AI/ML Synergy

EPIC Software Infrastructure Review

Cristiano Fanelli

On behalf of the EPIC Collaboration

Outline

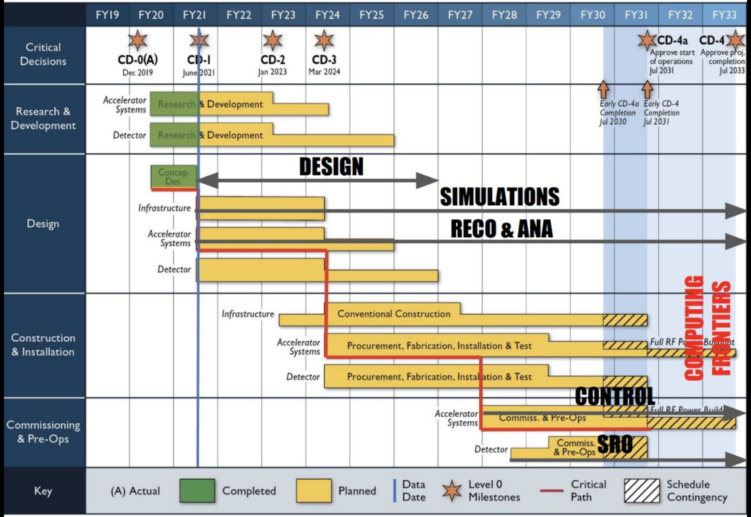
- AI/ML at EPIC picture (as of now)
- How AI/ML is folding into the SW planning
  - Forward-looking aspects of the EPIC SW favorable for AI/ML implementation
    - Concrete example and connections to previous talks
      - AI/ML in the design phase
      - AI/ML in SRO
      - Other
- AI/ML infrastructure aspects and planned discussion / events
  - Steps forward
  - AI/ML community at EIC
- Conclusions

AI: Artificial Intelligence; ML: Machine Learning; DL: Deep Learning

AI/ML at EPIC: present picture

EPIC is one of the first experiments to utilize AI since the design and R&D phases.

AI is anticipated to contribute to multiple aspects of EPIC for near real-time analysis, autonomous calibration, alignments etc.



AI/ML sessions at the 1st workshop on Artificial Intelligence for the Electron Ion Collider (AI4EIC), Sep 2021

https://eic.ai/workshops

Ongoing activities in EPIC

AI-assisted design, Fast ML for SRO, ML/DL for PID (e.g., muon-ID, low photons in ZDC, etc), DIS event-level analysis with DL, etc.

Some AI/ML references for EIC (collaborative efforts):

R. Abdul Khalek et al arXiv:2103.05419, Yellow Report, Chap 11

AI-optimized detector design for the future EIC: the dual-radiator RICH case - E. Cisbani *et al* 2020 *JINST* 15 P05009

AI-assisted Optimization of the ECCE Tracking System at the EIC - C. Fanelli et al, *arXiv:2205.09185* (2022)

