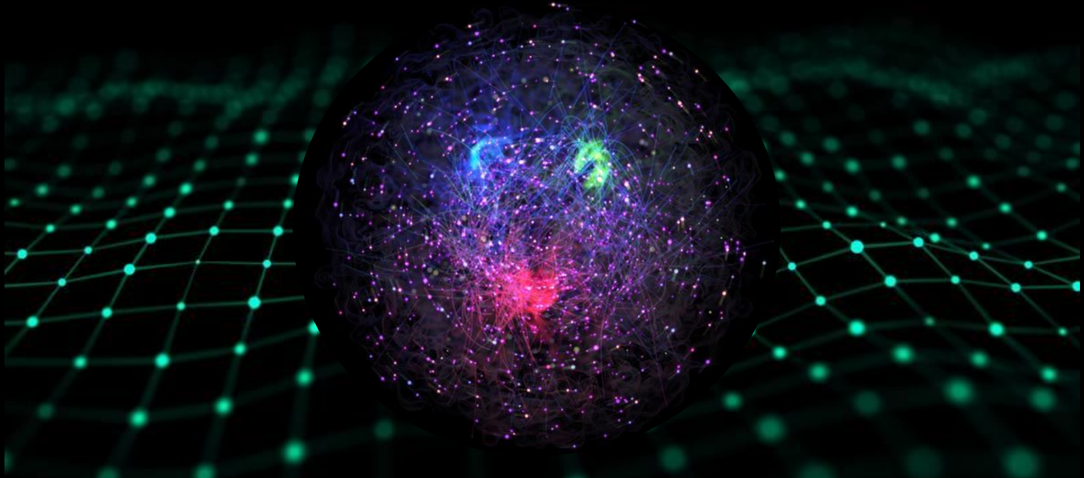
Machine Learning and Artificial Intelligence Applications for QCD Experiments

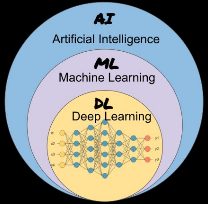


09/22/2022

Cristiano Fanelli

NSAC Long-Range Plan Town Hall Meeting on Hot and Cold QCD

Outline



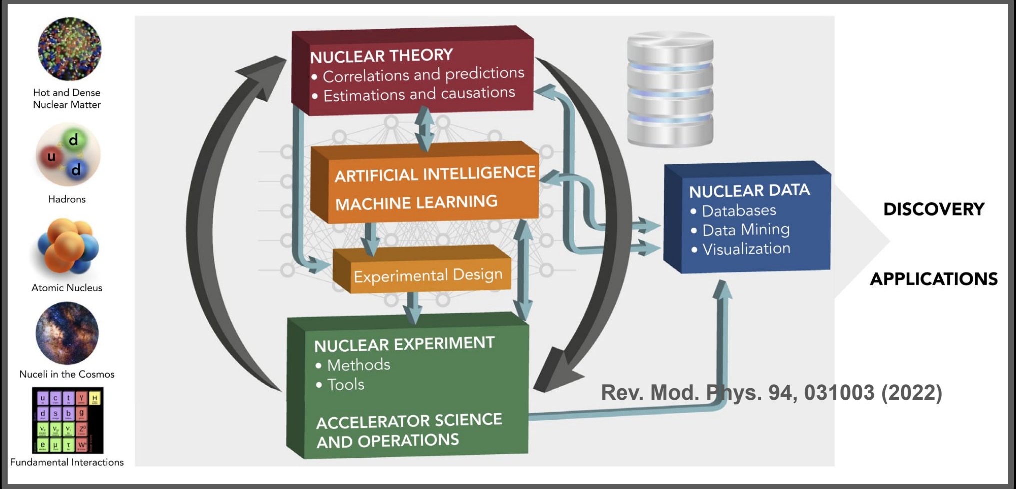
AI is the capability of a computer system to mimic learning, problem-solving and reasoning. Here is defined to broadly represent the next generation of methods to build models from data and to use these models alone or in conjunction with simulation and scalable computing to advance scientific research. These methods include (and are not limited to) Machine Learning (ML) — help the computer learn without direct instructions, Deep Learning (DL), Statistical Methods, Data Analytics, and Automated Control.

- AI/ML nearly everywhere in nuclear physics community
  - Experimental Applications in hot and cold QCD: a 10000m view
    - Multidimensional problems
    - Decisions in data streaming
    - Uncertainty quantification
- Future experiments (EIC): AI/ML from the beginning
  - Community (AI4EIC)
- Conclusions

This talk will cover AI/ML used for detector development and data analysis

AI/ML in QCD

- Several workshops have identified the scientific challenges and opportunities at the intersection between AI and data intensive science such as NP, highlighting the tremendous potential of AI for new insight and discoveries within NP research
- AI/ML techniques are now actively being used in multiple aspects of NP; they will be applied nearly in every system of next QCD frontier experiments like the EIC



*examples from [1]

JLab, sPHENIX, EIC

next slides

next slides

Particle Track Identification CLAS12

DL for calorimetry in GlueX FCAL

Jet Physics at EIC; Heavy-flavor tagging jets / interaction with QGP at STAR (with possible extensions to sPHENIX, EIC);

next slides

Deep Learning for Deep Inelastic Scattering reconstruction of kinematics, EIC

EIC

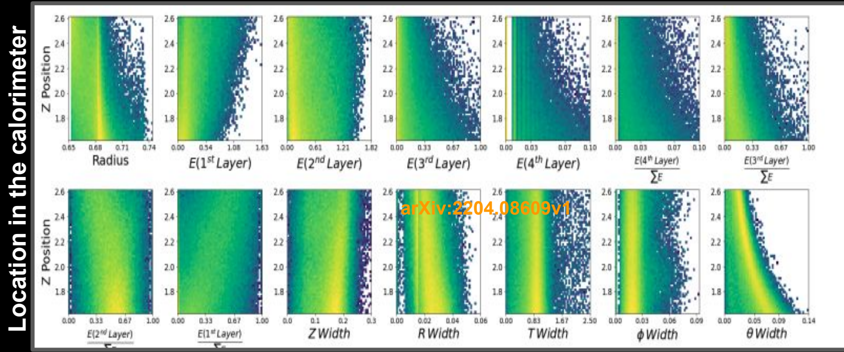
[1] A. Boehnlein et al., Machine learning in nuclear physics, Rev. Mod. Phys. **94**, 031003 (2022) and references therein

AI/ML in QCD

- AI/ML can cope with multi-dimensional problems, and can handle and capture complicated correlations. This, supported by the growth of computational power, is thriving research in directions previously unexplored due to complexity of problems: challenges and limitations for traditional/standard methods are often opportunities for AI/ML.

- AI gives the opportunity to include autonomous control and experimentation. This is highly relevant to accelerate science and drastically reduce the time between data taking and publication. Experiments are pushing for streaming readout and AI for this reason [SRO X workshop (2022)].

