Science and Technology Facilities Council
Hartree Centre

IBM

HARTREE NATIONAL CENTRE FOR DIGITAL INNOVATION

PROGRESS REPORT
MAY 2023

May 2023

Hartree National Centre for Digital Innovation: Progress Report

June 2021 – December 2022

Charlotte Glass, Lizahn van Gend, Neil Brown

Table of Contents

Executive Summary

Following the launch of the **Hartree National Centre for Digital Innovation** (HNCDI) in June 2021, STFC Hartree Centre and IBM have started to develop innovative solutions to industry challenges. The HNCDI programme utilises the powerful computing techniques, world class expertise and one-of-a-kind infrastructures available to UK businesses/public sector organisations at STFC Hartree Centre, in collaboration with IBM.

Delivered across six core strands of activity and supported by a central programme management team, HNCDI supports collaborative R&D projects (through Explore, Excelerate and Emerging Technology), delivers training and SME engagement (through Explain and the forthcoming HNCDI Hubs) and hosts a wide range of e-infrastructure.

This document presents the **first report** of the HNCDI programme as part of the monitoring and evaluation support provided by Technopolis. The data collected covers the period from the start of the programme, **up to and including December 2022**. Approximately a third of the way into the programme, HNCDI has established most of its core workstreams of activity and is overall making good progress against its original targets.

The programme has **launched 44 projects** as of December 2022 across the Emerging Technology, Explore and Excelerate workstreams, 9 of which have already completed. The programme has also placed more focus in these first 18 months on building a pipeline of projects to feed into the Excelerate workstream in future. Nonetheless these three workstreams have engaged **22 unique partner organisations** and already guaranteed **£157,080** from industry collaborations, including capability-build projects.[1] Explore, Emerging Technology and Excelerate projects have already begun producing codified knowledge, such as **ten peer-reviewed publications,** while **two patent submissions** are also in progress.

Through the Explain workstream – a dedicate platform of free training courses tailored to industry challenges – **544 individuals from 239 different organisations have attended.** Attendees come from a range of industry and public sector organisations. The courses have been well received, with around 80% of attendees reporting positive or very positive experiences about the relevance of the training provided and their high level of satisfaction with its delivery. In addition, the success of these courses and the lessons learnt from the implementation of the workstream have attracted interest from other centres across Europe interested to learn from the pioneering approach of HNCDI.

	Projects Launched			Explain attendees	
	Explore	Emerging Technology	Excelerate	People	Companies
End of programme target	104	20	32	5,000	500
Progress as of 18 months	33	7	4	544	239
% of target reached	32%	35%	13%	11%	48%

[1] In-kind contributions up to December 2022 were based on calculating the median value of staff time in consultancy x number of days reported across project master documents. Tracking commenced for further in-kind contributions.

Another measurable output of the programme is the increase in the STFC Hartree Centre's **e-infrastructure capacity**. The provision of public cloud to the HNCDI programme has now been set up and use is steadily increasing. The construction of the **supercomputing centre** began at Sci-Tech Daresbury in January 2023 and is on track for completion in the summer of 2024, after which the supercomputer will be installed and available from April 2025.[2]

In the coming months, the programme will also launch the **HNCDI Hubs**: a network of three Hubs across the UK established with the purpose of engaging their local or sectoral networks, facilitating the increased adoption of digital technologies such as AI, and up-skilling SMEs related to each regional area. In its first 18 months, HNCDI has therefore gained momentum in reaching its targets and impacting on the use and adoption of digital technologies across UK industry.

[2] https://www.hartree.stfc.ac.uk/news/2023/02/13/new-30-million-supercomputer-centre-at-daresbury-laboratory/

1 Introduction

1.1. Hartree National Centre for Digital Innovation

The Hartree National Centre for Digital Innovation (HNCDI) was launched in June 2021 to drive research into and increase the adoption of digital technologies including artificial intelligence (AI) and quantum computing in the UK. Based on a five-year collaborative agreement between the STFC Hartree Centre and IBM, it aims to support UK businesses and public sector organisations both large and small, in their journey to develop innovative solutions to common industry challenges.

The fields of AI and quantum computing have gained strategic importance in recent years.[3] Globally, the application of AI is estimated to add over US$3.5 trillion to the economy on an annual basis,[4] while the value of potential use cases for quantum computing are (conservatively) estimated at up to $700 billion by 2035.[5] The UK's science minister has expressed their intention for the UK to lead these fields. HNCDI is therefore a demonstrator of the key step taken by the UK to meeting that goal. Specifically, the programme seeks to overcome the practical barriers associated with adopting these technologies at industry level, by providing training and access to the computation facilities of both the Hartree Centre and the IBM Quantum Network, as well as access to the specialist knowledge and expertise of their staff.

There are five main workstreams delivered under HNCDI to address four key stages of digital adoption. These are:

- **Explain** – the provision of training courses that are accessible and application focussed. Courses span a range of topics in the digital economy, and are presented in a variety of formats, from self-paced to live sessions.
- **Explore** – collaborative projects that target the frontier of AI and digital technologies. Real-world problems are identified and proof-of-concept solutions with high potential value to industry are developed either in collaboration with external partners, or by HNCDI researchers with external validation.
- **Excelerate** – these collaborative projects aim to progress proof-of-concept and capability build projects, through to solutions and IP that are ready for adoption by specific industries.
- **Emerging Technology** – projects combining established high performance computing with emerging technologies like quantum computing, aiming to improve the understanding around their potential for industry application. These projects take place in collaboration with external partners or are led by HNCDI researchers with external validation.
- **SME engagement** – an awareness raising activity in which the HNCDI programme establishes three HNCDI Hubs in collaboration with delivery partners. These Hubs will engage with SME communities across the UK to start their digitalisation journey and deliver short projects/interventions as well as training.

[3] See for example UKRI's most recent strategy release. https://www.ukri.org/publications/ukri-strategy-2022-to-2027/ukri-strategy-2022-to-2027/

[4] https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20ai%20frontier%20applications%20and%20value%20of%20deep%20learning/notes-from-the-ai-frontier-insights-from-hundreds-of-use-cases-discussion-paper.pdf

[5] https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/quantum-computing-use-cases-are-getting-real-what-you-need-to-know

The above workstreams are underpinned by two further strands of activity:

- **E-infrastructure** – the procurement and construction of a new energy-efficient Data Centre at Daresbury Laboratory; the provision of a new AI-optimised High-Performance Supercomputer; procurement and delivery of a hybrid cloud platform.
- **Programme management** – support to programme delivery through project management, programme monitoring, evaluation, and communications activities. Cross-programme activities, including recruitment and running seminars and workshops outside of workstreams, have been included in this area.