AI4EIC Proceedings

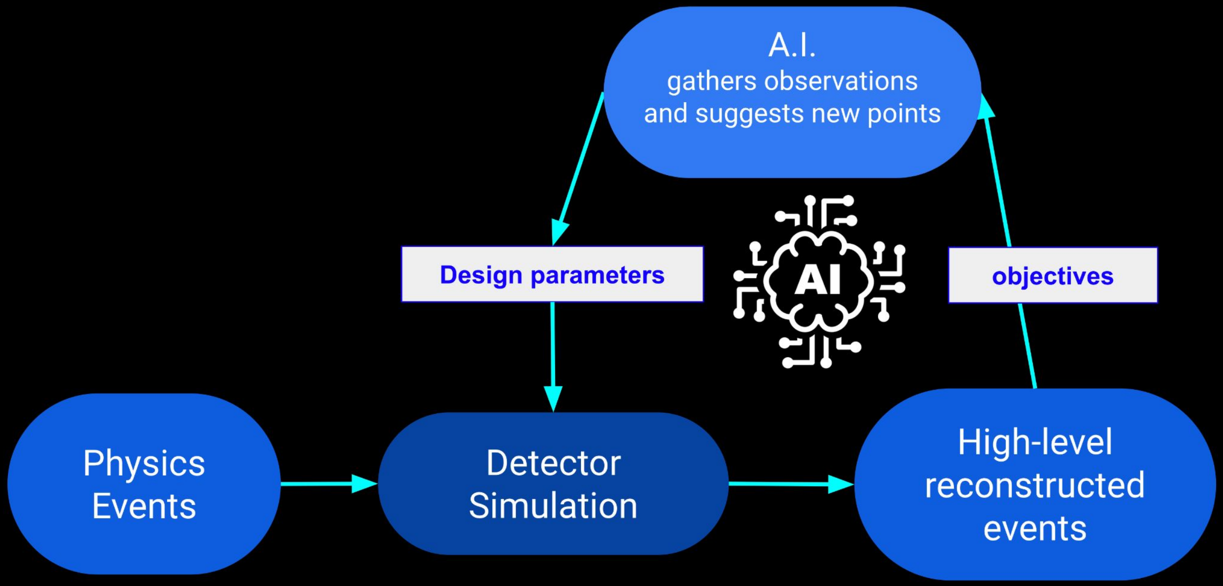
https://eic.ai/ai-ml-references

AI/ML activities are ramping up, and this trend will continue to grow in the next few years.

Low-hanging

AI-assisted Design as example

The AI-assisted design is a good example of how AI can be folded into the SW planning as it embraces all the main steps of the simulation/reconstruction/analysis pipeline



Agnostic to what is being optimized

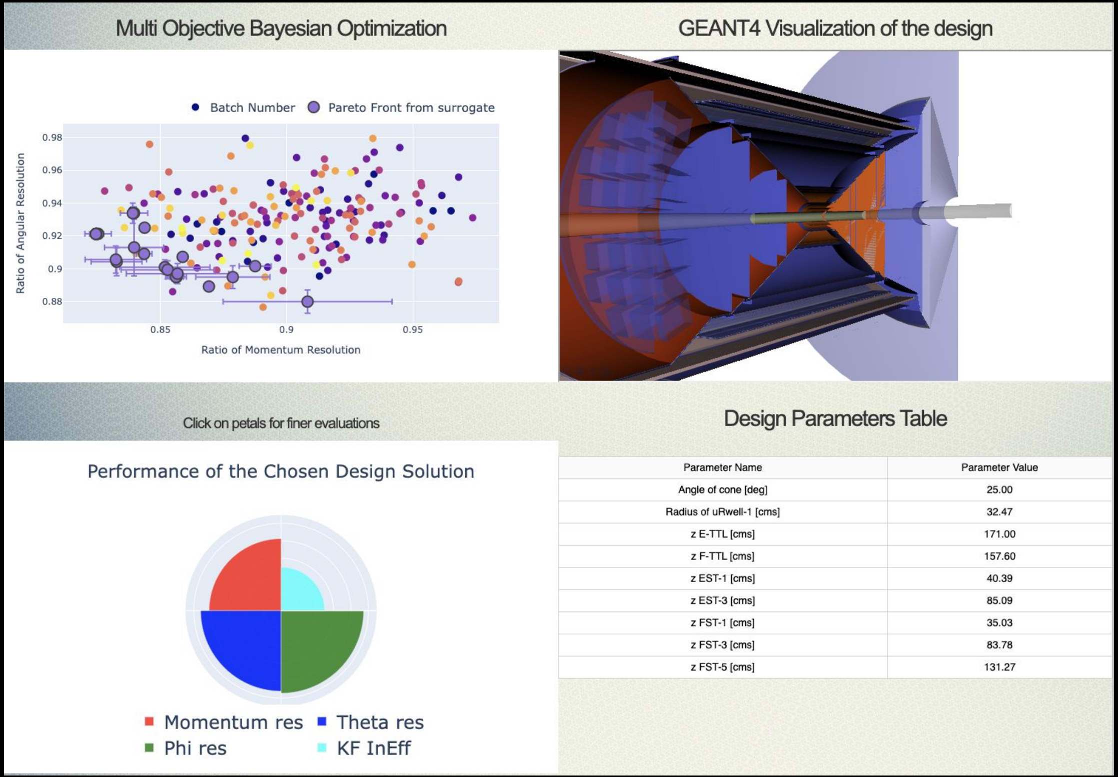
Leverages heterogeneous computing

Benefits from rapid turnaround time from simulations to analysis of high-level reconstructed observables

Needs production-ready SW stack throughout development and easy access to design parameters

