Summary of AI for PP exploratory meeting

SWIFT-HEP meeting, Warwick, 11-12 November 2024
On behalf of organizing committee team and all contributors

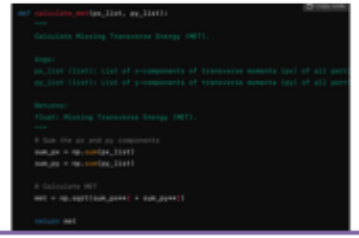
Davide Costanzo (Sheffield), Monica D'Onofrio (Liverpool), Rhiannon Jones (Sheffield), Biagio Lucini (Swansea), Tim Scanlon (UCL) & Alex Tapper (Imperial)

Purpose of the 1-day workshop

- Importance of AI
- Progressing further?
- Enabling AI in HEP
- Emerging technologies

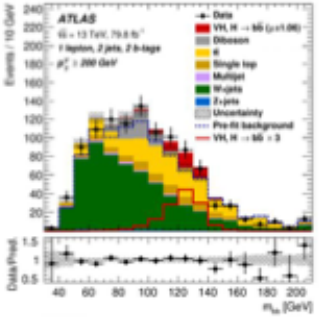
- AI is becoming increasing vital to a very wide range our physics programme
  - Can enhance almost all aspects of our experiments and some of theory:
    - Object reconstruction, event selection, simulation, event generation, data quality, detector control/monitoring, coding, trigger, hardware design, documentation….
  - Many examples where applying AI has been truly transformational

- AI has potential for further transformative change
  - Pushing the boundary:
    - Many (all?) students use AI tools to write code
      - Could envisage AI models that assist in the full analysis chain
        - *Focus more on physics rather than code frameworks?*



ChatGPT: Write a function to calculate MET

ChatHEP: 'Produce a data/MC comparison plot with latest theory/exp corrections for H->bb'



    - Feed raw data into Large Language Models which can reconstruct objects
      - *Solve CPU issues?*
    - Auto-encoders that can compress our data with little loss of information
      - *Assist with data storage issues?*
    - Marking all our exams/course work…..

Setting the stage

- From intro:

- Initially via Swift-HEP and ExaTEPP started an initiative for enabling AI in HEP
  - Proposed a work package on AI in Swift-HEP 2.0
    - Launched an AI Working Group in the interim
      - **Enabling AI in HEP experiment and theory**
        - Community-wide endeavour
  - Focus on challenges, barriers and opportunities in:
    - **Software, hardware and ML-Ops**
    - **Skills/Training and capacity building**
    - **Knowledge exchange and wider engagement**
  - This is the first exploratory workshop
    - Aim is to collect input on above areas and any additional topics missed
    - Decide on next steps for this initiative
    - Make links between AI focussed members of community

Structure

- https://indico.cern.ch/event/1450122/timetable/#20241001.detailed

Morning: overviews

Introduction
E3/E7, University College London

STFC's AI Strategy
E3/E7, University College London

ECFA AI Strategy
E3/E7, University College London

SciML Briefing
E3/E7, University College London

AI Funding Opportunities
E3/E7, University College London

Afternoon: expert-led discussion sessions

FastML Discussion — *Alex Martynwood et al.*
E3/E7, University College London 13:30 - 14:10

Skills/Training, Capacity Building and Knowledge Exchange — *Dr Timothy Paul Scanlon*
E3/E7, University College London 14:15 - 14:55

Tea/Coffee
E3/E7, University College London 15:00 - 15:30

Industry, AI and HEP — *Dr Gabriel Facini*
E3/E7, University College London 15:30 - 16:10

Software, Hardware and AI-Ops — *Andrew Chappell et al.*
E3/E7, University College London 16:15 - 16:55

Wrap-Up — *Biagio Lucini et al.*
E3/E7, University College London 17:00 - 17:05

Morning overviews

Introduction

E3/E7, University College London

STFC's AI Strategy

E3/E7, University College London

ECFA AI Strategy

E3/E7, University College London

SciML Briefing

E3/E7, University College London

AI Funding Opportunities

E3/E7, University College London

STFC AI Strategy

- Mark Wilkinson and Valerie Farr report

Why an AI Strategy

- Mission-led obligation to advance cutting-edge AI to carry on delivering world-leading science and innovation
- Latest AI has the potential to transform how we perform science and deliver discovery
- To make AI pervasive across STFC, need to take a strategic, organisation-wide approach to advancing AI
- Position ourselves as key contributors of sovereign AI capability
- Produce a well-informed Spending Review Bid, for the next multi-year SR

Key aims

- Galvanise STFC's communities to embrace potential of AI, as integral to delivering our mission and create mechanisms to achieve this.
- Help shift perceptions across UKRI and HMG about the nature of the relationship between STFC and AI
  - realise the transformative benefits of AI across STFC investments and advance the state of the art of AI.
- Enable potential funders to clearly see the value of advancing AI across STFC.
- Position STFC as credible delivery partner for HMG across the AI ecosystem.

