# UK Research & Innovation:
# Science case for UK Supercomputing

**Authors:** The UKRI Supercomputing Science Case Editorial Board

**Editor:** Mark Wilkinson

**Editorial Board:** Mark Wilkinson (Leicester; Chair), Ewan Birney (EBI), Nilanjan Chakraborty (NCL), Peter Coveney (UCL), David Ford (Swansea), Chris Johnson (EPCC), Bryan Lawrence (NCAS), Mark Parsons (EPCC), Andrew Prescott (Glasgow), David de Roure (Oxford), Debora Sijacki (Cambridge), Nick Trigg (STFC), Beth Wingate (Exeter).

DOI:10.5281/zenodo.4985325

21st October 2020

Version history

| | | |
|---|---|---|
| UKRI Supercomputing Science Case Submitted | 18/08/19 | Final submitted version |
| UKRI Supercomputing Science Case Submitted Revised 19.08.20 | 19/08/20 | MIW edits of introduction and executive summary; Edits to MCF section 7.5 |
| UKRI Supercomputing Science Case Submitted Revised 21.09.20 | 21/09/20 | Complete draft incorporating all proposed revisions – circulated to editorial board for final comments. |
| UKRI Supercomputing Science Case Final 21.10.20 | 21/10/20 | Final version of case submitted to EPSRC for publication. |

**Table of Contents**

# 1 Executive summary

This document presents the compelling science case for supercomputing in the UK, particularly with regard to UK Research & Innovation (UKRI) scientific research, drawing on examples spanning the full breadth of the UKRI programme. We are now at an auspicious juncture in the evolution of computing. Dramatic changes in computer architectures and how we use them can cause directional shifts both in the technology and the directions of human inquiry that rely on it. Supercomputing is now so essential to scientific research in the fields described here that UK researchers, both in academia and industry, can only remain internationally competitive if they have access to sufficient, appropriate computing resources and if support for associated activities (e.g. novel algorithm development and software development) is available. Over the coming decade, alongside the increasing demands from established computationally-intensive fields, significant growth in computing requirements from non-traditional fields, including Artificial Intelligence (AI), is expected.

Supercomputing investment leverages increased scientific productivity from investments in experimental facilities, enabling new science to be delivered with existing equipment and supporting the design and implementation of new experiments. It also supports research where the objects under study are so large, small, fast or slow, complex or difficult to study that experimental measurements are either impossible or must be supplemented by simulated data for their interpretation. Further, the research and development activities proposed in this document will contribute to the development of future computational technology, both hardware and software. UK participation in this co-design[1] process is critical for the economic growth aspirations outlined in the UK Industrial Strategy, much of which relies on technological developments (e.g. hand-held devices with AI capabilities) which are interrelated with innovations introduced in the process of supercomputing design.

As the case makes clear, technological advances, coupled with algorithm and software development, will offer an enormous opportunity to address questions which until now have been beyond our reach.

Examples of the resulting breakthroughs in research areas where the UK is world-leading include:

- **Expanding the frontier of fundamental sciences:**
  - the first realistic simulations of the observable Universe in which Milky Way-like galaxies are fully resolved, providing robust tests of the standard model of cosmology through comparison with observational data from the Large Synoptic Survey Telescope (LSST) and the Euclid satellite.
  - next generation Lattice Quantum ChromoDynamics (QCD) calculations which, together with the more precise experimental measurements from the high luminosity Large Hadron Collider (LHC), will for the first time uncover and exploit cracks in the Standard Model that may lead to the discovery of new physics.

[1] Co-design is “a methodology for scientific application, software and hardware communities to work together. […] The co-design strategy is based on developing partnerships with computer vendors and application scientists and engaging them in a highly collaborative and iterative design process well before a given system is available for commercial use.” (“On the Role of Co-design in High Performance Computing“, Transition of HPC Towards Exascale Computing (2013)). Successful co-design projects can result in systems which deliver significant performance gains for particular application classes, benefitting both the scientific research community and consumers.

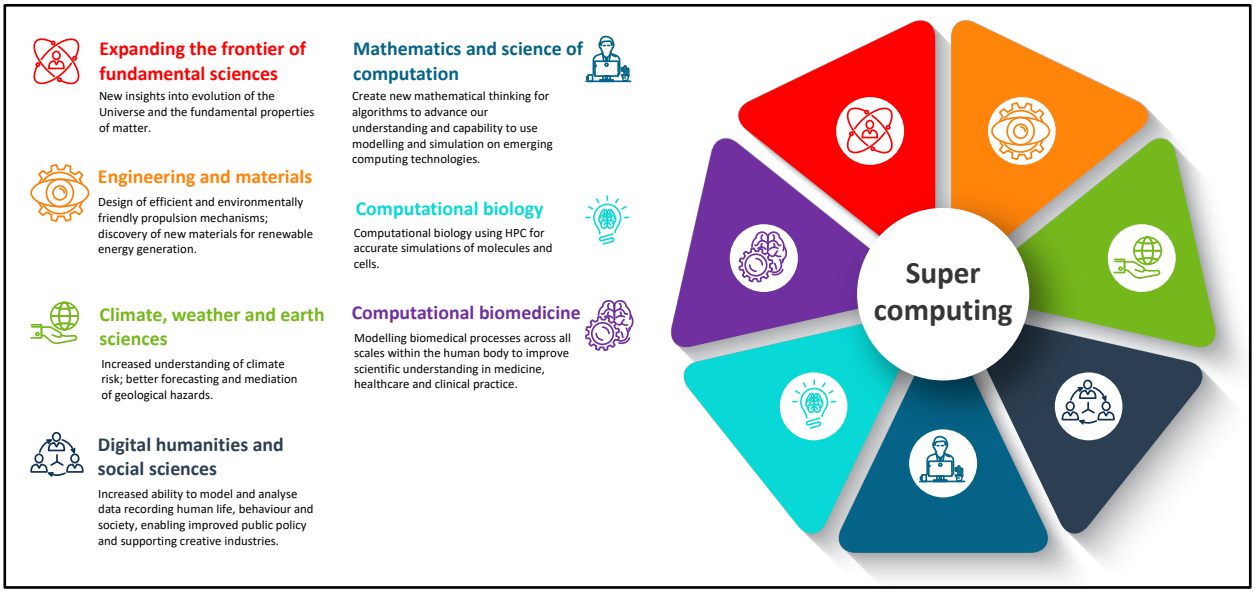



*Figure 1: Examples of science highlights from each theme in this science case.*

- **Climate, weather and earth sciences:**
  - increased understanding of climate processes and risk, by improving our understanding of the physical system, increasing fidelity in simulations, and supporting industry and society in adapting to the changing climate.
  - insights into fundamental geophysical processes, from mineral physics to earthquakes, providing routes to improved forecasting and mediation of geological hazards, and long-term carbon sequestration.
- **Computational biology:**
  - providing accurate and useful simulations of proteins and other molecules involved in living processes.
  - using combinations of Deep Learning, large scale statistics and simulation to understand biological systems.
  - exploiting genomic data at scale for insights into health and basic science.
- **Computational biomedicine:**
  - creation of a personalized quantitative mechanistic model of the entire human body at all scales from genome to organism (and beyond).
  - enhanced understanding of the origins of diseases and their treatment.
  - development and optimisation of healthcare solutions for clinical deployment in the context of personalised medicine.
- **Engineering and materials:**
  - heat and fluid flow and material modelling to test technologies which are essential for the design of efficient and environmentally friendly energy generation and propulsion mechanisms.
  - using Computational Fluid Dynamics (CFD) and Computational Materials Modelling (CMM) to reduce the cost of innovation and accelerate the design cycle of new products enabling, for example, the design of cars based on CFD without any experiments.

- development of new thermoelectric[2] and photovoltaic[3] materials, which play key roles in accelerating the alternative renewable energy generation methods in the UK.
- high fidelity simulations of whole systems, for example the creation of digital twins of gas turbine engines, such that virtual certification can be introduced.

- **Digital humanities and social sciences:**
  - step changes in the computational handling of music, creating innovative pipelines for production, analysis and retrieval, and exemplifying new commercial opportunities in creative industries.
  - major advances in corpus linguistics, facilitating language processing across a wide range of technologies from robotics to AI.
  - new forms of computational modelling to support improved decision making in public policy and economic areas, with analytics increasingly underpinned by rapidly growing data sources such as smart devices.

We have explicitly included a cross-cutting seventh theme entitled "**Mathematics and Science of computation**" to emphasise that in an era of new computing architectures, we need to nurture and develop the key skill sets that underpin successful supercomputing. We need to think about how to successfully develop and implement algorithms and applications at the scale of millions of heterogeneous parallel threads. The inclusion of this theme also highlights the fact that whether the application is in nuclear physics, climate science, the humanities or the economy, there is often a shared mathematical and computational structure and all require sound mathematical, algorithmic and software development strategies. What is more, development of these skill sets as an activity in and of itself also helps develop the skill sets that lead to inventive technology, a required component of the UK ecosystem if we are to meet the challenges of the UK Industrial Strategy.

