HENRY ROYCE INSTITUTE

Machine learning for materials science: Community requirements for building a UK capacity

Keith T. Butler, Alin M. Elena, Kim E. Jelfs, Reinhard J. Maurer

Introduction

The emergence of machine learning as a tool for scientific research has opened up new opportunities, but also poses a challenge to the existing infrastructures for teaching and scientific research. Recently, public funders in the UK have responded to these challenges with a number of eye-catching initiatives (such as AI hubs and CDTs focused on AI applications). These initiatives are welcomed, and we want to make sure that these and similar initiatives in the future are community-led and informed by the needs of those working in the field.

We convened a one-day workshop to canvas a wide range of practitioners from materials' science who have been early adopters of the use of machine learning (full list in Appendix A). The cohort consisted of individuals from academia, industry and the public sector, working in experimental and computational materials science, and spanning career stages. We set out to discuss the challenges faced in applying ML for materials science under five broad categories – Education/Training, Infrastructure, Ethics, Open Science and FAIR data, Community building and events. The structure of the workshop (discussed in Methods) allowed us to identify some clear requirements as well as draw out and explore tensions between requirements of different sectors of the materials science community.

Executive Summary

The focused discussions and round-table format allowed us to formulate key requirements for building a world-leading machine learning for materials science research capacity in the UK. Across the five themes of training/education, infrastructure, ethics, Open Science/FAIR data and community building we would advocate for the following:

- Establishing regular, sustainably funded training specifically for practitioners of ML in materials science
- A significant increase in the compute capacity for ML
- A dedicated, sustainably funded infrastructure for data storage and data and model sharing
- Establishing of an ethics framework around ML research in materials science
- A requirement for environmental impact statements in ML research proposals
- Explicit consideration of contributions to Open Science in research assessment exercises for all publicly funded organisations
- Establishing community-led, national-level bodies to act as a focal point of the ML in materials science community, in the form of learned body interest groups, AI hubs and/or an ML in Materials Science network