Key findings (Internal)

- We have depth and breadth of AI expertise: we must get better at exchanging science outcomes and skills across as dynamically as possible
- We need to use our decades worth of experience in designing and exploiting large scale infrastructure to drive national capability in order that our scientists have access to what they need
- Our success is dependent on the whole system pulling together to deliver the data and compute requirements and invest in people

STFC AI Strategy (2)

- From report

Strategic vision

Our vision is that STFC advances science-led AI, at scale and pace, to maximise the value of data across research and innovation system, thereby unlocking the potential of AI to accelerate scientific discovery, develop sovereign AI capability and resilience and create economic growth and impact across society.

Strategic Themes

4 core themes:

- Placing AI at the heart of STFC
- Advancing STFC science and driving sovereign AI capability
- Delivering mission-led AI solutions
- Environmental sustainability and efficiency

1 cross-cutting theme: Partnerships

Interdependencies: National compute and data capabilities/investment

Realising the vision

We will realise our vision by:

- Creating an optimum internal ecosystem.
- Working in partnership across an interconnected, AI-skilled R&I ecosystem, both within UKRI and internationally
- Combining a scientific approach to the development and application of AI with near-term impact will be critical to our endeavours.
- Translating our skills, expertise and capabilities and taking a leadership role in advancing AI across other data intensive and at-scale science domains within the UKRI portfolio and those that underpin core national missions.

Your input needed

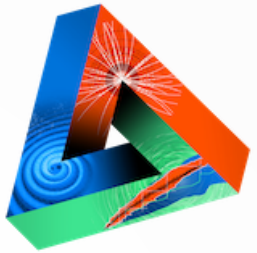
- STFC AI strategy is currently being developed
  - Provide examples of where AI is (or will be) essential to your work
  - Don't assume that someone else in your field is responding…..
- Comment on what you've seen in this talk
  - Highlight concerns/risks
- Provide feedback on strategy when circulated
- Engage with the delivery plan development process which will follow

In progress

Examples sent as representatives from various experiments/theory efforts

ECFA strategy: AI and computing

- Efforts at international level reported (MDO, talk)
- For computing needs (and AI) → JENAA: Joint ECFA-NuPECC-APPEC Activities
  - Started in 2022, a first **JENA computing** workshop was held on June 12-14, 2023 (Bologna)
  - https://indico.scc.kit.edu/event/3813/
- More specific for AI: EuCAIF → European AI for Fundamental Physics → Conference in 2024
  - EuCAIFCon2024 https://indico.nikhef.nl/event/4875/
- **Rationale:** e-group: eucaif-info@cern.ch
  - new European initiative for advancing the use of Artificial Intelligence (AI) in Fundamental Physics.
  - Members are working on particle physics, astroparticle physics, nuclear physics, gravitational wave physics, cosmology, theoretical physics as well as simulation and computational infrastructure.



JENAA
Joint ECFA-NuPECC-APPEC Activities

Aim of the event:
provide a **platform** for establishing new connections between AI activities across various branches of fundamental physics, by bringing together researchers that face similar challenges and/or use similar AI solutions.

Working group structure and mission

- **WG1: Foundation models.**
  - Mission: Facilitate research on large-scale foundation models for fundamental physics
  - Provide infrastructure, resources, data and models, connect researchers, define problems & metrics
- **WG2: AI-assisted co-design for detectors**
  - AI-assisted co-design of future ground- and space-based detectors - Related to the work of the MODE Collaboration
- **WG3: FAIR-ness and sustainability**
  - FAIR: findable, accessible, interoperable, and reusable data principles. Example (from UK) also here
- **WG4: ML and AI infrastructure –** Mission: understand status, guide funding and development efforts
  - Scalability and access to HPC Resources, Data Infrastructure and Infrastructure for Distribution, Integration of AI/ML into Existing Experimental Workflows, Large and Small Scale Models, Benchmarking and Standards
- **WG5: Building-bridges - community, connection**
  - Establish EuCAIF as a new brand for AI in physics (independent of CERN), eventually bringing in new grants
  - Aggregate and provide an overview of already existing structures and networks, put this on professional website

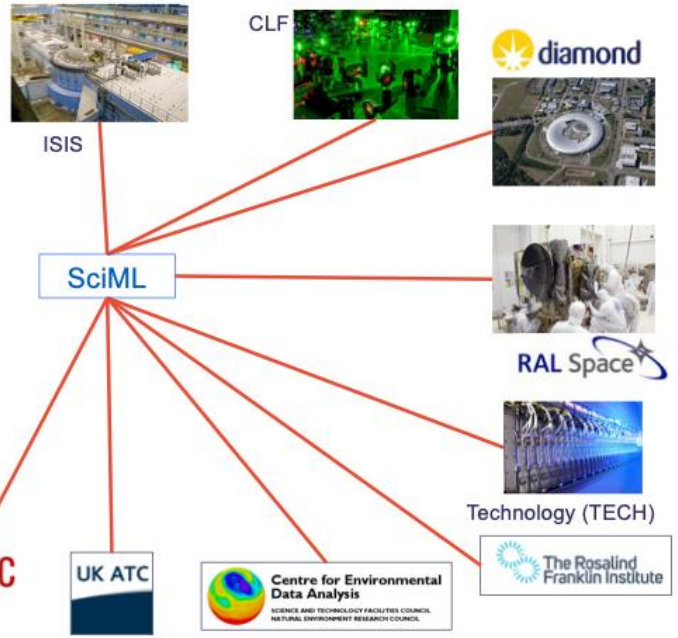
Overall, engagement from UK community but difficult to understand concrete impact in the short term. Experiences and good practice can be shared. Relevant for preparation of inputs to the European Strategy

