MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™ Part of the MACHINE LEARNING Series

ACCEPTED MANUSCRIPT • OPEN ACCESS

Roadmap on fast machine learning for science

To cite this article before publication: Sioni Summers *et al* 2026 *Mach. Learn.: Sci. Technol.* in press https://doi.org/10.1088/2632-2153/ae484b

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2026 The Author(s). Published by IOP Publishing Ltd**.

As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 4.0 licence, this Accepted Manuscript is available for reuse under a CC BY 4.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by/4.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 117.133.133.86 on 25/02/2026 at 03:41

Roadmap

Roadmap on Fast Machine Learning for Science

Sioni Summers[1,11,12], Alex Tapper[2,11,12], Thea Klæboe Årrestad[3], Chen Qin[4,] Karin Rathsman[5], Matthew Streeter[6], Charlotte Palmer[6], Jonathan Citrin[7], Changgang Zheng[8], Noa Zilberman[8], Alexander Titterton[9], Tobias Becker[10]

[1] *CERN, 1, Esplanade des Particules, Meyrin, Switzerland*
[2] *Deprtment of Physics, Imperial College London, London, United Kingdom*
[3] *ETH Zürich, Otto-Stern-Weg 5, Zürich, Switzerland*
[4] *I-X and Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom*
[5] *European Spallation Source ERIC, Lund, Sweden*
[6] *School of Mathematics and Physics, Queen's University Belfast, Belfast, United Kingdom*
[7] *Google DeepMind, London, United Kingdom*
[8] *Department of Engineering Science, University of Oxford, Oxford, United Kingdom*
[9] *Graphcore, 11-19 Wine Street, Bristol, BS1 2PH, United Kingdom*
[10] *Maxeler Technologies, a Groq company, 3 Hammersmith Grove, London, W6 0ND, United Kingdom*

[11] Guest editors of the Roadmap.
[12]Authors to whom any correspondence should be addressed.

E-mails: sioni@cern.ch; a.tapper@imperial.ac.uk

Abstract

The need for microsecond speed Machine Learning (ML) inference for particle physics experiments has emerged in recent years, in particular for the forthcoming upgrades to the experiments at the Large Hadron Collider at CERN. A community has grown around the need to develop the custom hardware platforms and tools required. The material presented in this report is drawn from the latest workshop held by the Fast ML for Science community and comprises of a collection of perspectives on the status of Fast ML in different scientific domains, and the supporting technology.

Keywords: machine learning for science, big data, particle physics, codesign, coprocessors, heterogeneous computing, fast machine learning, large language models

Contents

Accepted Manuscript

1 - Introduction

Sioni Summers, *CERN, 1, Esplanade des Particules, Meyrin, Switzerland*
sioni@cern.ch
Alex Tapper, *Imperial College London, London, United Kingdom*
a.tapper@imperial.ac.uk

The Fast Machine Learning for Science workshop was hosted by Imperial College London, from September 25th to 28th, 2023, the fourth edition of the workshop.

The genesis of the workshop series has been the need for microsecond speed inference for the High-Luminosity LHC detectors, in particular in the hardware trigger systems of the ATLAS and CMS experiments. This level of speed requires non-standard and generally custom hardware platforms, which are traditionally very challenging to program. In addition, while machine learning is becoming widespread in society, this ultrafast niche is not well served by commercial tools. Consequently the field of particle physics has developed tools and techniques and a community of people in this area.

The workshop brought together almost 200 scientists and engineers in a hybrid format, to discuss the latest developments in fast machine learning. Participation from students, including some undergraduates, and early career researchers was a particular feature of the workshop, alongside representation from key industry partners. A strong aim of the conference was to engage scientific communities outside particle physics and develop areas where the tools and techniques from particle physics can be game-changing for other scientific fields.

The interdisciplinary nature of the workshop, including fields such as particle physics, free electron lasers, nuclear fusion, astrophysics, computer science and biology, made for a varied and interesting agenda, with much to discuss in coffee breaks and elsewhere. In addition to particle physics, the attendees heard talks on how fast machine learning is being harnessed to speed up identification of gravitational waves, to improve multi-messenger astronomy, how in free electron laser experiments, machine learning is needed to handle high data rates and due to the fast turn around of experiments, how flexible and generic machine learning features are necessary to allow for optimal use of the limited time at the facilities. Speakers covered medical uses for fast machine learning, through the need for fast image processing and data analysis for diagnosis and treatment and topics in biology searching for known and unknown features in large, heterogeneous datasets, for example common features in different types of tissues. The use of machine learning in control systems and simulations was discussed in the context of laser accelerators and in nuclear fusion experiments. Even theoretical physics featured through the application of machine learning to solve the electron wave equation in condensed matter, working towards a detailed and fundamental understanding of superconductivity.

Key industry partners, including AMD, Graphcore, Groq and Intel contributed to the workshop, discussing current and future generation hardware platforms and architectures, and running tutorials on using their development toolchains. Groq and Graphcore presented their latest dedicated chips for artificial intelligence applications, such as transformers and graph-based processing, and showed they have interesting applications to science problems. AMD and Intel highlighted the flexibility of their FPGA platforms and explained how to optimise these for scientific machine learning applications.

Accepted Manuscript

Roadmap on Fast ML for Science

The following articles are a collection of perspectives on the status of Fast ML in different scientific domains, and the supporting technology.

Accepted Manuscript

2 - Fast Machine Learning at the Large Hadron Collider experiments

Thea Klæboe Årrestad, ETH Zürich, Otto-Stern-Weg 5, Zürich, Switzerland
Sioni Paris Summers, CERN, 1, Esplanade des Particules, Meyrin, Switzerland
thea.aarrestad@cern.ch, sioni.summers@cern.ch

Status

At the CERN Large Hadron Collider (LHC), bunches of trillions of protons are brought to collide in the center of the four particle detectors around the LHC ring: ATLAS, CMS, ALICE and LHCb. This generates showers of new outgoing particles that produce signals in the particle detectors located around the interaction point. Due to the high frequency of collisions, every 25 ns, and large number of read-out channels, $O(10)$ million, an enormous amount of data is generated. For the two hermetic general-purpose experiments, CMS and ATLAS, this is $O(100)$Tb/s. These signals are read out using custom electronics mounted on support structures inside the detectors. This large data rate can not be read out and stored as it would require a significant amount of electronics and power supply inside the detector, obstructing the path of generated particles and reducing the detector sensitivity. Rather, the data rate is reduced by a two-stage filtering system, the trigger. While data is buffered inside the detector on detector frontend electronics, a subset of the detector information is sent through high speed optical links to a radiation-shielded cavern located next to the detector. Here, the first stage of the event filtering system, the Level-1 trigger (L1T), is responsible for reducing the data rate by two orders of magnitude within the buffering time of $O(1)$ μs. Due to the strict latency constraints, this processing is done fully in firmware on $O(1000)$ field programmable gate arrays (FPGAs). The complexity of the data requires hundreds of reconstruction algorithms to run in parallel on several collision events simultaneously, which requires each algorithm to use minimal resources and time. The extreme latency and bandwidth requirements for the L1T system is unique and requires novel solutions in terms of IO and algorithms, as illustrated in Figure 1. Due to their ability to efficiently and accurately process and parametrise high dimensional input, Machine Learning (ML) algorithms are increasingly being explored as faster and better replacements of the classical algorithms currently in use in the L1T. The extremely low latency and resource requirements for ML algorithms deployed in the L1T system has led to the development of dedicated software libraries that facilitates the training and translation of ML models into efficient FPGA firmware; hls4ml [2] for deep neural networks and

Accepted Manuscript

Conifer [3] for decision trees. This has allowed for the first deep neural networks to be integrated into the L1T system, running inference within a few tens of nanoseconds.

Current and Future Challenges

With the rapid progress of hls4ml and Conifer, deep neural network and decision tree ensemble sub-microsecond inference in L1T systems is a reality. However, there are significant challenges ahead. In order to study increasingly rare physics processes in the LHC detectors, the LHC will be upgraded to its High Luminosity phase (HL-LHC), allowing the experiments to collect a factor of ten more data. This entails a substantial trade-off: The number of concurrent proton collisions will triple, leading to a notably increased event complexity, as illustrated in Figure 2. To cope with this, more granular detectors will be installed, demanding even more complex reconstruction algorithms due to the large increase in read-out channels. For the first time, the CMS L1T will also receive information from the inner tracking system. This will require the reconstruction of hundreds of charged particle tracks from thousands of detector hits, within a maximum latency of 5 μs. The input data rate to the L1T systems will increase by one order of magnitude, corresponding to around 5% of the total internet traffic [4]. This necessitates the design and implementation of significantly more advanced ML algorithms for the reconstruction, correction and selection of proton collision events.

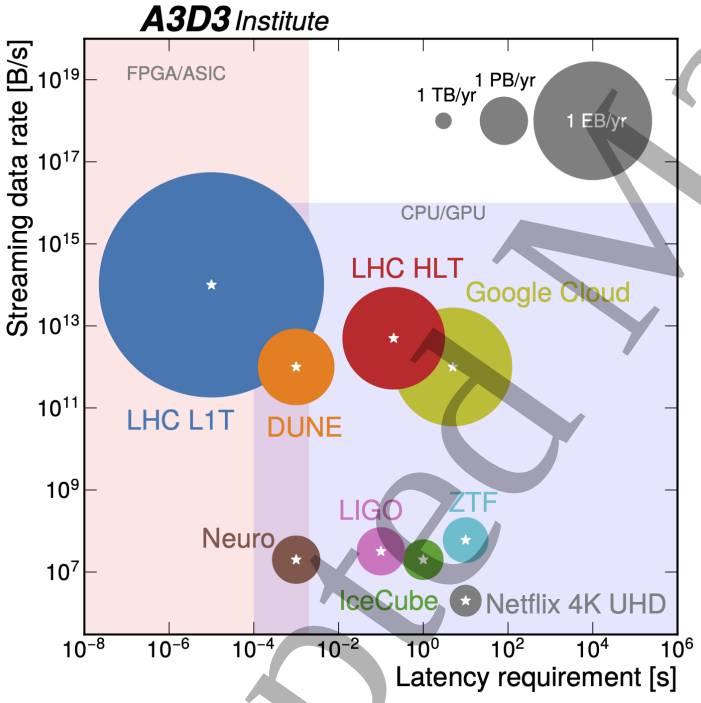


Figure 1. Typical data rate versus latency requirement of different science and industry projects. The size of each circle is proportional to the amount of data that is processed by the given system each year. The LHC L1T systems is shown in dark blue [1].

Advances in Science and Technology to Meet Challenges

In order to meet the challenges imposed by HL-LHC, there is a strong ongoing R&D effort within the particle physics community towards the development of powerful neural networks running on specialized hardware like ASICs, FPGAs or AI accelerators. Owing to the point cloud characteristics of particle physics data, which involve sparse and unordered hits within irregular geometry detectors, graph neural networks (GNNs) have become the preferred architecture for most reconstruction tasks faced by the HL-LHC. Using an interaction network architecture, a type of GNN, ML-based charged particle tracking on FPGAs has been successfully demonstrated with $O(100)$ ns latency [5,6]. For primary vertex finding and association of tracks to the vertex, end-to-end ML solutions are also being explored [7]. Additionally, GNNs on FPGAs have been successfully demonstrated for energy reconstruction for highly granular calorimeters [8]. Direct access to the particles will for the first time enable the grouping of particles into jets at L1T [9]. This has led to a significant effort focused on using particle information

Accepted Manuscript

to discern the flavor of the jet using ML, specifically with GNN-like architectures such as interaction networks [10] and transformers [11] or with set-based architectures [9]. In order to expand the reach of trigger systems beyond simple rule-based event selections, there has been a recent adoption of unsupervised learning for anomaly detection in the hardware trigger at both ATLAS and CMS [12][13]. Besides the deep neural networks mentioned earlier, decision tree ensembles are also extensively used in the L1T owing to their superior performance on tabular data and their low latency and resource requirements. These are not employed for reconstruction tasks, typically focused on mapping an input set (like a group of detector hits) to another set (such as reconstructed tracks). Instead, they are used for tasks such as regression [14], classification [15], and anomaly detection. To account for fill-level changes in detector conditions, techniques like Continual Learning (CL) are being explored in order to maintain model stability in the resource-constrained, low-latency setting of the L1T [16].

Finally, exploration is also underway into the use of machine learning in the front end electronics of detectors. This includes on-sensor data filtering using neuromorphic computing-based spiking neural networks [17], on-pixel track seeding [18] and data compression in detector front end electronics [19][20].

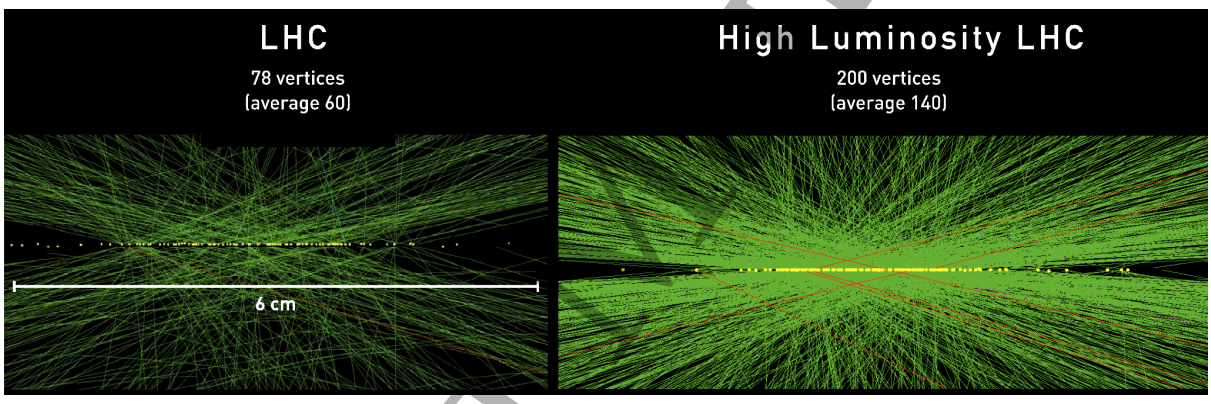


Figure 2. The number of simultaneous proton-proton collisions will increase by a factor of three as the LHC is upgraded to its high luminosity phase, HL-LHC. The data complexity will increase correspondingly.