1.2. This report and methodology

This document presents the first of three progress reports for the HNCDI programme. It reports on progress and emerging results for the period since the launch of the programme in June 2021 until December 2022. The methodology used to prepare this report has been aligned with the objectives set out in the Evaluation Framework Report (May 2022), and combines:

- **Desk research and review of programme documentation** – Programme monitoring information including HNCDI KPIs, project management documents (PMDs), and project surveys, were shared by the HNCDI team and reviewed by the study team.
- **Stakeholder interviews** – A series of 11 interviews with the programme delivery team were conducted. This included workstream leads from both the Hartree Centre and IBM, to better understand the progress of workstreams and the programme overall and to identify relevant case studies. Wherever necessary, interviews were also held with external project partners involved in the selected case studies. The list of interviewees is presented in 0 at the close of this report.
- **Case study development** – A set of four case studies focussing on projects and participants engaged in four of the programme's workstreams (Explain, Explore, Excelerate and Emerging Technology) to exemplify emerging outputs and outcomes of the programme. Each case study was developed using a combination of desk research, analysis of project monitoring information and interviews.
- **Regular conversations with the impact, evidence and evaluation manager at the Hartree Centre** – Fortnightly online meetings were held with the Hartree Centre's impact, evidence and evaluation manager to discuss programme progress across the targets.

The remainder of this report is structured as follows:

- **Section 2** presents an overview of the implementation of the HNCDI programme and emerging programme level benefits.
- **Section 3** presents the progress of the Emerging Technology, Explore and Excelerate workstreams and the emerging outputs and outcomes of the programme including an overview of the case studies presented in Appendix B.
- **Section 4** presents the HNCDI's training and engagement activities, Explain and the HNCDI Hubs.
- **Section 5** presents the progress and benefits emerging from the HNCDI's e-infrastructure.

Appended is a list of stakeholders interviewed for this first report (Appendix A) and full-length case studies (Appendix B).

Future reports will also draw from surveys completed by external partners and HNCDI project leads prior to project commencement, at project close and 6-9 months after project end. These surveys are capturing data and details of outputs and outcomes to beneficiaries.

However, as of December 2022, only 4 project closure surveys had been completed and no follow-up surveys, due to the number of projects that have gone through the programme in this timeframe. Given the small number of datapoints, it would not be appropriate to present aggregate analysis from the surveys at this stage.

2. Implementation of the HNCDI programme

It is clear that HNCDI has made substantial progress in establishing the activities and launching the various workstreams since it began in June 2021.

The **programme is on-track** and has **expanded the capacity of the Hartree Centre and IBM** with 45 new staff employed to deliver the programme's activities. Although there are current concerns globally about the competition for skills and recruitment, the HNCDI programme has so far employed 37 staff within the Hartree Centre and 8 within IBM Research UK. The global concerns around recruitment remain a tangential issue for the HNCDI team to monitor, in order to meet current targets by 2025.

The HNCDI programme has **strengthened the collaboration between the Hartree Centre and IBM**. Projects include both Hartree Centre and IBM staff, which has improved mutual understanding and increased appreciation of their complementary skills and capabilities, as well as refining the internal processes required for successfully delivering R&D projects. This has also been facilitated by the HNCDI's new management and governance structure, providing support to each workstream, and to cross-workstream coordination and knowledge sharing.

Interactions between the programme workstreams are beginning to emerge now that most are up and running. For example, team members from the Emerging Technology workstream are now working to produce an Explain training course to provide an introduction to quantum computing. This course will also build on IBM's materials and insights in this area. Reflecting on attendee feedback, the new quantum course is developing in part as a response to the higher-than-expected levels of interest in quantum computing. HNCDI is therefore engaging with and acknowledging industry demands across the entire programme.

In addition, the leadership and management of HNCDI has focussed its efforts on developing a pipeline of projects to move from Explain to Explore and Emerging Technology, through to Excelerate. This benefits from and builds upon the strength of the central management of the programme and collaborative working in coordination across the core workstreams.

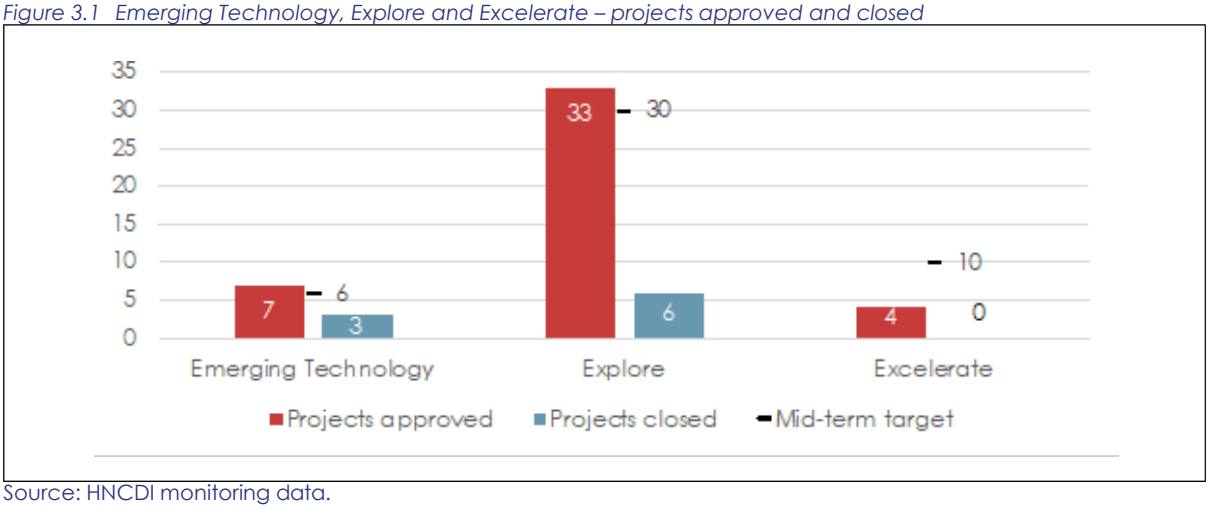
The programme is also enabling **new types of collaboration and support** to be delivered by the Hartree Centre, and in doing so has provided valuable insights and learnings for how to better support collaborators in future. For example, the Explain workstream is a novel initiative for both the Hartree Centre and for STFC more widely and in these first 18 months it has provided valuable input to identify the training needs of organisations interested in and working to adopt AI and other emerging digital technologies into their workflows.

In launching Excelerate projects, the Hartree Centre and IBM have worked hard to navigate the legal challenges relating to launching collaborative projects of this nature. In doing so, both organisations have adopted solutions for developing a strong **pipeline of potential projects and collaborators** to engage with HNCDI as the programme continues.