SciML: AI for Science at STFC

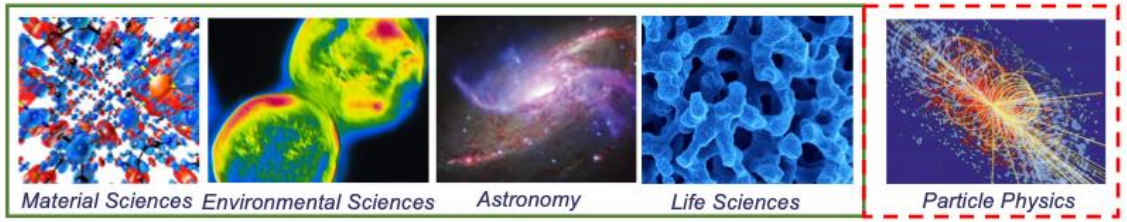
- Overview and opportunities presented by Jeyan Thiyagalingam

Our Role

- Work with national labs, facilities and STFC funded programmes
- Underpin the AI capabilities
- Our Remit: ***AI for Science***



Domain Coverage



Strands of AI for Science

Research to Advance AI

- Use AI for Science to improve AI itself
- Core research
- Attracts talents

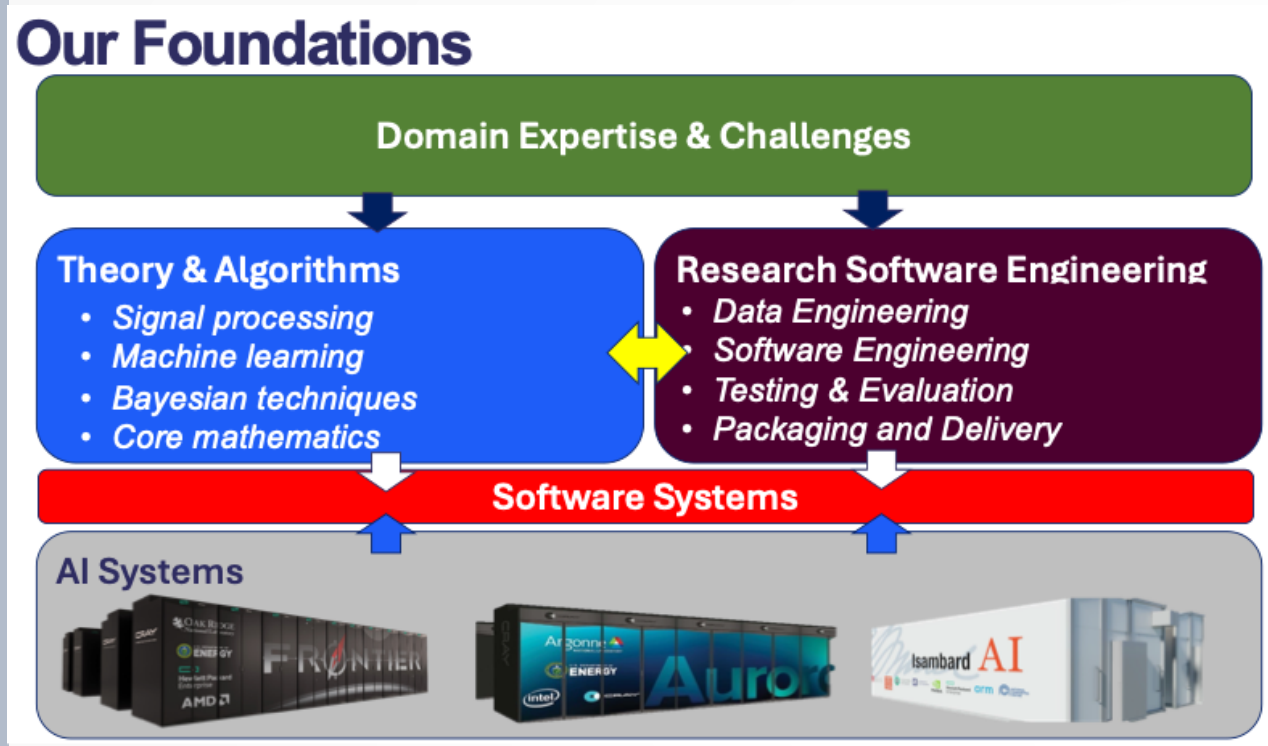
Use of AI in Science

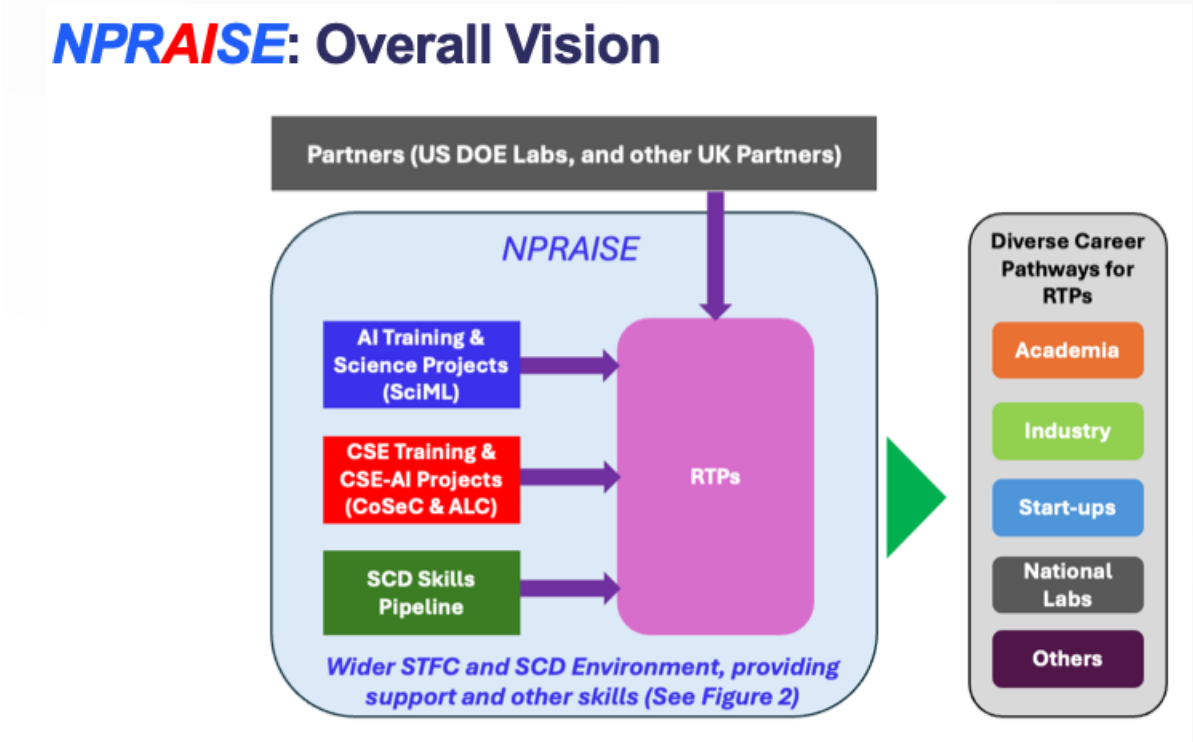
- Use AI to understand experimental results (from facilities)
- Rely on scientists for domain expertise
- Partially overcome first two challenges

Smart Facilities

- Smart facilities can improve science – high quality data, faster results etc.
- It is all about embedding AI at the heart of operations
- Rely on engineers for steer
- Partially addresses the first two challenges

SciML: AI for Science at STFC (2)





- Wide portfolio of activities, perhaps not sufficiently inclusive of experimental PP
- There are funded programmes: AI for Realistic Science(AIRS) by DSIT, Benchmarking for AI for Science (BASE-II) by EPSRC, etc.
- Only starting to grasp the surface of possibilities such as National Strategic Platform for RTPs on AI for Science and Engineering – NPRAISE