Concluding Remarks

In conclusion, the integration of Machine Learning algorithms into the Level-1 Trigger system of the LHC experiments marks a significant advancement in particle physics data processing. ML, particularly deep neural networks and decision trees, has enabled ultra-low latency and efficient data handling, making real-time event selection and reconstruction feasible. This achievement is exemplified by the hls4ml and Conifer libraries, which facilitate the deployment of ML models on FPGAs with sub-microsecond inference times. However, as the LHC transitions to its High Luminosity phase (HL-LHC) to collect even more data, new challenges emerge. The increased number of proton collisions and the introduction of more granular detectors demand advanced ML algorithms to handle the higher event complexity. The community is responding with ongoing research and development, focusing on powerful neural networks, particularly graph neural networks (GNNs), for tasks like particle tracking, energy reconstruction, and jet classification. Additionally, the adoption of ML extends beyond the trigger system and into the front-end electronics of detectors, enabling innovations such as neuromorphic computing-based data filtering and on-pixel track seeding. The pursuit of continual

Accepted Manuscript

learning techniques ensures model stability in the resource-constrained, low-latency environment of the L1T. The fusion of ML and particle physics is not just a technological leap but a testament to the collaborative efforts of scientists and engineers working towards a common goal of gaining a better understanding of our universe while taking advantage of cutting-edge technologies to meet the evolving challenges of particle physics experiments at the HL-LHC.

Acknowledgements

T.Å. is supported by the Swiss National Science Foundation Grant No.PZ00P2_201594.

References

[1] A3D3 Institute, 'AI to accelerate scientific discovery,' A3D3, 2024. [Online]. Available: https://a3d3.ai/about/. [Accessed: Jan. 10, 2024].

[2] J. Duarte, S. Han, P. Harris, S. Jindariani, E. Kreinar, B. Kreis, J. Ngadiuba, M. Pierini, R. Rivera, N. Tran, and Z. Wu, "Fast inference of deep neural networks in FPGAs for particle physics," JINST, vol. 13, no. 07, p. P07027, 2018, doi: 10.1088/1748-0221/13/07/P07027.

[3] S. Summers, G. Di Guglielmo, J. Duarte, P. Harris, D. Hoang, S. Jindariani, E. Kreinar, V. Loncar, J. Ngadiuba, M. Pierini, D. Rankin, N. Tran, and Z. Wu, "Fast inference of Boosted Decision Trees in FPGAs for particle physics," Journal of Instrumentation, vol. 15, no. 05, pp. P05026-P05026, May 2020. doi: 10.1088/1748-0221/15/05/p05026.

[4] CMS Collaboration, "The Phase-2 Upgrade of the CMS Level-1 Trigger," CERN, Geneva, Tech. Rep. CERN-LHCC-2020-004, CMS-TDR-021, 2020. [Online]. Available: https://cds.cern.ch/record/2714892.

[5] G. DeZoort, S. Thais, J. Duarte, V. Razavimaleki, M. Atkinson, I. Ojalvo, M. Neubauer, and P. Elmer, Charged Particle Tracking via Edge-Classifying Interaction Networks. Comput Softw Big Sci 5, 26 (2021). https://doi.org/10.1007/s41781-021-00073-z

[6] Elabd A, Razavimaleki V, Huang SY, Duarte J, Atkinson M, DeZoort G, Elmer P, Hauck S, Hu JX, Hsu SC, Lai BC, Neubauer M, Ojalvo I, Thais S, Trahms M. Graph Neural Networks for Charged Particle Tracking on FPGAs. Front Big Data. 2022 Mar 23;5:828666. doi: 10.3389/fdata.2022.828666. PMID: 35402906; PMCID: PMC8984615.

[7] C. E. Brown, A. Bundock, M. Komm, V. Loncar, M. Pierini, B. C. Radburn-Smith, A. Shtipliyski, S. P. Summers, J.-S. Dancu, and A. Tapper, "Neural Network-Based Primary Vertex Reconstruction with FPGAs for the Upgrade of the CMS Level-1 Trigger System," Tech. Rep. CMS-CR-2022-018, CERN, Geneva, 2023. [Online]. Available: https://cds.cern.ch/record/2801638. doi: 10.1088/1742-6596/2438/1/012106.

[8] Y. Iiyama, G. Cerminara, A. Gupta, J. Kieseler, V. Loncar, M. Pierini, S. Rukh Qasim, M. Rieger, S. Summers, G. Van Onsem, K. A. Wozniak, J. Ngadiuba, G. Di Guglielmo, J. Duarte, P. Harris, D. Rankin, S. Jindariani, M. Liu, K. Pedro, N. Tran, E. Kreinar, and Z. Wu, "Distance-Weighted Graph Neural Networks on FPGAs for Real-Time Particle Reconstruction in High Energy Physics," Frontiers in Big Data, vol. 3, 2021, doi: 10.3389/fdata.2020.598927. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2020.598927.

Accepted Manuscript

[9] S. Schaefer, C. Brown, D. Hoang, S. Summers, and S. Wuchterl, “Advancing the CMS Level-1 Trigger: Jet Tagging with DeepSets at the HL-LHC,” 2025. [Online]. Available: arXiv:2509.24371.

[10] Z. Que, M. Loo, H. Fan, M. Pierini, A. Tapper, and W. Luk, "Optimizing Graph Neural Networks for Jet Tagging in Particle Physics on FPGAs," in Proc. of the 32nd International Conference on Field-Programmable Logic and Applications, Aug. 2022, doi: 10.1109/FPL57034.2022.00057.

[11] F. Wojcicki, Z. Que, A. D. Tapper, and W. Luk, "Accelerating Transformer Neural Networks on FPGAs for High Energy Physics Experiments," in Proc. of the 2022 IEEE International Conference on Field-Programmable Technology (ICFPT), Dec. 2022, doi: 10.1109/ICFPT56656.2022.9974463.

[12] A. Gandrakota, “Real-time Anomaly Detection at the L1 Trigger of CMS Experiment,” 2024. Available: arXiv:2411.19506

[13] M. M. Cohen et al., ATLAS Collaboration, “GELATO: a Generic Event-Level Anomalous Trigger Option for ATLAS – Slides,” 2025. Available: https://cds.cern.ch/record/2938881

[14] CMS Collaboration, 'Electron Reconstruction and Identification in the CMS Phase-2 Level-1 Trigger,' 2023. [Online]. Available: https://cds.cern.ch/record/2868782.

[15] D. Reikher, “Leveraging Machine Learning for Enhanced FPGA- Based Tau Triggering and Combined H→(bb/cc) Analysis in the ATLAS Experiment,” 2025. doi: 10.17181/93sta-ptr08.

[16] CMS Collaboration, "Continual Learning in the CMS Phase-2 Level-1 Trigger," 2023. [Online]. Available: https://cds.cern.ch/record/2859651.

[17] S. R. Kulkarni, A. Young, P. Date, N. R. Miniskar, J. S. Vetter, F. Fahim, B. Parpillon, J. Dickinson, N. Tran, J. Yoo, C. Mills, M. Swartz, P. Maksimovic, C. D. Schuman, and A. Bean, "On-Sensor Data Filtering using Neuromorphic Computing for High Energy Physics Experiments," 2023. [Online]. Available: arXiv:2307.11242 [cs.NE].

[18] J. Dickinson, R. Kovach-Fuentes, L. Gray, M. Swartz, G. Di Guglielmo, A. Bean, D. Berry, M. B. Valentin, K. DiPetrillo, F. Fahim, et al., "Smartpixels: Towards on-sensor inference of charged particle track parameters and uncertainties," 2023. [Online]. Available: arXiv:2312.11676 [hep-ex].

[19] G. Di Guglielmo, F. Fahim, C. Herwig, M. B. Valentin, J. Duarte, C. Gingu, P. Harris, J. Hirschauer, M. Kwok, V. Loncar, Y. Luo, L. Miranda, J. Ngadiuba, D. Noonan, S. Ogrenci-Memik, M. Pierini, S. Summers, and N. Tran, "A Reconfigurable Neural Network ASIC for Detector Front-End Data Compression at the HL-LHC," in IEEE Transactions on Nuclear Science, vol. 68, no. 8, pp. 2179-2186, 2021, doi: 10.1109/TNS.2021.3087100.

[20] A. Yue, H. Jia, and J. Gonski, “Variational autoencoders for at-source data reduction and anomaly detection in high energy particle detectors,” MLST, 2025, doi: 10.1088/2632-2153/adf0c0

Accepted Manuscript

3 – Deep learning for fast MR imaging and analysis

Chen Qin, I-X and Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, UK
[c.qin15@imperial.ac.uk]

Status

Magnetic resonance imaging (MRI) is a leading diagnostic modality for a wide range of exams and is one of the most useful and important imaging techniques in hospitals, due to its lack of ionising radiation and ability to probe various aspects of the physiology. MRI has been widely applied in neuroimaging, cardiovascular and musculoskeletal systems, and more than 50,000 scanners are estimated to be in use worldwide [1]. A typical MRI workflow ranges from patient scheduling and acquisition to image analysis and prognosis (Figure 1) [2]. Conventionally, the MRI workflow can be very slow due to both the fundamentally limited acquisition speed of MRI and the laborious manual labelling for human analysis and interpretation. Recently, deep learning (DL) techniques have opened the possibility to accelerate this considerably, where they can impact all aspects of the MRI workflow. For instance, in cardiovascular MR (CMR) imaging, several acquisition-related aspects of CMR examination such as the automatic view planning can be automated or substantially shortened using DL [2,3]. For MRI acquisition and reconstruction, DL has also been emerging as a popular and powerful alternative to the generic compressed sensing techniques. State-of-the-art DL-based methods have proposed to leverage routinely performed MR scans as training data to exploit prior knowledge as well as incorporating the physically acquired raw data in the reconstruction process [4,5]. They can not only achieve high reconstruction quality but can also offer high efficiency in terms of reconstruction speed, which makes the clinical deployment feasible [2].

DL in MR image analysis and interpretation has also been widely investigated. DL in MRI segmentation has shown great potential in achieving human-level performance while significantly speeding up the annotation process [6]. For instance, a DL method [6] only takes 2.2 seconds to analyse images of a single subject at two time points of the cardiac cycle, whereas it typically takes a trained expert 20 minutes. DL also enables tasks that are impossible for a human to label, such as estimating dense correspondences between images or tracking the motion in sequences. They also offer much faster inference speed (more than 10x) compared to conventional optimisation-based approaches [7,8]. According to the statistics in [9], automated image interpretation using DL has the potential to reduce the time radiologists spend on reviewing images by 20%.

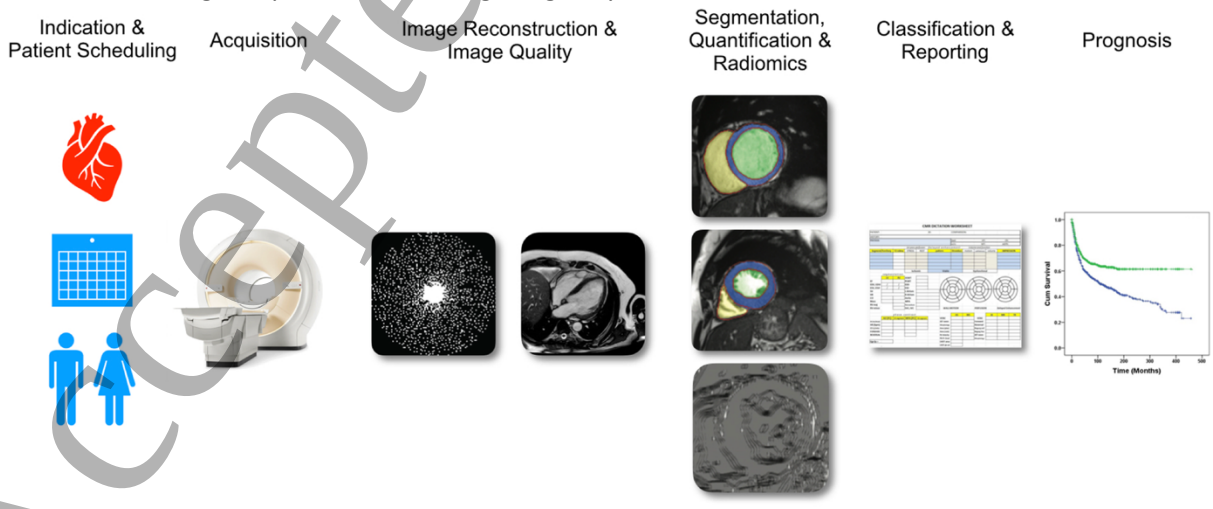


Figure 1. A cardiovascular magnetic resonance imaging workflow from patient scheduling to image analysis and prognosis [2].

Accepted Manuscript

Current and Future Challenges

Despite the remarkable performance and the tremendous promise that DL has shown in the field, there are several challenges and limitations that need to be mentioned. One of the major challenges is the limited data availability. DL methods are known to be data-hungry techniques, which rely heavily on the availability of training data. However, medical data is often expensive to obtain and the labelling of them can also be imperfect. Most current datasets for MRI images are typically in the range of a few hundreds, which may lead to a high risk of DL models being overtrained on the training data, and moreover, most likely there are more healthy subjects than diseased patients in the datasets, resulting in highly imbalanced classes. In addition, the sharing of data between institutions has also been challenging due to privacy regulations.

A relevant challenge that is related with the data limitations is the model bias and generalisation. Distribution shifts across sites, scanners, and acquisition protocols can significantly reduce model performance on unseen data [10]. Although there are some large population datasets available in recent years such as UK Biobank [11], there also exist biases towards groups with certain demographics. DL models trained on those can also lead to model bias towards certain groups of subjects and therefore result in concerns of fairness issues in the domain.