3. Collaborative R&D Projects

To execute collaborative R&D projects, the HNCDI programme supports three key areas – **Emerging Technology, Explore and Excelerate** – each of which is focussed on technologies at different stages of development. Across these three workstreams, the HNCDI programme has **launched 44 projects** as of December 2022. For Emerging Technology and Explore, the number of projects that have been approved already exceeds the targets set for the programme's mid-term point (June 2023). The Excelerate workstream is currently behind its mid-term target, having approved four of the ten projects originally expected. However, the numbers here are expected to increase in future as the pipeline evolves and Explore projects continue to close.

Figure 3.1 Emerging Technology, Explore and Excelerate – projects approved and closed



Source: HNCDI monitoring data.

By December 2022, the three collaborative R&D workstreams have engaged **22 unique partner organisations**, including two academic institutions. This comprises organisations from a range of sectors and of a range of different sizes. The three workstreams have received approximately **£157,080** of **company contributions** from these partner organisations demonstrating the value of collaborative R&D carried out through HNCDI.[7]

The HNCDI programme continues to support the production of new codified knowledge, evidenced for example by the execution of **ten peer-reviewed publications**.

In addition and only 18 months into the programme, there are already **two patent submissions** in progress from HNCDI-supported projects, as well as two recorded instances of **technology transfer** into UK businesses.

As well as building capabilities in emerging technologies and knowledge assets across a variety of Explore projects, HNCDI also aims to **increase the confidence in technological solutions for digital transformation and innovation** of its partner's. This continues to be systematically captured through the planned surveys and will therefore be reported on in future reports, however there are already examples of such benefits flowing through to partner organisations.

[7] Estimate based on staff time. HNCDI monitoring data.

REPROCELL used HNCDI to accelerate the process of generating trained machine learning algorithms used in precision medicine strategies. The project involved working with the STFC Hartree Centre and IBM to develop an AI-powered platform capable of simplifying complicated health datasets. Through this type of collaboration, pharmaceutical companies can streamline drug development processes and identify more effective treatment for those suffering from Inflammatory Bowel Disease (IBD). This work has the potential to be expanded out into cancer therapies and could transform drug development, clinical trial timescales and efficiencies.

Short overview of Excelerate case study. Full case study in appendix B.

In launching **Excelerate** projects, the Hartree Centre and IBM have worked to navigate the legal challenges relating to launching collaborative projects of this nature. In doing so, both organisations have developed solutions for adapting a strong **pipeline of potential projects and collaborators** to engage with HNCDI as the programme continues. Though the workstream has launched fewer projects than originally forecast at the outset of the programme, the management teams have reprofiled the phasing and timeline of the workstream. This new profile better reflects the pipeline of potential projects the programme is building through Explore and the forthcoming **HNCDI Hubs** (detailed in Section 4).

In addition, interest in the **Emerging Technology** strand has been far greater than originally anticipated, with a wide range of organisations expressing particular interest in quantum computing. This was attributed in part to the widening publicity and increased discussion around quantum computing, stoking interest both in the technology and the rapid pace of its development in this area.

Emerging Technology projects have already begun to successfully demonstrate the potential applications and benefits, of quantum computing, including for example in the application to drug discovery. See for example a short overview of the case study for *Emerging Technology: Quantum Machine Learning enabled drug discovery*.

The discovery process for new drugs is known within the healthcare industry to be notoriously time-consuming and expensive.

Through the Emerging Technology workstream, the HNCDI team investigated the potential of quantum computing techniques to speed up the drug discovery process and bring down costs through Ligand-Based Virtual Screening, a computational technique that screens digital databases of molecules to identify structures most likely to bind to a drug target. To undertake the rapid screening, the project demonstrated that application of the Quantum Support Vector Classifier – available through IBM's open-source software developing kit Qiskit – often performed better than the equivalent classical algorithm and in some instances greatly outperformed the deep learning methodologies that are currently state of the art in drug discovery.

Training traditional machine learning algorithms can be prohibitively resource intensive due to the amount of information available on molecule properties, so quantum-enabled machine learning offered a potential advantage by supporting calculations at exponentially higher-orders of complexity to increase efficiency and accuracy.

Short overview of Emerging Technology case study. Full case study in appendix B.

The **Explore** workstream is also actively supporting the development of new capabilities and application prototypes (proofs of concept). Explore has two core objectives: to identify early-stage digital technologies with strong industry potential, and to develop stacks of complementary projects with solutions that when used together offer value greater than the sum of their parts. In delivering against these objectives, Explore works in synergy with Excelerate to develop a pipeline of early-stage technologies with strong industry potential to be translated into a commercial product.

To identify early-stage technologies with the potential for wider industry applications, Explore runs a series of calls for proposals in which STFC, IBM and a sector representative collate promising projects within STFC and IBM to compile an industry challenge statement.[8] Interested collaborators are then invited to match their particular problems to the early-stage solutions identified within the challenge statement, which in turn ensures alignment of projects to real use cases. Each call and its external representative is specific to a given sector, such as the materials and automotive sectors.

Explore has so far developed three stacks, bringing together projects also supported by the Excelerate workstream, including:

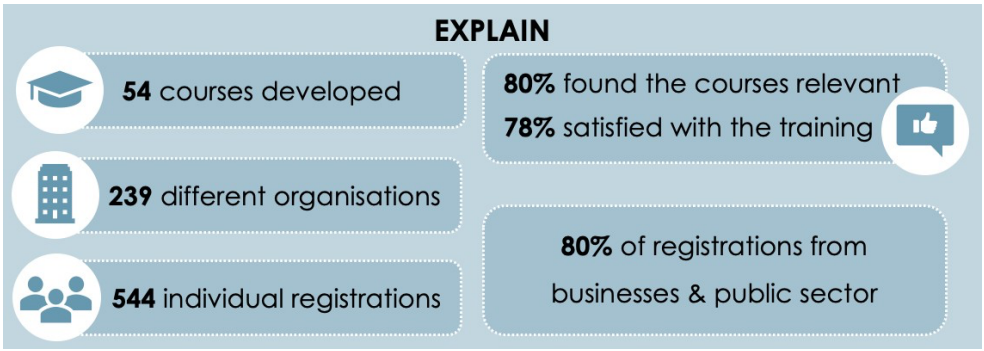
- A materials simulation platform to automate the materials discovery process, from identification of potential candidates through to simulating material properties,
- A microbiome/multi-omics stack to support personalised medicine which also includes a collaboration with REPROCELL and
- A geospatial stack. A platform with the ability to map climate change events and predict the impact of weather events (see case study in appendix B *Explore: AI flood detection*).

[8] See for example https://www.hartree.stfc.ac.uk/news/2022/11/17/materials-call/ and https://www.hartree.stfc.ac.uk/explore/2023/02/28/automotive-call/

4. Training and Engagement - Explain and HNCDI Hubs

To develop training and engagement, the HNCDI programme supports two parallel workstreams: Explain and the HNCDI Hubs.

The **Explain workstream** develops training targeted at mid-career skills development, with the aim of delivering this to **5,000** individuals by the end of the programme.