Example of multi-dimensional feature space: Shower reconstruction in GlueX Barrel Calorimeter

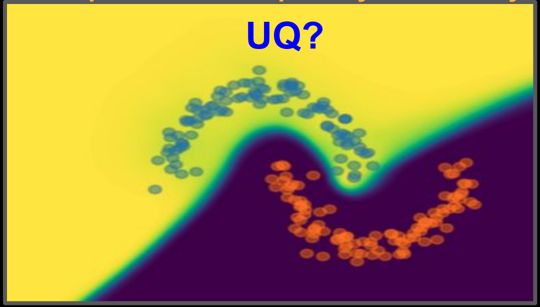


Multiple reconstructed features

- We, as a community, have the opportunity to take advantage of the full potential of AI/ML: this can a have tremendous impact, e.g., 3D imaging of quarks and gluons in the nucleon

- The above can result in a paradigm shift for NP if we understand the uncertainties and biases in the approach. There is a breadth of topics in this area and our requirements are quite unique and are typically not being solved by industry [topical meeting on UQ at AI4EIC].

Example of need to quantify uncertainty



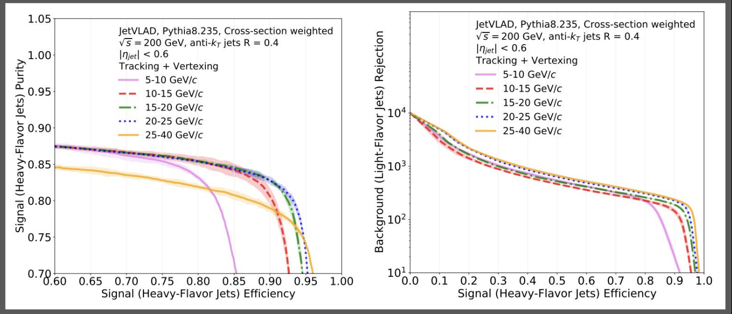
Jefferson Lab
Exploring the Nature of Matter

Problem-specific

JetVLAD

Courtesy of R. K. Elayavalli (Vanderbilt)

Mimic ResNet family, width the same as output of NetVLAD



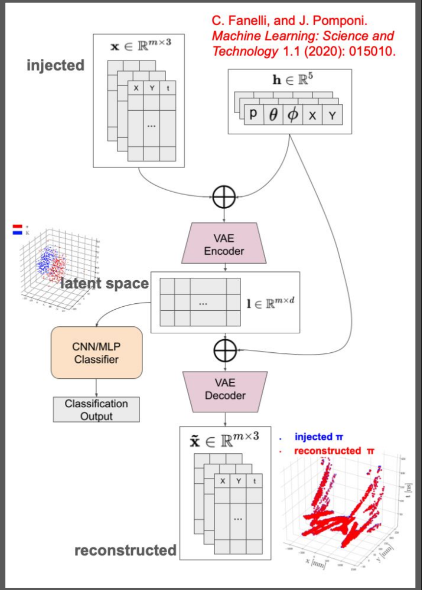
Studied for STAR (potential application @sPHENIX, EIC)

- Focused on identifying jets originating from heavy quarks such as b and c, as opposed to lighter quarks and gluons. Trained on jets produced with PYTHIA.
- JetVLAD takes charged jet constituents with varying quantities as input and aggregates to a descriptor vector which can then be used to compare different jet populations. This offers a feature space with improved classification performance.
- At increased jet momenta found that signal purity ~constant with increased background rejection. Studies highlight the importance of a precision vertex detector for HF.
- Work on extending JetVLAD to use meson tagging as opposed to quark tagging (reduce dependence on simulation fragmentation). Other ongoing projects on simulation-based inference given a jet structure (JETSCAPE Coll)

[1] J. Bielíková et al, "Identifying heavy-flavor jets using vectors of locally aggregated descriptors", 2021 JINST 16 P03017

DeepRICH

Cherenkov detectors are the backbone of PID @EIC



- Need to speed-up simulations
- Complex hit patterns, sparse data, response as a function of the kinematics – DIRC detector produce the most complex hit patterns — need accurate and fast reconstruction
- DeepRICH: Deeply Learning the Reconstruction of Imaging Cherenkov detectors Possibility to learn at the event-level rather than at the track/particle level.
- Can learn to generate hit patterns (also trained on high purity sample from real data) — calibration, alignment
- Same performance of best performing reconstruction algorithm with ~4 orders of magnitude speed-up in inference time on GPU

[1] C. Fanelli, J. Pomponi, "DeepRICH: learning deeply Cherenkov detectors", Mach. Learn.: Sci. Technol., 1.1 (2020): 015010
[2] S. Joosten (ANL), Bottlenecks in classical simulations:: where AI can help? AI4EIC, 2021

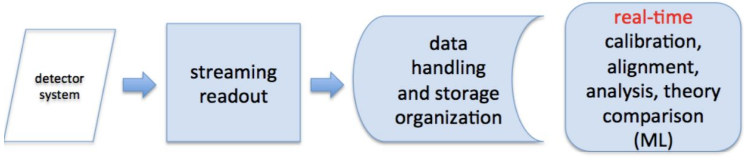
Jefferson Lab
Exploring the Nature of Matter

AI/ML in SRO

The development of streaming readout (SRO) for the NP driven by research initiatives:

- Streaming Grand Challenge [1] and the facility for “Innovation in Nuclear Data Readout and Analysis” (INDRA) at JLab
- BNL LDRD "High Throughput Advanced Data Acquisition for eRHIC, Particle Physics and Cosmology Experiments"
- PHENIX, STAR and sPHENIX (BNL), KM3NeT(INFN), BDX (JLAB) and CBM (FAIR)

SRO Grand Challenge [1]



Aim to remove separation of data readout and analysis
take advantage of modern electronics, computing, and analysis

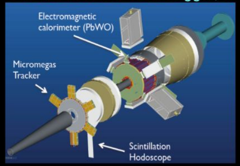
SRO for next generation electron scattering [2]

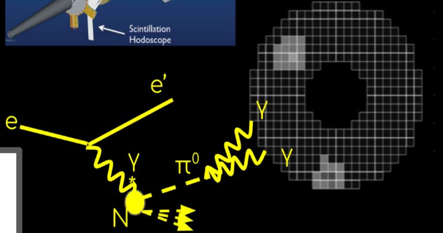
ML deployed on stream of real data

CLAS + EPSCI @JLab

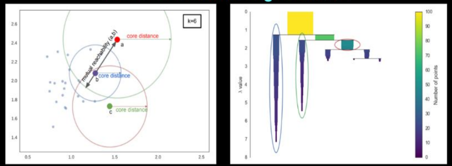
- CLAS12 SRO setup
- TriDAS SR back end
- JANA2 reconstruction framework

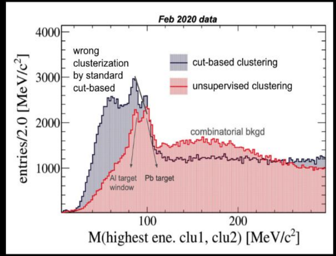
The CLAS12 Forward Tagger, JLab

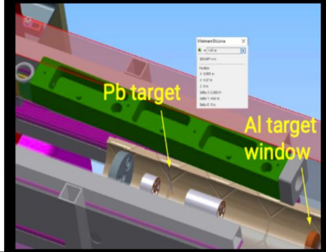




Hierarchical clustering in JANA2







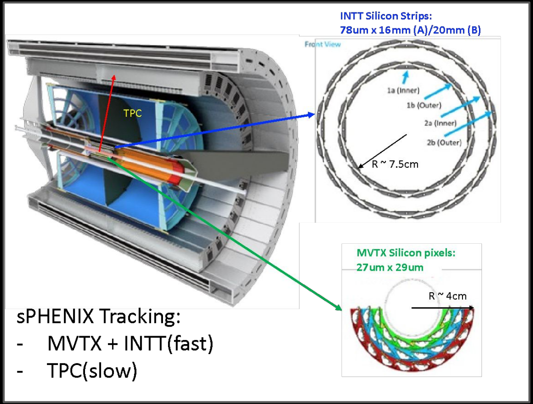
Hierarchical clustering VS traditional clustering of energy deposited by photons; AI robust against variations in experimental conditions* (uncalibrated data in SRO)