*AI/ML can potentially enter in all the steps of the design pipeline

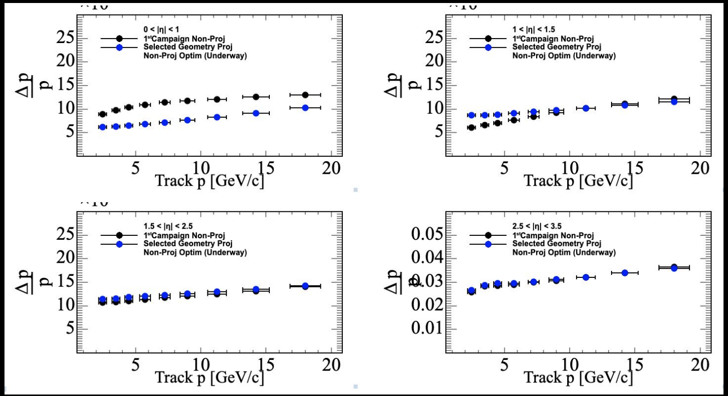
Integrate Modern Data Science tools



arXiv:2205.09185

The whole idea of the AI-assisted design is that of determining trade-off optimal solutions in a multidimensional design driven by multiple objectives

For an interactive visualization:
https://ai4eicdetopt.pythonanywhere.com



Leverage Geometry and Detector Interface

- Geant4 will continue to be the standard for detector simulations; **See M. Diefenthaler's talk**
  - Compute intensive simulations
    - AI-assist the design and in achieving optimality reduce usage of computing resources
    - Great interest in speeding-up known bottlenecks (e.g., calorimetry and Cherenkov) and have both full and fast simulations [1,2] (see, e.g., FastCalo GAN in ATLAS AtlFast3 [3])
- Geometry implementation via data source (DD4Hep uses ROOT TGeo) makes transparent the coupling of AI to the software stack design parameters; minimal changes needed to run different optimization pipelines
- Modularity of geometry description reduces complexity of parametrization and therefore computational complexity
- Other automated feature desirable for AI-assisted design, e.g., checking overlaps

[1] S. Joosten, Bottlenecks and limitations in classical simulations: where can AI help? 1st Workshop on AI4EIC, Sep 2021
[2] B. Nachman, Generative ML applications for simulations in colliders 1st Workshop on AI4EIC, Sep 2021
[3] G. Aad, et al., AtlFast3: the next generation of fast simulation in ATLAS, Computing and Software for Big Science 6.1 (2022): 1-54