As of December 2022, Explain has developed and delivered 54 different training programmes. Courses are offered across five technology themes (HPC and Exascale, Data Science, Software Engineering, Hybrid Computing and AI Modelling) and three skill levels (Introductory, Learner and Intermediate). Explain's focus on mid-career training reflects international demand for high performance computing (HPC) up-skilling to practitioner level as well as for senior management.[9] The latter is crucial for decision-maker buy-in; training that demonstrates the applicability of HPC and HPC skills to their use-cases can nudge companies toward investing in HPC capability.

Table 4.1 Explain training courses developed

Technology theme	**Level**				**Total**
	Introductory	Learner	Independent user	Practitioner	
HPC & Exascale	3	3	5	0	11
Data Science	4	4	2	0	10
Hybrid Computing	4	5	1	0	10
AI Modelling	7	4	2	0	13
Software Engineering	2	5	3	0	10
Total	*20*	*21*	*13*	*0*	*54*

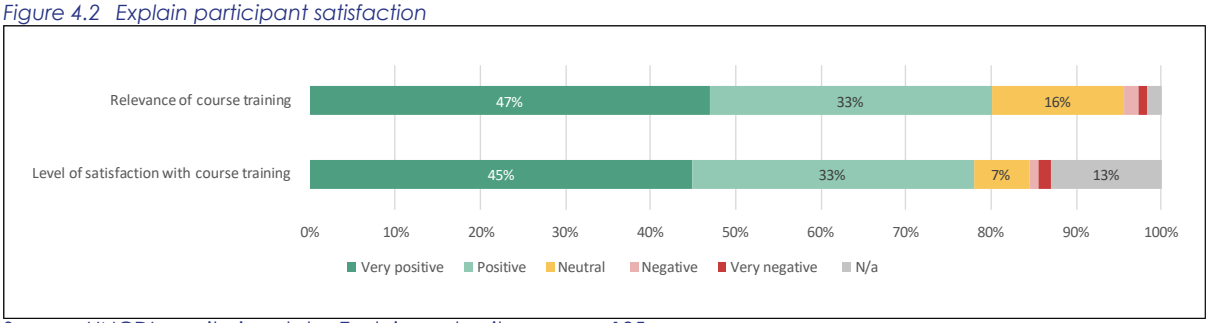
Having launched the first courses in September 2021, **544 individuals** have attended the courses as of December 2022. Though this is only 18% of the original mid-term (June 2023)

[9] https://www.scientific-computing.com/news/high-demand-supercomputing-skills-ichec

target of 3,000 set at the outset, the uptake of training is expected to increase in future with the roll-out of the online learning management system, the introduction of self-paced learning and the launch of the HNCDI Hubs (presented below). Further expansion and development of the Explain workstream has also led to the planning of more practitioner-level courses for future training. Explain targets companies for training, aiming to present their courses to 500 unique organisations. The workstream has reached nearly half of this target already, having registered the attendance of **239 organisations** (95% of the mid-term target).

Individual training attendees to Explain courses also reported high levels of satisfaction overall. 80% were very positive about the relevance of the course and 78% were satisfied with the training (see Figure 4.2).

Figure 4.2 Explain participant satisfaction



Source: HNCDI monitoring data: Explain post-exit survey. n=185.

As an example, Explain is providing the glass manufacturer NSG with training from a trusted source, that has benefitted a range of teams, from materials science through to legal teams. This has also increased the confidence of NSG staff in framing, communicating and engaging with data science projects.

NSG Pilkington manufactures and processes glass, offering specific glass technology in the architectural and automotive sectors. To add value to its portfolio, the company collaborated with the STFC Hartree Centre on two materials discovery projects, which inspired a company-wide recognition of potential data science applications. Consequently, NSG piloted a programme to train their staff in data science through the HNCDI Explain training programme, acknowledging the Hartree Centre as a reputable and trusted source working at the cutting edge of digital technologies. NSG Pilkington were drawn to the flexible delivery and range of experience levels, which ensured consistency in approach and terminology.

From materials science teams to legal, staff have since registered for 61 courses across both introductory and advanced levels, reporting improved confidence in framing and communicating their requirements to data science experts. Other feedback suggested the courses have improved readiness to tackle internal data science challenges. Pilot participants also recommended Explain to new data science recruits, demonstrating the usefulness and wide application areas of the range of courses on offer.

Short overview of Explain case study. Full case study in appendix B.

At the start of the programme, Explain invested heavily to adapt content for businesses and non-academic audiences. Though it was no small challenge to develop courses aimed at end-users while maintaining technical integrity, the result has been a success, with a variety of skill-levels catered to. Close to 80% of Explain course registrations are non-academic users from businesses and the public sector, in contrast to comparable HPC courses which attract mostly

academic participants. For example, the HPC Centre Stuttgart (HLRS), which has traditionally offered training tailored to academics, launched its Supercomputing Akademie for industry in 2018.[10] HLRS have sought advice from the Explain team to reach similar levels of business participation. Further demonstrative of Explain's success, the training offering has also attracted interest from representatives from supercomputing centres in Finland, who want to learn how their courses can appeal to industry users.

A key facilitator for Explain's future activities was the launch of the learning management system (LMS) in October 2022. Content has been adapted for the online LMS to allow some courses to be taken in a self-paced format – particularly the introductory courses – which is more practical for both the attendees and the staff who present these courses. The LMS is also planned to deliver training packages of pre-selected courses that cater to particular industry challenges. In combination, these additions will drive Explain's accessibility and relevance even further.

Adjacent to Explain and linked to the activity of the HNCDI Hubs, the programme also hosts workshops and seminars relating to business development. A target of 500 unique companies in attendance at these events has been set, and as of December 2022, attendance has reached 410 which surpasses the initial timeline. This also suggests that there is a clear link between the offerings of HNCDI and business needs across various industry sectors.

Alongside the establishment and delivery of the Explain workstream, there has been a robust selection process for the **HNCDI Hubs,** which ran until January 2023 through a competitive call. Three HNCDI Hubs have since been selected and will be launched in the coming months. They will deliver support to organisations within their local or sectoral networks, facilitate engagement with the HNCDI's Explain training courses and engage in more collaborative R&D projects.

5. E-infrastructure – Cloud, On-Premise Cloud and the Supercomputing Centre

Finally, HNCDI supports 3 strands of activity related to the Hartree Centre's compute capacity.

The public cloud provision is online, and use is steadily scaling up as more HNCDI projects begin. The HNCDI programme has also established a new process for securing access to public cloud platforms such as Google and AWS to support the delivery of HNCDI projects.

The procurement of the **on-premise cloud** for the HNCDI programme has completed and it is now fully operational. It builds the capabilities of HNCDI projects so that they can be applied across multiple domains and has been used across a wide range of Explore and Excelerate projects up to December 2022. It will continue to benefit HNCDI projects across the programme's lifecycle.