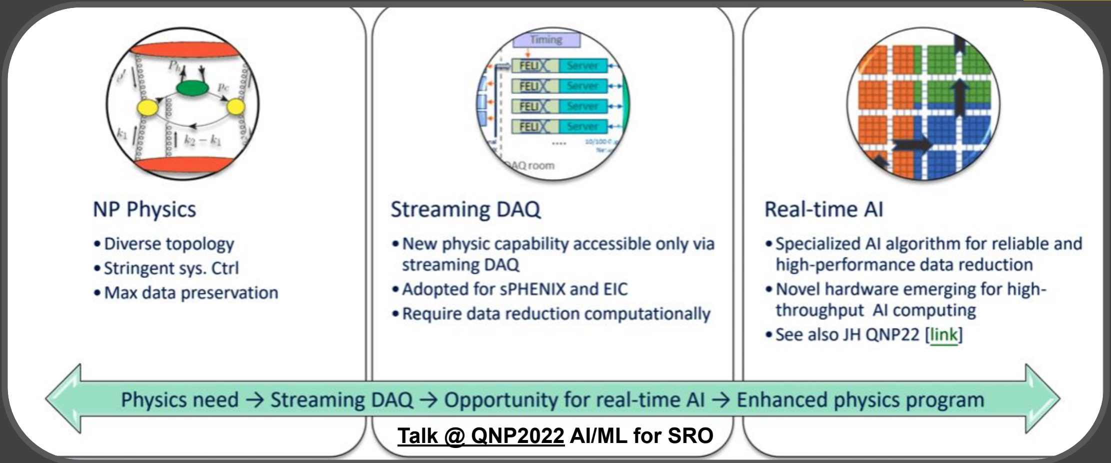
Courtesy of M. Battaglieri (JLab)

Many active projects regarding SRO at JLab: INDRA/ASTRA [3], AIEC (AI for Experimental Control) [4], Hydra (Online monitoring) [5], SRO with ML on FPGA [6]

[1] A. Boehnlein, R. Ent, and R. Yoshida, Grand Challenge in Readout and Analysis for Femtoscale Science, 2018
[2] F. Ameli, et al., Streaming readout for next generation electron scattering experiments, Eur. Phys. J. Plus, 2022
[3] M. Diefenthaler et al., Diefenthaler, Markus, et al. Evaluation & Development of Algorithms & Techniques for Streaming Detector Readout. No. 2020-LDRD-LD2014. 2020.
[4] T. Jeske, et al. "AI for Experimental Controls at Jefferson Lab." JINST 17.03 (2022): C03043. — AI4EIC proceedings
[5] T. Britton, B. Nachman. "Accelerator and detector control for the EIC with machine learning." JINST 17.02 (2022): C02022. — AI4EIC proceedings
[6] S. Furletov et al., Machine learning on FPGA for event selection — AI4EIC proceedings

AI/ML in SRO





Courtesy of J. Huang (BNL) [1]

FastML: Fast Data Processing and Autonomous Detector Control for sPHENIX and Future EIC Detectors

Intelligent Experiment Through Real-Time AI
(DOE FOA funded 2022-2023)

Collaboration of NP, HEP and CS:
LANL, MIT, FNAL, NJIT, ORNL, UNT, CCNU

Identify D/B hadrons with real-time ML

- Topology of D/B decays
- Monitor collision vertex
- Feedback for improvement

The challenges:

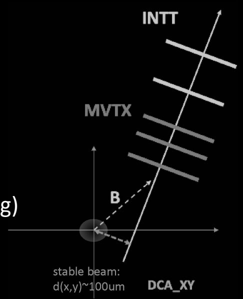
Very high p+p collision rate: ~3MHz

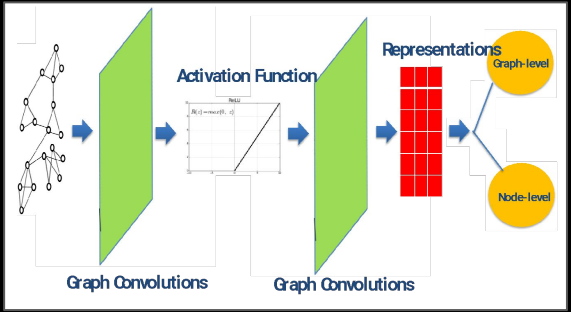
Low rate of rare signals: ~150Hz (beauty for eg)

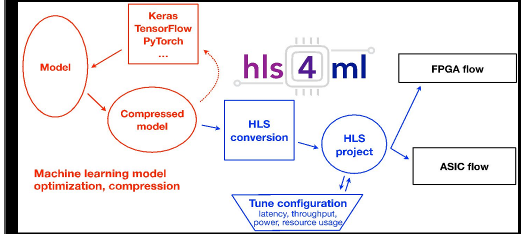
Limited DAQ trigger bandwidth: ~15 kHz

(or 0.5% of p+p collisions)

No effective conventional triggers available







Courtesy of Ming Liu (LANL)

[1] Huang, Yi, et al. "Efficient Data Compression for 3D Sparse TPC via Bicephalous Convolutional Autoencoder." 2021 20th IEEE (ICMLA). IEEE, 2021.
[2] F. Fahim, et al., “HLS4ML” arXiv:2103.05579 (2021)

Leitmotif in AI/ML

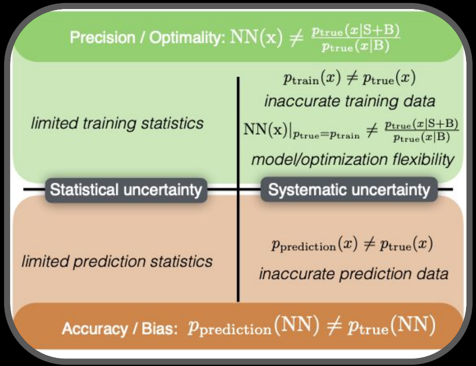
Courtesy of B. Nachman (LBNL)

Uncertainty Quantification

statistical (aleatoric) / systematic (epistemic)

decrease with more events

model bias



"If the network architecture is not flexible enough it may be that the likelihood ratio is not well-approximated. This means that the procedure will be suboptimal and will not achieve the best possible precision. However, if the classifier is well-modeled by the simulation, then p-values computed from the classifier may be accurate, which means that the results are unbiased. Conversely, a well-trained network may result in a biased result if the simulation used to estimate the p-value is not accurate."