Collectively, the cases presented below demonstrate that supercomputing resources will underpin the future success of UKRI-supported research and they make a compelling argument for increased provision of supercomputing resources on a range of scales. Small-scale, experimental systems are needed to support the development of innovative new algorithms which can harness the power of novel hardware to tackle new research problems. Larger-scale systems providing tens to hundreds of PetaFlops coupled with significant storage volumes will make it possible to carry out rigorous uncertainty quantification studies of computationally expensive models which are impossible using the resources currently available in the UK. Finally, the largest, grand challenge calculations described here will require the computational power of an exaflop system. In preparation for a UK exascale service, pre-cursor systems are urgently needed to allow the algorithm and software development which will ensure that the UKRI community can take full advantage of the scientific opportunities presented by exascale supercomputing.

The case presented here constitutes the current vision from the UKRI research community of the key science questions for the next 5 years. However, as each field advances, these questions will evolve as new challenges (or indeed research areas) arise and grow in significance. We therefore anticipate that this case will be revised on a 2-3 year timescale, to adhere to the guiding principle that the UKRI e-infrastructure should be research-driven. At all times, robust peer review based on scientific excellence and technical appropriateness for UKRI resources will be used to ensure that only the highest quality research is carried out on our national supercomputing systems.

Although this case is focused on academic research, access to UKRI supercomputing resources will be open to high quality research from both academia and industry. The benefits to UK industry will

[2] https://doi.org/10.1088/2515-7639/ab16fb

[3] https://doi.org/10.1021/acs.chemmater.6b03944

include access to the results of academic research performed on UKRI resources, direct access to those resources where appropriate and an increased supply of highly-skilled researchers who have received training on the UKRI supercomputing services.

Delivery of the science presented in this document will require a significant uplift in e-infrastructure investment in the UK. Across all research areas, there currently exists significant unmet demand: the cases presented below note that factors of 10 to 100 increases in computing power are needed just to deliver immediate science goals. Increased supercomputing provision is thus urgently required both to help meet these existing needs and to open up entirely new opportunities for supercomputing applications that have so far been outside the reach of UK researchers.

The assembly of this science case is the first step in the process which will ultimately deliver specific recommendations on the technical requirements for the various supercomputing services comprising the UKRI e-Infrastructure. Following peer review of the science case, a technical design process will be initiated to determine and quantify the high-level computing requirements for each component. These high-level technical requirements will also be peer reviewed. At that point, co-design work with supercomputing industry partners will be undertaken to translate the high-level requirements into specific technical designs. This approach to service design was recommended in the UKRI supercomputing white paper and will maximise the scientific productivity and impact of the supercomputing services that are ultimately deployed to deliver the UKRI science programme.

# 2 Introduction

The recently completed White Paper, “UKRI National Supercomputing Roadmap 2019-30”, describes the supercomputing ecosystem required to allow UK researchers in academia and industry to deliver world-class research throughout the next decade. The underlying principles proposed for this ecosystem are that it will be:

1) Research driven;
2) Assessed based on scientific and industrial productivity;
3) Supported by parallel investments in software and algorithms.

In order to ensure that the ecosystem reflects the true needs of the research community, the White Paper recommended the urgent assembly of a peer-reviewed UKRI science case for supercomputing. This science case, intended to be as inclusive of all research areas as possible, would then be used to define the required range of supercomputing services at all scales and would underpin future business cases. This document is the community response to this recommendation.

**Definition of supercomputing and scope of this case**

In 2018, as input to the development of the e-Infrastructure chapter of the UKRI Infrastructure Roadmap, the UKRI e-Infrastructure Expert Group was commissioned to develop a set of seven white papers. These were completed by April 2019 and cover all aspects of the computing ecosystem needed to support research in the UK, both in academia and industry, namely: Supercomputing; Research Data Infrastructure and High Throughput Computing (HTC); Cloud; Network; Software & Skills; Authentication, Authorisation & Accounting Infrastructure (AAAI); Security; Industry. This UKRI science case for supercomputing is the first of a series of such cases. It will be complemented, in due course, by science cases underpinning the other aspects of the UKRI e-Infrastructure. Collectively, these will provide the scientific justification for the business cases which will deliver the funding needed to implement the roadmap recommended.

Following the approach taken in the white paper, we define supercomputers as systems with tight coupling between all the processors (supporting large amounts of communication) and between the processors and the storage system, so that algorithms can be executed efficiently (in resource terms) and at a satisfactory speed (to meet science requirements). This tight coupling, delivered by specialised interconnects and often with specialised storage, is what defines the difference between supercomputing and other large-scale computing platforms such as commercial cloud computing or some forms of high throughput computing (e.g. that needed to support real time data processing) suitable for problems which can be split into sub-calculations that can be performed almost independently of each other. Both HTC and commercial cloud were considered in separate white papers. For most of the supercomputing workflows presented here, commercial cloud is not currently a cost-effective or technically competitive solution. Further, for reasons of both national security and national energy security, it is important that we nurture the creative and technical expertise to design and develop supercomputers, as well as to exploit them, rather than relying solely on the expertise of the companies that supply them. However, commercial cloud will naturally form part of the UKRI research computing ecosystem and its relevance for supercomputing will be kept under review.

**The role of supercomputers in modern science**

Supercomputing has become so fundamental to scientific understanding that, in many fields, the delivery of world class research depends on it. In such fields, science without supercomputing would be akin to attempting to carry out astronomy research without telescopes or biology without microscopes. Supercomputers are scientific instruments, often used to help answer scientific questions for which the objects under study are so large, small, fast or slow, complex or difficult to study that measurements of their properties must be supplemented by simulated data for their interpretation. Supercomputers therefore form the fabric that allow us to compare theory with

experiment. Today, supercomputers are also being used for large-scale Artificial Intelligence (AI) and Machine Learning challenges and this convergence between AI and supercomputing is a trend we expect will continue and deepen, as we discuss below. They are a core tool of modern science and underpin world-leading research and innovation across the entire UKRI remit.

Supercomputers have traditionally been associated with modelling and simulation. However, as the following sections will demonstrate, supercomputing is increasingly important for data analytics, data modelling and data fitting, collectively referred to as data intensive computing. The range of supercomputing applications described below requires resources on a range of scales from small proof-of-concept systems to the largest, capability systems.

Modelling & Simulation is increasingly a core research activity across all UKRI research areas as well as in industry: the scale and flexibility of UK supercomputing provision must reflect this. In addition to supporting scientific leadership across flagship computational science fields, the full commercial exploitation of UK-led discoveries requires supercomputing resources capable of delivering all phases of development from proof-of-concept through to “whole system modelling”.

The computational requirements of theoretical modelling are continually evolving, and will continue to become increasingly demanding due to the combined drivers of:

- increased resolution: running models incorporating existing mechanisms or processes but at finer scales.
- increased complexity: introducing new features into models to reflect progress in theoretical understanding – this is often needed to match resolution increases.
- coupling of models: multi-feature and multi-scale modelling; the ultimate goal is “whole system modelling”.
- data integration & high performance data analytics: scientific insights in many areas (e.g. systems biology, astronomy) demand the simultaneous modelling and analysis of multi-source, multi-scale data sets which present significant computational challenges.
- verification, validation & uncertainty quantification (VVUQ) – making not only predictions, but also using large ensembles of simulations and intrusive UQ methods to provide robust statistics, uncertainties and quantified risk – engendering trust in modelling and simulation and making models “actionable”.

The provision of sufficient supercomputing resources is essential for maintaining scientific leadership and ensuring that the UK benefits fully from investments in experimental/observational facilities. The ability to analyse data and/or run associated simulations is vital for guaranteeing that UK researchers reap the scientific rewards from data sets obtained from projects with significant UK investment. As noted in the white paper, the traditional, hierarchical view of the scales of supercomputing is not an appropriate representation of the UK ecosystem in which all current scales of supercomputing deliver world-class science. Figure 2, reproduced from the white paper, instead presents the various “tiers” of supercomputing provision as segments of a balanced ecosystem.

This science case is focussed on Leadership Class systems (Tier 0/Exascale) and Discipline/Programme Specific National Facilities (Tier 1). The Tier 1 systems will support codes and workflows that require specialist hardware (“Customised simulation hardware”, for example large amounts of memory or storage), or which have modes of operation which do not mix well with general usage (“Customised usage”, e.g. the need to execute jobs, such as a weather forecast, in a particular time window), or which require managing and/or moving so much data that they do not mix well with more computation-based tasks. The science case for Instrument/Facility-based and Entry-level/Exploratory (Tier 2) systems will be made elsewhere, although we note that these services are an essential complement to the larger supercomputing resources and significant innovation in algorithm and system development will continue to emerge from the Tier 2 deployments (see e.g.