Public cloud provision is in place to support skills development, the packaging and deployment of HNCDI developed solutions, and to provide access to a range of cloud services. In the first year of the programme, the uptake of the public cloud resource has not been as high as originally expected as the demand for this is, for the most part, tied to the delivery of Excelerate projects. In the coming year, the HNCDI programme will train further staff

[10] https://www.hlrs.de/news/detail/the-supercomputing-akademie-a-new-hpc-training-program-opens

to manage and support the adoption of the public cloud resources while formalising the processes for onboarding once a pipeline of projects for Excelerate is established.

The access to public cloud provision provided through HNCDI is a new capability for the Hartree Centre and it has already raised interest from teams and projects beyond the HNCDI programme. This indicates that this new route for accessing public cloud resource, and the knowledge within the HNCDI team in delivery and supporting access, will have further spill-over benefits to projects and organisations not directly involved in the HNCDI programme. In addition, the provision has also created opportunities for wider team members within The Hartree Centre to begin exploring the public cloud platforms.

The centralised management and delivery of public cloud provision through the HNCDI programme has provided the Hartree Centre with better insights into what users need and what they are implementing within projects. A better understanding of what is used, when and why, supports stronger management of the compute needs of projects. This further improves the profile of the use of technical resources within the Hartree Centre's portfolio of projects. Historically, it has been very challenging, if not practically impossible, for the Hartree Centre to accurately estimate the compute needs for proposed tasks. This new understanding and the growing evidence base that will be collected as more HNCDI projects make use of the cloud resource will help to improve estimations of compute resource needs for future iterations, improving the productivity and management of R&D carried out at the Hartree Centre.

Construction of the supercomputing centre also began at Sci-Tech Daresbury, the home of the Hartree Centre, in January 2023 and is on track to complete by summer 2024, with the supercomputer installed and available for April 2025.

Wider contextual factors have had a significant impact on the design and delivery of the centre. In reaching Stage 3 of the design process for the building, the war in Ukraine and fluctuations in steel and construction prices meant that costs increased sharply. To limit and mitigate the impact of this, the Hartree Centre and the delivery partners applied value engineering principles to critically assess and review the design requirements for the supercomputing centre. As a result, some aspects of the initial design have been sacrificed.

Nevertheless, the new supercomputing centre will provide expanded capacity and flexibility for the Hartree Centre in the longer run. The final design also includes allowances for improvements and additions in future, for example for the addition of solar panels or a ground source water cooling system. Additional considerations have also been given to the materials for the construction of the supercomputing centre, including using low carbon cement and an easy to deconstruct steel frame for the building for reuse after the end of its life. The procurement of the supercomputing centre has therefore been considered with its environmental impact at the forefront of its planning and despite the significant challenges and delays, delivery is only six months behind schedule.

To conclude, this report has outlined the progress, successes, and challenges for building up the Hartree National Centre for Digital Innovation in its first 18 months. Overall, significant progress has been made meaning that the programme's momentum can now continue. Further effort to monitor and evaluate the emerging benefits of HNCDI will be carried out as per the Evaluation Framework (2022). See the appendices below for a more detailed summary of the interviewees relating to this report and fuller case studies for each workstream.

Interviewees

Area of expertise	Name	Organisation	Role
HNCDI Staff			
STFC Hartree Centre	Katherine Royse	STFC	Hartree Centre Director
Programme management	Sergio Malo-Peces	STFC	HNCDI Technical Programme Manager
Programme management	Michael Bradley	IBM	HNCDI Programme Director
	Peter Waggett	IBM	IBM UK Director of Research
Explore	Richard Anderson	STFC	Head of Hartree Centre Chemistry and Materials Explore Lead
	Edward Pyzer-Knapp	IBM	AI Enriched Modelling and Simulation Lead Explore Lead
Excelerate	Jonathan Smith	STFC	Business Development Manager
	Chris Gibson	IBM	UK Programme Manager Excelerate Lead
Emerging Technology	Luke Mason	STFC	High Performance Software Engineering Lead Emerging Technology Lead
	Jason Crain	IBM	Senior Technical Staff IBM Research Europe Emerging Technology Lead
Explain strand	Nia Alexandrova	STFC	Training and Events Manager; Explain Lead
HPC and Cloud Procurements & supercomputing centre	David Cable	STFC	Associate Director, Infrastructure and Compliance
HPC and Cloud Procurements & supercomputing centre	Mark Mawson	STFC	Technical Programme Manager
Interviewees contributing to case studies			
Explain case study	Rory Back	NSG	R&D Incubator Technical Manager
	Graham Siddons	NSG	Team Leader, Simulation
	Jonathan Bracken	NSG	Advanced Technologist
	Mark McLachlan	NSG	Technical Manager, Systems and Control
Explore case study	Louise Butcher	STFC	Senior Data Scientist
	Paolo Fraccaro	IBM	Staff Research Scientist
Excelerate case study	Karen Bingham	REPROCELL	COO
	Graeme Macluskie	REPROCELL	Director of Precision Medicine

2.1 Explore: AI flood detection for the UK.

Workstream: Explore	
Project: AI-powered multi-source flood detection for the UK	**Timeframe:** Aug 21 – Dec 22
Partners involved: n/a	

Mapping the risk of climate events such as flooding requires the analysis of massive datasets, usually done through a time-consuming, semi-manual process.

Through HNCDI, the STFC Hartree Centre and IBM have developed a machine learning algorithm capable of identifying past and current flood events using satellite data. The team trained the algorithm using suitable open-access ground-truth data. The capability was then verified against a set of known flood events.

The application of AI significantly accelerates the process of identifying and labelling flood event data, which can then be fed into a suite of related modules developed under HNCDI Explore, all hosted on a cloud platform also developed through HNCDI, called the Geospatial Discovery Network (GeoDN). When used in combination, the platform can predict which areas are at high risk of flooding and be used to inform decisions around infrastructure development, maintenance plans or crisis management.

The project suite is highly versatile and can be adapted to other climate events, like wildfires, or to use data from optical, unmanned aerial vehicles (UAVs) or other radar sources.

Introduction

Severe flood events have increased in recent years, and the trend is expected to continue for as long as climate change persists.[11] These events pose a major threat to human life and to critical infrastructure, with over 29 million people affected and US$82 billion in losses worldwide in 2021.[12] The ability to predict the occurrence of floods could save many lives and impact livelihoods. The adoption of digital technologies such as machine learning to combat this challenge has a significant part to play given the surface area which needs to be covered and the number of natural and man-made systems that interact to produce floods.