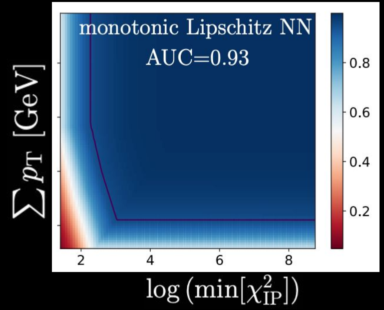
inference/uncertainty-aware approaches

[1] B. Nachman, "UQ for ML Applied to Data Analysis", talk at AI4EIC Meeting on Uncertainty Quantification

[2] B. Nachman, *How to achieve optimality and account for uncertainty,* arXiv:1909.03081

Courtesy of M. Williams (MIT/IAIFI)

Robustness

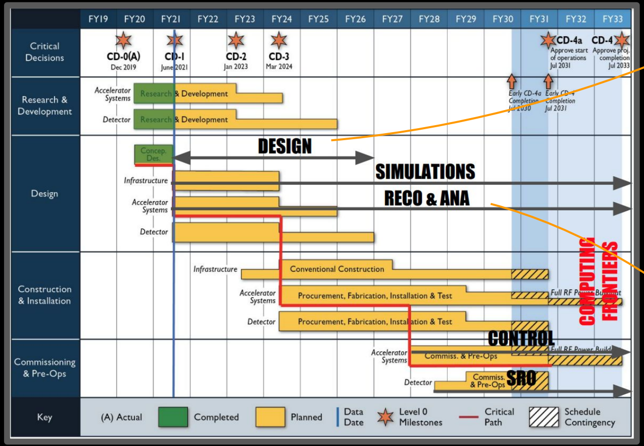


- The Lipschitz constant of the map between the input and output space represented by a neural network is a natural metric for assessing the robustness of the model.
- This new method constrains the Lipschitz constant of dense DL models (can also be generalized to other architectures). The method relies on a simple weight normalization scheme during training that ensures the Lipschitz constant of every layer is below an upper limit specified by the analyst.
- The algorithm was used to train a powerful, robust, and interpretable discriminator for heavy-flavor decays in the LHCb realtime data-processing system.
- LHCb has adopted this for the major selection algorithms, and looking at it for PID, fake-track killers.

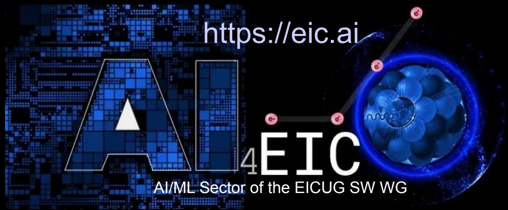
[1] O. Kitouni, N. nolte, M. Williams "Robust and Provably Monotonic Networks", arXiv:2112.00038

AI since the beginning: EIC

AI considered since the very beginning in EIC, cf. [1]

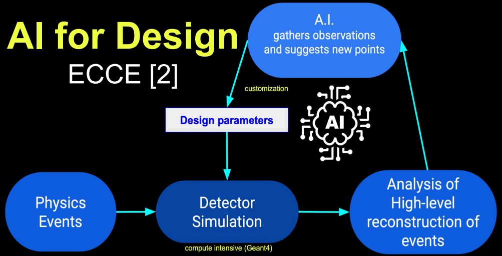


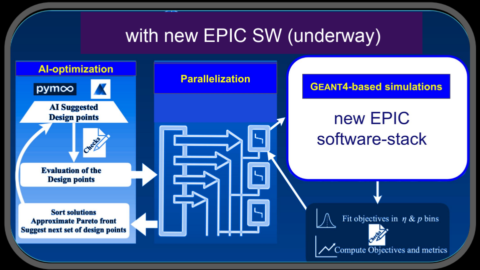
(EIC schedule shown at 1st AI4EIC Workshop, 2021)



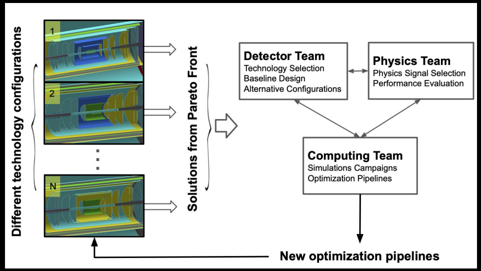
AI for Design

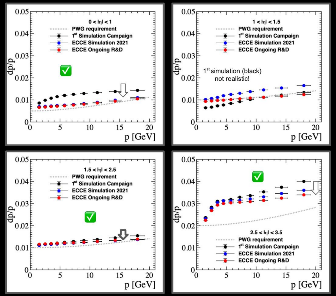
ECCE [2]





Adaptive Multi-objective Optimization of the EIC Detector Design





ePIC SW stack [3]

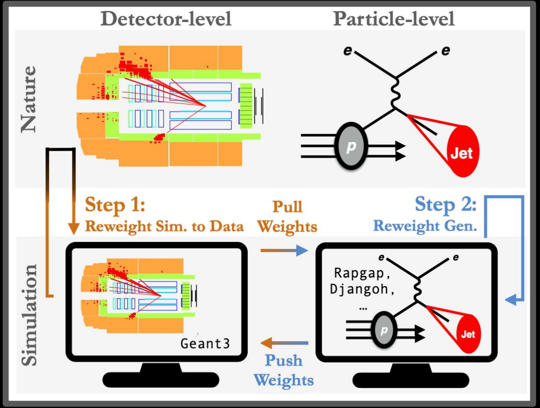
The ePIC Collaboration is developing a modern SW stack that embraces the EIC SW statement of principles, with forward-looking aspects favorable for AI/ML implementation and utilization of heterogeneous computing

[1] R. Abdul Khalek, et al. “EIC yellow report." Nuclear Physics A 1026 (2022): 122447.--- Chap. 11.12 on AI for EIC
[2] C. Fanelli, et al. (ECCE), "AI-assisted Optimization of the ECCE Tracking System at the Electron Ion Collider." arXiv:2205.09185 (2022).
[3] W. Deconinck et al., “The EIC Software Stack: Designing a Scientific Software Environment for the 2030s”, APS Meeting, NP Division, Fall 2022

Unfolding and "data-driven" learning

Courtesy of B. Nachman

Unfolding



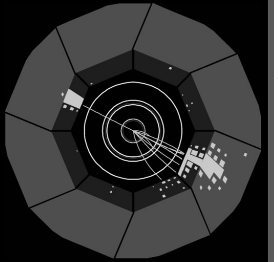
OmniFold [1]

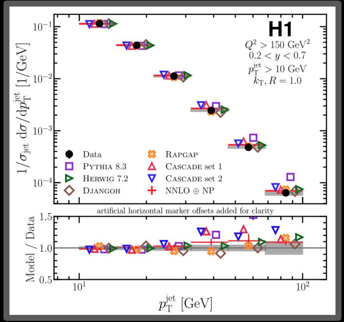
Using ML for differential cross section measurements (OmniFold and otherwise). These tools for recent measurements with DIS from HERA data and the same tools could be used at the EIC.

A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, and J. Thaler "OmniFold: A Method to Simultaneously Unfold All Observables" Phys. Rev. Lett. 124, 182001 2020

Courtesy of M. Arratia (UCR), B. Nachman

Lepton-jet correlation in DIS at H1 [1]





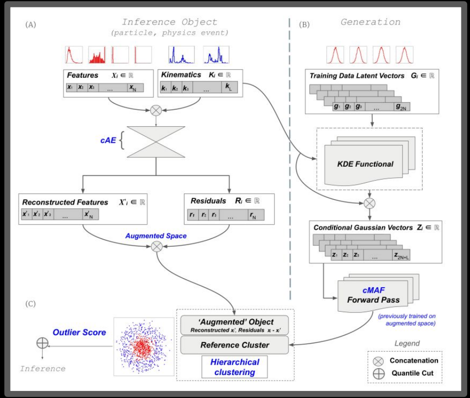
- First example of ML-assisted unfolding (MultiFold method): enables simultaneous and unbinned unfolding in high dimensions.
- This development will allow us to do unbinned cross-section measurements
- Similarly, this could be applied at EIC

[1] V. Andreev et al. (H1 Collaboration), "Measurement of Lepton-Jet Correlation in Deep-Inelastic Scattering with the H1 Detector Using Machine Learning for Unfolding" Phys. Rev. Lett. 128, 132002

In the "opposite" direction, it could be exciting thinking about data-driven learning that relies less on simulations, with tools like, e.g., one-class classification / anomaly-detection [1] and weak supervision / topic modeling [2].