Possible ways for funding

- Davide made some considerations about funding strands

Possible funding opportunities

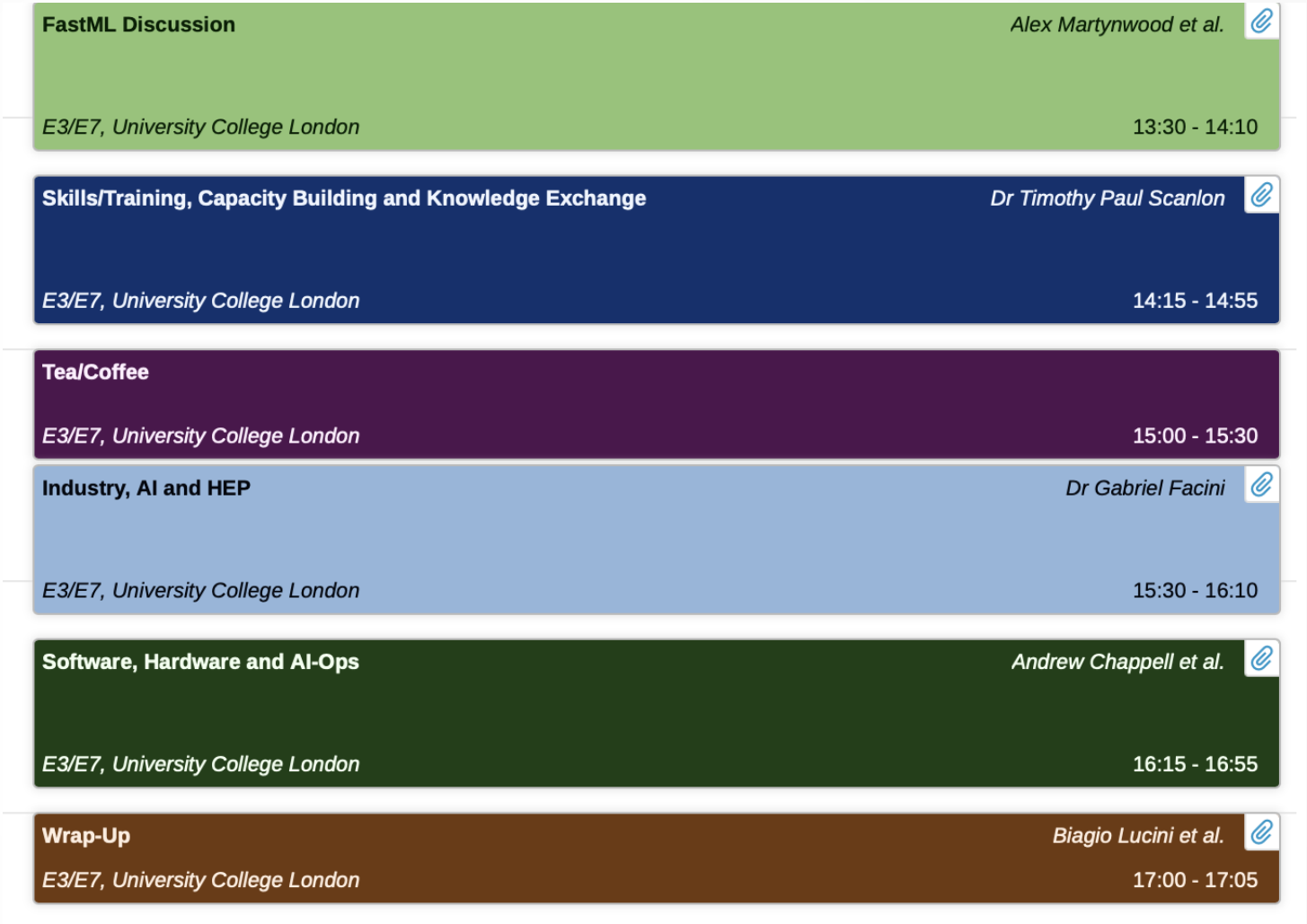
Let's separate things into two categories

- Funding applications that can be specific for particle physics
  - SWIFT-HEP v2.0 (if approved…)
  - ExaTEPP - or a future evolution
  - CoSeC CCP call (Roadmap writing, some RSE time, some CoSeC RSE)

- Applications that are of wider context (e.g. Network+, UKRI grants, etc)
  - We should make sure Particle Physics is represented
  - Typically led by non-HEP people, often competitive among us (eg CDTs, EU calls, etc)
  - We, as a community, should support these and get reports back
  - This forum could be a place to discuss these as openly as we can

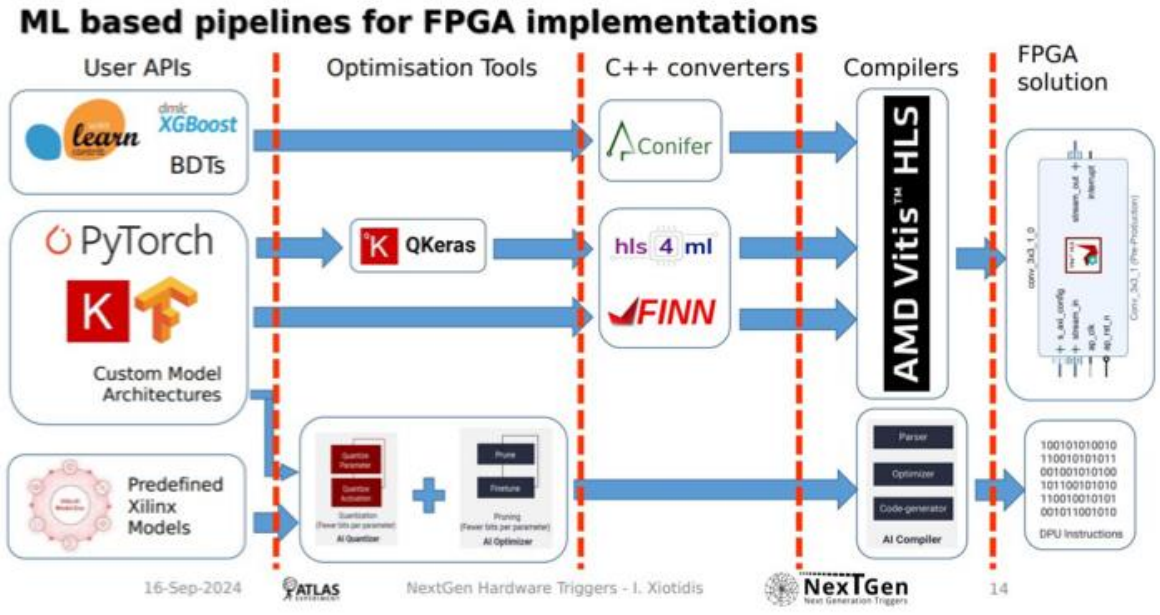
ExaTEPP

Expert-led discussion sessions



Fast-ML/FPGA

- **What are the HEP use cases for fast-ML/FPGAs and what is the current**
  - Trigger/DAQ applications: O(us) latency – “traditional” use on custom hardware
  - Wider use as accelerator for CPU based workloads – emerging – less UK involvement?