A critical step toward prediction is identification; insights about future flooding behaviour must be based on real-time and past data. Satellite imagery is a crucial source of intelligence about prior flood events, but the datasets are too vast for manual searches so that many floods go unrecorded in the public domain where predictive AI can analyse them. To address this, the AI-powered multi-source flood detection for the UK project, launched in 2021 under the

[11] https://www.metoffice.gov.uk/research/climate/understanding-climate/uk-and-global-extreme-events-heavy-rainfall-and-floods

[12] See https://www.itij.com/latest/news/flooding-drove-global-economic-losses-natural-catastrophe-2021 and https://www.statista.com/statistics/273897/natural-disaster-victims-by-continent-and-type-of-catastrophe/

Explore workstream, aimed to develop an AI module that can identify flooding events in IBM's satellite image database.

This project was part of a series of interlinked flood projects within HNCDI's Explore portfolio that together continue to produce this predictive capacity (among other things). The AI capabilities using satellite data in this project feed into the other flooding-related projects within the portfolio. This includes an AI flood risk project which uses the detection module's output developed in this project, alongside other geographical data to identify areas at high risk of flooding. It also feeds into the Geospatial Discovery Network (GeoDN) cloud-based platform, which has been developed to host and run geospatial and climate-related analysis modules (including those related to flooding) in a way that is configurable and scalable.

Project activities

The first step in the delivery of the machine learning model was to establish a ground-truth; the team used an open access database of known flood events and mapped them onto satellite data from IBM's Geospatial Analytics platform. The model was trained using this ground-truth and then validated by running it against known floods events in the UK.

A key enabler to project delivery was the diversity of skills available to the project team because of the STFC/IBM collaboration. Among the disciplines consulted were climate scientists, geography experts, data scientists and cloud software engineers. A challenge to the delivery of this particular project was its interdependence with the other HNCDI flood projects, most notably the GeoDN platform. While GeoDN was developed, an interim solution was found by taking advantage of other supercomputing capabilities available through HNCDI. In this instance Scafell Pike enabled the development of the model until its transfer to the custom-built GeoDN.[13]

Project impact

The project has successfully produced an algorithm that can identify flood events with a great degree of accuracy while utilising available satellite image datasets. Crucially, the code developed works effectively on data external to what it was trained on, highlighting its generalisability and readiness for real-world use.

The project's potential impact is enhanced by its inclusion in the greater flood risk portfolio of projects. For example, by running it on the GeoDN platform which streamlines the connections between the AI module and the datasets it draws from, the model can be easily customised and deployed. The flood events identified by the detection module can also be used to validate the output of the flood risk module, generating flood risk maps and further demonstrating the potential for reusability. Lastly the same code base used in the project to develop the AI model could also be applied to other types of remote sensing data (UAV, for example) to focus on floods at a higher resolution, or other relevant climate hazards. Given the potential of the solution developed here, both in isolation and in combination with the larger flood portfolio, the Hartree Centre and IBM are currently in conversation with several end users about scaling up and commercialising the end product.

In addition to the module's capability, the project has developed the skills of both the Hartree Centre and IBM teams, as well as allowing for the cross-pollination of skills between data science and AI. Both teams are now more experienced in the use of satellite data, a skill which can be applied to a range of use cases including other climate events like wildfires.

[13] https://www.hartree.stfc.ac.uk/technologies/high-performance-computing/

Wider impacts

As of February 2023, there remains an ongoing conversation around the industrial adoption of this model, and with potential end-users to take this capability forward. Prospective applications in the future are numerous. In addition to flood monitoring, the suite of projects developed under the HNCDI Explore workstream can be applied to other climate events such as wildfires once the necessary ground-truths have been established.

In isolation, the code developed by the flood detection project can be used to generate insight to help infrastructure developers optimise their maintenance plans. In near-real time, it can be used by emergency service providers and non-profit organisations to assess damages and prepare crisis responses in remote areas. Similarly, when used in combination with the flood risk module, governments and non-profit organisations can optimise their crisis response capacities or plan their infrastructure to avoid high risk areas.

As satellite image databases are growing and becoming more readily available for public use, AI-based solutions such as this are the best alternative to current event identification. The benefits coming out of HNCDI in its early phases are therefore paramount to improving the prediction processes which are at present time consuming and limited in their ability to analyse and interpret data.

HNCDI value added

The delivery of the flood detection model required a considerable amount of knowledge, access to resources like satellite image databases, and compute-time. The HNCDI programme connected complementary skills and facilities within the STFC and IBM teams to produce a toolkit of resources that were essential in completing the project objectives. Furthermore, interest in the detection aspect as well as the combined detection and risk-mapping platform may yield continued mutual work and benefits for the teams involved as a result of the programme.

Sources

- Interview with: Louise Butcher, Senior Data Scientist at the STFC's Hartree Centre
- Interview with Paolo Fraccaro, Staff Research Scientist at IBM Research Europe
- Project Master Documents
- Project Benefit Profile
- Project Closing Presentation

2.2 Explain training: NSG

Workstream: Explain	
Project: Explain training courses in data science	**Timeframe:** June 2021 to present
Participant organisation: NSG	
NSG Pilkington manufactures and processes glass, offering specific glass technology in the architectural and automotive sectors. To add value to its portfolio, the company collaborated with the STFC Hartree Centre on two materials discovery projects, which inspired a company-wide recognition of potential data science applications. Consequently, NSG piloted a programme to train their staff in data science through the HNCDI Explain training programme, acknowledging the Hartree Centre as a reputable and trusted source working at the cutting edge of digital technologies. NSG Pilkington were drawn to the flexible delivery and range of experience levels, which ensured consistency in approach and terminology. From materials science teams to legal, staff have since registered for 61 courses across both introductory and advanced levels, reporting improved confidence in framing and communicating their requirements to data science experts. Other feedback suggested the courses have improved readiness to tackle internal data science challenges. Pilot participants also recommended Explain to new data science recruits, demonstrating the usefulness and wide application areas of the range of courses on offer.	

Introduction

Pilkington Group Ltd, now NSG, has been in the business of glassmaking for nearly two centuries. Producing glass solutions from glass manufacturing and processing to merchanting, they offer specific glass technologies for the architectural and automotive sectors. Part of the Japanese owned NSG Group, it is based in Lathom, Lancashire and employs around 3,000 people across the UK. To maintain their market position for a product as ubiquitous as flat glass, NSG invests heavily in materials discovery and materials property simulations. To accelerate the development of its glass products, NSG has been collaborating with the Hartree Centre and the University of Liverpool's Rosseinsky Group as part of two research projects to support the application of high performance computing to materials discovery since 2015.[14]

NSG first became aware of HNCDI's Explain training programme through these collaborations. Recognising the value the Hartree Centre's data scientists added to their projects and in the context of a wider expectation that AI and data science methods would become increasingly ubiquitous across the company, NSG was keen to bolster their internal capability to maximise the benefits of its projects with the Hartree Centre and receive training directly from the source. As a result, NSG launched a pilot programme in external data science training, and NSG staff were among the first to receive instruction under Explain. The training was expected to help technical staff understand the techniques and limitations of the field, to generate useful data, and eventually apply data science techniques themselves.