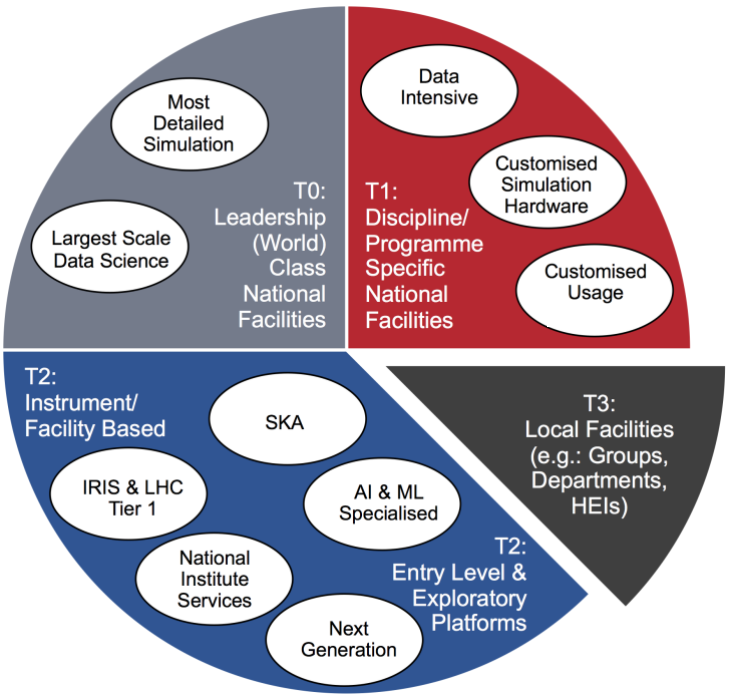



*Figure 2: The components of a balanced supercomputing ecosystem. The traditional labels of tiers (Tier 0, Tier 1, Tier 2, Tier 3) are kept for consistency, but their definitions are explicitly listed. [Image Credit: Lawrence & Jenner; reproduced from the white paper "UKRI National Supercomputing Roadmap 2019-2030".]*

Section 0). Tier 2 and Tier 3 systems are included in Figure 2 to emphasise the interdependence of systems on all scales, and their importance for the scientific productivity of the UK.

**The importance of algorithms and software as a cross-cutting theme**

The current rapid increase in supercomputing hardware diversity, while providing significant opportunities for increasing scientific productivity, will inevitably set demanding challenges for most software communities. In particular, old algorithms which no longer run efficiently on new hardware (largely due to the massive increase in parallelism of modern systems) must be replaced in production software. New, flexible software infrastructure must be made available so that applications can use the latest computational mathematics algorithms as they become available. Because of the evolution of computer architectures, it is increasingly rare for codes to be able to use a single, unmodified algorithm over many years. If we are to be world leaders in computational science, software and algorithm development must be considered as an evolving process. Software engineering and algorithm development effort is therefore vital to ensure that our research software is increasingly agile and able to deliver optimal performance on the latest hardware without damaging scientific productivity by diverting research funding and/or researcher effort. Investment in programmes of research software engineering and mathematics at scale is the most cost effective and efficient way to achieve this important goal in the coming years as technology becomes increasingly exotic. Without this development work, our research programmes will become prohibitively expensive and irreparably damaged.

Given the current shortage of research software engineers (RSEs), in particular those with strong domain expertise, investment will be needed to create a larger RSE pool. These people will be highly prized in industry and many RSEs are likely to move into industry careers – investment in RSEs thus contributes to UK economic growth. However, there is also a pressing need to develop better career pathways for RSEs within research institutions. The inclusion of a Unit of Assessment for computational science in the next Research Excellence Framework (REF) would be a significant step towards this.

**The international context**

The lack of sufficient supercomputing provision in the UK is placing the UKRI research community at a competitive disadvantage relative to international partners. Several countries (e.g, China, Japan, USA, Germany France, etc.), including those which did not traditionally have a track record of computational research (e.g. Saudi Arabia), have recently invested heavily in supercomputing infrastructure. The supercomputers based at King Abdullah University of Science and Technology, Saudi Arabia (Shahin II, which is 45th in the June 2020 Top500[4] list), and Kyoto University (89th in the June 2020 Top500 list, but not one of the top 10 fastest systems in Japan) are faster than the fastest accessible computer based in the UK and these are only two examples out of several possible ones. As a result, researchers based in countries such as China, Japan, USA and several EU countries (e.g. Germany, France, Spain and Switzerland) have access to faster supercomputers than UK researchers. These enable them to routinely run computationally intensive, multi-physics, multi-scale simulations without limiting simplifications and empiricisms (e.g. Direct Numerical Simulations (DNS) of turbulent flow with high Reynolds number, DNS of turbulent reacting flows with detailed chemical mechanisms, materials modelling simulations with more molecules), which are either impossible in the UK, or only possible as a one-off calculation linked to specific funding bids (e.g. Leadership calls on ARCHER[5]).

Examples of calculations which are currently accessible to our competitors but which are beyond the capabilities of UK resources include:

- In climate science, international competitors are currently able to deploy 10-100x the compute capacity available to UK researchers, allowing higher resolution (10x, e.g. [1]), and/or much larger ensembles (10x, e.g. [2]). This allows both increased understanding of the earth system and better quantification of climate risk - delivering information of more use to industry (e.g. insurance) and society (e.g. for adaptation).
- By melding AI techniques and traditional supercomputing algorithms, researchers are transforming the solution of large systems of partial differential equations - a staple task of many numerical algorithms. Researchers in Japan have trained AI systems to precondition solution matrices for the well-known conjugate gradient algorithm reducing time to solution massively for large scale earthquake simulations and demonstrated the benefits on the USA's SUMMIT system. The approach is applicable in many areas of modelling and simulation.
- PRACE (Partnership for Advanced Computing in Europe) recently awarded time for a set of large-scale galaxy formation simulations requiring 135TB of memory and 40M core-hours (on 36000 cores). These simulations will have 100 billion particles and will be an order of magnitude higher-resolution than the current state-of-the-art. Further, PRACE awarded resources for large, full radiation-hydrodynamics simulations of the Epoch of Reionization (see recently awarded SPHINX project) requiring 54 million core-hours to resolve more than 15 thousand galaxies which will make definitive predictions for the James Webb Space Telescope (JWST) and Square Kilometre Array (SKA). In the UK there are currently no supercomputing facilities to perform such calculations.
- Direct Numerical Simulations of turbulent reacting flows involving 35 million CPU hours on 10,000 cores were hero simulations in 2009 [3] but can now be routinely done by some US and Japanese groups (e.g. the same group in [3] reported simulations involving 50 million CPU hours on 72,000 cores in a recent publication [4]). UK-based researchers working on magnetic confinement fusion

[4] https://en.wikipedia.org/wiki/TOP500#TOP_500

[5] ARCHER (Advanced Research Computing High End Resource) is the UK national facility which provides supercomputing resources for the Engineering and Physical Sciences Research Council (EPSRC) and Natural Environment Research Council (NERC) communities.

currently must maintain access to the Italian 19 PF MARCONI system in order to remain internationally competitive. The UK-based researchers using global gyro-kinetic codes can only complete this work due to collaborations with colleagues in Switzerland allowing them access to the 21 PF Piz Daint system.

- Simulations of blood flow offer the potential to aid medical decisions and improve clinical outcomes. Patient-specific, clinically-oriented simulations of the entire cardiovascular system require the exascale computing systems which will soon be available in other countries and code development work for these calculations must already be tested on resources outside the UK.

Mathematics as an underpinning research and development activity required for science and engineering computations has been recognised by ETP4HPC[6] in a sequence of reports between 2013 (very little mathematics), 2016 (more mathematics is added), and 2017 (mathematics and algorithms recognised as a key activity required for success in supercomputing). The strategic Research Agenda issued by ETP4HPC has been used by the European Commission to define the Future and Emerging Technologies HPC (FETHPC) Work Programmes during the Horizon 2020 programme. In the UK, there is no coordination of "mathematics and algorithms at scale for computing". At present, researchers in the UK can only participate in this endeavour through standard grants to EPSRC Mathematical Sciences that supports the entire remit of mathematical sciences, or through the EU's Horizon 2020 programme which in future will be implemented by the EuroHPC Joint Undertaking. The UK is not a member of EuroHPC and this funding will not be available after Horizon 2020 ends.

As our competitors deploy pre-exascale and exascale systems over the next two years, the UK risks becoming unattractive as a destination for researchers in computational science domains and losing leadership of fields in which the UK has traditionally been a key global leader. As access to PRACE facilities will in future be associated with EuroHPC membership, access to European computing facilities for UK researchers is expected to end following the departure of the UK from the EU. UK-led computational projects will therefore not be able to make use of the pre-exascale or exascale systems which will be deployed across Europe under the auspices of EuroHPC.