Flux+Mutability [1]

cAE + cMAF + HDBSCAN



Same architecture applied to n/γ showers reconstruction in GlueX and BSM dijet signatures at LHC

[1] C. Fanelli, J. Giroux, and Z. Papandreou. "Flux+ Mutability": A Conditional Generative Approach to One-Class Classification and Anomaly Detection." arXiv:2204.08609 (2022).
[2] M. LeBlanc, B. Nachman, and C. Sauer. "Going off topics to demix quark and gluon jets in α_S extractions." arXiv:2206.10642 (2022).

Jefferson Lab
Exploring the Nature of Matter

AI Community in QCD

- The “A.I. for Nuclear Physics” workshop (2020) and report [1], along with a hackathon of 8 teams each with 4 participants, contributed to create a proto-community around AI for NP; this has been followed by the AI4NP winter school (369 registered participants), and the 1st AI4EIC workshop (2021) (243 registered participants). All huge successes.

https://eic.ai

- Starting from the Yellow Report [1], and as clear from the 1st AI4EIC workshop, AI is being integrated in all aspects of the EIC

- AI4EIC (https://eic.ai) is a working group of the EICUG dedicated to AI for the EIC community; good forum to address important cross-cutting aspects (accelerator, detector, theory, DS/CS)

- It organizes regular monthly meetings (typically topic-oriented), annual workshops, hackathons and data challenges, tutorials and schools; it contributes to disseminate AI in the EIC community

AI/ML Sector of the EICUG SW WG

- Upcoming 2nd workshop — October 10-14, 2022, William & Mary; the workshop will have sessions on accelerator/detector design, theory/experiment connections, reconstruction/PID, AI/ML infrastructure and frontiers, streaming readout; it will also host tutorials (experts from academia, industry, national labs) as well as an (international) hackathon event. More info at https://indico.bnl.gov/e/AI4EIC

[1] P. Bedaque, et al. "AI for nuclear physics." The European Physical Journal A 57.3 (2021): 1-27.
[1] R. Abdul Khalek, et al. “EIC yellow report." Nuclear Physics A 1026 (2022): 122447.--- Chap. 11.12 on AI for EIC

Conclusions

AI is a perfect fit for experimental applications in hot and cold QCD:

- Need support for interdisciplinary research and develop multi-disciplinary workforce:
  - engage with data science community; providing FAIR dataset; collaborations in HPC exascale systems and AI/ML; take full advantage of exciting possibilities offered by new HW and SW and AI/ML within the NP community through educational and training activities.
- Take full advantage of SRO and AI using heterogeneous computing. This can improve near real-time analysis and control (e.g., “intelligent” and automated detectors).
  - A common theme is applying AI-methods with well-understood UQ (both systematic and statistic). If we understand the uncertainties and biases, near real-time analysis with SRO can result in a paradigm shift for NP with faster turnaround time to produce scientific results.
- Transitioning from prototyping to deployment in production environments — How do solutions/prototyping from LDRD projects end up in production environments in our experiments? E.g. Fast simulations; SRO.
  - AI/ML Infrastructure: looking ahead, we shall adopt actual MLOps (end-to-end pipelines CI-CD-CT-CM); this is connected to Data Management, particularly provenance and reproducibility; another important topic is distributed strategies for training.
- Need for problem-specific tools: the most interesting challenges that can be approached in NP and AI will require approaches that go beyond industry standard tools.
- Other cross-cutting themes: Robustness, Explainability, also very important features for applications in our field.

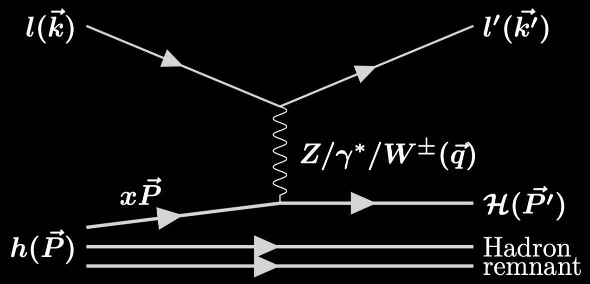
[1] Computational Nuclear Physics and AI/ML Workshop, 6-7 Sep 2022
[2] P. Bedaque, et al. "AI for nuclear physics." The European Physical Journal A 57.3 (2021): 1-27.

Backup

AI/ML applications: DIS

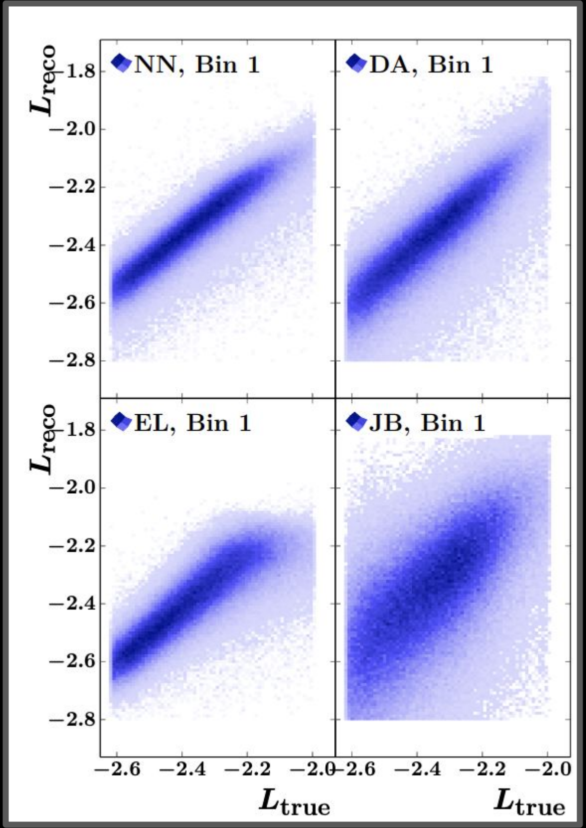
Deeply Learning DIS

Courtesy of M. Diefenthaler (JLab)



(Born level)

DIS fundamental process @EIC



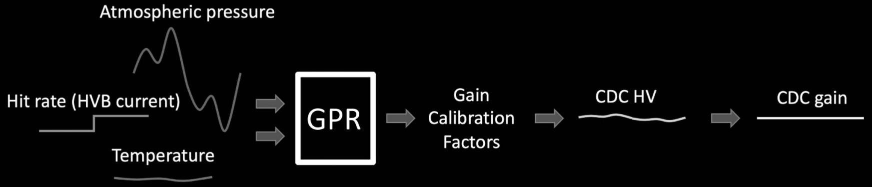
- Use of DNN to reconstruct the kinematic observables Q^2 and x in the study of neutral current DIS events at the ZEUS experiment at HERA.
- The performance compared to electron, Jacquet-Blondel and the double-angle methods using data-sets independent from training
- Compared to the classical reconstruction methods, the DNN-based approach enables significant improvements in the resolution of Q^2 and x

[1] M. Diefenthaler, et al. “Deeply Learning DIS Kinematics” arXiv:2108.11638
[2] M. Arratia, et al., “Reconstructing the kinematics of DIS with DL”, NIM-A 1025 (2022): 166164

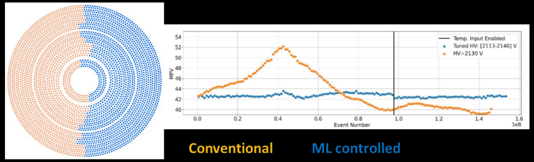
AI/ML for Control

T. Britton, D. Lawrence, K. Rajput (JLab)

AIEC: AI for Experimental Control [1]



Verification Results:
Cosmic Rays
(no beam)



Most probable value from Landau fit to experimental data as function of event number for the GP-controlled (blue) and constant 2125 HV (orange) sections of the CDC."