A further major challenge that currently limits the translational potential of such DL models in clinical practice is the trustworthiness [12], which encompasses model transparency, explainability, robustness, and safety. Although DL methods can achieve promising performance, there is often a lack of consideration of their reliability, e.g., they may show instabilities and lack of generalisability when encountered with perturbations and new domains [13]. On the other hand, the 'black-box' nature of DL approaches also often leads to the decision-making process being opaque, whereas it is crucial to make informed decisions in the safety-critical healthcare applications.

Computational scalability also represents a practical barrier. High-resolution volumetric imaging and multi-modal datasets demand substantial computational resources, making traditional pipelines slow and often impractical. Fast machine learning approaches can address this challenge, enabling efficient training and real-time inference, which are critical for deploying DL models in time-sensitive clinical workflows.

Advances in Science and Technology to Meet Challenges

Learning with limited or imperfect labels remains a challenge. Supervised learning with manual annotations is often constrained by that, and therefore methods that are beyond supervised learning are needed to tackle the challenge. For instance, self-supervised learning approaches [14] to learn useful representations without any human intervention can be investigated, such as via exploring the intrinsic structure within data itself or via contrastive learning to learn general features of datasets without labels. Unsupervised generative modelling is also another opportunity, where the generative power of DL models such as generative adversarial networks or diffusion models [15] can be leveraged to generate synthetic labelled datasets for training. The self-supervised or unsupervised learnt representations are an effective way of exploiting unlabelled data and can be potentially leveraged to transfer knowledge to relevant new tasks where labelled data is sparse. Recent advancement of foundation models is a successful case of that, where models are trained in a self-supervised fashion on broad data and then be adapted to a wide range of use cases. There have also been explorations

Accepted Manuscript

of foundation models in MRI [16], which mitigate the need of large data collection and labelling for each specific task. Such foundation models also have the potential to address the model generalizability and imbalanced dataset issues due to the large-scale datasets that are used. [16] To protect data privacy while sharing the multi-cohort knowledge, federated learning [17] can be leveraged to enable distributed learning across multiple institutions without central data sharing.
DL trustworthiness in medical imaging especially in MRI computing is also an important topic to tackle. It is necessary to equip the DL models with the ability of knowing what they do not know and inform the potential risks of failures. To achieve this, uncertainty modelling [18] can serve as a mechanism to communicate the knowledge boundary of DL systems and provide an additional reference for visualising and interpreting prediction reliability. The quantified uncertainty could also be leveraged to handle cases outside distribution and improve model confidence and robustness, potentially through test-time adaptation and uncertainty reduction techniques. Furthermore, DL trustworthiness in fast MR imaging and analysis could also be enhanced by improving model explainability either by design or using external techniques [19], such as with attention-based mechanisms or prototype learning methods. Causal analysis [20] is also a promising emerging field that can provide more insights within models and strengthen their clinical translation potential. Fast machine learning approaches can also complement these advances by making high-resolution, uncertainty-aware, and interpretable analyses computationally feasible. Efficient architectures, model compression, and hardware-optimized pipelines ensure that DL models can operate in real-time, enabling both clinical scalability and reliable deployment.

Concluding Remarks

This section provides a brief introduction on the recent advancement of deep learning (DL) technologies for fast MR imaging and analysis. DL has the potential to improve the entire MR imaging workflow and has been shown to have greatly improved the efficiency of radiology workflow, allowing time for clinicians to focus more on patients' care. The section also identifies a few current and future research challenges that are limiting the DL development in medical imaging and hindering their clinical translation potential, which consist of issues related with data scarcity, model generalizability and model trustworthiness. Recent advancement in DL technologies that can meet the above challenges have also been discussed, including self-supervised/unsupervised learning approaches, foundation models and uncertainty modelling. The development of DL in medical imaging will continue to rise, and it would be expected to have a widespread impact on the future of fast MR imaging and analysis, ultimately bringing significant benefits to patients and the healthcare industry.

Acknowledgements

C. Qin acknowledges the support by the UK's Engineering and Physical Sciences Research Council (EPSRC) (grants EP/X039277/1 and EP/Y002016/1).

References

[1] Rinck, P. A. (2019). Magnetic resonance in medicine: a critical introduction. The Basic textbook of the European magnetic resonance forum. BoD Book on Demand.

[2] Leiner, T., Rueckert, D., Suinesiaputra, A., Baeßler, B., Nezafat, R., Išgum, I., & Young, A. A. (2019). Machine learning in cardiovascular magnetic resonance: basic concepts and applications. Journal of Cardiovascular Magnetic Resonance, 21, 1-14.

Accepted Manuscript

[3] Blansit, K., Retson, T., Masutani, E., Bahrami, N., & Hsiao, A. (2019). Deep learning–based prescription of cardiac MRI planes. Radiology: Artificial Intelligence, 1(6), e180069.

[4] Qin, C., Schlemper, J., Caballero, J., Price, A. N., Hajnal, J. V., & Rueckert, D. (2018). Convolutional recurrent neural networks for dynamic MR image reconstruction. IEEE transactions on medical imaging, 38(1), 280-290.

[5] Qin, C., Duan, J., Hammernik, K., Schlemper, J., Küstner, T., Botnar, R., Prieto, C., Price, A.N., Hajnal, J.V. and Rueckert, D. (2021). Complementary time-frequency domain networks for dynamic parallel MR image reconstruction. Magnetic Resonance in Medicine, 86(6), 3274-3291.

[6] Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M. and Zemrak, F. (2018). Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance, 20(1), 1-12.

[7] Qin, C., Bai, W., Schlemper, J., Petersen, S. E., Piechnik, S. K., Neubauer, S., & Rueckert, D. (2018). Joint learning of motion estimation and segmentation for cardiac MR image sequences. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11 (pp. 472-480). Springer International Publishing.

[8] Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., & Kamen, A. (2019, May). Unsupervised deformable registration for multi-modal images via disentangled representations. In International Conference on Information Processing in Medical Imaging (pp. 249-261). Cham: Springer International Publishing.

[9] Topol, E. (2019). The topol review. Preparing the healthcare workforce to deliver the digital future, 1-48.

[10] Glocker, B., Robinson, R., Castro, D.C., Dou, Q. and Konukoglu, E. (2019). Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. Medical Imaging meets NeurIPS Workshop.

[11] Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., ... & Allen, N. E. (2020). The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nature communications, 11(1), 2624.

[12] Kondylakis, H., Osuala, R., Puig-Bosch, X., Lazrak, N., Diaz, O., Kushibar, K., Chouvarda, I., Charalambous, S., Starmans, M.P., Colantonio, S. and Tachos, N. (2025). A review of methods for trustworthy AI in medical imaging: The FUTURE-AI Guidelines. *IEEE Journal of Biomedical and Health Informatics.*

[13] Antun, V., Renna, F., Poon, C., Adcock, B., & Hansen, A. C. (2020). On instabilities of deep learning in image reconstruction and the potential costs of AI. Proceedings of the National Academy of Sciences, 117(48), 30088-30095.

Accepted Manuscript

[14] Shurrab, S., & Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Computer Science, 8, e1045.

[15] Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., & Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis, 102846.

[16] Zhang, S., & Metaxas, D. (2023). On the Challenges and Perspectives of Foundation Models for Medical Image Analysis. Medical Image Analysis, 102996.

[17] Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., ... & Braren, R. (2021). End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence, 3(6), 473-484.

[18] Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence, 1(1), 20-23.

[19] Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. Journal of imaging, 6(6), 52.

[20] Vlontzos, A., Rueckert, D., & Kainz, B. (2022). A review of causality for learning algorithms in medical image analysis. Machine Learning for Biomedical Imaging, 2022.

Accepted Manuscript

4 – Fast Machine Learning for Accelerator Controls

Karin Rathsman,
https://orcid.org/0009-0005-0715-8905

karin.rathsman@ess.eu
European Spallation Source ERIC
(Box 176, SE-221 00, Lund)

Status

Accelerator-based research facilities include some of the world's most complex human-made systems. Operating a user facility, such as the European Spallation Source (ESS) in Lund, is a challenge since scientific users will only have access to the facility during short periods in a tight schedule. This implies high availability demands, which for ESS is 95%. Other challenges and limiting factors for accelerator-based research are operational costs for electrical power, staff, maintenance and waste management. Given the recent progress in AI in general, there are high expectations that AI can address these challenges to get more science out of the accelerator-based facilities at reduced costs for the society and the planet.

To illustrate how existing machine learning algorithms can make a difference for automation and fault detection, we can imagine a subsystem with inputs, outputs, feedback loops and a goal for the subsystem to fulfil. If the output of the subsystem can be forecasted from the input using for example recurrent neural networks (RNN), then the subsystem could respond pre-emptively and accurately to changes in the input. If values of measured and predicted output would differ it could be an indication of an issue with the subsystem. This illustration also highlights similarities between control system networks and neural networks, which is not a coincidence since control systems and nervous systems share a similar task, namely, to regulate complex systems.

AI has been used for decades on the experimental side of accelerator-based facilities. Only recently it started to be applied to the accelerators themselves. The first workshop for AI in high-energy and nuclear physics was arranged in 1990 [1], compared to the first workshop for machine learning for accelerators in 2018 [2]. However, AI was the topic of 21 out of the 218 contributions at the international accelerator and experimental physics controls conference ICALEPCS in 2021 [3]. The applications of AI in these articles can mainly be categorized into fault detection, automation, optimization, augmentation, as well as machine learning operation platforms for control systems.

Notable progress in fast machine learning for accelerator controls includes the development of a system to regulate magnetic power supplies at the Fermilab Booster [4]. This system uses reinforcement learning with feed-forward control from an RNN-based surrogate model and it is implemented on an FPGA. The authors of reference [5] designed a fast feedback system based on reinforcement learning to suppress micro-bunch instabilities in the electron beam at KIT. Reference [6] provides an overview of how Bayesian optimizers can be applied online for efficient parameter tuning during real-time beam control, and the authors of reference [7] report a successful beam-tuning experiment at KEK using Bayesian optimization.

Accepted Manuscript

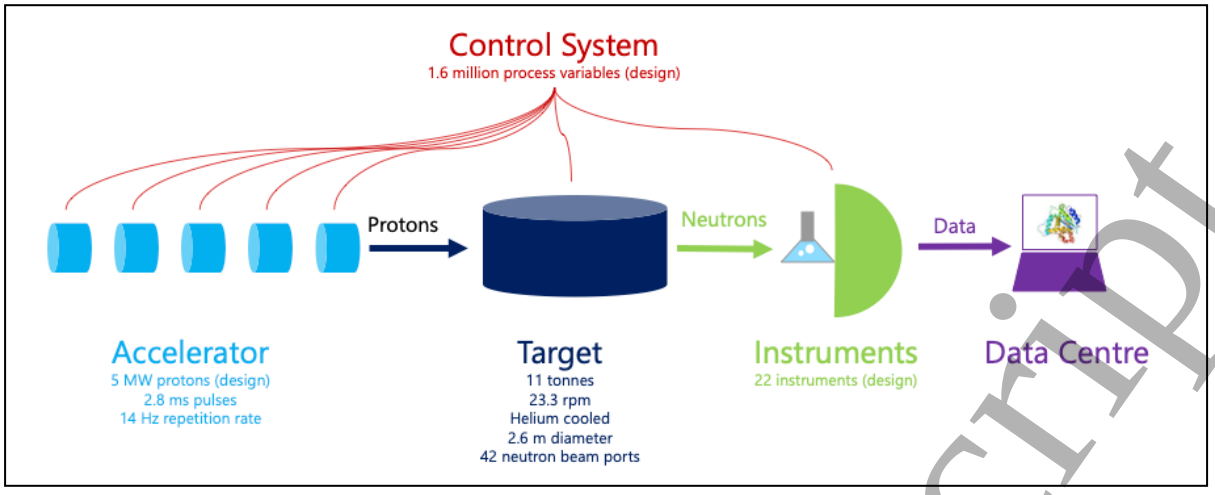


Figure 1. Schematic layout of the ESS machine with design parameters. The accelerator and target are used to create neutron beams, which are used for neutron scattering experiments in a variety of applied science fields.

Current and Future Challenges

Learning dynamics from data is in general straightforward compared to simulating the dynamics, which often require solving partial differential equations with complicated boundary conditions. Surrogate modelling is however a challenge for real-time accelerator applications due to highly nonlinear and fast dynamics, often in the kHz ranges and above.

In general, accelerator controls require fast hardware, such as FPGAs and time resolution below one microsecond. Since conventional PLC-based control systems lack the flexibility, scalability, and temporal precision needed, accelerator facilities develop their control systems in-house or rely on open-source controls communities. These communities have the potential to develop control system to meet the needs imposed by a higher level of automation. However, the engagement is limited. For example, the largest controls community, EPICS [8], currently has fewer than ten core developers. As a result, existing applications for storing and retrieving control system data offer limited functionality, and tools for life-cycle management of machine-learning models within the control system are absent.

It is well known that data quality, i.e., how well data serves its intended purpose, is crucial for machine learning. In accelerator controls, maintaining high data quality is particularly challenging since accelerators are both fast and highly complex, with millions of process variables. The situation is worsened by a tendency to oversample slower process variables [9].

Despite the abundance of data within control system networks, the parameter space can remain sparsely populated due to the high dimensionality. Sparsity can also arise when certain process variables change only infrequently and require longer periods to adequately fill the parameter space. The coexistence of slow and fast processes therefore requires datasets that span long durations while still maintaining a high sampling rate.

High-power accelerators have high damage potential and require failsafe and fast *O(μs)* protection systems that can intervene within microseconds. Deploying fast machine learning models in this environment introduces additional risks, for example that a trained real-time machine learning model might exploit imperfections in the fail-safe protection system and drive the machine in a critical state.

Accepted Manuscript

Finally, machine learning for accelerator controls is foremost a research field in automatic controls and computer science applied to accelerators. However, external collaboration is time-consuming and requires efforts to document and transfer data from protected control systems, which involves both practical and organisational issues.