ROYCE

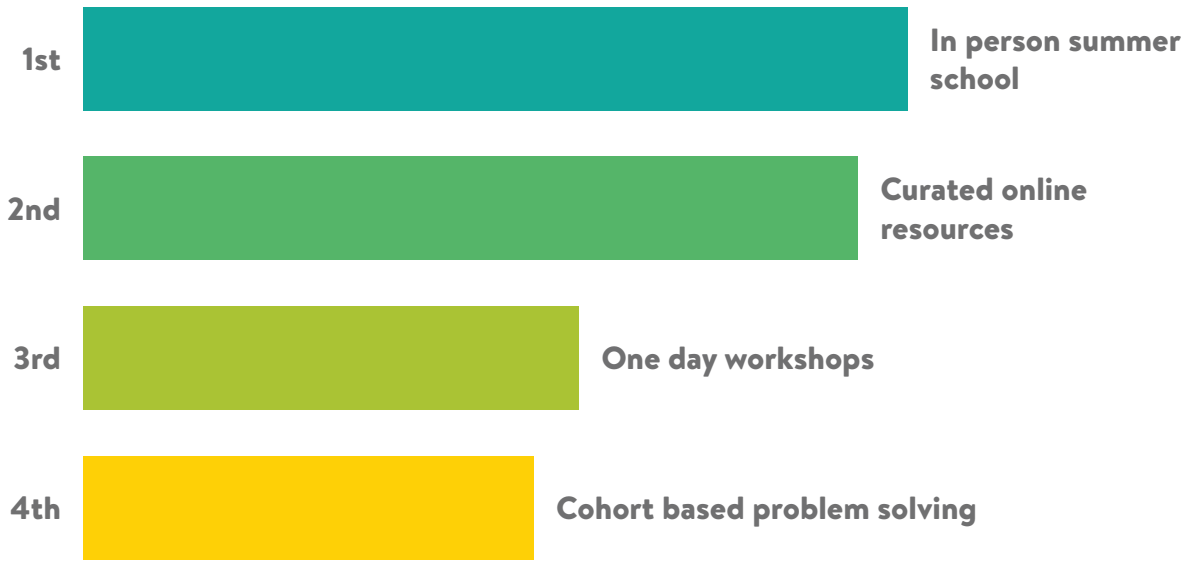
Training/Education

It was clear from our discussions that there is an acute skills shortage across all the sectors involved in our workshop. There is a need for undergraduate (and earlier) curricula to be updated to cover more statistics/probability and linear algebra. There is a demand for curated repositories of trusted online resources. Companies often value specialism and there is a demand for focussed 'seasonal schools' for graduates. Best practice in writing and reading code and dealing with data is severely lacking in graduates at all levels and needs to be urgently addressed.

Our participants identified a wide range of gaps in existing training and education, these are largely summed up by a comment from one participant 'science courses are not numerate enough and computer science courses are not scientific enough' emphasising a disconnect in skills between communities. A common theme was the lack of training in both statistics/probability and linear algebra for undergraduates in physical sciences subjects, ability in these areas is seen as key to being able to critically engage with the latest toolchains in AI and ML. Critical engagement is also identified by the recurrent mention in different groups of the importance of learning to understand the limitations of ML.

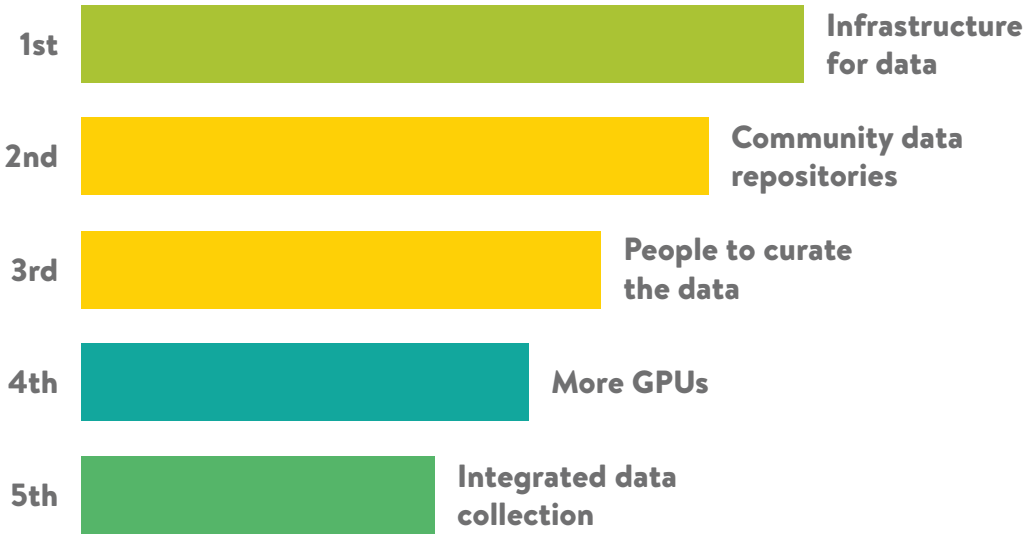
There was a recurring theme of access to training datasets across the groups. In all cases this was framed within the context of learning to apply ML techniques to problems from the materials science sector, rather than from learning on generic datasets. There are exciting materials science datasets, but very few of these are in any kind of format that would be amenable to use for training, or Kaggle-style competitions (which was one suggestion for a training route). The number of datasets that are actually consistent, large and well-curated enough to be used for training purposes are surprisingly small, for example some properties in the Materials Project or AFLOW datasets. This ties in closely with another training need that was identified, which is teaching practitioners about how to deal with and think about data, and spans several of our other themes including, infrastructure, ethics and Open Science/FAIR data.