- Status
  - First BDTs and NNs running in ATLAS and CMS hardware triggers deployed
  - Large range of examples planned for HL-LHC upgrade
  - Development pipelines for ML on FPGAs – established but further development needed
  - Next Gen trigger project at CERN platform for development (within ATLAS UK) of ML pipelines for FPGA

Fast-ML/FPGA (2)

- **What are the challenges/barriers that we are facing in this area (e.g. software, training, skills-capacity, hardware, etc.)?**
  - Steep learning curve – albeit not as steep as VHDL
  - Understanding and tools e.g. hls4ml, pruning, quantisation
  - Bottleneck is (skilled, technical) effort – technical PhD funding, DRD funding ....
  - Consolidate UK skills across experiments – currently quite siloed
  - Evaluate industry directions e.g. Versal AI engines vs custom electronics (DRD7 etc.)
  - Challenges with obsolescence and software tools support
  - Explainability a key topic for confidence in tools and algorithms – in general but especially so in real time where data is lost
- **What are the opportunities in this area for HEP (e.g. enhanced outcomes, wider connections, funding, industry engagement, knowledge exchange etc.)?**
  - Highly trained people useful to science and UK generally – many people go on to data science and elsewhere e.g. CERN
  - Niche in AI – not well covered or supported by industry in general
  - Need to prepare case(s) to allow access to wider funding – examples from science & beyond
  - Industry connections – good connections in this area to key players and UK SMEs
  - Community tool hls4ml – UK contributes, plans driven by US towards formal basis

Skills/Training, Capacity building and KE

- Skills/Training – Essentials:
  - trained researchers in AI, expertise on latest developments, ability to identity and exploit latest developments
- Currently lacking a coherent and concrete approach to AI training in HEP across the UK

Online courses and material
→ lack depth or necessary HEP focus, little (so far) from the HSF on AI
University courses and 1st year PhD courses
→ can lack depth or necessary HEP focus
STFC's **DIS Summer School** (also international HEP-ML summer schools)
→ overall very useful
STFC's **Centres for Doctoral Training (CDT)** in Data Intensive Science (DIS)
→ perhaps most successful way to train next generation of researchers

General consensus on the need to have a better AI training focused on HEP

What form could the training take and how could it be provided?
- Available for PhD, PDRAs, academics, technical staff?
- Expand CDT programme and make sure benefits are available across the field
- In-depth courses available to PhD students at beginning of PhD?
- Should we collaborate internationally and if so what is the best conduit (HSF)?
- Hold in person training in software carpentry style to build national cohorts?
- Make DIS summer school into something more solid?
- Augment training with interactions with foundational/industry experts?

Skills/Training, **Capacity building** and KE

- **Capacity building** → Once trained, large fraction of experts leaves for industry
  - Good and healthy, but need to ensure _some_ stay in HEP
    - Need options for career progression and job security
    - Training (interactions with industry/foundational-AI experts)
    - Ability to stay at the cutting-edge (software/hardware)
    - Ensuring interesting work environment

Questions and discussions around the following
Ideas shared, some more difficult to realise than others

- Do you need some dedicated AI experts, general uplifting of AI skills or a good mix of both? What is the right balance?
- How do we ensure we retain enough AI expert practitioners:
  - Career paths, opportunities and job security?
    - See RSE pools e.g. ARC at UCL? Are there similar AI expert pools yet?
    - Based at institutes/labs, centrally or a mix, what is the right balance?
    - Career progression at institutes becomes less research-focussed 'up the ladder'
  - Staying at the cutting-edge and providing compelling opportunities?
    - Create posts split 60/40 with industry or in a pool working across areas
  - Providing sufficient hardware/software tools
  - Are there any other factors important to expert retention?

Skills/Training, Capacity building and **KE**

- **Knowledge exchange** → several layers to be considered
- Perhaps there isn't a preferred options but all ways should be explored
- **Training comes back as very important**

- Built into HEP - train fully fledged expert AI partitioners or hire AI-expert practitioners directly
  - Effective bridges between the AI and HEP communities (e.g. CDT)
- Across HEP - Areas enriched in ML expertise, encourage KE:
  - PPTAP identified this as a key area to build upon
- Foundational AI experts (e.g. Computer Science, Statistics….)
  - Many keen to engage providing a unique perspective
  - Better engage SciML and other expertise in Ada Lovelace Centre (see earlier talk)
- Industry (see next discussion item)
  - Big possible gains from engaging with industry
  - Training, hardware access, tools/techniques, funding
    - Hartree and CDT programmes plus direct engagement already ongoing
  - Good examples to build upon and grow community-wide

Effective sharing

- AI forums
- Interdisciplinary PhDs or co-supervision of PhDs (CDTs)
- Develop common tools and training courses
- Build stronger cohorts (CDT programmes, common training programmes, summer schools)
- Mapping expertise across the community, matching need to expertise
- Other ideas?