Advances in Science and Technology to Meet Challenges

Addressing the data quality challenge begins with establishing data-governance practises [9] and adapting a data-centric approach [10] to systematically and continuously improve data quality.

The issue of sparse data is to some extent also a data-quality problem, driven by the large number of process variables. Currently the number of process variables at ESS exceeds 10 million, which is considerably higher than the design value of 1.6 million. This can be partially mitigated by pruning the controls network and constraining operational ranges.

To address challenges in machine learning operations (MLOps) [11], the focus needs to be slightly shifted from machine learning modelling to the full workflow to develop, deploy, monitor and maintain machine learning models. Several initiatives have been started to develop platforms for MLOps, for example at CERN [12].

Accelerator-based research facilities have challenges and data that researchers and students are interested in, and therefore there are strong motivations for collaboration. To make data easily accessible for external parties, accelerator-based research facilities should share control system data [13]. For example, the dataset from Fermilab which was used in reference [4] is publicly available and well documented in reference [14]. Likewise, in reference [15] a publicly available dataset from ESS is described.

Finally, trained machine learning models will not be deployed without safety analysis and they will only be accepted for non-safety critical parts of the facility until they have proven safer than the operators, for example to detect potential faults. However, providing operators with a better understanding of the machine, with improved alarms, graphical representation of dependencies, virtual diagnostics and other augmentation techniques based on machine learning can help them in their daily work to operate the machine, if this information is accurate.

Accepted Manuscript

Figure 2. Control racks at ESS. Some racks are empty since the high energy end of the accelerator, from where the picture is taken, has not yet been installed. The image gives an impression of the size of control systems at accelerator-based research facilities.

Concluding Remarks

Machine learning and other advanced methods will likely change the way accelerators are operated. Indeed, there are mathematical, structural, organisational, educational and safety-critical challenges to take on along the road towards autonomous accelerators. However, Accelerator-based research facilities offer unique environments to study and develop machine learning for process controls.

References

[1] S. Gleyzer, F. Carminati, G. Wonju and D. Perret-Gallix, “The rise of deep learning,” CERNCourir, 9 July 2018. [Online]. Available: https://cerncourier.com/a/the-rise-of-deep-learning/. [Accessed 12 January 2024].

[2] PSI, “2nd ICFA Mini-Workshop on Machine Learning for Charged Particle Accelerators,” Indico, 26 February 2019. [Online]. Available: https://indico.psi.ch/event/6698/. [Accessed 01 2024].

[3] 18th Int. Conf. on Acc. and Large Exp. Physics Control Systems, Geneva: JACoW Publishing, 2022.

[4] J. St. John, C. Herwig, D. Kafkes, J. Mitrevski, W. A. Pellico, G. N. Perdue, A. Quintero-Parra, B. A. Schupbach, K. Seiya, N. Tran, M. Schram, J. M. Duarte, Y. Huang and R. Keller, “Real-time artificial intelligence for accelerator control,” *Phys. Rev. Accel. Beams,* vol. 24, no. 10, p. 104601, 2021.

Accepted Manuscript

[5] T. Boltz, M. Brosi, E. Bründermann, B. Haerer, P. Kaiser, C. Pohl, P. Schreiber, M. Yan, T. Asfour and A.-S. Müller, "Feedback design for control of the micro-bunching instability based on reinforcement learning," in *CERN Yellow Reports: Conference Proceedings*, CERN, Geneva, 2020.

[6] R. Roussel, A. L. Edelen, T. Boltz, D. Kennedy, Z. Zhang, F. Ji, X. Huang, D. Ratner, A. S. Garcia, C. Xu, J. Kaiser, A. F. Pousa, A. Eichler, J. O. Lübsen and N. M. Isenberg, "Bayesian optimization algorithms for accelerator physics," *Phys. Rev. Accel. Beams,* vol. 27, no. 8, p. 084801, 2024.

[7] G. Mitsuka, S. Kato, N. Iida, T. Natsui and M. Satoh, "Machine-learning approach for operating electron beam at KEK electron/positron injector linac," *Phys. Rev. Accel. Beams,* vol. 27, no. 8, p. 084601, 2024.

[8] "EPICS Home page," [Online]. Available: https://epics-controls.org/.

[9] C. Källström, "Data Quality and Quantity for Machine Learning at the European Spallation Source," 2025.

[10] M. H. Jarrahi, A. Memariani and S. Guha, "The Principles of Data-Centric AI," *Association for Computing Machinery,* vol. 66, no. 8, p. 84–92, 2023.

[11] D. Kreuzberger, N. Kühl and S. Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," *IEEE Access,* vol. 11, pp. 31866-31879, 2023.

[12] J.-B. de Martel, R. Gorbonosov and N. Madysa, "Machine Learning Platform: Deploying and Managing Models in the CERN Control System," in *18th Int. Conf. on Acc. and Large Exp. Physics Control Systems*, Shanghai, 2021.

[13] P. Runeson, T. Olsson and J. Linåker, "Open Data Ecosystems — An empirical investigation into an emerging industry collaboration concept," *Journal of Systems and Software,* vol. 182, p. 111088, 2021.

[14] D. Kafkes and J. B. St. John, "A Dataset for Accelerator Control Systems Data," *https://doi.org/10.3390/data6040042 ,* vol. 6, no. 42, 2021.

[15] S. W. Mogensen, K. Rathsman and P. Nilsson, "Causal discovery in a complex industrial system: A time series benchmark," in *Proceedings of the Third Conference on Causal Learning and Reasoning*, Los Angeles, 2024.

Accepted Manuscript

5 – Fast machine learning for laser-plasma accelerators

Matthew Streeter, Queen's University Belfast, UK
m.streeter@qub.ac.uk
Charlotte Palmer, Queen's University Belfast, UK
c.palmer@qub.ac.uk

Status

Laser plasma accelerators (LPAs) use ultra-short laser pulses focused to a high intensity ($\gtrsim 10^{18}$ W/cm^2) to accelerate charged particles within a plasma. This field has developed rapidly over the last decade, demonstrating the capability to accelerate electron and ions beams to high energies over short (micron to centimetre) distances. A key metric of accelerators is the maximum achievable energy within a given distance. For laser-driven ion acceleration, the state of the art is acceleration of $\sim$100 MeV protons using a high energy 200 J interacting with a $\sim$100 nm thick plastic foils[1]. For laser-driven electron acceleration, electron energies up to 8 GeV [2] have been reached using a 20 cm long low-density plasma channel and a 31 J laser pulse through so called Laser Wakefield Acceleration (LWFA).

Several proof-of-concept experiments have demonstrated the utility of these sources for near term applications. For example, LPA proton beams have been used for studies of radiobiology at ultra-high dose-rates[3], and for materials science[4]. X-rays generated in LWFAs by the accelerating electron beam have been used for radiography[5], tomography[6] and x-ray absorption spectroscopy[7]. Recent experiments have also demonstrated the potential of LWFA electron beams for use as compact drivers for x-ray free electron lasers (FELs)[8], [9], [10]. Owing to the multi-modal and highly tuneable nature of LPAs, there are many more applications which are being pursued at laser labs around the world [11].

With a longer-term view, plasma accelerators are an exciting prospect for "compact" future high-energy particle colliders. This is due to the extremely high accelerating gradients that can be sustained in a plasma, making acceleration lengths many orders of magnitude shorter than "conventional" radio-frequency accelerators. For example, GeV electrons have been accelerated in centimetre scale LPAs, in comparison to the 100 m scale conventional machines that achieve equivalent electron energies. In the long term, incorporating LPAs may offer a route to future colliders with a significantly reduced overall size [12], which is currently a major obstacle for extending the high energy frontier.

The dominant laser technology currently used in LPAs have repetition rates of 1-10 Hz at $\sim$10 J per pulse, with a wall plug efficiency of $\lesssim$ 0.1 %. Development of alternative laser technology promises to achieve this energy level at >1 kHz repetition rates and >20% efficiency[13] and is expected to dramatically improve the attractiveness of LPAs for many applications. This step change in repetition and data rates will present new opportunities and challenges for research in the field.

Current and Future Challenges

The non-linear physics at the core of extreme plasma accelerators presents challenges for control systems. The accelerator properties are highly sensitive to initial conditions and so development must focus on the stabilisation of laser drivers and plasma conditions. A demonstration experiment, using

Accepted Manuscript

a specially stabilised laser system, has shown the potential to run an LWFA for >24 hours continuously with electron beam energy stability on the few percent level[14]. The large datasets generated by modern multi-Hz facilities have also been used to reveal the sensitivity of electron beam parameters to the natural variation of the laser system[14], [15]. As the field matures, laser facilities are adopting automation into their operations and using real-time Bayesian Optimisation to efficiently optimise electron[16], [17], x-ray[16] and proton beams[18] from LPAs. These techniques have been mostly demonstrated on experimental systems operating at a few-Hz, due to the limitations of the laser systems. However, it will be necessary to increase the speed of the automation and feedback systems for optimal control of kHz plasma accelerators. The requirement for system stability and reliability has been recently highlighted by a project to develop a LPA injector for PETRA IV at DESY[24].

The challenging quality requirements on the outputs of LPAs motivates the need for active stabilisation of control parameters as well as the use of optimisation algorithms to tune the machines and identify stable regimes of operation [19]. Deep learning techniques have been used to construct statistical models of LPAs[15], [20]. Even on few-Hz systems, such models have shown promise for predictive control systems which react to measurements taken at ~ 1 kHz to adjust for environmental conditions and correct the pointing of the laser system. Extending this to additional laser parameters (e.g. laser energy, wavefront, spectral phase) and for higher repetition rate laser systems will require fast diagnostic analysis and model inference on the sub-millisecond timescale. Future models may also be required to adapt for drift in the machine states and correlations between controls to make long term operation more stable, robust, and reproducible. For instance, diagnosing spatiotemporal couplings (of critical importance in laser-plasma systems [21]) is complex and currently requires combinations of measurements and the use of machine learning models for inference [22]. This was reported to take on the order of 0.1 ms with a normal computer. Accelerated machine learning would provide much faster inference and enable such measurements at $\gtrsim 1$ kHz.

Operation of high power laser experiments is typically achieved through the independent control of multiple sub-systems with varying degrees of automation [23]. The next generation of high energy $\gtrsim 100$ Hz laser facilities will need to be more autonomous. High power laser systems are normally operated near their damage thresholds, which is achieved by careful balance of non-linear optical phenomena. It is relatively easy for a laser to enter states which cause significant damage to the system resulting in expensive repairs and lost beamtime. Therefore, it is paramount to incorporate reliable and robust control systems which can react fast enough to minimise this risk.

Advances in Science and Technology to Meet Challenges

Augmenting expert human operations with autonomous systems is likely required to ensure safe operations of future kHz high power laser systems. Indicators of problems in laser operation can manifest over just a few shots, which at high repetition rate would be far beyond the reaction time of a human operator. These autonomous systems will need to be reliable and obey both known and unknown constraints to prevent the overall machine, or any of its components, entering dangerous states. Such systems may also benefit from the capability to make predictions of damage before it occurs, monitoring for anomalies and responding before irreversible damage occurs, thereby minimising costly replacements and system downtime.

Accepted Manuscript

It will be necessary to adopt networked control systems, as is commonplace in large scale accelerator facilities, to fully automate experiments. To achieve this, the heterogeneity of existing systems will need to be overcome. Both the hardware and software systems will need to be capable of dealing with large data rates, performing data reduction and analysis, and communicating with remote high-performance computing for physics modelling. Automation will require fast device control to minimise disruption to the machine outputs while performing optimisation and stabilisation tasks. Developments of radiation and electromagnetic pulse resistant components will be required, to ensure that systems are robust enough to survive the hostile environments created by high-intensity laser-plasma interactions.

Finally, LPAs are generally not fully diagnosed due to the difficulty of accessing information about the ultrafast microscopic accelerator dynamics as well as the destructive processes that occur when making use of the LPA outputs (e.g. loss of information of the incident electron spectrum in strong-field QED experiments[20]). ML modelling of such systems can suffer from aleatoric and epistemic uncertainties due to the random variation and drift of unobserved parameters. Accurately incorporating this into ML models will be required to maximise their utility. Approaches such as 'continual learning' may offer a way of maintaining optimal control of LPAs under such conditions, by endlessly adapting to changes to the machine state, e.g. degradation of optical components or thermal drifts in the laser system.

Concluding Remarks

Laser plasma acceleration is an emerging technology which is moving into the realm of high-repetition rate operation (kHz) and high data-rates (TB/s). Many techniques in automation can be adapted from developments at large scale accelerator facilities, but there are some unique characteristics that will require bespoke treatment. In particular, the highly non-linear physics of LPAs and of the laser systems themselves present challenges for autonomous control, especially in the presence of significant uncertainty about the machine states. Furthermore, the variety in the outputs of LPAs and within the end-user applications means that automation systems will have to incorporate configuration changes as efficiently as possible and in a user-friendly manner. The benefits of increasing the automation in this field, and incorporating ML for predictive and adaptive control, could be transformative in the utility of this technology by improving efficiency, stability, output quality and machine safety.

Acknowledgements

MJVS acknowledges support from the Royal Society URF-R1221874.

References

[1] A. Higginson *et al.*, 'Near-100 MeV protons via a laser-driven transparency-enhanced hybrid acceleration scheme', *Nat. Commun.*, vol. 9, no. 1, Art. no. 1, Feb. 2018, doi: 10.1038/s41467-018-03063-9.

[2] A. J. Gonsalves *et al.*, 'Petawatt Laser Guiding and Electron Beam Acceleration to 8 GeV in a Laser-Heated Capillary Discharge Waveguide', *Phys. Rev. Lett.*, vol. 122, no. 8, p. 084801, Feb. 2019, doi: 10.1103/PhysRevLett.122.084801.

[3] F. Kroll *et al.*, 'Tumour irradiation in mice with a laser-accelerated proton beam', *Nat. Phys.*, vol. 18, no. 3, Art. no. 3, Mar. 2022, doi: 10.1038/s41567-022-01520-3.