However, there is an opportunity to enable UK-based scientists and engineers to remain internationally competitive and address increasingly challenging problems in science and engineering, through investment in the proposed UKRI supercomputing ecosystem. This will accelerate innovations and technological step changes, with positive impacts on energy efficiency, environmental friendliness, wealth creation, industrial competitiveness and the development of highly-skilled personnel in the UK.

Finally, we note that the UK computational communities are fully engaged with international partners. UK groups have leadership roles in many international collaborations and UK researchers actively work with European partners to make use of PRACE resources. The international nature of collaboration within the computational community has been highlighted during 2020 by the many activities undertaken in support of COVID-19 related research: for example, UKRI membership of the US-led COVID-19 HPC Consortium (www.covid19-hpc-consortium.org) delivering computing resources and research software engineering to the effort. Such outward-facing activities will continue to be an essential aspect of computational research in the UK with new partnerships in the area of co-design expected (see Section 9).

[6] The European Technology Platform for High Performance Computing (ETP4HPC: https://www.etp4hpc.eu/) is the European consortium whose main task is to define research priorities and action plans in the area of high performance computing hardware and software development.

**Supercomputing and AI – complementarity not competition**

In recent years, Artificial Intelligence (AI), machine learning and deep learning have attracted significant levels of attention as new hardware developments (for example Graphics Processing Units [GPUs]) have made it feasible to implement algorithms which were previously intractable. While often represented as competing computational approaches, we consider both supercomputing and AI approaches to be complementary aspects of modern scientific computing. The efficiency of AI algorithms can sometimes be increased by means of supercomputing approaches to optimization or by re-engineering the algorithms to run on supercomputing hardware. Likewise, AI algorithms can deliver automated computational steering of simulations, decreasing the time to science for computationally intensive projects, and can facilitate the replacement of simple, parametric "sub-grid" models in multi-scale simulations with more informative, "surrogate" models. We expect that future supercomputing systems will increasingly be "dual-use", able to support both large-scale simulations and AI calculations, sometimes within a single workflow.

Simulations provide data with which to test complex models through confrontation with observations or experimental data. Increasingly, simulations are also used to train AI algorithms which can later be used to interpret new data in the context of known models, thereby testing the veracity of the models and enabling new understanding to be developed through the analysis of model failures.

AI algorithms can also be used to monitor the progress of simulations and identify those which are unlikely to deliver results which match known data. This automated computational steering can greatly reduce the number of simulations required to achieve a pre-determined quality of agreement with existing data and can therefore allow much more complex models to be matched to experimental data in a reasonable time.

The key development needed in this area is of so-called explainable AI, where sufficient information about the intermediate steps of an AI decision-making process are stored to allow researchers to identify the features which led to one model being favoured over another. Unless an AI black-box can be explored in this way, the benefits to the world of simulations will remain limited to reducing the number of dead-end simulations which are performed, rather than contributing directly to the creation of new scientific knowledge.

Finally, we note that the white paper on the UKRI supercomputing ecosystem does not include the needs of AI researchers in either the scale of services proposed or the funding envelope requested. The convergence of supercomputing and AI resources will lead to many research benefits, but in this case the scale of supercomputing provision (and hence funding) required would necessarily be larger in order to support both the AI and supercomputing research programmes in full.

Within this science case, the use of AI is implicit in many of the workflows discussed, and is not therefore presented as a standalone theme. The UKRI white paper "UKRI AI Review: our approach to AI" (in prep.) will contain more details about UKRI AI-related activity.

**Industrial applications**

The huge returns on investment associated with supercomputing are well documented. While the case presented here is focussed on academic research, progress in industrial research and innovation increasingly relies on supercomputing to deliver competitive advantage, contribute to increased productivity and enhance economic growth. The UKRI supercomputing ecosystem will support greater access from industry as well as academia, with opportunities anticipated for at least four categories of collaborative innovation projects:

1. Access to academic systems to develop ideas;
2. Understand scaling - exploring what large and/or novel systems can deliver if a company were to have access to such a system;

3. Develop new research results by developing new models and simulations or methods of exploiting data, in collaboration with academic departments.
4. Access to academic systems which are beyond the scale that any one company can own in order to answer specific research and innovation questions.

Industry will also benefit from the enhanced pool of skilled researchers who will receive training in advanced computing skills within the UKRI e-Infrastructure. In addition to contributing to a reduction in the digital skills gap, increased migration of researchers between academia and industry will undoubtedly facilitate greater levels of engagement between these communities.

**Structure of the case**

In the remainder of this document, we present the research goals of the UKRI research community grouped according to seven broad themes. The introduction to each theme provides some context and describes common features of the individual research cases presented.

**Acknowledgments**

We are grateful to the many colleagues who have contributed to the various cases presented below. This case could not have been written without their effort.

**References**

[1] *Neumann et al 2019: https:doi.org/10.1098/rsta.2018.0148*
[2] *Mizuta et al 2016: https://doi.org/10.1175/BAMS-D-16-0099.1*
[3] *J. H. Chen, et al., 2009, Comput. Sci. Discov., 2, 015001.*
[4] *A. Gruber, et al., 2018, Phys. Rev. F, 3, 110507.*

# 3 Expanding the frontiers of fundamental sciences

*Editors: Debora Sijacki[1], Mark Wilkinson[2] ([1]Cambridge, [2]Leicester)*

Simulation and modelling drive scientific discovery across all areas of theoretical astrophysics, particle physics, cosmology and nuclear physics. The research questions discussed in this section span length scales varying by over 50 orders of magnitude, from the largest structures in the Universe down to distances inside subatomic particles. Common themes include: precision calculations using existing theory for stringent tests against experiment or observation; quantitative understanding of the consequences of new physics; and the development of novel simulation and analysis techniques to facilitate exploration of new windows onto the Universe at both ends of the length-scale spectrum. The over-arching goal is to take these science challenges to a new level. In some cases, this means removing approximations and extrapolations; in others, it means including physics previously ignored because it was numerically too challenging. Examples are the inclusion of the effects of electromagnetism in simulations of the strong force inside the proton and neutron, or the inclusion of complex baryonic processes in simulations of galaxy formation and evolution. Achieving these aims already demands more than an order of magnitude increase in both CPU and storage above current resources (the DiRAC[7] facility currently provides just under 4PF). With major code re-engineering already underway, the capability calculations are moving towards exascale: the largest particle physics and cosmology massively parallel calculations will reach scales of 250PF in 2021 and 1EF from 2022 requiring significant uplift in the size of individual HPC facilities..

The next generation of experimental and observational facilities, many of them with significant UKRI (STFC – Science and Technology Facilities Council) investment, will generate unprecedented volumes of data. This section presents novel plans to exploit these vast data sets, extracting information from them through confrontation with detailed theoretical calculations and simulations. We will ensure that the UK reaps the scientific rewards of previous investments in major new facilities including: Advanced LIGO, ALICE, ALMA, ATLAS, CMS, DES, eBOSS, Gaia, LHCb and Planck and pave the way for future UK-led science from upcoming facilities including: CHEOPS, CTA, Euclid, FAIR, JWST, LHC-upgrades, PLATO, SKA, WEAVE, LISA and IPTA.

Increased supercomputing infrastructure will be critical to ensuring that the UK community continues to deliver step-changes across all STFC Frontier Science Challenges. Many valuable synergies already exist between the various strands of research within the DiRAC community, facilitating the cross-fertilisation of ideas between different fields: we will continue to exploit these links and foster new ones, including with partners across UKRI. For example, the hunt for dark matter brings together cosmologists, astronomers and particle physicists as it requires consistency between cosmological simulations of galaxy formation, detailed modelling of individual galaxies and determinations of the properties of dark matter candidates in theories Beyond the Standard Model, as well as determination of the probability of dark matter interacting with our detectors on Earth. We finally note that, in many areas, UK teams are working to develop multiple, complementary approaches, facilitating cross-validation of methods and verification of results.

[7] The DiRAC facility (Distributed Research utilising Advanced Computing; dirac.ac.uk) is the UK national facility which provides supercomputing resources for the theory communities in Particle Physics, Astrophysics, Cosmology and Nuclear physics.

## 3.1 The Extreme Universe: Black Holes and Gravitational Waves

*Contributors: Hannam[1], Figueras, Hawke, Lim, Reynolds, Sperhake, Witek ([1]Cardiff University)*

**Vision** – The first direct detection of gravitational waves (GWs) in 2015 revolutionised astronomy. The Advanced LIGO/Virgo detectors have already observed ten black hole (BH) binary systems, and hundreds more BH observations are expected over the next decade, as detector sensitivities increase. In addition, observations of binary-neutron-star mergers provide, in conjunction with electromagnetic (EM) observations, rich data on matter under the most extreme conditions ever measured. GW astronomy also provides the first strong-field tests of Einstein's general theory of relativity, and constraints on alternative theories of gravity.