Leverage Code repository, CI and Containerization

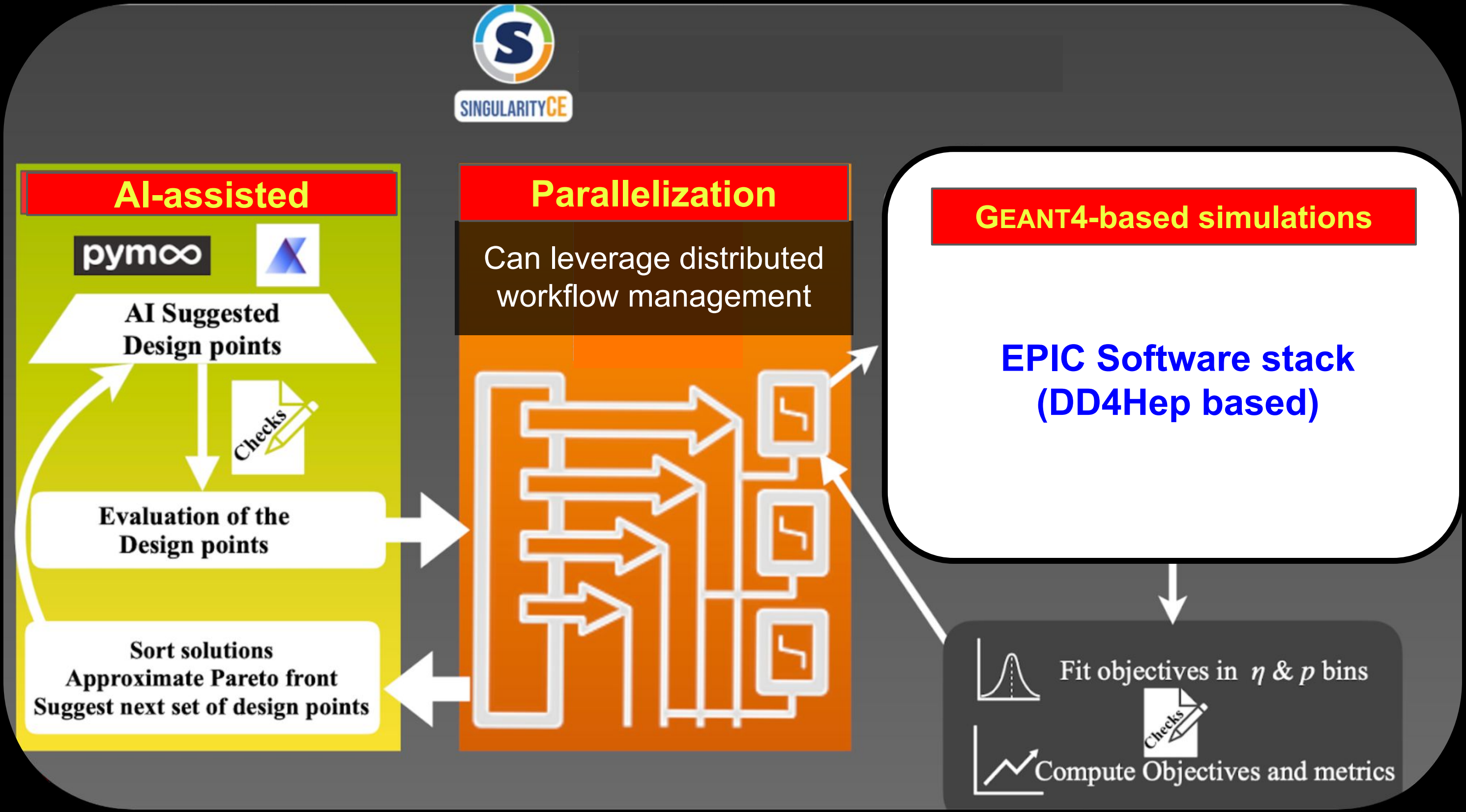
- In general AI/ML-related projects will follow best practices model for the repository (open and public; external packages not be forked/cloned to the eic organization and modified unless under exceptional circumstances).
- For the AI-assisted design:
  - CI/CD is mostly about keeping up-to-date with the EPIC simulation framework: it is needed when relevant updates are made to the simulation or a newer approach for optimization is adopted.
  - Containerization is being used in EPIC and previously in the proto-collaborations. Using singularity is typically preferable since it does not need elevated privileges to install additional packages/frameworks, which may make it easy to bundle AI/ML packages. Singularity can be integrated with the filesystem while preserving security restrictions.
- In general, when it comes to deploy / maintain ML models in production reliably and efficiently, Github Actions serve as a preliminary solution (accompanied to, e.g., platforms like wandb.ai https://github.com/wandb/wandb). Looking ahead, we shall adopt actual MLOps (end-to-end pipelines CI-CD-CT-CM) — see, e.g., MLFlow https://mlflow.org/

See W. Deconinck's talk

AI-assisted Design: Approaches

- AI-assisted approach satisfies EIC SW principles:
  - Open source and accessible to the whole community
  - Aim at reproducibility
  - User-centered design
  - Leverage heterogeneous computing
  - Not reinvent the wheel, aimed to build / extend existing efforts in wider scientific community
    - E.g., allows to engage and collaborate with Meta/Facebook Open Source (Ax: adaptive experimentation platform supported by Meta/Facebook)
      - Integrate cutting-edge data science built-in features for database backend to store experiments, visualization/interpretation, and presentation of results
      - See 2nd workshop on AI4EIC (tutorial by Meta), https://indico.bnl.gov/e/AI4EIC
- Design of increased complexity can take advantage of distributed computing. [**see J. Osborn's talk**]

Improving Design Workflow



Leverage Data Model Solution

- Open, simple, self-descriptive data formats. Flat data model in general allows flexibility for AI/ML applications. Data can be written to other ROOT, HDF5 files, etc
  - Collaboration with other scientists outside NP and HEP; among podio core features, it provides easy use interface to users, treating python as first class citizen (interface via pyROOT) [1]
  - Heterogeneous computing works best on flat data.
  - LHC Olympics for Anomaly Detection on 2020 stored events as pandas dataframes and saved to compressed HDF5.
- In the talk on data model, it has been pointed out how Standardized Data Model allow swapping different alternative as long as they adhere to the data model interface.
  - Example of clustering algorithm
  - An additional level of abstraction/portability is provided by unsupervised clustering, in that it is agnostic to the objects being clustered, as long as a metric distance can be defined to identify similar properties and form clusters
  - HDBSCAN currently being tested for caloriemtry in EPIC
- Supports for truth information in MC — useful for training