Accepted Manuscript

[4] M. Barberio *et al.*, 'Laser-accelerated particle beams for stress testing of materials', *Nat. Commun.*, vol. 9, no. 1, Art. no. 1, Jan. 2018, doi: 10.1038/s41467-017-02675-x.

[5] A. E. Hussein *et al.*, 'Laser-wakefield accelerators for high-resolution X-ray imaging of complex microstructures', *Sci. Rep.*, vol. 9, no. 1, p. 3249, Dec. 2019, doi: 10.1038/s41598-019-39845-4.

[6] J. M. Cole *et al.*, 'Tomography of human trabecular bone with a laser-wakefield driven x-ray source', *Plasma Phys. Control. Fusion*, vol. 58, no. 1, p. 014008, Jan. 2016, doi: 10.1088/0741-3335/58/1/014008.

[7] B. Kettle *et al.*, 'Single-shot multi-keV X-ray absorption spectroscopy using an ultrashort laser wakefield accelerator source', *Phys. Rev. Lett.*, vol. 123, no. 25, p. 254801, Dec. 2019, doi: 10.1103/PhysRevLett.123.254801.

[8] W. Wang *et al.*, 'Free-electron lasing at 27 nanometres based on a laser wakefield accelerator', *Nat. 2021 5957868*, vol. 595, no. 7868, pp. 516–520, July 2021, doi: 10.1038/s41586-021-03678-x.

[9] M. Labat *et al.*, 'Seeded free-electron laser driven by a compact laser plasma accelerator', *Nat. Photonics*, vol. 17, no. 2, Art. no. 2, Feb. 2023, doi: 10.1038/s41566-022-01104-w.

[10] S. K. Barber, 'Greater than 1000-fold Gain in a Free-Electron Laser Driven by a Laser-Plasma Accelerator with High Reliability', *Phys. Rev. Lett.*, vol. 135, no. 5, 2025, doi: 10.1103/vh62-gz1p.

[11] C. N. Danson *et al.*, 'Petawatt and exawatt class lasers worldwide', *High Power Laser Sci. Eng.*, vol. 7, p. e54, ed 2019, doi: 10.1017/hpl.2019.36.

[12] B. Foster, R. D'Arcy, and C. A. Lindstrøm, 'A hybrid, asymmetric, linear Higgs factory based on plasma-wakefield and radio-frequency acceleration', *New J. Phys.*, vol. 25, no. 9, p. 093037, Sept. 2023, doi: 10.1088/1367-2630/acf395.

[13] L. Kiani *et al.*, 'High average power ultrafast laser technologies for driving future advanced accelerators', *J. Instrum.*, vol. 18, no. 08, p. T08006, Aug. 2023, doi: 10.1088/1748- 0221/18/08/T08006.

[14] A. R. Maier *et al.*, 'Decoding Sources of Energy Variability in a Laser-Plasma Accelerator', *Phys. Rev. X*, vol. 10, no. 3, p. 031039, Aug. 2020, doi: 10.1103/PhysRevX.10.031039.

[15] M. Kirchen *et al.*, 'Optimal Beam Loading in a Laser-Plasma Accelerator', *Phys. Rev. Lett.*, vol. 126, no. 17, p. 174801, Apr. 2021, doi: 10.1103/PhysRevLett.126.174801.

[16] R. J. Shalloo *et al.*, 'Automation and control of laser wakefield accelerators using Bayesian optimization', *Nat. Commun.*, vol. 11, no. 1, Art. no. 1, Dec. 2020, doi: 10.1038/s41467-020-20245-6.

[17] S. Jalas *et al.*, 'Bayesian Optimization of a Laser-Plasma Accelerator', *Phys. Rev. Lett.*, vol. 126, no. 10, p. 104801, Mar. 2021, doi: 10.1103/PhysRevLett.126.104801.

Accepted Manuscript

[18] B. Loughran *et al.*, 'Automated control and optimisation of laser driven ion acceleration', *High Power Laser Sci. Eng.*, pp. 1–11, Mar. 2023, doi: 10.1017/hpl.2023.23.

[19] A. Döpp, C. Eberle, S. Howard, F. Irshad, J. Lin, and M. Streeter, 'Data-driven science and machine learning methods in laser–plasma physics', *High Power Laser Sci. Eng.*, vol. 11, p. e55, Jan. 2023, doi: 10.1017/hpl.2023.47.

[20] M. J. V. Streeter *et al.*, 'Laser wakefield accelerator modelling with variational neural networks', *High Power Laser Sci. Eng.*, vol. 11, p. e9, Jan. 2023, doi: 10.1017/hpl.2022.47.

[21] A. Jeandet *et al.*, 'Survey of spatio-temporal couplings throughout high-power ultrashort lasers', *Opt. Express*, vol. 30, no. 3, pp. 3262–3288, Jan. 2022, doi: 10.1364/OE.444564.

[22] S. Howard *et al.*, 'Single-shot spatiotemporal vector field measurements of petawatt laser pulses', *Nat. Photonics*, vol. 19, no. 8, pp. 898–905, Aug. 2025, doi: 10.1038/s41566-025-01698-x.

[23] S. Feister *et al.*, 'Control Systems and Data Management for High-Power Laser Facilities', *High Power Laser Sci. Eng.*, pp. 1–31, June 2023, doi: 10.1017/hpl.2023.49.

[24] I. Agapov *et al.*, *The Plasma Injector for PETRA IV: Enabling Plasma Accelerators for Next-generation Light Sources. Conceptual Design Report*. Hamburg: Deutsches Elektronen-Synchrotron DESY, 2025.

Accepted Manuscript

6 - Fast ML for fusion simulation, optimization, and control

Jonathan Citrin, Google Deepmind, London, UK
citrin@google.com

Status

Fusion energy promises high-energy-density electricity and heat production with essentially limitless fuel reserves, inherent nuclear safety, and negligible impact on the environment [1]. The key societal importance of clean abundant energy, coupled with recent progress in the field, is currently fuelling intense interest and fusion investment in both the public and private spheres

Multiple approaches exist to achieve fusion energy. Laser fusion, a sub-branch of inertial confinement fusion (short fusion pulses at high plasma density), has recently achieved the historic milestone of net fusion gain, generating more fusion energy out than energy deposited into the fuel, in individual pulses at the National Ignition Facility in the US [2]. In magnetic confinement fusion, fusion occurs in lower density plasmas over long (ideally continuous) pulses. Various 2D and 3D magnetic configurations are studied, including stellarators, field-reversed-configurations, and tokamaks. Tokamaks are the most mature technology of this class. The next generation of tokamak experiments aims for net fusion power gain for the first time [3,4]. While the challenges and simulation technologies outlined here share commonalities across the range of fusion approaches, the remainder of this review focuses on tokamaks

Previous generations of tokamak device and experimental design have primarily relied on empirical scalings of plasma behaviour for determining device and plasma scenario configuration. However, continuous improvements in tokamak plasma simulation capabilities are bridging discrepancies between simulations and measurements, known is the s im2r eal gap. For example, as shown in figure 1, recent record-breaking fusion performance in the JET tokamak DT campaigns was successfully predicted by integrated modelling beforehand [5]. These advances increase confidence in theory-based extrapolations to performance in next-step devices, and enable a plethora of new simulation-based applications. In general, tokamak simulation has multiple use-cases. Interpretive simulations obtain information unavailable from measurements or help constraint measurements, for example calculating radial distributions of plasma heating and fuelling, and currents. Accurate simulations can provide physics understanding, validating theory-based models and untangling cause-and-effect in experiments through complex multi-physics simulations [6, 7]. The costs and risks of tokamak operation can be reduced by using simulation for experimental preparation (including inter-shot), performance optimization including new scenario discovery, model-based controller design, and reactor design itself. In pulse-planning for next generation devices, pre-simulation is considered a requirement. ML can play a key role in accelerating simulation while maintaining accuracy, enabling these complex use-cases.

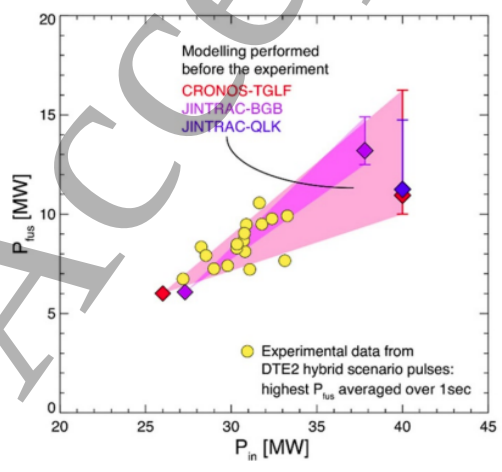


Figure 1. Comparison between predicted and measured fusion power for the "hybrid" plasma operating scenario in the JET DTE2 campaign. The measured fusion power is within the envelope of the predict-first simulation campaign. Reproduced courtesy of IAEA. Figure from [5]. © EURATOM 2023.

Accepted Manuscript

Current and Future Challenges

Tokamak integrated modelling is inherently multiscale and multiphysics, with multiple orders of magnitude in spatiotemporal scales between relevant physics processes. These include: magnetic equilibrium; transport arising from plasma collisions, magnetohydrodynamics and turbulence; heating and fuelling; line radiation arising from atomic and molecular interactions; and plasma material interaction. While high-fidelity direct numerical simulation is possible, this requires coupling multiple high-fidelity models at extreme (exascale) expense. Such workflows can provide ultimate ground truth for numerical verification of reduced models, but are not pragmatic for most use-cases.

A central challenge in tokamak simulation is thus bridging the conflicting constraints of model fidelity and tractability. Tokamak integrated modelling consists of coupled sets of transport PDEs, with reduced physics models calculating the various PDE coefficients [8]. The fidelity level of integrated modelling depends on the quality of these reduced models, with a trade-off between accuracy and tractability. Accurate but fast reduced physics models for these coefficients (related to plasma turbulence, heating, etc) are required. Multiple spatial domains must be simulated and coupled, comprising the plasma core (nested magnetic flux surfaces) which is amenable to a 1D radial approximation, the plasma edge region intersecting with the containment wall and requires a 2D description, and models for the plasma-facing-materials. In practice, workflows consisting of a single domain at a time are commonly used, although self-consistent coupled simulations are desirable. In general, this approach is suitable for experimental interpretation, gaining physics understanding, and extrapolation to future scenarios. However, accurate simulations of this type are typically too slow for many-query use cases such as pulse-design, optimization, and controller-design.

Additional challenges arise within the context of realtime control. Fast and accurate simulation is required for theory-based nonlinear models in classical optimal control techniques such as Model Predictive Control. Fast and accurate simulation is also required for realistic environments for agents in reinforcement learning approaches. Accurate state estimation for controllers would improve with higher availability of realtime diagnostics and low-latency event detection.

Advances in Science and Technology to Meet Challenges

A key method for improving tokamak multiphysics modelling speed is incorporating learned ML surrogates of integrated modelling physics components. These act as drop-in replacements of physics models incorporated in the transport PDEs. Furthermore, surrogates, e.g. based on neural networks, are differentiable, enabling their use within differentiable simulation for gradient-driven scenario optimization and parameter identification. Surrogates are made using supervised learning techniques, learned from simulation databases generated in relevant parameter space using large-volume compute. General function approximators such as neural networks have the added advantage of being analytically differentiable functions, allowing incorporation into differentiable simulators for nonlinear PDE solvers and gradient-driven optimization use-cases [9]. Prior knowledge of the input-output mapping characteristics of the physics models can be incorporated into the loss functions used for model training, improving evaluation accuracy. Numerous ML-surrogates for tokamak simulation have recently emerged. These include surrogates for neutral beam injection and current drive models [10], edge-transport-barrier formation [11], reduced turbulence models [11, 12, 13], and equilibrium [14]. These models were developed with different tokamak parameter ranges of validity, and cannot

Accepted Manuscript

necessarily be incorporated into a single simulation. A concerted community effort to combine datasets and efforts, enabling a wide selection of ML-surrogates to be used in a single simulation for a given device, is desirable.

Taking advantage of the fact that in surrogate model generation, the primary computational burden is relegated to the data generation phase, ML-surrogates can be developed of models at higher fidelity than those typically used in integrated modelling, leading to simulation both faster and more accurate than the state-of-the-art [15]. High-fidelity models themselves can also be accelerated by ML techniques, for example by on-the-fly surrogate generation in narrow parameter space in simulation workflows [16], developing surrogates of internal expensive physics [17], or accelerated solver methods [18]. These approaches would enable generation of higher quality datasets for surrogate generation, and/or enable more routine high-fidelity verification of reduced models.

For physics components where theory-based models cannot deliver a sufficiently small s im2r eal gap, or are too expensive to generate sufficiently informative datasets for surrogate model generation, a hybrid data and model-driven approach is necessary [19]. Wider access to tokamak data would significantly accelerate development, as often access to quality data is the bottleneck in successful ML applications. Rapid advances in generative AI may also significantly improve data-driven predictions, e.g. by learning latent structures, diffusion methods, and sequence-to-sequence models such as LLMs and structured state spacemodels .

For realtime control applications, fast and accurate simulation enabled by ML surrogates can form environments used for training control policies with reinforcement learning, for example extending a recent approach pioneered for tokamak magnetic control, to multiphysics [20]. For state estimation, the range of available realtime diagnostics can be extended by using ML for accelerating inference, e.g. tomographic inversion, collisional-radiative calculations, or constrained equilibrium reconstruction. Reduced latency of these calculations would make previously privileged information available to controllers and supervisory systems, improving the efficacy of control policies.

Concluding Remarks

ML can play a key role in the development and control of future fusion reactors. Learned surrogates can provide fast and accurate simulation, needed for pulse design and optimization, as well as simulation environments for learning control policies with reinforcement learning. A combination of high-volume compute, adaptive sampling and workflow automation is required to develop the range of necessary surrogates. Open data practices at facilities is key for simulation validation, and incorporation of data-driven approaches for simulation components which have an insufficient theory-driven sim2real gap. Classical high-fidelity simulation is still required to refine reduced-order-models and constrain missing physics. ML can extend the range of realtime diagnostics available for state estimation, providing more information that can be used by control policies. Together, these approaches enable physics understanding, optimal pulse design, general control policies, all on the path towards the operational goals required for achieving fusion energy.