Still some open questions:
- Can we effectively and consistently engage with AI foundational experts?
- How do we make this happen in practice, at the least cost?

Industry and other engagements

- Related to the previous KE aspect
- Current status of engagement healthy, but also reliant on hit&miss programmes (e.g. studentships, projects)
- **Industry leads, we follow** … perhaps not the best and most effective way
- Questions we should try to answer:

- Advanced Tools/Algorithms Developed by Industry
  - pyTorch: Torch @ Idiap Research Institute (EPFL) -> pyTorch @ META -> Linue foundation
  - DeepMind plasma control w/ RL EPFL -> COSY injection optimised
  - **FastML open question – do we hit their level of complexity?**
- Computing Technology and Infrastructure
  - Cloud Computing Services - CERN utilizes Azure for scalable computing resources and Fermilab uses AWS (neutrino program?)
  - NVIDIA blog on LHC physics
- Quantum computing i.e. IBM Quantum Computing and CERN
- Data Challenges:Higgs boson, TrackML, CMS collisions
- Industry projects internship – not related to research, skills based

- Why should industry care about us?
  - Our research, more Higgs bosons, has zero street value
    - Honorary titles have some value for some poeple
  - Time is money and industry does not waste time (to first order)
- When are we useful?
  - **Our skills are in need – let me tell you about how I know.**
  - **Opportunity to upskill themselves**
  - Interests can allign, but difficult to find. Hard requirement.

Examples made:
- Projects through CDTs
- Undergraduate projects
- Placements at industry

Very good feedback on contributions from us overall, but is that enough?

Industry and other engagements (2)

What are the challenges/barriers that we are facing in this area?

- **Research interest?** Curosity driven, not business driven (not enough)
- **Spreading their infulence?** Some platform companies will try hard to get you to use their platforms (esp new quantum companies).
  - Some will offer personal to help with problems if commercial gains are high enough to warrant their time (via payout from IP)
- **Lasting links?** Industry is dynamic. People move and situations change i.e. economic downturns
- **Generating income?**
- **Paperwork?** Yes! University paperwork and company paperwork.
- After have project, main challenge: demonstrating enough value that cost of ramp-up time is worth the effort.

What are the opportunities in this area for HEP

- **It is easier to make projects with no requirement on alignment**
  - More alignment, higher bar to satisfy due to needs in HEP
- **Enhanced outcomes**: yes, for student careers
- **Access to new tools/techniques/hardware**: case dependent (quantum?)
- **Co-creation of new tools**: maybe, see alignment above
- **Increased funding opportunities**:
  - I think so – 1st year PhDs have value
  - "Longer in academia, harder to break bad habits"
- **Knowledge exchange**: likely (but useful?)
- **Cultural exchange**: likely and useful!
- **Impact**: yes

More questions

- What is the most important gain from industry?
- How do we ensure we capture that?
- How do we build upon these examples and grow community-wide, to everyone can benefit?
- Can we share relationships?
- Can we get to the a level where industry comes looking for us?

My two cents: It is clearly a challenge to find sustained pattern of collaborations with obvious return from industry to research,…

Software, hardware and ML-Ops

- Good collection of information from large part of the community – LHC experiments, DUNE, LZ, Mu3e. Perhaps more aspects to be considered
- Hardware Trends, utilization models, software evolution (PyTorch), and Practices

- **Training**
  - Universal adoption of GPUs for model training across high-energy physics (HEP) experiments.
  - Some use of specialised accelerators like IPUs (DUNE, LHCb)
- **Inference**
  - CPU still predominant, but clear shift from CPU-based inference to accelerators like GPUs, FPGAs, and IPUs, e.g., CMS rewriting software trigger code to run on heterogeneous computing
  - Use of accelerators depends on specific use cases and availability, with some experiments exploring GPUs for offline inference and FPGAs for real-time processing.

- **Traditional and Advanced Models**
  - Boosted Decision Trees (BDTs), Deep Neural Networks (DNNs), Graph Neural Networks (GNNs), Convolutional Neural Networks (CNNs), including Sparse CNNs.
  - Bayesian Neural Networks (BNNs), Point Clouds, HyperGraphs.
- **Transformer Architectures**
  - Transformers (mainy the attention part, but also tokenization) and Large Language Models (LLMs) are being explored (e.g., LIPS)

- **Experiment Tracking and Reproducibility:**
  - Use of tools like Comet (ATLAS) for tracking experiments allows for reproducibility and sharing of results; CMS runs trainings through the Spotify luigi-based workflow management but also sees need for preservation of trainings and metadata for reproducible results (e.g. Comet, Hugging Face?)
- **Deployment Challenges**
  - Common practice involves manual copying of model files (e.g., ONNX files) to deployment directories.
  - Highlights a need for improved, centralized deployment tools to streamline processes.
- **Adoption of Industry Practices**
  - Opportunity to integrate industry-standard ML-Ops tools and practices to improve efficiency.
  - Emphasis on automating deployment pipelines and utilizing advanced tracking systems for better maintenance and scalability.