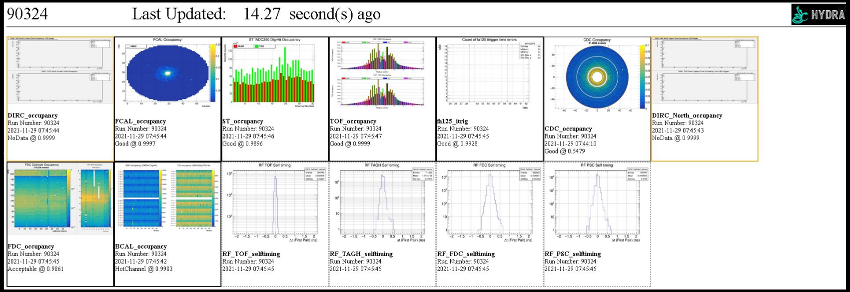
Use GP regression and Uncertainty to make an action

AI predicted Gain Correction Factors compared to existing GCFs for 2018 and 2020. Able to predict the existing GCFs using input features readily available via EPICS system during data taking.

N. Jarvis (CMU) T. Jeske, D. McSpadden (JLab)

Hydra: Online Monitoring Tasks [2]

- Take off-the-shelf ML technologies and deploy in near real-time monitoring tasks for GlueX in Hall D.
- It was the online monitoring coordinator's job to sift through hundreds of images produced in the previous 24 hours, looking for missed anomalies. This "human-in-the-loop" method prone to errors.
- Hydra was created to tackle these challenges. Hydra is an AI system that leverages Google's Inception v3 for image classification.



It uses for training the collection of monitoring plots that GlueX had previously recorded. A webpage was created to label the collected images and the entire system is driven by a database. Hydra is able to spot problems missed by humans and has been shown to perform better than humans at diagnosing problems.

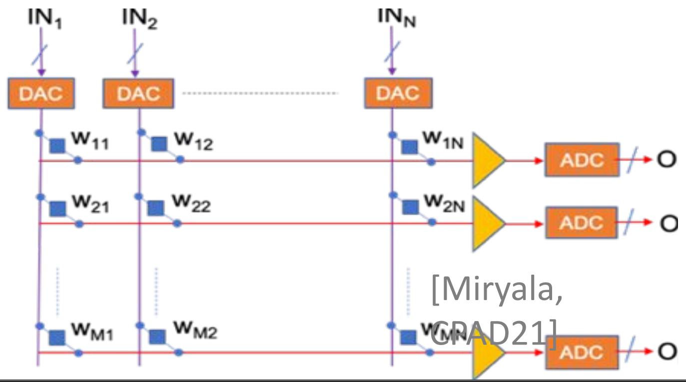
[1] T. Jeske, et al. "AI for Experimental Controls at Jefferson Lab." JINST 17.03 (2022): C03043. — AI4EIC proceedings
[2] T. Britton, B. Nachman. "Accelerator and detector control for the EIC with machine learning." JINST 17.02 (2022): C02022. — AI4EIC proceedings

Streaming DAQ and real-time AI

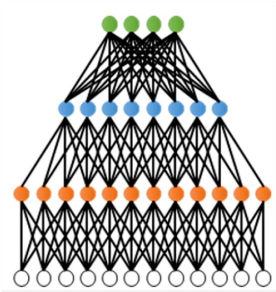
Courtesy of J. Huang (BNL)

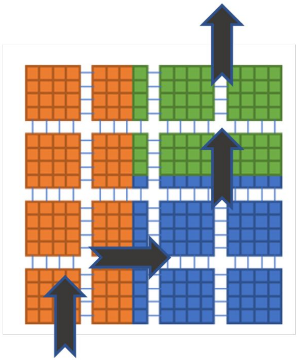
- NP Physics studies diversified event topology with stringent systematics control. ⟶ Streaming DAQ; example adoption in sPHENIX and EIC
- Streaming DAQ require large data reduction computationally
  ⟶ Opportunity for real-time AI, e.g. feature extraction, lossy compression
- Multiple effort in building specialized AI algorithm for reliable and high-performance data reduction, and testing on emerging hardware for high-throughput AI computing, examples:

In-memory computing at ASIC | DNN on FPGA | AI-chips w/ non-von-Neumann Architecture



[Miryala, CPAD21]