Accepted Manuscript

References

[1] E.R. Sadik-Zada, A. Gatto, and Y. Weißnicht. "Back to the future: Revisiting the perspectives on nuclear fusion and juxtaposition to existing energy sources." Energy 129150, 2023.

[2] A.B. Zylstra, O.A. Hurricane, D.A. Callahan, A.L. Kritcher, J.E. Ralph, H.F. Robey, et al. "Burning plasma achieved in inertial fusion." Nature 601, no. 7894, p. 524-548, 2022

[3] B. Bigot, "Preparation for assembly and commissioning of ITER." Nuclear Fusion 62, no. 4, p.042001, 2022

[4] A.J. Creely, M.J. Greenwald, S.B. Ballinger, D. Brunner, J. Canik, J. Doody, et al. "Overview of the SPARC tokamak." Journal of Plasma Physics 86, no. 5, 865860502, 2020

[5] J. Garcia, F. J. Casson, L. Frassinetti, D. Gallart, L. Garzotti, H-T. Kim, et al. "Modelling performed for predictions of fusion power in JET DTE2: overview and lessons learnt." Nuclear Fusion 63, no. 11, 112003, 2023

[6] M. Marin, J. Citrin, L. Garzotti, M. Valovic, C. Bourdelle, Y. Camenen, et al., "Multiple-isotope pellet cycles captured by turbulent transport modelling in the JET tokamak". Nuclear Fusion, 61(3), p.036042, 2021

[7] P. Rodriguez-Fernandez, A.E. White, N.T. Howard, B.A. Grierson, G.M. Staebler, J.E. Rice, "Explaining cold-pulse dynamics in tokamak plasmas using local turbulent transport models", Physical Review Letters, 120(7), p.075001, 2018

[8] *F.M. Poli. "Integrated Tokamak modeling: When physics informs engineering and research planning," Physics of Plasmas, 25(5), 2018.*

[9] S. Van Mulders, F. Felici, O. Sauter, J. Citrin, A. Ho, M. Marin, K.L. van de Plassche, "Rapid optimization of stationary tokamak plasmas in RAPTOR: demonstration for the ITER hybrid scenario with neural network surrogate transport model QLKNN". Nuclear Fusion, 61(8), p.086019, 2021

[10] M.D. Boyer, S. Kaye, and K. Erickson, "Real-time capable modeling of neutral beam injection on NSTX-U using neural networks." Nuclear Fusion, 59(5), p.056008, 2019

[11] O. Meneghini, S.P. Smith, P.B. Snyder, G.M. Staebler, J. Candy, E. Belli, et al, "Self-consistent core-pedestal transport simulations with neural network accelerated models", Nuclear Fusion, 57(8), p.086034, 2017

[12] K.L. van de Plassche, J. Citrin, C. Bourdelle, Y. Camenen, F.J. Casson, V.I. Dagnelie, F. Felici, et al, "Fast modeling of turbulent transport in fusion plasmas using neural networks", Physics of Plasmas, 27(2), 2020

[13] A. Ho, J. Citrin, C.D. Challis, C. Bourdelle, F. J. Casson, J. Garcia, et al. "Predictive JET current ramp-up modelling using QuaLiKiz-neural-network." Nuclear Fusion 63, no. 6, 066014, 2023

[14] S. Joung, Y.C. Ghim, J. Kim, S. Kwak, D. Kwon, C. Sung, et al., "GS-DeepNet: mastering tokamak plasma equilibria with deep neural networks and the Grad–Shafranov equation." Scientific Reports, 13(1), p.15799, 2023

[15] J. Citrin, P. Trochim, T. Goerler, D. Pfau, K.L. van de Plassche and F. Jenko, "Fast transport simulations with higher-fidelity surrogate models for ITER." Physics of Plasmas, 30(6), 2023

[16] P. Rodriguez-Fernandez, N.T. Howard, and J. Candy, "Nonlinear gyrokinetic predictions of SPARC burning plasma profiles enabled by surrogate modeling." Nuclear Fusion, 62(7), p.076036, 2022

[17] S.T. Miller, N.V. Roberts, S.D. Bond and E.C. Cyr, "Neural-network based collision operators for the Boltzmann equation." Journal of Computational Physics, 470, p.111541, 2022

[18] D.N. Tanyu, J. Ning, T. Freudenberg, N. Heilenkötter, A. Rademacher, U. Iben, and P. Maass, "Deep learning methods for partial differential equations and related parameter identification problems.", Inverse Problems, 39(10), p.103001, 2023

[19] P.W. Hatfield, J.A. Gaffney, G.J. Anderson, S. Ali, L. Antonelli, S. Başeğmez du Pree, et al., "The data-driven future of high-energy-density physics." Nature, 593(7859), pp.351-361, 2021

[20] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, et al., "Magnetic control of tokamak plasmas through deep reinforcement learning." Nature, 602(7897), pp.414-419, 2022

Accepted Manuscript

7 - In-Network Machine Learning: Inference at the Speed of Data

Changgang Zheng and Noa Zilberman, University of Oxford
changgang.zheng, noa.zilberman@eng.ox.ac.uk

Background and Status

Machine learning (ML) was widely adopted and extensively utilized by data-intensive applications such as stream processing and cybersecurity. However, the high demands of ML workloads pose challenges that local accelerators (e.g., GPUs, FPGAs) struggle to address. To cope with the scale and volume of the data to process, distributed ML is often used.

Networking and ML have a complex relationship. ML services rely on the network to carry data and connect accelerators, but often the exchange of data between accelerators is a bottleneck in distributed ML [10, 14]. At the same time, network devices can carry more data than an accelerator card can process, since a single network switch can process more than 100 terabits per second (Tbps), two orders of magnitude more than current accelerators. One way to bridge the mismatch between networking and ML is in-network ML. To explain it, a brief introduction to network programmability is provided.

High end network devices are designed to process as many packets as possible, as quickly as possible, exceeding 10 billion packets per second (Bpps) with sub-microsecond latency. Over the last decade, many network devices have become programmable [3], combining the processing pipeline (referred to as data plane) with reconfigurable match tables (RMT). The most popular general architecture is called PISA, Protocol Independent Switch Architecture, shown in Figure 1(c). The domain specific language commonly used is P4.

As shown in Figure 1, the PISA architecture has three components: i) a programmable parser for extracting header information from incoming packets, ii) a match-action pipeline that looks up keys (e.g., headers) in tables and matches them with actions (e.g., output port), and iii) a deparser for reconstructing packets. Current programmable network devices range from software switches to switch-ASICs, FPGAs, SmartNICs, and DPUs.

The programmability of network devices enabled in-network computing, the offloading of applications normally running on a host to run within network devices. Examples of In-network computing applications include telemetry, caching [7] and gradient aggregation [14]. It demonstrated order of magnitude improvement in throughput, latency, and power efficiency [18]. Accelerating ML using in-network computing is one promising direction. This includes [21] both Network-Assisted ML, meaning using the network to accelerate workloads running on traditional targets, and in-network ML.

In-network ML offloads the entire ML inference process to programmable network devices [21]. It trains the ML model on servers or standard accelerators, and then maps the trained model to the device's data plane for inference. While resource utilization varies, programmable switches often use only half of their resources for basic network functions [23], enabling in-network ML to leverage remaining resources. Models scale up to 20 trees and 55 features, achieving high classification accuracy [24]. The three main benefits of in-network ML are location, latency and load (3Ls) [24]. Network devices are already part of the infrastructure that data goes through (location), they can infer data closer to its source (latency) and inference performance scales with network bandwidth, while reducing load on end-hosts (load). In-network ML can be deployed either on network devices that are already on route from the data source, or as a network-attached accelerator. The first requires no additional cost or latency, while the second benefits from more available processing resources.

Accepted Manuscript

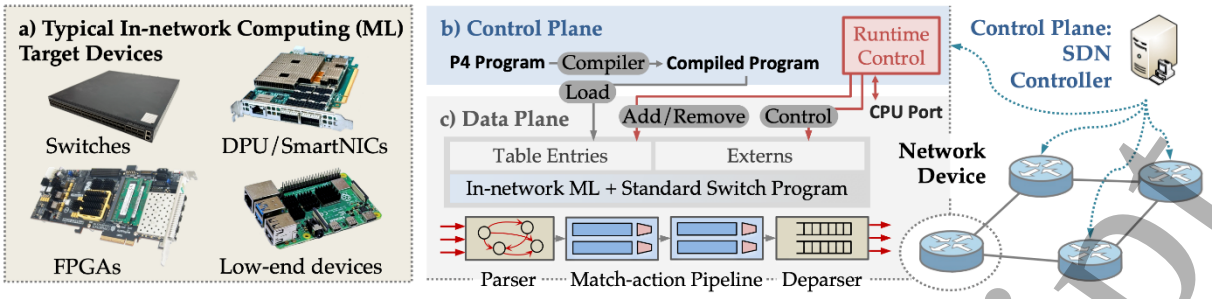


Figure 1. a) Commodity programmable network devices commonly used for in-network ML. b) The control plane of a network devices, e.g., switch-box CPU. c) A PISA based data plane of a programmable network devices, combining in-network ML and standard network functions.

Meeting the Exascale Challenges

In-network ML inference process has two major components: feature extraction, and mapping of a trained ML algorithm to the data plane.

Feature Extraction. The extraction of features from incoming data can be done on various levels of granularity. Packet-level features are extracted directly from incoming packets, and are stateless, meaning the data from one packet does not affect later packets. Flow-level features consider the data flowing between a source-destination pair, and are stateful, meaning that information is accumulated over time, such as the traffic volume of the flow. Aggregation level features are also stateful, but combine information based on different metrics, e.g., the statistics of combining multiple ports. File-level feature extraction is possible but limited to certain types of files (e.g., text) and may have reduced performance, for example due to recirculation of packets within the pipeline. While packet-level features can be extracted in the parser, more complex feature extraction and feature engineering requires match-action pipeline stages and memory resources. One of the challenges for feature extraction is that packets arrival may be unordered, and information may travel through multiple routes.

ML Model Mapping. To maximise packet processing rate, network devices are constructed with limitations on programmability. For example, mathematical operations such as multiplication, and data types such as floating-point, are unavailable. Moreover, these are billions-of-transistors devices, therefore, to make their manufacturing feasible, user-available resources are limited in comparison with CPUs. For example, memory on a switch-ASIC is in the range of 100's of megabits, and the number of processing stages is e.g., 12 or 20. There is no native support for loops, and the closest equivalent, packet recirculation, has throughput implications. While some works have addressed these challenges through changes to the hardware [17] or by focusing on NICs [16], other successfully mapped inference models to off-the-shelf devices without loss of performance [21]. The mappings of ML models to the data plane can be divided into three groups [25]: Direct Mapping, where every step of the inference is translated into a stage in the pipeline, Lookup-based, where lookup tables are used for complex calculations, and Encode-based, where the feature space is sliced and encoded.

In-network ML Implementation. Implementing mapped ML models on target devices, and porting between different devices, requires significant effort. To solve this problem, Planter [25] offers rapid prototyping of in-network ML. Planter supports a range of ML models, architectures and target devices. It is modular, so users can extend the functionality. The framework also supports automatic parameter tuning, and several example use cases. Targets range from low-cost Raspberry Pi, through

Accepted Manuscript

Alveo FPGA, to high performance Intel Tofino switch-ASIC. The demonstrated performance using ensemble tree models was 4.8 billion decisions per second.

Advances in Science and Technology to Meet Challenges

While in-network ML achieves high system performance, the size and complexity of the models that can be deployed are limited by resource constraints. This problem was identified by Microsoft as a key challenge for in-network computing [2], and Microsoft developed a resource elasticity solution [9] to address it.

As often only small or medium size models can be deployed on a network device [1, 5, 8, 12, 13, 25], ML performance can be lower than an unlimited size model running in a data centre. One solution is to distribute the model across multiple network devices, as demonstrated by DINC [23]. It has shown that ML models can be distributed across a data centre network or a wide area network, while coexisting with network functionality. A different solution, proposed in IIsy [24], is a hybrid deployment of in-network ML: deploying a small model on a network device, and a large model at the backend. This way, high confidence inference decisions are taken within the network, while only a fraction of the transactions needs to be inferred in a traditional manner. This enables both high ML performance and high system performance, while saving compute resources.

Data shift and model updates further challenge production deployments. P4Pir [20] suggests solving this problem by introducing a continuous update solution, focused on in-network ML for IoT.

Most in-network ML research focused so far on network traffic classification and cybersecurity [1, 8, 16, 17, 21, 24]. Examples that go beyond these use cases include load-balancing [22], predicting future stock prices [6], and image classification [15]. Still, the greatest challenge for in-network ML is wide adoption. This calls for extending its range of use cases, with data intensive and time-sensitive applications being the immediate candidates. Real time processing of sensor data, for example, contextual streams from the GOTO telescope [19], from particle detectors at CERN's Large Hadron Collider [4], and from modern multi-Hz laser wakefield acceleration facilities [11], holds a significant potential.

Concluding Remarks

In-network ML is an emerging technology, providing notable system performance benefits for machine learning inference. Available solutions enable deployment on commercial devices, supporting a range of ML models. Still, resource constraints remain a challenge, with hybrid and distributed deployments suggested as a solution.

There are opportunities for breakthrough research in the applications of in-network ML, benefiting the scientific community across disciplines, and now is the time to seize it.

Acknowledgements

This work was partly funded by VMware and Innovate UK (project 10056403) as part of the SmartEdge EU project (grant agreement No. 101092908). We acknowledge support from Intel and NVIDIA. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

References

[1] Aristide Akem Tanyi-Jong, Michele Gucciardo, and Marco Fiore. Flowrest: Practical flow-level inference in programmable switches with random forests. IEEE INFOCOM 2023-IEEE Conference on Computer Communications 2023.