Challenges and barriers

Training

- Reliance on personal/university/national GPUs (institute HPC, AWS accounts, perlmutter)
- Some centralized GPUs exist (e.g. CERN lxplus-gpu, SWAN)
  - Not enough to satisfy demand, maybe have higher barrier for entry than local resources
  - In general not suitable for extensive development and heavy training jobs
- Resource management is tricky: need to ensure fair and efficient use of hardware, interactive jobs vs queues for longer jobs

Deployment

- Generally models are trained using Python and then serialised/exported for use in a C++ environment
- This process is difficult, time consuming and prone to errors (e.g. interfacing with experiment data models)
- Many different export libraries available, lwtnn, libtorch, ONNX → currently the most common
- Lack of feature parity between Python/C++ implementations (notably with ONNX), e.g. serialisation of custom operators
- Potential solution: Inference-as-a-Service (e.g., LBNL/CMS Triton inference service called SONIC)
  - Example of using detector reconstruction hits as input to a torch_geometric model deployed in NVidia Triton
  - Example of Triton image with torch_geometric: https://hub.docker.com/r/fastml/triton-torchgeo
  - Example from DUNE / ProtoDUNE-SP using SONIC: https://doi.org/10.3389/fdata.2020.604083

Library and Compatibility Issues

- Once models are deployed in production, updating library versions can be difficult
  - e.g. deployed model breaks with library update, reproducibility not there to fix
- This leading to dependency on outdated library versions and prevent the deployment of newer models
  - e.g. DUNE with LibTorch 1.4

Challenges and barriers (2)

Reproducibility and Robustness

- Challenges in ensuring reproducibility of trainings, and allowing for maintenance of deployed models
- Little use of industry-standard tools (MLFlow, Avalanche), "deploy and forget"

Software Maintenance

- General code standards can be lower than industry, lack of documentation and standardised tooling
- Code can break with updates in Python libraries; slow troubleshooting
- Need for robust CI/CD pipelines (e.g. Salt with 90% test coverage)
  - CERN's GPU-enabled GitLab CI runners could enhance reproducibility and coverage

Knowledge and Skills

- Increasing reliance of methods developed by industry, rather than homegrown development of methods
- Need to ensure ML talent is nurtured and retained in the field
  - Tutorials, skills building workshops, documentation all essential

But also opportunities ..

Enhanced Physics Performance:

- Significant performance gains with ML observed across experiments, from BDTs for the Higgs discovery to GenAI to accelerate simulations for future highly granular detectors
- Potential for improved physics outcomes and increased efficiency across the field

Inter-Experiment Collaboration:

- Shared tooling, hardware resources, and training frameworks among experiments
- Standardized deployment approaches while supporting ongoing R&D
- Collaborate on Foundation Models?

Centralized/Democratized Hardware Access:

- Proposal for experiments to allocate budget portions to ML compute resources
- Mitigate the "institute lottery" by better utilising national resources
- Create a central website to advertise national computing resources and how to use them

Standardized Tooling and Frameworks:

- Either expand homebrew tools like Salt (ATLAS) or b-hive (CMS) to other experiments or move more to industry solutions? - (we should not "reinvent the wheel")
- Release models (automatically?) and datasets on platforms like HuggingFace, HEPData, and Zenodo
- Incorporate industry knowledge and best practices into HEP workflows

Public Challenges and Open Datasets:

- Promote HEP as an ideal testbed for cutting-edge ML (e.g., trackML challenge)
- Develop **standardized datasets and models** accessible to all experiments

Cost-Effectiveness and Funding Opportunities:

- Highlight the cost benefits of using ML in achieving physics goals
- Prepare compelling pitches for funding calls and compute cluster allocations (e.g. Isambard AI, Dawn, DiRAC)
- Leverage success stories to secure digital research infrastructure funds

Many of these resonating also with other discussion presented before

Summary

- Overall, very good brainstorming around topics which are crucial to ensure development, deployment and exploitation of AI for HEP
- Training is crucial at all level – investing on this is important
- KE within HEP plus capacity building are key, as well as collaboration with industry, the latter perhaps not as tailored toward HEP as we would like
- Sharing of tools and identification of funding streams that can be used to support are key for a successful strategy – ie engagement with STFC to implement their strategy is key
- A lot to discuss still, so a document is now in its final stage of preparation for the UK community, plus examples of AI in HEP have been collected for the AI Strategy document by STFC
  - That is also in its final stage of preparation and should be distributed soon