**Key research challenges:** Accurate theoretical models of GW signals are essential to decode observations and measure source properties. Numerical simulations are necessary to construct these models. The UK plays a leading role in this work. For example, the only generic-binary model used to analyse all the detections to date, was produced using simulations performed on DiRAC2. Multimessenger astronomy also requires neutron star (NS) merger simulations, to link GW signals with EM observations and probe, for example, gamma-ray-burst models. Understanding the accretion of matter onto both isolated and binary BHs is essential for understanding EM counterparts to GW sources. Current studies fail to capture even the basic phenomenology of well-studied BH systems such as state-transitions and jet-cycles, presumably due to the limited numerical resolution, which prevents realistic implementation of important microphysics. Increased UK resources are essential to launch a new generation of accretion models that can be validated on well-studied BHs, then applied to GW counterparts. The GW breakthrough enables novel observational probes of fundamental physics including the dark matter quest, open questions in cosmology and modifications of general relativity that are well motivated by quantum gravity paradigms. Identifying their signature in GW data relies on accurate waveform models that are still scarce.

**Computing Demand:** Extending the accuracy and validity of these models is urgent, requiring over 100 million CPU hours per year over the next three years, and a factor of 2-3 increase is expected in the following three years. In addition, NS simulations, which are more computationally challenging and expensive than BH simulations, require a computational power increase of 10-20 times with respect to DiRAC2.5, with the inclusion of more realistic models of radiative transport and non-ideal magnetic effects being some of the immediate challenges. Furthermore, to produce the accurate waveform models the main numerical targets are (i) BH binaries and ultralight fields to probe for dark matter candidates and beyond-standard model particles; (ii) BH dynamics in modified gravity beyond proof-of-principle simulations that provided the strongest observational constraints to-date; (iii) GW signals generated by the merger of axion or boson dark matter stars; (iv) collapse and formation of compact objects beyond general relativity. This last work collectively requires additional ~50M CPU hours per year in 2019-22, and ~200M CPU hours per years through 2023-6.

**Track Record**

[1] *B. P. Abbott et al. (Virgo, LIGO Scientific), PRL, 116, 061102 (2016)*
[2] *M. Hannam et al., Phys. Rev. Lett. 113, 151101 (2014)*
[3] *E Berti et al, CQG, 32, 243001 (2015)*

## 3.2 Dark Matter, Dark Energy and Large-Scale Structure

*Contributors: Frenk[1], Bolton, Cole, Haehnelt, Jenkins, Kawata, Massey, Read, Baugh, Koyama, Lahav, Li, Nichol, Peacock, Pearce, Smith, Zhao, Shellard, Baldauf, Challinor, Efstathiou, Fergusson, Lewis, Sherwin ([1]Durham University)*

**Vision** – The two most important outstanding problems in cosmology and astrophysics today are the identity of the dark matter and the nature of the dark energy. The solution to these two problems is not only key to understanding the origin of structure in the Universe but will also have profound consequences for other disciplines, e.g. particle physics, gravitational physics and astronomy. Indeed, dark matter and dark energy provide the scientific motivation for ambitious (and expensive) experimental and astronomical science missions planned for the next decade, from direct dark matter searches in the laboratory (including CERN) to ESA's Euclid space telescope. Supercomputer simulations will play a central role in achieving the scientific goals of these missions.

**Key research challenges:** The current standard model of cosmology assumes that the dark matter consists of cold elementary particles (Cold Dark Matter - CDM) and that the dark energy is Einstein's cosmological constant, $\mathbf{\Lambda}$. In spite of its spectacular success in predicting key observables, the model rests on shaky foundations. Indeed, there are a number of well-motivated alternatives to CDM, such as sterile neutrinos, and to $\mathbf{\Lambda}$, such as modifications of Einstein's gravity. Cosmological simulations serve two roles: (i) formulate theoretical predictions for the properties of cosmic structures for different assumptions for the dark matter and dark energy; (ii) construct ``mock'' catalogues, based on (i), that include all experimental details. These are indispensable for the design of experiments and surveys, the interpretation of the data and the extraction of their physical significance. As an example, the next generation international $\boldsymbol{\gamma}$-ray observatory (with UK participation), CTA, will search for the tell-tale annihilation radiation of CDM; yet we do not have reliable theoretical predictions: the best simulations fall short in mass resolution by orders of magnitude.

**Computing Demand:** Over the past 5 years, UK researchers have developed new cosmological hydrodynamics simulation codes which are over 10 times faster than the best current codes and show near perfect scaling. Using 100,000 cores in the short term, and a million when available, it will be possible to achieve the mass resolution of $10^{-9}$ Solar masses needed to make the required theoretical predictions. Similarly, accurate predictions for gravitational lensing observations with LSST (2021) and Euclid (2022), which could, in principle, rule out CDM, are possible on a system delivering hundreds of PF.

DESI (Dark Energy Spectroscopic Instrument; 2020), LSST and Euclid will collect data for hundreds of millions of galaxies and quasars enabling clustering statistics to be measured with 1% precision. Simulations (and associated mocks) of comparable precision are essential to make sense of the data requiring over 500,000 cores for a few representative dark energy models. The associated data storage will reach tens of PB. This is an exascale computational programme, vital for the exploitation of the new astronomical facilities.

**Track Record**

[1] *Schaye, J. et al. 2015, MNRAS 446, 521*
[2] *Springel, V. et al. 2005, Nature 435, 629*
[3] *Navarro, J. F., Frenk, C. S., White, S.D.M. 1997, ApJ, 490, 493*

## 3.3 Galaxy and Structure Formation within Our Universe

*Contributors: Slyz[1], Bolton, Bower, Crain, Davé, Devriendt, Frenk, Fumagalli, Gibson, Haehnelt, Iliev, Jenkins, Kay, Khochfar, Kobayashi, McCarthy, Meiksin, Nayakshin, Pearce, Pontzen, Pritchard, Read, Sijacki, Theuns, Thomas ([1]Oxford University)*

**Vision** – In light of the forthcoming plethora of revolutionary observational datasets from ALMA, Gaia, MUSE, Euclid, LSST, JWST and SKA, understanding galaxy physics at the level needed to interpret the observations will require a multi-faceted theoretical approach: careful implementation of next generation "sub-grid" physics models in cosmological simulations of representative volumes complemented by highly resolved simulations of individual galaxies featuring detailed interstellar medium (ISM) modelling. The next generation of supercomputers will be key in pursuing such an ambitious goal. Indeed, they will allow (i) larger volumes to be simulated, to better pin down the role played by large-scale environment; (ii) higher resolutions to be reached in individual galaxies to better capture how internal processes shape global properties; and (iii) more of the relevant physical processes to be modelled *ab initio*.

**Key research challenges:** The emergence of the first galaxies during the epoch of reionisation is the new frontier. Modelling the interplay of ionising radiation and gas on the fly is poised to dramatically advance the field. As such, multi-Petascale and ultimately Exascale resources are critical not only to include radiative transfer physics but also to resolve the ISM through which photons must stream on their way to escape galaxies. Supercomputing facilities with 50-100PF would permit the inclusion of a non-equilibrium, multifrequency UV background mode to pin down the timing of reionisation and the relative contribution of stars and quasars to the ionising emissivity. Such simulations will provide invaluable direct 21cm predictions for future radio surveys. They would be complemented by extremely challenging radiative transfer cosmological re-simulations with sub-pc resolution to quantify the escape of ionising radiation from black hole accretion disks and star clusters, a key to JWST data interpretation. In parallel with the reionisation strand, the exploration of magneto-genesis scenarios through (radiative) magneto-hydrodynamic cosmological simulations becomes feasible with greater supercomputing resources. Indeed, solving the puzzle of what process magnetized the Universe to its current levels in different environments, a key science driver of SKA, requires enough resolution to capture the turbulent dynamo in galaxies. This constitutes a truly formiddable goal, especially when coupled with the requirement of simulating volumes large enough to track the pollution of the intergalactic medium (IGM).

**Computing Demand:** With expertise gained through hydrodynamical simulations of large-scale cosmological volumes (Eagle, Horizon, Illustris, Simba) and with major ongoing work to re-design codes for the next generation HPC facilities, such as SWIFT (also thanks to the DiRAC RSE effort), the UK community is ideally positioned to scale up the effort.. A ~ $(300\text{Mpc})^3$ volume, requiring 100TB of RAM, 50M core hours, and producing 1PB of legacy data on DiRAC-3 would increase by a factor ~8 with a 50-100PF uplift. An Exascale facility would deliver a volume comparable to the Euclid survey whilst resolving Milky-Way sized galaxies, which is crucial to calibrate the matter power spectrum and thereby constrain the nature of dark energy. It would further provide a goldmine for the interpretation of SKA data which will open an almost entirely unexplored window into neutral hydrogen at intermediate redshifts, allowing us to trace the joint evolution of gas and stellar mass over cosmic time. Identifying a Milky-Way like galaxy in this volume for re-simulation to the resolution level of forthcoming Gaia data (~ a billion stars), would also become possible.