Accepted Manuscript

[2] Victor Bahl. Unlocking the potential of in-network computing for telecommunication workloads. Website. https://azure.microsoft.com/en-us/blog/ unlocking-the-potential-of-in-network-computing-for-telecommunication-workloads/ [Online; accessed November 2023].

[3] Pat Bosshart, Glen Gibb, Hun-Seok Kim, George Varghese, Nick McKeown, Martin Izzard, Fernando Mujica, and Mark Horowitz. Forwarding metamorphosis: Fast programmable match-action processing in hardware for SDN. *ACM SIGCOMM Computer Communication Review*, 43(4):99–110, 2013.

[4] CMS Collaboration, The Phase-2 Upgrade of the CMS Level-1 Trigger, CERN, Geneva, Tech. Rep. CERN-LHCC-2020-004, CMS-TDR-021, https://cds.cern.ch/record/2714892, 2020.

[5] Busse-Grawitz Coralie, Roland Meier, Alexander Dietmüller, Tobias Bühler, and Laurent Vanbever. pforest: In-network inference with random forests. arXiv:1909.05680, 2019.

[6] Xinpeng Hong, Changgang Zheng, Stefan Zohren, and Noa Zilberman. LOBIN: In-Network Machine Learning for Limit Order Books. In *2023 IEEE 24th International Conference on High Performance Switching and Routing (HPSR)*, pages 159–166. IEEE, 2023.

[7] Xin Jin, Xiaozhou Li, Haoyu Zhang, Robert Soule, Jeongkeun Lee, Nate Foster, Changhoon Kim, and Ion Stoica. NetCache: Balancing Key-Value Stores with Fast In-Network Caching. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 121–136, 2017.

[8] Lee Jong-Hyouk and Kamal Singh. SwitchTree: in-network computing and traffic analyses with Random Forests. *Neural Computing and Applications*, pages 1-12, 2020.

[9] Daehyeok Kim, Zaoxing Liu, Yibo Zhu, Changhoon Kim, Jeongkeun Lee, Vyas Sekar, and Srinivasan Seshan. Tea: Enabling state-intensive network functions on programmable switches. In *Proceedings of the 2020 ACM SIGCOMM Conference*, pages 90–106, 2020.

[10] Liang Luo, Peter West, Jacob Nelson, Arvind Krishnamurthy, and Luis Ceze. Plink: Discovering and exploiting locality for accelerated distributed training on the public cloud. Proceedings of Machine Learning and Systems 2 (2020): 82-97.

[11] Kirchen Manuel, Sören Jalas, Philipp Messner, Paul Winkler, Timo Eichner, Lars Hübner, Thomas Hülsenbusch et al. Optimal beam loading in a laser-plasma accelerator. *Physical review letters*, 2021.

[12] Friedman Roy, Or Goaz, and Ori Rottenstreich. Clustreams: Data plane clustering. In *Proceedings of the ACM SIGCOMM Symposium on SDN Research (SOSR)*, pages 101-107, 2021.

[13] Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. Can the network be the AI accelerator?. *Proceedings of the 2018 Workshop on In-Network Computing*. 2018.

[14] Amedeo Sapio, Marco Canini, Chen-Yu Ho, Jacob Nelson, Panos Kalnis, Changhoon Kim, Arvind Krishnamurthy, Masoud Moshref, Dan Ports, and Peter Richtarik. Scaling distributed machine learning with In-Network aggregation. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, pages 785–808, 2021.

[15] Hisham Siddique, Miguel Neves, Carson Kuzniar, and Israat Haque. Towards network-accelerated ML-based distributed computer vision systems. In *2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 122–129. IEEE, 2021.

[16] Giuseppe Siracusano, Salvator Galea, Davide Sanvito, Mohammad Malekzadeh, Gianni Antichi, Paolo Costa, Hamed Haddadi, and Roberto Bifulco. Re-architecting Traffic Analysis with Neural Network Interface Cards. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 513–533, 2022.

[17] Tushar Swamy, Alexander Rucker, Muhammad Shahbaz, Ishan Gaur, and Kunle Olukotun. Taurus: A Data Plane Architecture for Per-Packet ML. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 1099–1114, 2022.

[18] Yuta Tokusashi, Huynh Tu Dang, Fernando Pedone, Robert Soule, and Noa Zilberman. The Case for In-network Computing on Demand. In *Proceedings of the Fourteenth EuroSys Conference 2019*, pages 1–16, 2019.

[19] U F Burhanudin, J R Maund, T Killestein, K Ackley, M J Dyer, J Lyman, K Ulaczyk, R Cutter, Y-L Mong, D Steeghs, D K Galloway, V Dhillon, P O'Brien, G Ramsay, K Noysena, R Kotak, R P Breton, L Nuttall,

Accepted Manuscript

E Pallé, D Pollacco, E Thrane, S Awiphan, P Chote, A Chrimes, E Daw, C Duffy, R Eyles-Ferris, B Gompertz, T Heikkilä, P Irawati, M R Kennedy, A Levan, S Littlefair, L Makrygianni, D Mata-Sánchez, S Mattila, J McCormac, D Mkrtichian, J Mullaney, U Sawangwit, E Stanway, R Starling, P Strøm, S Tooke, K Wiersema, Light-curve classification with recurrent neural networks for GOTO: dealing with imbalanced data, *Monthly Notices of the Royal Astronomical Society*, Volume 505, Issue 3, pages 4345–4361, https://doi.org/10.1093/mnras/stab1545, 2021

[20] Mingyuan Zang, Changgang Zheng, Lars Dittmann, and Noa Zilberman. Towards Continuous Threat Defense: In-Network Traffic Analysis for IoT Gateways. *IEEE Internet of Things Journal*, 2023.

[21] Changgang Zheng, Xinpeng Hong, Damu Ding, Shay Vargaftik, Yaniv Ben-Itzhak, and Noa Zilberman. In-Network Machine Learning Using Programmable Network Devices: A Survey. *IEEE Communications Surveys & Tutorials*, 2023.

[22] Changgang Zheng, Benjamin Rienecker, and Noa Zilberman. QCMP: Load Balancing via In-Network Reinforcement Learning. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Future of Internet Routing & Addressing*, pages 35–40, 2023.

[23] Changgang Zheng, Haoyue Tang, Mingyuan Zang, Xinpeng Hong, Aosong Feng, Leandros Tassiulas, and Noa Zilberman. DINC: Toward Distributed In-Network Computing. *Proceedings of the ACM on Networking*, 1(CoNEXT3):1–25, 2023.

[24] Changgang Zheng, Zhaoqi Xiong, Thanh T Bui, Siim Kaupmees, Riyad Bensoussane, Antoine Bernabeu, Shay Vargaftik, Yaniv Ben-Itzhak, and Noa Zilberman. Ilsy: Hybrid In-Network Classification Using Programmable Switches, *IEEE/ACM Transactions on Networking*, pages 2555-2570, 2024.

[25] Changgang Zheng, Mingyuan Zang, Xinpeng Hong, Liam Perreault, Riyad Bensoussane, Shay Vargaftik, Yaniv Ben-Itzhak, and Noa Zilberman. Planter: Rapid Prototyping of In-Network Machine Learning Inference, *ACM SIGCOMM Computer Communication Review*, pages 2-21, 2024.

Accepted Manuscript

8 – Accelerating Traditional HPC using Artificial Intelligence: A Selective Overview

Alexander Titterton, Graphcore (11-19 Wine Street, Bristol, BS1 2PH, United Kingdom)
[alexandert@graphcore.ai]

Status

Historically, High-Performance Computing (HPC) applications in fields such as high-energy particle physics, medical research and quantum chemistry have relied upon performing extensive calculations. We present a brief overview of some of the new applications in the intersection between Artificial Intelligence and HPC, along with some of the challenges these aim to address.

Whilst modern supercomputing clusters typically feature a large number of processing cores, allowing for such applications to be highly parallelised, complex workloads can still take weeks or even months to complete. Furthermore, such applications often consist of a large codebase comprising thousands of lines of C++ code, in turn introducing the non-trivial challenge of code optimisation.

These challenges and the trade-off between simulation speed and accuracy motivate a new approach; a promising candidate for such being the introduction of Artificial Intelligence (AI)/Machine Learning (ML) models into traditional HPC workflows. Experimentation with surrogate models – ML models which replace a numerically and computationally intensive part of an HPC workload – is becoming commonplace in various fields such as weather forecasting and high-energy particle physics.

Since an AI/ML model is simply learning to emulate the behaviour of a system, by producing an output which aims to better satisfy certain conditions in order to minimise the training loss, it need not be as computationally intensive as a simulation or analysis program written by hand. Furthermore, the mathematical operations which constitute these ML models are often well-suited to run on dedicated accelerator hardware in a massively-parallel fashion.

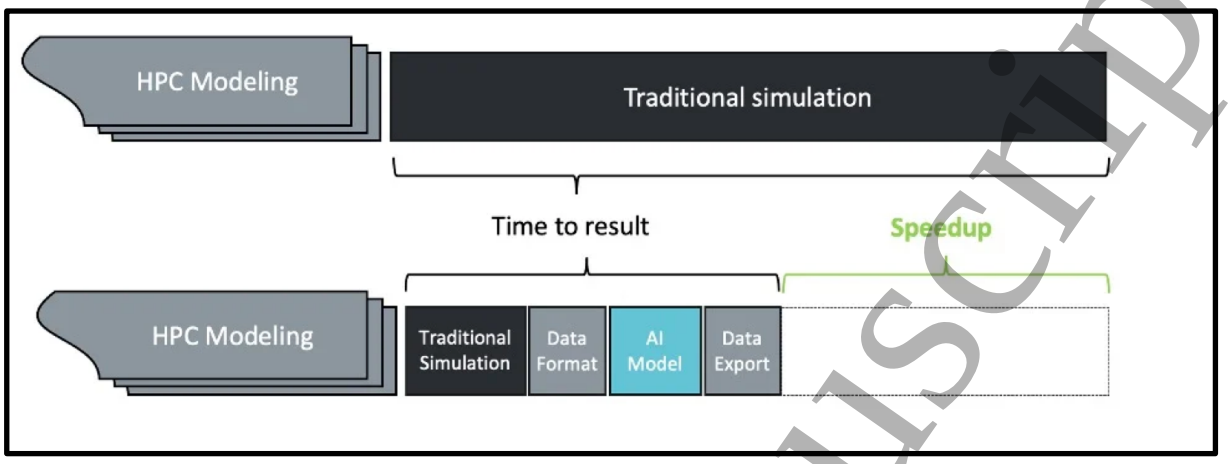
Mainstream Python-based Deep Learning (DL) frameworks such as PyTorch and TensorFlow support ML accelerators such as IPUs and GPUs natively. This has two key benefits: firstly, an application can easily and efficiently target dedicated hardware, offering huge speedups as well as better power efficiency (for example performance per Watt). Secondly, the amount of low-level code development required in order to produce a performance-optimised application is often dramatically reduced compared with more traditional HPC applications, thus allowing for more portable and reusable codebase which can more easily be shared and collaborated upon.

Using ML models in scientific HPC workloads is not only beneficial in terms of computational efficiency, however. There are many physical processes for which observation is possible, but we lack sufficient understanding of the underlying mechanisms to develop an accurate simulation. In such cases, ML models may be able to learn to emulate the behaviour of a system without requiring upfront an explicit description or set of rules which govern said system.

As shown in Figure 1, AI-driven surrogate models may be used to reduce bottlenecks in HPC workloads by replacing a computationally expensive part of the process with an AI model, whilst preserving other

Accepted Manuscript

elements of the pipeline. Other parts of the workload, for example data pre/post-processing and other simulation or analysis components are typically kept, rather than replacing the entire end-to-end process with an AI model, allowing for more fine-grained control over the outputs from each stage. In general, this surrogate model approach appropriately describes the implementation of each of the applications presented in this chapter.



gure 1. An AI-driven solver can accelerate the time-to-result compared with traditional HPC simulation by emulating the output of a computationally intensive function, alleviating bottlenecks.

Current and Future Challenges

As the scientific community advances, the ever-increasing complexity of data analysis, simulation and modelling drives the requirement for higher computing power. Since CPU clock speed has plateaued in recent years [1] parallel programming has become more popular, as has the use of dedicated hardware accelerators – though not without their own challenges and limitations. As was described in "The Hardware Lottery" [2], often a proposed software solution is not that which best befits the problem one is trying to solve, but that which best suits the hardware one has on hand.

In the field of experimental high-energy particle physics, we see an enormous demand for computing resources corresponding to the sheer scale at which data is produced by Large Hadron Collider experiments. At the CMS experiment alone, the Phase-2 upgrade introduces a 4x increase in the recorded event size and a 7x increase in the event acceptance rate, leading to an overall increase of 20x in terms of the rate at which data is written to disk [3]. In order to ensure the event selection is efficient and robust, this in turn requires faster and more complex "online" event reconstruction and analysis tools. Such an increase in overall observed data also corresponds to a need to generate sufficient Monte Carlo simulated data, introducing further computing challenges.

The challenges in the field of medical drug discovery in the healthcare industry, however, are quite different. The average cost to develop a new drug lies in the region of $2.6 Billion, often taking more than 10 years. Worse still, only a small percentage of new drugs are successful in moving from clinical trials to market [4]. Contrary to the case in high-energy particle physics, where accurate simulation of particle interactions is possible, the mapping of a protein's structure to its function in medical treatment is not always well understood, and therefore an entirely new approach is required.

Nuclear fusion has potential to provide clean energy without consumption of fossil fuels. However only as recently as 2022 [5] has a fusion device been able to achieve a gain factor of greater than 1;

Accepted Manuscript

i.e. produce more energy than was required to maintain the plasma steady state. Efficient power generation using a reactor such as a Tokamak requires a control system which is able to predict the evolution of the plasma state over time, which involves solving non-trivial Partial Differential Equations (PDEs). Solving such PDEs numerically is computationally intensive, yet must be performed in real time in order to be able to maintain a stable plasma.