[14] https://gtr.ukri.org/projects?ref=EP%2FN004884%2F1

Training activities

NSG generates a wealth of data from its manufacturing and simulation activities, so several of their staff were already familiar with the applications of data science at a high level. As part of the pilot, NSG invited those staff with the most potential for data science and machine learning in their work to participate in various Explain courses, with no limit placed on the number that any one person could attend. These ranged from introductory courses to more hands-on training involving opportunities to interact with real data and work through examples. By January 2023, the company had **61** course registrations with Explain.

Included among those who attended Explain courses were members of NSG's legal IP team. For them, machine learning is of relevance given the prominence of questions on AI, originality and inventorship.[15] In future, NSG's IP lawyers are expected to be involved at the early stages of R&D concerning AI, and therefore need to understand the consequences for their approach to NSG's current and future IP.

Explain impact

Explain attendees intending to improve their general understanding of the applications of data science and potential gaps in their teams' skillsets reported increased confidence in communicating with data experts and defining their problems in technical terms. The introductory courses provided a solid starting point on their data science journey and an overarching understanding of how to approach AI and data science related challenges. One attendee noted the introductory training had increased their ability to collaborate with academic partners, allowing them to better understand and articulate their problem and project, to better communicate with a PhD student and their supervisor.

Explain attendees with prior experience in AI and data science tools noted that whilst they have not yet had the opportunity to fully implement new knowledge or skills gained through the courses, it has contributed to their overall readiness to tackle future data science challenges.

Positive engagement with Explain by participants of NSG's pilot data science programme has led to recommendations to train newly recruited NSG data scientists under Explain as well. This further demonstrates the applicability of Explain training across a range of backgrounds and how the HNCDI programme has built up a reputation for leading application focussed training.

HNCDI value added

Although NSG also offers internal training programmes, these are not directly related to data science. Explain training has allowed NSG staff to access a variety of targeted courses, with different levels of focus on hands on activities, and delivered for the most part through the participants' browser for maximum flexibility. In particular, representatives from NSG noted the value in having a broad range of people receiving training at different levels from the same source, ensuring a degree of consistency in the approach and terminology across different teams and departments.

In addition, NSG representatives noted the value of receiving training from practitioners at the cutting edge, increasing their confidence that the tools and approaches they're learning are both leading and reputable.

Notably, the NSG pilot external training programme coincided with the first wave of Explain courses, prior to implementation of the Explain online learning management system. With the

[15] https://www.carpmaels.com/patenting-ai-the-nature-of-inventorship-and-the-mechanics-of-ownership/

integration of this new system, future Explain attendees are now also able to undertake modular, self-paced learning which can offer greater utility and tailored training packages to suit their needs.

Sources

- Interview with Rory Back, R&D Incubator Technical Manager at NSG
- Interview with Graham Siddons, Simulations Team Leader at NSG
- Interview with Mark McLachlan, Systems & Control Technical Manager at NSG
- Interview with Jonathan Bracken, Advanced Technologist at NSG
- Explain monitoring data
- https://www.pilkington.com/en-gb/uk/about
- https://www.liverpool.ac.uk/digital-innovation/working-with-us/pilkington-case-study/
- https://www.materialschemistry.org.uk/case-studies/prioneering-materials-discovery-1

2.3 Excelerate: Precision medicine with REPROCELL

Workstream: Excelerate	
Project: Integrating multi-omic data and demographic data with explainable machine learning to identify precision medicine strategies - Inflammatory Bowel Disease (IBD)	Timeframe: Apr 22 – May 23
Partners involved: REPROCELL	
The objective of this project is to deliver an AI platform which generates unparalleled insight into drug efficacy in different patient populations. Without AI, current datasets can be too large and complex for researchers to fully benefit from them. To this end, HNCDI, in partnership with REPROCELL, are developing a product platform that aims to simplify and speed up the process of generating trained machine learning algorithms from these complex datasets, that can then be utilised in precision medicine strategies. The current iteration of the project can be used to help pharmaceutical companies to streamline their drug development processes and could help to identify more effective treatment for those suffering from IBD. Future iterations could be expanded to include cancer therapies and could be utilised across multiple stages of drug development and/or clinical trial design.	

Introduction

Medical treatments are intended to relieve symptoms and/or cure conditions, however, different people respond in different ways to these treatments. Many experience adverse reactions, 72% of which are avoidable.[16] Precision medicine provides an opportunity to tailor treatment to individuals based on their clinical profiles, their demographic data, and genetic and molecular data (omics).

REPROCELL is a contract research organisation (CRO) providing research services across the entire workflow of pre-clinical and early stage-clinical drug development, with labs in Glasgow and Durham, UK, as well as in Japan and the USA. REPROCELL's European division is particularly focussed on providing predictive data on the efficacy, safety and pharmacokinetic profiles of new drug therapies, and performs research in human living tissues that are residual to surgery or transplant, with the express consent of donors. Importantly, the responses to drugs observed in REPROCELL's tests in human living tissues reflect the variation in drug response observed in the general population, where the effectiveness of drugs can vary markedly from patient-to-patient. It is believed the answer to these differences arises from a combination of "nature and nurture".

Whilst REPROCELL has access to deidentified electronic health records ("nurture") and genetic data ("nature"), it has been a challenge to fully realise the benefits of its human tissue research and enable precision medicine strategies, due to the size and complexity of the datasets generated.

[16] HNCDI monitoring data.

To explore the opportunities of using these datasets to advance precision medicine, REPROCELL partnered with the Hartree Centre and IBM as part of the Innovation Return on Research programme (IROR), a precursor to HNCDI. The IROR project resulted in a proof-of-concept for REPROCELL, demonstrating a solution that offers insights into inflammatory bowel disease (IBD) drug responses that were previously unachievable. This research also resulted in the joint publication of an academic paper for IBM and REPROCELL, with ongoing discussions for further joint publications in future.[17]

Now, within HNCDI's Excelerate workstream, REPROCELL is further developing this solution into a product that can be added to the REPROCELL portfolio of research and clinical trials services. The success of this project will streamline drug development processes and timelines, and ultimately benefit those patients who receive better treatment because of improved applications of precision medicine.

Project activities

Through HNCDI, REPROCELL are more involved in product development, providing crucial input to the STFC and IBM teams to inform the design of the AI platform, its user interface and the outputs that are most relevant to medical researchers to ensure an industry-ready solution.

Under the project, the HNCDI team and REPROCELL are also developing a commercialisation roadmap for the products that will arise from this collaboration. This plan will further detail how REPROCELL could adopt the product into their organisation and how to best incorporate it into their service offering to clients.