In terms of training provision there is a clear preference for two options, albeit that they are rather different and potentially complementary approaches. The first approach is to establish traditional international Summer (/Season of your choice) schools, targeted at graduate level students and intended to complement the offering from any CDTs and AI Hubs that will be established in the coming round of funding(see the success of CCP5 Summer School). The second approach is to offer a service that curates existing online training materials and allows materials science practitioners to build training specific to their requirements, but with the assurance that the content has been vetted by the curation service. A popular idea was the synthesis of these two approaches with seasonal schools running and recording material, which would be recorded and archived in a central service, which ensures evolving content that is agile enough to keep up with the latest developments in AI/ML as well as providing fundamentals for those who desire that.



Infrastructure

The discussion on infrastructure generally split into two concerns: data and compute resource. In both cases, the current infrastructure provides strong limitations to research and innovation. Interestingly in the open discussion at the end, there was more discussion of compute resource, but more attendees identified data services as the important consideration. Commitments at the length of at least 10 years are required for both compute and data-management services to bring the UK in line with other countries with similar AI ambitions.



All groups criticised the lack of top of the class AI/ML compute resource in the UK. While the largest academic cluster in the UK has 320 A100s(current best performing general hardware for AI/ML), the JUWELS machine in Germany has 3744, while Summit and Permultter in the US have more than 34800 GPUs. In the commercial sector Google A3 cluster has 26000 H100 Hopper GPU cards while Microsoft's OpenAI has 285000 CPU and 10000 GPU. The UK's academic GPU provision is orders of magnitude behind competitors who also see themselves as world-leading. This is impacting research in the UK – one attendee who oversees compute allocation for their institution spoke of having to cut all applications by 80% of their requests, due to lack of resource, this experience was echoed by others in similar positions and regarding Tier 2 machines. Interestingly, inference GPU nodes on research clusters are reported as being *under* utilised. The lack of centralised AI/ML compute facilities is leading to a proliferation of small group level clusters; this is reminiscent of the HPC situation in the 1990s and is extremely sub-optimal from an efficiency and green perspective.

Collecting, storing and moving data was identified as a blocker by several of the small groups during discussions. While there is a proliferation of places to share data, discoverability is lacking. It was recognised that the Physical Sciences Data Infrastructure (PSDI) is making moves to address some of these problems, but as with compute infrastructure there is a lack of long term funding for maintainable resources to build up this kind of capacity.

Ethics

There may be an impression that ethics is not something that needs too much attention in ML for materials science, however our attendees spoke at length and placed great importance on a number of issues that come under the heading of ethics. There was discussion on culpability in AI, environmental impacts of research and the reproducibility problem. Reproducibility will be discussed in the Open Science/FAIR data section.

Building from the discussion on infrastructure, the proliferation of small local compute resources is leading to wasteful research. In computational science in general there is often little consideration given to the environmental impacts of performing research. The notion of virtual research leads the perception of minimal physical waste. However, with increasing compute capacity, this is patently wrong. According to the carbontracker tool, chat-GPT training took the same amount of energy as 126 homes for one year[1].

Explainable AI and AI with uncertainty quantification were identified as critical factors for industrial

1 https://medium.com/techtalkers/artificial-intelligence-contributes-to-climate-change-heres-how-405ff919186e

applications of AI/ML. Explainability is critical for human-in-the-loop applications, where it is vital that the operator be able to trust the predictions generated by the algorithm. In materials design for safety critical applications, such as aerospace, there is also the requirement of having culpability which will require aspects of explainability in results. Closely linked to this is the issue of uncertainty quantification (UQ), a greater focus is required on developing ML approaches that do not fail whilst providing wildly over-confident predictions.

Linked to all of these points and the earlier discussion in training is the capacity for critical self-reflection of AI/ML practitioners. There was a generally popular suggestion that research for AI/ML should require ethics statements and carbon emissions estimates at the proposal stage, to enforce a culture of reflection. The required contents of such a statement should be developed by further consultation with the research community.