See W. Deconinck's talk

[1] EDM4hep and podio - The event data model of the Key4hep project and its implementation, F. Gaede et al., talk at vCHEP 2021

Use of HEP-supported packages

- Example of Acts, an experiment-independent toolkit for tracking, is free software, implemented in modern C++, and is currently being used or considered by ALICE, Belle II, CEPC, EIC, FASER, PANDA and sPHENIX, among others. [1]
- The project has three overarching goals:
  - Preserve current tracking approaches while enabling new developments
  - Serve as an algorithmic testbed for research in track reconstruction
  - Enable realistic development of new tracking detectors
- The framework includes the ONNX, an open-source AI ecosystem that empower developers to choose the right tools and frameworks to develop and deploy their Neural Network. [2]

[1] L.-G. Guignon, Machine Learning for track reconstruction at the LHC, JINST 17 C02026, AI4EIC Proceeding, 2022
[2] Open Neural Network Exchange, https://onnx.ai

AI/ML in Reconstruction Framework

- JANA2 framework handles streaming data in online triggerless environments.
- The core framework of JANA2 is written in modern C++ but includes an integrated Python interface — which facilitates integration of ML/DL applications [1]
- The first AI-based application in SRO using real data actually been realized in [2] using JANA2…

See D. Lawrence's talk

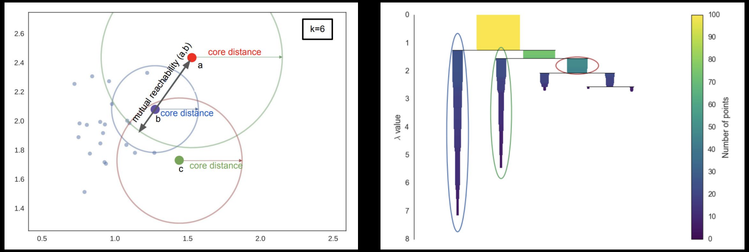
[1] D. Lawrence, A. Boehnlein, N. Brei, JANA2 Framework for Event Based and Triggerless Data Processing EPJ Web Conf. Volume 245, 2020
[2] F. Ameli, et al Streaming readout for next generation electron scattering experiment, 2022 (accepted on EPJP)

ML in Streaming Readout

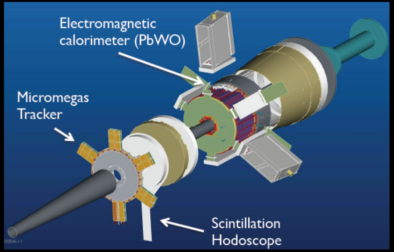
F. Ameli, et al Streaming readout for next generation electron scattering experiment, 2022 (accepted on EPJP)

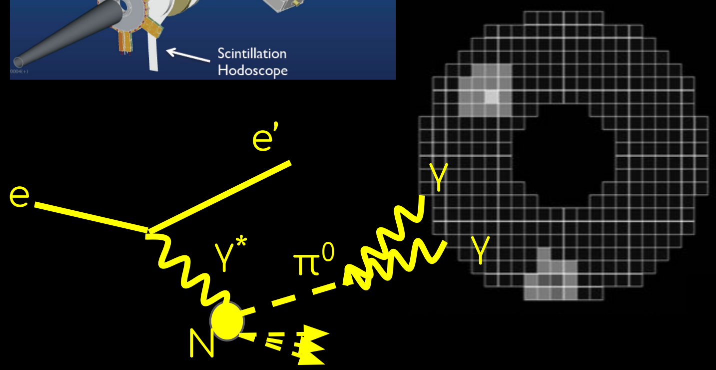
- CLAS12 SRO setup
- TriDAS SR back end
- JANA2 reconstruction framework