Advances in Science and Technology to Meet Challenges

Whilst ML models can be more computationally efficient compared with HPC methods, achieving faster performance typically requires accelerator hardware. GPUs have long been repurposed for AI applications, but only recently has dedicated AI-first hardware begun to surface. An example is Graphcore's IPU, which consists of a large number of independent processing cores along with on-chip memory [6], offering an alternative architecture to GPUs. Such alternatives can offer researchers new possibilities for more efficiently enhancing their workloads, whilst crucially supporting popular ML frameworks such as TensorFlow and PyTorch.

A variety of AI-driven approaches have been employed by the high-energy particle physics community. The use of generative models, for example Generative Adversarial Networks (GANs), enables accurate simulated data to be generated very quickly. A GAN comprises two ML models, with one trained to distinguish between generated and real data and the other trained to generate realistic data to fool the former. GANs are investigated for simulating di-jet production in [7], with the performance compared between CPUs, GPUs and IPUs. Recent work has shown further advances in the use of diffusion models for accelerated generation of simulated collision data [8]. Diffusion models are able to synthesize data by learning to iteratively reverse a fixed stochastic process that progressively corrupts a data sample into pure noise. Furthermore, following the release of ChatGPT, transformers have become popular in the AI community. The work in [9] introduces ParT (Particle Transformer) as a new benchmark for ML-driven jet tagging models.

In pharmaceutical research, *de novo* drug discovery is split into two main task categories: predictive (mapping protein structure to function) and generative (designing a protein to perform a desired function), with AI relevant in both cases. Transformer-based models are also proving popular here, with researchers starting with BERT [10], a model designed for English language, and replacing the vocabulary with one based upon the four bases of DNA sequences (A C G T). This approach has demonstrated the ability to learn protein structural and functional properties [11]. However recent developments such as ProteinBERT [12] demonstrate how modifying models can better suit protein-related tasks.

An approach for solving PDEs is Physics-Informed Neural Networks (PINNs), whereby a PDE may be solved by training an ML model whose loss functions represent the PDE and its initial and boundary conditions. This type of application has potential to produce fast numerical solvers for PDEs which cannot be solved analytically, such as those which govern the plasma instability in a tokamak. PINNs are being investigated for such applications in nuclear fusion, however optimisation presents a challenge. The loss landscape for a PINN is often tumultuous, often resulting in the model getting stuck at a local optimum. To overcome such challenges, researchers in [13] investigate a hybrid methodology of training a PINN using labelled data and fine-tuning using an unsupervised approach.

Accepted Manuscript

Concluding Remarks

As has been seen in this brief overview, the use of surrogate ML models to augment HPC applications in scientific research is an exciting and fast-progressing field. The ever-increasing complexity of computational challenges is clearly out-pacing the computational resources available, suggesting that in many cases in the near future traditional non-ML workloads simply may not suffice.

There remain numerous challenges, such as the explicability of a "black box" neural network, and the need to build up relevant technical skills in the scientific community, particularly for students and junior researchers. In light of such challenges, close collaboration between academia and industry is more important than ever.

References

[1] Markov, I. L., "Limits on fundamental limits to computation", *Nature*, vol. 512, no. 7513, pp. 147–154, 2014.
[2] Hooker, S., "The hardware lottery", *Communications of the ACM*, 64, 58-65, 2021.
[3] CMS Collaboration, "The High-Level Trigger for the CMS Phase-2 Upgrade", *PoS ICHEP2022*, 209, 2022.
[4] Deng, J., Yang, Z., Ojima, I., Samaras, D. and Wang, F., "Artificial intelligence in drug discovery: applications and techniques." *Briefings in Bioinformatics* 23, no. 1, 2022.
[5] Bishop, B., "Lawrence Livermore National Laboratory achieves fusion ignition", https://www.llnl.gov/archive/news/lawrence-livermore-national-laboratory-achieves-fusion-ignition (accessed 11/1/24).
[6] Jia, Z., Tillman, B., Maggioni, M., and Scarpazza, D. P., "Dissecting the Graphcore IPU Architecture via Microbenchmarking", *arXiv e-prints*, 2019.
[7] Mohan, L.R., Marshall, A., Maddrell-Mander, S., O'Hanlon, D., Petridis, K., Rademacker, J., Rege, V., & Titterton, A., "Studying the potential of Graphcore IPUs for applications in Particle Physics", *Comput.Softw.Big Sci.* 5, 2021.
[8] Kita, M., Dubiński, J., Rokita, P. and Deja, K., "Generative Diffusion Models for Fast Simulations of Particle Collisions at CERN", *arXiv e-prints*, 2024.
[9] Qu, H., Li, C., and Qian, S., "Particle Transformer for Jet Tagging", *International Conference on Machine Learning*, PMLR, 2022.
[10] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *North American Chapter of the Association for Computational Linguistics*, 2019.
[11] Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Fatema Rajani, N., "BERTology Meets Biology: Interpreting Attention in Protein Language Models", *arXiv e-prints*, 2020.
[12] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M., "ProteinBERT: a universal deep-learning model of protein sequence and function", *Bioinformatics (Oxford, England)*, *38*(8), 2102–2110, 2022.
[13] Gopakumar, V., Pamela, S., and Samaddar, D., "Loss Landscape Engineering via Data Regulation on PINNs", *Machine Learning with Applications*, *12*, p.100464, 2022

Accepted Manuscript

9 – Inference Speed is Key to Unleashing the Potential of Large Language Models

Tobias Becker, Maxeler Technologies, a Groq company, 3 Hammersmith Grove, London, W6 0ND, United Kingdom

tbecker@maxeler.com

Status

Large Language Models (LLMs) are a class of artificial intelligence (AI) models that are trained on a vast amount of text data to generate language outputs that are coherent and natural-sounding. These models have become increasingly popular in recent years due to their ability to generate text that is often indistinguishable from human-written text. LLM models represent a significant breakthrough in natural language processing and have the potential to revolutionise many areas by providing a simple human language interface to computer systems that hold large amounts of data. Potential use cases include language translation, text summarisation, content generation, coding assistant, question answering, sentiment analysis, and many more.

LLMs are trained on massive datasets of text data, often from various sources such as books, articles, and websites. During training, the model learns to predict the next word in a sequence of text given the previous words. This process allows the model to learn the patterns and structures of language, such as grammar, syntax, and semantics.

The technology behind LLMs is based on a type of artificial neural network called a transformer [1], one of the most important innovations in the field of AI in recent years. Central to the transformer architecture is a technique called attention that allows the model to focus on different parts of the input text when generating each output word. This allows the model to capture long-range dependencies in the input text and generate coherent and natural-sounding output. Essentially, the attention mechanism offers a solution to the challenge of extreme combinatorial complexity when computing a longer sequence of words, therefore making it possible for the first time to generate high-quality language with computers. Despite this innovation, running inferences on an LLM is still computationally very demanding due to the very large size of the model which often has billions to hundreds of billions of parameters.

LLMs have gained widespread attention with the release of ChatGPT in November 2022 [2] and the field currently undergoes rapid innovation in many areas. Open-source models such as LLaMa have been released [3], and various techniques have been developed to customise LLMs without having to retrain the entire model from scratch [4]. While LLMs are currently limited in speed and accessibility, further advances in models and computer hardware are expected to make LLMs much more available to everyone.

Current and Future Challenges

There are several challenges associated with LLMs and many of those relate to the sheer size of the models themselves. Training LLMs is extremely compute intensive and requires large amounts of high-quality training data. GPT-4, a very large model with over one trillion parameters that was created by OpenAI, is estimated to have taken 90-100 days of training time on a cluster with 25,000 Nvidia A100 GPUs at an estimated cost of over $100M [4]. Training also requires access to a large number of data sources with high-quality text. Training LLMs is therefore an activity that is only accessible to a small

Accepted Manuscript

number of large institutions, and further increasing model sizes will increase this challenge further. It is also unclear if ever increasing model sizes are necessary. Recent work suggests indeed that increasing the model size will lead to better performance [6] while other research disagrees and indicates that some large models will perform worse at certain tasks then smaller ones [7]. Another challenge lies in measuring the quality of an output given by an LLM to which there is currently no single agreed metric.

Another significant challenge with LLMs relates to the fact that these models inevitably have limited domain knowledge. Even if a large amount of training data is used, the model will have limited knowledge of specialised areas such as law or medicine. Furthermore, the information embedded in the model is static and does not capture any new information. LLMs can therefore be limited in their ability to produce usable outputs in specialised use cases or to generalize to new, unseen data such as giving answers related to recent events. Due to the high cost of training, retraining an entire LLM for a specialised use case or new data will not be feasible. Finally, the last challenge relates to the compute complexity of the LLM inference, and the speed needed for practical use cases. For example, LLaMa 2 70B can require, depending on implementation parameters, over 100GFlops to produce a single output token (roughly the equivalent of a word). At the same time, the computation cannot be heavily parallelised because of the sequential dependencies between the tokens in the model. This means that real-world performance of current LLM systems is often sluggish, with single digit to tens of tokens per second being produced.

Advances in Science and Technology to Meet Challenges

The entire field of LLMs currently undergoes rapid development and innovation to address the various challenges mentioned above. One area of research addresses the complexity of the training process with the goal to reduce the overall compute time needed for training [8], but the overall effort for training remains very high. There are also several techniques to reduce the compute complexity during inference such as quantisation, knowledge distillation, or pruning. A large body of research addresses modifications and optimisations of the transformer architecture itself. For example, mixture of expert models [9], multi-head attention for decoding [10], or sparse transformer models [11] all address the compute complexity in the transformer model during the inference process.

There are several techniques that address the limitations of fixed, pre-trained LLMs. Fine-tuning is a training technique that enables adapting a pre-trained model to a specific task or specialised domain, e.g. language translation, question answering, medical knowledge. It is a much more efficient and lightweight process than training a full model from scratch and involves feeding a small amount of task-specific data to the model and adjusting some of the model's parameters. Low rank adaptation (LoRA) is an example of an efficient fine-tuning technique [12]. Other approaches incorporate additional data sources into the LLM. Retrieval-augmented generation (RAG) extends a pre-trained LLM with retrieval model that can feed in relevant information from a separate database and augment the generation process. This is a way of adding new or specialised information that is not embedded in the pre-trained LLM [13].

In order the run LLMs are the speeds needed for practical applications, suitable compute hardware is needed. At present, most LLM services run on GPU hardware resulting in inference speeds of tens of tokens per seconds. However, due to the sequential properties of LLMs, it is difficult to optimise throughput by parallelising across many GPUs. As an alternative, more specialised hardware can be

Accepted Manuscript

used. The AWS Inferentia2 inference accelerator can run various Llama 2 models cost efficiently but faces similar parallelisation challenges as GPUs [14]. Recently, much higher throughput of several hundreds of tokens per second have been reported when using GroqChip, an AI inference accelerator that can handle the sequential properties of LLMs much better than conventional CPUs or GPUs. GroqChip uses a large amount of on-chip SRAM which enables fast, low-latency inference. A dedicated chip-to-chip interconnect between the multiple GroqChips offers scalability for running large models with very high throughput. Results are rapidly evolving and point towards continued improvement on inference speeds across various vendors [15].

Concluding Remarks

LLMs have the potential to provide many benefits across many different sectors and application domains by providing a human language interface to powerful computers that have access to large amounts of data. Adding a human language interface to computers could be similarly transformative to the introduction of graphical user interfaces many decades ago that made computers much more usable by a wider audience. However, current LLM technology faces many challenges and many relate to the sheer size of these models, most notably the high computational effort to train and run LLMs. Training is extremely costly and running inference on LLMs is also costly and too slow for many practical applications. At present, the field is rapidly evolving to find solutions to these challenges. New fine-tuning techniques reduce the need for full model training. Advances in models and techniques combined with new specialised AI accelerator hardware are improving inference speeds. Together, these developments will overcome current limitations and soon lead to LLM performance good enough for a wide range of real-world applications.

References

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[2] Wu, Tianyu, et al. "A brief overview of ChatGPT: The history, status quo and potential future development." IEEE/CAA Journal of Automatica Sinica 10.5 (2023): 1122-1136.

[3] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

[4] Dodge, Jesse, et al. "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping." arXiv preprint arXiv:2002.06305 (2020).

[5] https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

[6] Kaplan, Jared, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).

[7] McKenzie, Ian R., et al. "Inverse Scaling: When Bigger Isn't Better." arXiv preprint arXiv:2306.09479 (2023).

[8] Shoeybi, Mohammad, et al. "Megatron-lm: Training multi-billion parameter language models using model parallelism." arXiv preprint arXiv:1909.08053 (2019).

[9] Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." arXiv preprint arXiv:1701.06538 (2017).

[10] Cai, Tianle, et al. "Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads." arXiv preprint arXiv:2401.10774 (2024).

Accepted Manuscript

[11] Ding, Jiayu, et al. "Longnet: Scaling transformers to 1,000,000,000 tokens." arXiv preprint arXiv:2307.02486 (2023).
[12] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).
[13] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.
[14] https://huggingface.co/blog/inferentia-llama2
[15] https://github.com/ray-project/llmperf-leaderboard

Accepted Manuscript

Acknowledgements

The Guest Editors of this Roadmap would like to thank the workshop scientific committee.

Fast Machine Learning Workshop 2023 Scientific Committee:
Thea Aarrestad (ETH Zurich)
Javier Duarte (UCSD)
Phil Harris (MIT)
Burt Holzman (Fermilab)
Scott Hauck (U. Washington)
Shih-Chieh Hsu (U. Washington)
Sergo Jindariani (Fermilab)
Mia Liu (Purdue University)
Allison McCarn Deiana (Southern Methodist University)
Mark Neubauer (U. Illinois Urbana-Champaign)
Jennifer Ngadiuba (Fermilab)
Maurizio Pierini (CERN)
Sioni Summers (CERN)
Alex Tapper (Imperial College London)
Nhan Tran (Fermilab)

Accepted Manuscript