A.I./ML research**
  - scientific, systematic approach to applying A.I. / ML approaches to NP problems
  - activation functions, network design particular to NP applications
  - building efficient networks no more complex than necessary
  - NP analysis:
    - want to extract information from **all** the data and find correlations / common features
    - key difference with respect to HEP
  - **need to trust A.I. / ML and AI**
    - drive for *explainable AI* and *uncertainty quantification*
    - human interaction could be applied with great benefit to better understand the requirements and dynamics of such criteria in the NP domain
    - debatable whether *explainable* is a useful criterion for a ML model. We don't have the words for theories we haven't discovered yet

Summary: Data science is about data

- **reference datasets for A.I. / ML development in NP**
  - always an issue to get access to big real datasets
    - amount of training data required often unknown
    - often two orders more data for simulations / training required than data
  - important to cultivate ML development
  - but, always difficult to understand from outside the experiment what the data means
    - common struggle for analysis preservation
    - project by the library community to address open data: Open Science Framework
- question to NP community: **Can we as a group figure out what datasets to ask for?**
  - the data was paid for by the DOE in the first place, after all
  - we have to ask, it's not going to magically just appear on the web
  - we need data to make progress
- **pose open challenges and run contests**
  - this has really worked to draw in new young people and new ideas
  - give prizes!
  - Can we think about benchmark problems?



Summary: Simulation challenges

Identify challenges coming over the next few years? pick our top problems?

- **Accelerating simulations**
  - calorimeter in particular
  - but also PID, e.g., Cerenkov detectors
- **A.I driven detector design** Bayesian optimization for EIC detector R&D
- **HPC utilization**
  - Experimental NP, HEP have few or no payloads appropriate to the LCF/Exascale which are accelerator based.
  - ML is the best near term prospect for using them effectively.
  - Can we find the ML payloads? Do they use substantial processing resources?
- **ML for event generators**
  - replace models with ML as we do in detector simulations (e.g., LUND string model)

Thank you very much for the discussions