Open Science/FAIR Data (OS-FAIR)

The need for more Open Science and better adoption of FAIR data policies was echoed almost universally by our attendees. Although industrial actors cannot necessarily make their data open they were keen to stress how valuable academically generated open data is to them. To take the analogy from structural biology, this is a field that has had a culture of openness for decades and that is what facilitated AlphaFold, the Protein Data Bank is estimated to facilitate research worth in excess of $1 billion annually, with a return on investment estimated at $8 billion over 30 years[2]. There is evidently value in driving a culture change in the relationship between materials scientists and data, it needs to be driven by educators, funders and employers.

Our attendees had a wide range of suggestions of how to change the data culture in materials science. Several groups mentioned the importance of better training in best practice and tools for OS-FAIR. This could be made a compulsory element of any CDT funding for subjects likely to generate large amounts of data, this is not restricted to AI/ML related programmes. It was suggested that we should develop a Plan-S for Open Science. Such a plan for data would need to be sector specific, but communities can work to develop standards of openness expected for computer code and data arising from research that is publicly funded. There is also a need for employers to recognise the value of OS-FAIR outputs – production of which is highly time-consuming and often ill-rewarded. Writing OS-FAIR criteria into job adverts and promotion rubrics can signal the commitment of institutions to this culture change. Funders could incentivise institutions by including OS-FAIR outputs in the next REF.

Community Building

There was a clear desire from all attending to develop a recognisable UK ML for Materials community. There are currently many disparate groups, companies and institutes working in this area and an abundance of excellent research and innovation, providing coherent focus for these efforts will be critical for realising the full potential of the UK. It is hoped that the outcomes of the AI Hubs for Scientific Data call from the EPSRC will provide a starting point and that whichever consortia prevail will put a strong emphasis on engaging the full breath of this community. There is a clear route to community building through learned society interest groups and perhaps a role for an activity such as a Network grant in ML for molecular and materials science. It is also recognised that these hubs are just a small start and continued sustained funding for research, training and infrastructure will be essential to ensure world leading status for the UK ML for Materials community.

Do we want an interest group for this community?

As a short-term outcome of our workshop, we identified a quick route as establishing a special interest group for ML in materials science. This group will hopefully be hosted across learned bodies, specifically the IOP and the RSC, reflecting the make-up of the community at our meeting. There is a need to develop regular events to provide a focal point for the community. Given the broad nature of materials science, the audience felt that a balance of small focussed one day events, with regular annual events with a wider scope would be the best solution.

2 https://cdn.rcsb.org/rcsb-pdb/general_information/about_pdb/Economic%20Impacts%20of%20the%20PDB.pdf

Acknowledgments

The organisers thank the Henry Royce Institute for sponsorship of the event.

Appendix A: Attendees

Gabor Csanyi; University of Cambridge
Linjiang Chen; University of Birmingham
Jon Rowe; University of Birmingham
Tom Whitehead; Intellegens
James Cumby; University of Edinburgh
Joao Fonseca; University of Manchester
Alessandro Troisi; University of Liverpool
Jochen Blumberger; UCL
Matthew Foulkes; Imperial College
Chris Pickard; University of Cambridge
Laura Ratcliff; University of Bristol
Pui-Wai (Leo) Ma; UKAEA
Rob Akers; UKAEA
Mark Storr; AWE
Ricardo Grau-Crespo; University of Reading
Matt Probert; University of York
Samantha Kanza; University of Southampton
Jeyan Thiyagalingam; STFC
Simon Coles; University of Southampton
Tim Albrecht; University of Birmingham
Ed Pyzer Knapp; IBM
James Kirkpatrick; Deep Mind
Sergei Dudarev; UKAEA
Chris Race; University of Manchester
Volker Deringer; Oxford University
Kim Jelfs; Imperial College
Keith Butler; Queen Mary University of London
Alin Marin-Elena; STFC
Reinhard Maurer; University of Warwick

ROYCE