**Track Record**

[1] *Bolton et al. 2017, MNRAS, 464, 897*
[2] *Rosdahl et al. 2018, MNRAS, 479, 994*
[3] *Katz et al. 2019, MNRAS, 484, 2620*

## 3.4 Star and Planet Formation

*Contributors: Alexander[1], Bate[2],* B*onnell, Clark, Clarke, Madhusudhan, Mayne, Nelson, Ogilvie, Owen, Paardekooper, Rice, Stamatellos ([1]University of Leicester, [2]Exeter University)*

**Vision** – In the last decade we have discovered thousands of new planetary systems, but most exoplanets look very different to anything we see in the Solar System. Understanding the formation of planets has become one of the major challenges in modern astrophysics.

**Key research challenges:** Planets form in cold discs of dust and gas around young stars. But historically, star formation and planet formation have generally been treated as separate processes. This is largely because the discs around young stars where planets form have been difficult to resolve observationally. However, the unprecedented resolution and sensitivity of several new instruments (particularly ALMA (Atacama Large Millimeter/submillimeter Array) and VLT/SPHERE (Spectro-Polarimetric High-contrast Exoplanet REsearch)) are bringing these two fields together by allowing us to study the structure of discs around young stars, including those that are still accreting most of their mass. Similarly, although many physical process are involved in both star and planet formation (e.g. gravitational and fluid dynamics, magnetohydrodynamics (MHD), radiation, astrochemistry), modelling the large and small spatial scales together has been prohibitively expensive. However, recent supercomputers are starting to provide the computational power necessary to study star formation and disc formation and evolution simultaneously.

**Computing Demand:** The key for both fields is to span large dynamic ranges of $>10^6$ in space and $>10^9$ in time, making both star and planet formation two of the most technically challenging and computationally expensive problems in astrophysics. Compute power in the range 1.5-15PF is needed to allow systematic simulation of the dominant physical processes on global time-scales for physically relevant parameter ranges. Existing codes are efficient for individual calculations up to ~14k cores (~1PF); in 1-2 years, with additional code development, higher resolution calculations including additional physics could be carried out but would then require ~15PF of compute resource for individual simulations. In star formation, calculations of how star clusters form would have the resolution to study disc formation and evolution in detail, allowing us to understand the initial conditions for planet formation. We also need to perform larger calculations that enable us to study the formation of clusters containing thousands of stars in which massive stars form alongside low-mass stars. The relative rarity of stars with masses above 10 solar masses means that current calculations of star cluster formation usually do not produce any massive stars, although these have the greatest impact on the energetics and chemical evolution of galaxies. We would be able to study the effects of the radiation from massive stars on the discs of low-mass stars (e.g. disc truncation and evaporation). For both star formation and protoplanetary discs, increased computational power is necessary to improve our treatments of physical processes, such as including non-ideal MHD processes and improving the accuracy of radiative transfer in multi-dimensional simulations. In the planet formation context, time-consuming calculations of dust-gas dynamics and gas-grain chemistry in protoplanetary discs are required to determine the thermal properties of the discs. These are necessary to produce the realistic thermo-chemical models necessary to understand observations made with ALMA, and new facilities such as JWST. For the dynamics of protoplanets in discs, such an increase in computational power would allow long-duration simulations of planet-disc interactions to move from 2-D to 3-D, opening the door to detailed statistical simulations of exoplanet populations, for comparison to future exoplanet surveys such as VLT-ESPRESSO (Echelle SPectrograph for Rocky Exoplanets and Stable Spectroscopic Observations) and the European Space Agency (ESA) PLATO mission (PLAnetary Transits and Oscillations).

**Track Record**

[1] *Haworth et al., 2016, MNRAS 463, 3616*
[2] *Bate 2014, MNRAS, 442, 285*

## 3.5 Solar system magnetism

*Contributors: Arber[1], Hood[2], Hughes[3], Tobias[3] ([1]Warwick, [2]St Andrews, [3]Leeds)*

**Vision:** Develop a comprehensive model for how the Sun's magnetic field is generated, how it is transported and dissipated and what effect it has on planetary atmospheres. Through this programme maintain the UK's leading role in MHD simulations of solar system magnetism.

**Key research challenges**: The Sun generates a magnetic field, with a 22-year cycle, which is transported from its formation inside the Sun, through the convective outer layer and into the visible corona. During this journey it forms sunspots and active regions (ARs) on the solar surface and the magnetic field interacts with the overlying coronal plasma, heating it and generating flares, coronal mass ejections (CMEs) and the solar wind. These latter dynamic events interact with planetary magnetic fields. Major research challenges exist at all stages of this magnetic field generation and evolution. The UK maintains a leading international role in the study of this magnetism through magnetohydrodynamic (MHD) simulations, coupling fluid motion of the ionised atmospheric plasma with the magnetic field. The major research challenges, in order from the solar interior outwards, are:

***Solar interior:*** The dynamic events observed in the solar atmosphere arise as a result of magnetic activity – driven by a hydromagnetic dynamo. Replenishment of the magnetic field generates sunspots and regulates the solar cycle that critically affects space weather. The discovery of the *tachocline* (i.e. the strong shear transition region between the radiative interior and the differentially rotating outer convective zone) has prompted a wide range of fundamental theoretical questions, such as: the location of the solar dynamo; the possibly different mechanisms of small- and large-scale field generation at the turbulent conditions that pertain to the Sun; the instability and emergence of the magnetic field; and the formation and maintenance of the tachocline.

***Solar atmosphere:*** The energy generated in the interior can be transported to the solar atmosphere through either (i) observed waves, (ii) flux emergence or (iii) slow convective motions. In the partially ionized chromosphere, the waves steepen and form shocks that help to heat this layer and, in the corona, detailed models are required to determine how the waves are generated, propagate and dissipate their energy and how they contribute to the heating of the corona. An alternative coronal heating mechanism assumes that the magnetic energy is continuously stored in the coronal field and released in multiple locations through reconnection. The modelling of an AR magnetic field, over several weeks, requires following its evolving equilibria, using a flux transport model and a relaxation method. The aim over the next 5-10 years is to provide a systematic model of the evolution of ARs, coronal heating models and the prediction of eruptions that may generate space weather events.

***Planetary atmospheres:*** A key challenge is to understand how the magnetospheres couple to their surrounding environment, and how mass, energy and momentum is transferred through the system. Previous studies have developed 3D MHD codes to model planetary magnetospheres. Over the next 5-10 years increased supercomputing will enable a new generation of world-leading magnetospheric simulations that will provide an essential counterpoint to *in situ* and astrophysical observations. Previous studies have shown that magnetospheres are dynamically coupled to planetary ionospheres by currents flowing between the two. A significant leap forward will be possible by coupling deep convective models of giant planet interiors to high-resolution simulations of their atmospheres. On terrestrial planets, the surface topography plays an important role in controlling phenomena such as the mineral dust cycle, but current global models are woefully under-resolved to take advantage of the very high resolution topographic maps now available; the power afforded by a 10-20 fold increase in supercomputing will allow these datasets to be properly exploited.

**Computing demand**: All the MHD simulations above involve many different length-scales and time-scales and model plasmas with widely varying densities and temperatures. Most leading edge simulations are in 3D so a 16x increase in compute power is needed to double resolution up to

$(2048)^3$ – with the resulting calculations requiring 32M core hours each. For explicit codes, these higher resolution simulations could be carried out immediately, with the potential by 2022 for a further doubling of resolution or increasing the domain size at the same resolution. For spectral and implicit codes, additional development work is required. While seemingly incremental, such resolution increases will provide sufficient scale separation in the simulations to compare with asymptotic regimes and thereby definitively chose between competing theories. The large simulation volumes allowed will permit models coupling layers of atmospheres with magnetic field structures which match observations.

**Track Record**


[1] *C. S. Brady and T. D. Arber, ApJ, 829, (2016) 80*
[2] *Hood, A.W., Cargill, P.J., Browning, P.K. and Tam, K.V., ApJ, 817, (2016)*
[3] *P. Pagano, D. H. Mackay, A. R. Yeates, J SWSC, Submitted on: 16/05/2017*


## 3.6 Stellar Hydrodynamics, Interstellar Medium, Astro-chemistry and Computational spectroscopy of exoplanets


*Contributors: Baraffe[1], Hirschi[2], Dobbs[1], Viti[3],* Yurchenko[3], *Tennyson, Mant, Al-Rafaie, Chubb, Coles ([1]Exeter University, [2]Keele University, [3]UCL)*


**Vision:** Stellar evolution models are fundamental to nearly all fields in astrophysics, from exoplanet characterisation, star formation, to galactic and extra-galactic research. However, many fundamental stellar (magneto-)hydrodynamics processes are still poorly understood and require multi-dimensional approaches to improve their description. Furthermore, to date our understanding of star formation has been limited by both the huge dynamic range of scales, and the complex physical processes involved, particularly stellar feedback. Around many stars, thousands of extra-solar planets have been discovered in recent years and we still know little about their nature. From the observational perspective, interpretation of exoplanetary atmospheres places immense demands on the spectroscopic data required to characterise these new worlds.