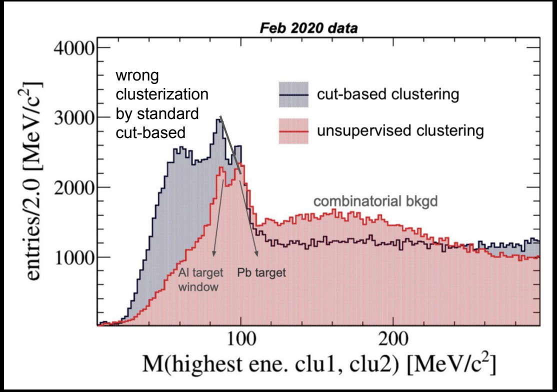
Hierarchical clustering in JANA2

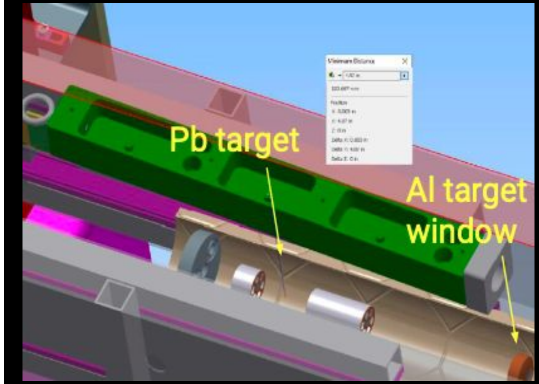


The CLAS12 Forward Tagger, JLab









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

Further improvements

Typical Data Science Pipeline

- Typical data science/ML pipeline has peculiar aspects not in common with standard analysis pipelines

Data Source
- Real or synthetic
- Quality
- Dimensionality
- Format
- Density
- Size

Data Preparation
- Data cleaning
- Data restructuring
- Correlations
- Dynamics
- Visualization

ML Applications
- Classification
- Regression
- Clustering
- Feature Extraction

Training Tools
- Cross-Validation
- HPO

Results
- Predictions
- Confidence Level
- Explainability

- This reflects, for example, in “Data and Analysis” Preservation

Infrastructure for AI/ML

- Machine learning lifecycle MLOps:
    - Models, hyperparameters, training datasets (depending on approach), etc…
    - How to manage experimentation, reproducibility, deployment, and a central model registry.
        - Record and query experiments: code, data, config, and results
        - Package data science code in a format to reproduce runs on any platform
        - Deploy machine learning models in diverse serving environments
        - Store, annotate, discover, and manage models in a central repository
- Deploy automated workflows to optimize neural networks

Scalability, Distributed vs Collaborative

- Federated computing architecture was deployed by proto-collaborations and a WLGC style architecture is envisioned [see **J. Osborn's talk**]
    - Rapid turnaround of raw data to online/offline productions; compatibility with Streaming Readout and near real-time physics ready productions; enabled distributed workflows HTC/HPC
- Distributed strategies may become necessary in AI pipelines working with big data: training time exponentially increases, scalability cumbersome, other limitation factors (e.g., algorithm computational complexity outpaces the main memory)
- Discussion on required infrastructure for next generation AI architectures will take place at the AI4EIC workshops and monthly meetings; with discussion on modern approaches, e.g.:
    - Distributed Learning - multi-node ML based on centralized data and distributing the model training
    - Collaborative Learning - multiple users collaboratively train a centralized model, with decentralized data and training

Community

The AIWG will serve as an entry point to AI applications and will organize workshops, tutorials, and Kaggle-like challenges.

https://eic.ai/workshops

2nd General Workshop on Artificial Intelligence for the Electron Ion Collider (October 10-14, 2022)

https://indico.bnl.gov/e/AI4EIC

- 4 sessions (one dedicated to infrastructure)
- Tutorials (one dedicated to lifecycle: MLflow)
- hackathon

1st workshop on Experimental Applications of Artificial Intelligence for the Electron Ion Collider (September 7-10, 2021)

https://eic.ai/events

Monthly meetings typically topic-oriented (UQ, Design, AI/ML in SRO, continual learning, etc)

https://eic.ai/community

Help organize educational events (tutorials, lectures) and collect documentation useful to disseminate AI/ML in the EIC community

Conclusions

- The recently formed EPIC collaboration is quite active in AI/ML:
  - EPIC detector can be one of the first experiments to be designed with the support of AI
  - The number of AI/ML activities is anticipated to grow in the next few months (e.g., reconstruction, PID); in the long-term, AI/ML will likely permeate and contributed to multiple aspects of near real-time analyses
- Lots of work has been recently done on the EPIC SW stack for the collaboration (DD4Hep, data model, JANA2), a fundamental step towards the CD2/3a
  - The EPIC SW embraces several forward-looking features that allow for AI/ML applications and utilization of heterogeneous resources.
- EPIC has a unique opportunity to integrate AI/ML in the SW from the beginning (and from an AI perspective)
  - Large-scale AI/ML applications entail considerations on scalability and specific infrastructure needs that require additional discussion — ML lifecycle; distributed training; etc
- The EIC community is engaged in AI/ML activities, and the AI4EIC WG is a good forum to address these important aspects. More info on meetings and workshop in https://eic.ai/events