Project impact

As of February 2023, the project is on track to deliver a minimum viable product that will enable REPROCELL to demonstrate this drug response insights service to their clients, with immediate potential to generate revenue. If successful, REPROCELL aim to increase the scope and capacity of their precision medicine offering, including in-house adoption of digital technologies.

Wider impacts

The current iteration of the platform has been designed for pre-clinical research, with a focus on IBD treatment as proof of concept. IBD affects nearly half a million people in the UK[18], up to 40% of whom may experience adverse drug reactions[19]. The platform can also be adapted for clinical trials – especially in defining patient populations – and has potential for use in a range of therapeutic areas, including oncology.

Once commercialised, the solution would provide pharmaceutical companies with a service for faster and less costly drug development. Downstream, the benefits of this would then follow through to patients who could be offered treatments that are more likely to yield positive clinical outcomes and subsequent health benefits.

HNCDI value added

The collaboration between REPROCELL and HNCDI has enabled the development of a novel product by accelerating the adoption of AI through knowledge transferred, and computing facilities accessed which, in their absence, would have presented a significant barrier to this innovation. Should the commercial demonstration prove a success, REPROCELL hopes to

[17] https://doi.org/10.1371/journal.pone.0263248
[18] https://ampersandhealth.co.uk/myibdcare/resources/ibd-statistics-2022-crohns-and-ulcerative-colitis/
[19] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8053178/

continue refining and expanding the platform with the assistance of the Hartree Centre and IBM.

Sources

- Interview with Karen Bingham, REPROCELL COO (Europe)
- Interview with Graeme Macluskie, REPROCELL Director of Precision Medicine (Europe)
- Project Master Document
- Pre-project survey
- Project Benefit Profile
- Project Mid Term Review

2.4 Emerging Technology: Quantum Machine Learning enabled drug discovery

Workstream: Emerging Technology	
Project: Quantum Machine Learning strategies for accelerated Ligand-Based Virtual Screening.	Timeframe: July 2021 – April 2022
Partners involved: n/a	
The drug discovery process is notoriously expensive and time-consuming. Computational techniques like Ligand Based Virtual Screening offer opportunities to expedite the process and bring costs down by screening digital databases of molecules to identify structures most likely to bind to a drug target. This technique could employ machine learning to undertake rapid screening but given the amount of information available on the derived and experimental properties of molecules, training such algorithms can become prohibitively resource intensive. It is here that Quantum enabled Machine Learning offers a potential advantage. This proof-of-concept project between the Hartree Centre and IBM aimed to investigate and demonstrate that quantum computing, able to support calculations at exponentially higher-orders of complexity, offers a potential solution and opportunity to increase the efficiency and accuracy of Ligand Based Virtual Screening. The project has demonstrated that the application of the Quantum Support Vector Classifier – available through IBM's Qiskit – often performed better than the equivalent classical algorithm and in some instances greatly outperformed the deep learning methodologies that are the current state of the art in drug discovery. A paper presenting the results was published in February 2023 and the HNCDI programme are currently scoping potential follow-on projects in collaboration with UK pharmaceutical companies to take this forward.	

Introduction

The drug discovery process is notoriously expensive and time-consuming. Estimated R&D costs for new drugs vary widely, ranging from $161m to $4.54bn.[20]

To expedite the process and bring costs down, drug discovery has employed virtual screening techniques for some time now. Virtual screening is a computational technique that searches through libraries of small molecules to identify structures which are most likely to bind to a drug target. However, this depends on good structural knowledge about the drug target site. In the absence of this information, drug discovery instead looks at other molecules known to bind to the biological target site of interest (aka ligands). Known as Ligand Based Virtual Screening (LBVS), this technique could employ machine learning to undertake rapid screening. However, given the amount of information available on the derived and experimental properties of molecules, training such algorithms can become prohibitively resource intensive. There is therefore a trade-off between the speed of training, and the accuracy of the outputs, in employing Machine Learning for the LBVS.

It is at this point that quantum machine learning algorithms could provide a potential solution, providing an alternative to speed up calculations for problems which are not computed efficiently classically. The rationale being that quantum computing is able to support

[20] https://link.springer.com/article/10.1007/s40273-021-01065-y

calculations of higher-orders of complexity. In this case, quantum computing can better handle the breadth of features relating to each molecule of interest and contained within the molecule databases.

Project activities

Within this project, IBM and the Hartree Centre investigated a potential quantum advantage – a novel, practical use of quantum computing to solve a particular challenge to Ligand Based Virtual Screening. The main objective of this Emerging Technology project was to test (and prove) that quantum machine learning algorithm, with access to classical data, could outperform the equivalent classical methodology.

The project was organised into three work packages:

- WP1 Digitalisation of classical molecular descriptors – to explore how to pass real data to the quantum algorithm and what features or parameters should be incorporated into the formula.
- WP2 Algorithm workflow – to build the workflow to integrate the Quantum Support Vector Classifier (QSVC) – available through IBM's Qiskit - with the classical algorithm.
- WP3 Virtual screening of a digital database of molecules – to apply and test this new workflow (QSVC) against other classical algorithms for which performance is known (CSVC) as applied to a database devised specifically to benchmark AI and ML methods.

Project impact

This is the first project to propose and trial a new concept and approach to integrating quantum computing within a drug discovery pipeline. The proof-of-concept project demonstrated that it is possible to successfully train hybrid classical-quantum machine learning algorithms to identify new possible therapeutics within a digital database of molecules.

The project found that the QSVC algorithm performed better than, or even in the worst case equivalent to, classical SVC algorithms. In addition, the project also showed that, in some instances, the QSVC method greatly outperformed deep learning methodologies that are the current state of the art in drug discovery. The project found that whilst the model performed less well when applied to small datasets, it consistently improved in accuracy when increasing the size of the dataset and confirmed that, as is the case with CSVC, the performance of QSVC depends on the quality of the dataset.

The paper presenting the project's findings were published in February 2023.[21] The next steps for the project will be to understand how QSVM performs with a more realistic quantum backend and to run the simulation on real hardware.

Wider impacts

This work is expected to be of interest to pharmaceutical companies, and the HNCDI delivery team are already engaging in follow-on work to scope further partnerships to apply this algorithm to further real-world problems. In the longer-term, and with further development, increasing the efficiency and accuracy of LBVS will have downstream benefits for identifying and developing new drugs faster and in a more cost-effective manner. This is especially valuable for new and emerging diseases for which the target sites have not yet been profiled.

This work could also be applied to other areas of chemistry. For example, there is potential for the Quantum Machine Learning approach developed to be applied to the identification of

[21] Stefano Mensa et al 2023 Mach. Learn.: Sci. Technol. 4 015023

novel materials for climate mitigation, through identification of photocatalysts, improved solvents or battery electrolytes.

Sources

- Project Benefit Profile
- Project Master Document
- QCML Drug Discovery presentation slides