**Key research challenges**: The stellar hydrodynamics community has now entered a new era with the development of state-of-the-art numerical tools based on sophisticated algorithms and accounting for complex physics that enable 3D numerical simulations of key stellar physics processes. These require at least two orders of magnitude increase in computational power to revolutionise the field. Large scale supercomputing facilities with 30-100PF will enable multiscale and multiphysics simulations of star formation which will answer questions such as how large scale structure influences star formation, the role of gas dispersal on cluster evolution, and the initial conditions needed not just for individual stellar clusters and OB associations, but also for numerous co-existing young stellar groups in large complexes. The use of molecules to determine the physics and chemistry of the dense gas in our own as well as in external galaxies requires the determination of the sensitivity of astrochemistry to the range of physical parameters that determine the appearance of the different types of galaxies. In computational spectroscopy, present atmospheric retrievals are severely impacted by the lack of corresponding spectroscopic data. Even for relatively simple molecules, the opacity functions are generally complicated as they vary strongly with both wavelength and temperature. A comprehensive list of spectroscopic transitions, or line list, for a single molecule can contain significantly more than ten billion lines which puts an enormous pressure on the computational and storage resources. The UCL ExoMol team is world leader in providing molecular line lists for atmospheres of hot gas giants and lava planets. Exactly these types of hot planets will be the likely targets of NASA's JWST (launch 2021) and ESA's ARIEL (Atmospheric Remote-sensing Infrared Exoplanet Large-survey; launch 2028, led by UCL).

**Computing demand**: In stellar hydrodynamics, high resolution 3D simulation of the convective envelope of a solar-like star requires 50M core hours to cover one single convective turn-over timescale, while tens are required for statistically significant results. To follow the Silicon-burning phase of a 15 solar-mass star, just prior supernova explosion, requires ~100M core hours just for this phase. The generalisation of full 3D studies to a large range of stellar masses and covering relevant statistical times will only be feasible with systems delivering hundreds of PF. This field is also particularly active in developing strategies to overcome scaling and parallelization issues and will be dramatically boosted by exascale computing capabilities. Traditionally, astrochemistry has always been dominated by trial and error grid-based analysis combined with simple statistics. However, in the era of extremely large, complex and heterogeneous datasets (e.g. ALMA, JWST, SKA) a paradigm shift in the way we interpret astrochemical models, involving Bayesian, high statistics Monte Carlo (MC) simulations as well as Machine Learning techniques, is needed. The generation of databases consisting of millions of models that extend over a large parameter space, covering the physical conditions across the different types of galaxies will require an increase of current DiRAC compute power by a factor of 50. In computational spectroscopy, within the next 5 years (2019-2024), we will produce line lists containing nearly trillion of transitions for larger molecules (up to 10 atoms) and more complex (open-shell) species than have been computed before. To complete this ambitious task, for which the individual line lists will become at least a factor 5 more computationally expensive than at present, requires large-scale supercomputing resources able of supporting 1000s of calculations which individually use 2k cores or more and with individual RAM footprints of up to 12TB. The delivery of line lists for 5 molecules during 2021-24 will consume more than 30M core hours p.a., between 5 and 10 times what can be provided by existing DiRAC services. Software development effort will also be required to ensure that the code scales efficiently to these larger core and memory footprints.

**Track Record**

[1] *Baraffe et al. 2017, ApJL, 845, L6*
[2] *Dobbs et al. 2017, MNRAS, 464, 3580*
[3] *S.N. Yurchenko, at al, A&A., 605, A95 (2017)*

## 3.7 Quantum ChromoDynamics (QCD) in extreme environments

*Contributors: Aarts, Allton, Buividovich, Hands, Langfeld, Lucini, Rago*

**Vision** – Simulations of lattice QCD will yield a quantitative description of matter in the hadronic and quark-gluon plasma phases, to complement the experimental accelerator programmes, and provide necessary theoretical understanding of the strong interaction under extreme conditions.

**Key research challenges:** Numerical simulations of lattice QCD have firmly established the thermodynamic properties (e.g. pressure, entropy, fluctuations, etc.) of QCD at non-zero temperature $T$, for physical quark masses in the continuum limit, providing first answers to questions such as *What is the nature of nuclear and hadronic matter?* and *How do the laws of physics work when driven to the extremes?* Experimental progress in heavy-ion collisions (HICs) at the LHC and the future FAIR facility (Facility for Antiproton and Ion Research) will drive the research agenda in the coming years, which concerns the detailed quantitative understanding of spectral features (masses, widths, dissolution and melting, transport, connection to chiral symmetry) at vanishing and small baryon chemical potential $\mu$. The FASTSUM collaboration, using DiRAC resources, has led internationally in this area, with studies of the survival/melting of quarkonia (charmonium, bottomonium), of hyperons and chiral symmetry, and with the first calculation of a transport coefficient, the electrical conductivity $\sigma$ and the related charge diffusivity $D$. A flagship project will be a precise computation of hadron masses in thermal QCD using (near-)physical quark masses; their dissolution at high temperature will provide indispensable information for HIC phenomenology. The

medium-term goal (up to 2024) is the determination of the QCD spectrum, including transport of energy-momentum, captured by the shear and bulk viscosity, for light and heavy quarks at all temperatures probed in the heavy-ion programme, using physical quark masses; the ambitious long-term goal (up to 2030) is to achieve this in the continuum limit.

Following on from the current Beam-Energy Scan (BES) at RHIC (Relativistic Heavy Ion Collider), the next generation of low-energy nuclear colliders, the Nuclotron-based Ion Collider fAcility (NICA) at the Joint Institute for Nuclear Research (JINR; from 2020) and FAIR at GSI (from 2022), will focus on the region of the QCD phase diagram at lower temperature but with nonzero baryon density. To support this experimental activity, it is necessary to include a nonzero chemical potential μ. FASTSUM is leading this endeavour by including corrections in spectroscopy using an expansion in $\mu/T$, building on pioneering QCD thermodynamics work by the Swansea-Bielefeld collaboration in the early 00s. The medium-term goal here is to establish the response of hadronic properties and transport at first and second nontrivial order in the $\mu/T$ expansion. At larger chemical potentials and lower temperatures, relevant for nuclear matter and neutron stars, alternative methods not relying on standard importance sampling are required, due to the sign problem. Various pioneering proposals, including the complexification of the degrees of freedom and constrained sampling of the density of states, are being explored. The examination of various approaches will continue at short and intermediate stages; a breakthrough will deliver the determination of the QCD phase diagram as a long-term goal.

**Computing demands:** In spectral studies of the hadronic and quark-gluon plasma phases, the interest is in temporal evolution rather than bulk thermodynamic behaviour, which is accessible using highly anisotropic lattices, with a finer temporal than spatial lattice spacing ($a_t \ll a_s$). Consequently, two independent continuum limits, $a_t \to 0$ and $a_s \to 0$, exist. Moreover, a range of temperatures both below and above the deconfinement transition is needed, at fixed lattice spacings and quark masses, which calls for a fixed-scale approach. And while most thermodynamic studies use the staggered-fermion formulation, improved Wilson quarks are better suited for spectroscopy on lattices of finite temporal extent, while being more expensive to simulate. To obtain predictions of spectral quantities and transport coefficients at physical quark masses and fixed lattice spacing, we estimate that systems of order 10PF are required. To reach the continuum limits, $a_{t,s} \to 0$, via controlled simulations at a sequence of lattice spacings at fixed physical volume and at physical quark masses, will require systems at least an order of magnitude larger (~100PF).

**Track Record**


[1] *Aarts et al., 2014, JHEP 07 [097]*
[2] *Aarts et al., 2019, PRD99 [074503]*
[3] *Aarts et al., 2015, JHEP 02 [186]*
[4] *Aarts et al., 2014, JHEP 09 [087]*
[5] *Langfeld et al., 2014, PRD90 [094502]*


## 3.8 Hadron Spectroscopy and Structure

*Contributors: Bouchard, Davies, Horsley, McNeile, Rakow, Thomas*

**Vision** – **"***How do quarks and gluons form hadrons?***"** is a key STFC Science Challenge. Lattice QCD (LQCD) provides the only systematically-improvable way to perform wide-ranging *a priori* computations of the mass spectra and properties of hadrons, and so answer this question and enable experimental results to be fully exploited. UK Groups, as part of international

**Key research challenges:** **Spectroscopy of lower-lying hadrons.** Accurate computations of the masses of lower-lying hadrons in each flavor sector are important benchmarks of LQCD. The HPQCD consortium (High Precision QCD) has accurately computed the spectrum of stable mesons

including light, s, c, and b quarks, using DiRAC resources – LHCb (Large Hadron Collider beauty) and CMS (Compact Muon Solenoid) recently discovered the $B_c'$ meson in agreement with HPQCD's prediction. QCDSF/UKQCD is performing complementary studies of baryons – current results support LHCb's 2017 observation of the doubly-charmed $\Xi_{cc}^{++}$ state over the candidate found by SELEX in 2002, but to definitively rule out the SELEX result, and make precise predictions for other doubly-charmed baryons, requires a reduction in statistical errors and a finer lattice. Results for low-lying hadron masses are reaching a precision where isospin-breaking effects (both $m_u \neq m_d$ and QED) are becoming important. HPQCD plan to include isospin breaking in the calculation of meson masses with heavy quarks and QCDSF/UKQCD will continue its programme to include electromagnetic corrections to hadron structure. The long-range nature of QED means that a wide range of different volumes and good theoretical understanding of UV and IR electromagnetic effects are needed, and so a step change in computational power is required. For example, doubling the lattice size would require a factor of at least 20 in computer time, or about 4 EF days.