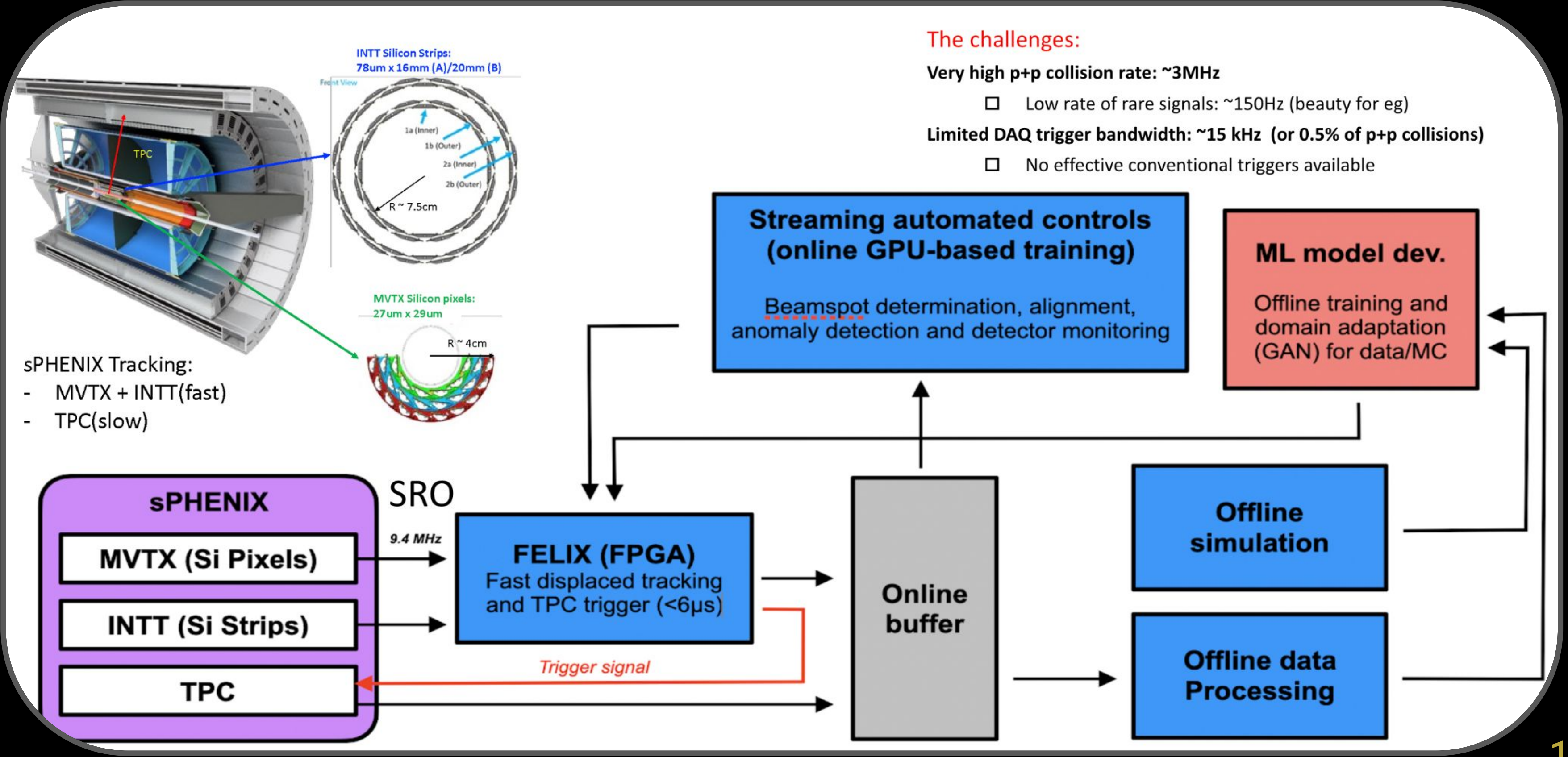




[Cerebras web doc.]

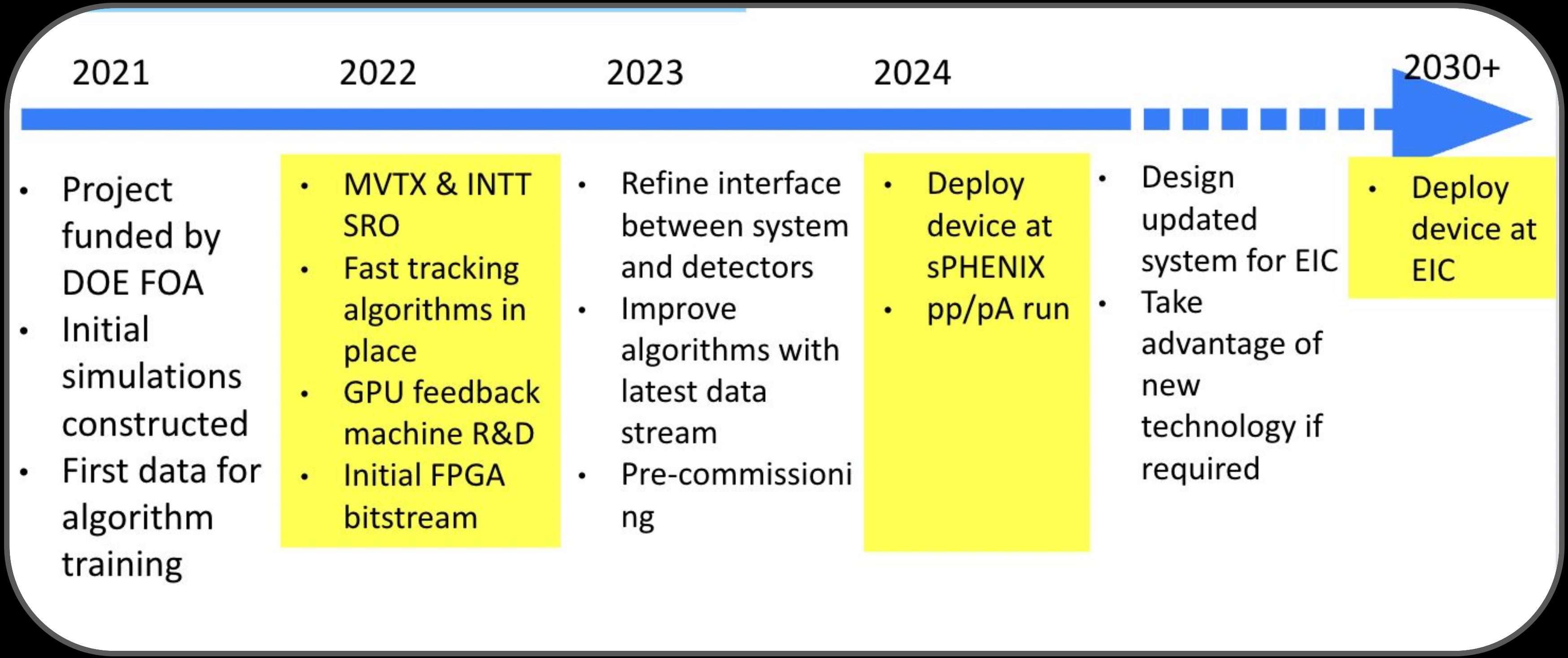
Intelligent Experiment Through Real-Time AI: (DOE FOA funded 2022-2023) Fast Data Processing and Autonomous Detector Control for sPHENIX and Future EIC Detectors

Courtesy of Ming Liu (LANL)



Timeline and Outlook

Courtesy of Ming Liu (LANL)



Trigger AI Algorithm R&D

Courtesy of Ming Liu (LANL)

Implemented several models to solve the trigger detection problem:

- **Directly applied GNN model to trigger detection problem (GNN)**
- Added a global vector to the GNN model to represent some global feature (VPGNN)
- DiffPool model (DiffPool)
- VpGNN + DiffPool (GNNDiffPool)
- ParticleNet , Giorgian

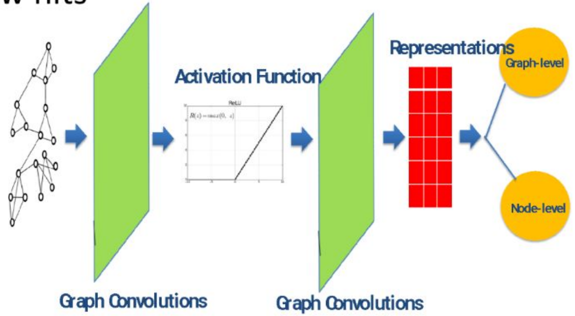
Another model we tried: Set2Graph (Affinity Matrix Prediction)

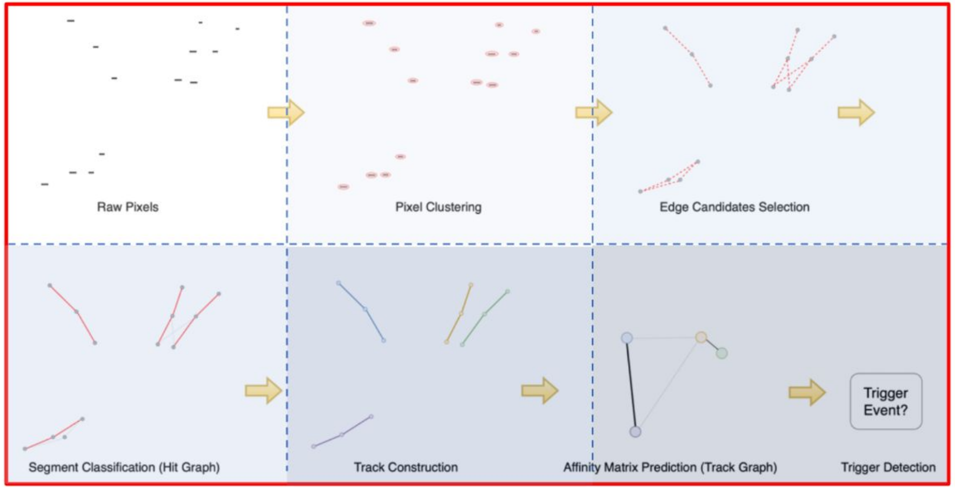
True_tracklets:

1) 90% BG rejection, Sig_eff ~ 90%
2) 99% BG rejection, Sig_eff ~ 40%

Inputs:

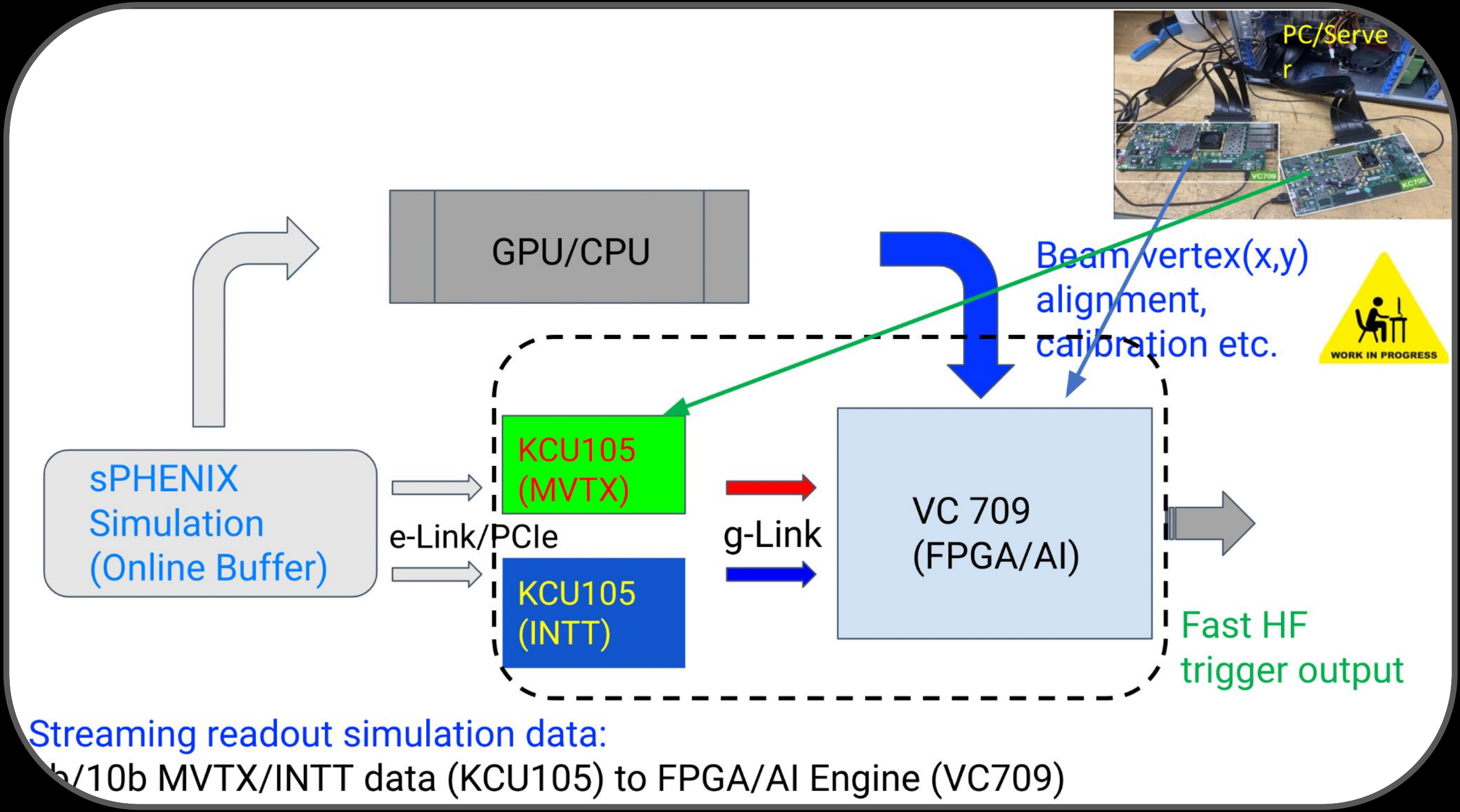
-raw hits





A Toy Model Hardware Implementation

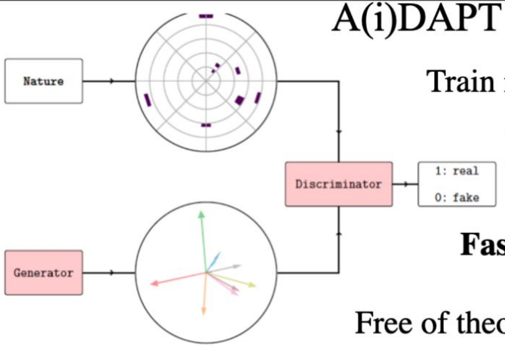
Courtesy of Ming Liu (LANL)



AI for Data Analysis and Preservation

Courtesy of M. Battaglieri, A. Hiller Blin (JLab)

A(i)DAPT — applied to JLab physics program



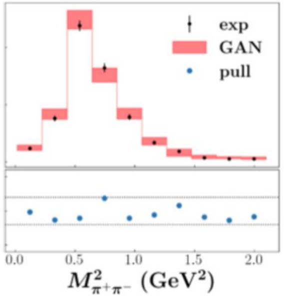
Train machine learning based event generators (**MLEG**) on experimental data.

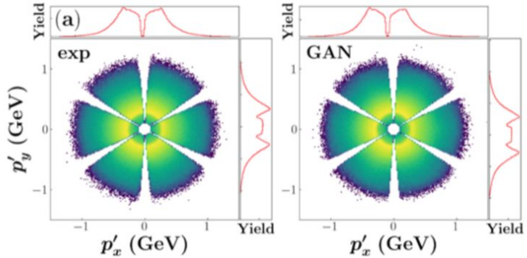
Y. Alanazi, T. Alghamdi, P. Ambrozewicz, M. Battaglieri, G. Constantini, A. Hiller Blin, E. Isupov, T. Jeske, Y. Li, L. Marsicano, W. Melnitchouk, V. Mokeev, N. Sato, A. Szczepaniak, T.Viducic

Faster than conventional Monte Carlo event generators (MCEG) — **high statistics** achievable.

Free of theory assumptions — highly efficient and **minimum bias interpolators**.

Means to **preserve data** in a compact form — **no histogramming/binning** needed!





Physics analysis — explore universality in kinematics and production channels.

Implement **detector effects** vs extract **physics at vertex level** — folding vs unfolding.

Validation and **uncertainty quantification** metrics — closure tests.

[1] Y. Alanzi et al.,"ML-based event generator for e-p scattering" arXiv:2008.03151
[2] Y. Alanzi et al., "A survey of ML-based physics event generation", arXiv:2106.00643