**Excited hadron spectroscopy.** Experimental discoveries of a number of resonances which do not appear to fit the conventional pattern, e.g. "*X,Y,Z*'s", charged charmonium and bottomonium-like structures, have created intense interest in excited mesons. Further results are expected from Jefferson Lab (JLab), the LHC, Beijing Spectrometer Experiment III (BESIII) and the planned PANDA experiment. The HadSpec collaboration has made extensive state-of-the-art lattice calculations of hadron spectra and developed techniques to study resonances (unstable hadrons). Examples using DiRAC include studies of $D\pi$, $D\eta$, $D_s\bar{K}$ scattering (the first coupled-channel scattering calculation involving charmed mesons), potential candidates for exotic charmonium-like tetraquarks, and the $b_1$ resonance. HadSpec are now performing a systematic study of channels relevant for the enigmatic "*X,Y,Z*'s" – these involve many coupled channels and so are challenging and computationally expensive. The pioneering calculations were performed with heavier-than-physical pions and only one lattice spacing. Quantifying the systematic effects will require calculations on more lattice ensembles, including larger volumes, which gives a huge increase in the computational power and storage required. HPQCD are planning numerically very challenging studies of exotic hybrid mesons (with an excited gluonic field) – definite results require a significant increase in resources.

**Hadron Structure.** Form factors (FFs) provide probes of hadron structure, e.g. charge and spin distributions and magnetic moments. HPQCD will determine EM FFs for $\pi$ and K mesons, giving predictions for JLab experiments and testing how the non-perturbative physics at low 4-momentum transfer ($q$) evolves into perturbative physics at high $q^2$. Calculations up to $|q^2|$ = 6 GeV$^2$ for a 'pseudo-pion' made of *s* quarks show that the FF here is a long way from the perturbative result. Future calculations will push to higher $q^2$ – a factor of 10 increase in CPU power is expected to enable us to probe up to $|q^2| \approx 20$ GeV$^2$. Mapping the large-$q^2$ behaviour of the neutron and proton elastic FFs will be a focus at JLab. QCDSF/UKQCD has developed a method giving access to higher $q^2$, but current results are limited to heavy pion masses and so improvement is essential.

CalLAT (California Lattice) plan to compute nucleon FFs which are vital for fully leveraging results from neutrino experiments (e.g. the Deep Underground Neutrino Experiment (DUNE), T2K, HyperK). Improved calculations yielding $g_A$ with better than 0.5% precision are needed to distinguish between discrepant experimental values for the neutron lifetime and the proton charge radius. Beyond this, further reductions in uncertainty will be essential to the future experimental neutrino program and require exascale resources. Based on current CalLAT calculations using an INCITE (Innovative & Novel Computational Impact on Theory & Experiement) programme allocation on Summit which is expected to achieve 0.5% precision for nucleon $g_A$, 5% for the nucleon axial form factor over a range of up to $|q^2| \approx 1$ GeV$^2$, and a few % for nucleon charge radii, the computational cost of the next generation of calculations will be 10-20 times higher than the current allocation of 200K node hours on Summit.

Scattering processes at the LHC, crucial for testing the Standard Model (SM) and constraining new physics, depend on non-perturbative QCD physics encoded in parton distribution functions (PDFs). The only *ab initio* way to compute PDFs is using LQCD but previous determinations were restricted to their lowest moments, limiting their usefulness. QCDSF/UKQCD plan to use recent theoretical advances to determine complete PDFs but, in order to impact experiments, the calculations need large volumes with physical quark masses, requiring substantial additional computing resources. Using a finer lattice than is currently possible, to allow for larger momentum transfer and approaching the physical quark mass, would initially require a factor of 50 in computer time, or about 10 EF days. To maintain international leadership, such calculations should begin as soon as possible.

**Computing demands:** Quantifying some of the systematic effects in investigations of excited hadrons will require calculations on more lattice ensembles which requires a huge increase in the computational power (~40 million core-hours) and storage (~200 TB, along with ~1 TB local storage on each node or other appropriate temporary storage). Precision studies of exotic charmonium and bottomonium hybrids (10 MeV errors at 4 lattice spacings between 0.15 fm and 0.06 fm) will require ~70M core-hours, with an additional ~1400M core-hours to perform calculations which allow for their unstable nature. Pushing calculations of the 'pseudo-pion' form factor to higher $Q^2$ will require ~10-20M core-hours. Improving precision in studies of nucleon form factors will require exascale resources: a system delivering a sustained 1 EF of compute by 2021/22 would make it possible to build on the INCITE programme and allow physical pions with a continuum extrapolation.

**Track Record**


[1] *Horsley et al (QCDSF/UKQCD), J. Phys. G43 10LT02 (2016),* arXiv:1508.06401
[2] *Moir et al (HadSpec), JHEP 1610, 011 (2016),* arXiv:1607.07093
[3] *Chambers et al (QCDSF/UKQCD), Phys. Rev. Lett. 118, 242001 (2017),* arXiv:1703.01153
[4] Chang et al (CalLAT), Nature 558, no.7708, 91-94 (2018), arXiv:1805.12130
[5] *Koponen et al (HPQCD), Phys Rev D96, 054501 (2017),* arXiv:1701.04250


## 3.9 Beyond the Standard Model (BSM) physics

*Contributors: Boyle, Cossu, Del Debbio, Drach, Garron, Jüttner, Lucini, McNeile, Piai, Portelli, Rago, Sachrajda, Skenderis*

**Vision** – Notwithstanding its tremendous successes, the Standard Model leaves many crucial questions completely unanswered that BSM extensions are trying to address. These include the origin of the particle spectrum, the origin of dark matter, the origin of the matter/antimatter asymmetry, and more generally how to provide a satisfactory theory that is predictive at arbitrarily short distances.

**Key research challenges:** The “naturalness” problem of the Higgs boson highlights the need of a mechanism capable of generating an isolated light scalar particle. Such a scenario can be embedded in composite Higgs theories, in which the Higgs is a bound state of a yet unknown strong interaction, whose lightness can be interpreted in terms of a Nambu-Goldstone phenomenon. The field of lattice BSM physics has the main objective of identifying theories that yield parametrically large-scale separations, and hence can be used to construct BSM models with realistic mass hierarchies without violating existing constraints from precision measurements (STFC science challenges[8] **C:1** and **C:9**). Furthermore, lattice investigations can shed light on the nature of the electroweak phase transition in composite Higgs models, identifying whether such models can allow for electroweak baryogenesis and thereby explain the baryon asymmetry of the universe (**C:8** and **A:1**). Group-theoretical classifications of the four-dimensional fermionic gauge theories providing an ultraviolet completion of composite-Higgs models exist: however, any first principles study of these theories requires lattice

[8] https://stfc.ukri.org/research/science-challenges/

calculations. These theories are usually characterized by a rich matter content featuring fermions in multiple representations and various realizations of the gauge symmetry. The technology has only recently been developed, with substantial involvement of UK groups: simulations are starting to be attempted.

Dark matter is an essential ingredient in current models of the evolution of our universe. Aside from its gravitational interactions, the precise nature of particle dark matter has remained elusive, but some properties have now been strongly established: Dark matter must be stable, electrically neutral and effectively neutral with respect to the standard model. The required stability suggests the existence of some appropriate symmetry in the dark sector which prevents or greatly suppresses dark matter decay, while the weak interaction with the standard model motivates an early-universe connection which is stronger than gravity. These observations lead naturally to the idea of strongly-coupled composite dark matter (**A:3** and **C:4**). A dark sector with non-Abelian gauge interactions and fundamental constituents charged under the standard model can give rise to a stable, neutral dark matter candidate which can interact more strongly in the early universe. Such a sector can arise quite naturally from composite Higgs boson extensions or could simply arise as an additional dark sector. Lattice gauge theory allows for precise, quantitative study of many interesting quantities, including the full spectrum of composite states and matrix elements important for direct and indirect detection, relic abundance, self-interactions, and collider production. As for the previous case simulations are starting to be attempted and require large scale computations.

From a complementary perspective, Lattice QCD calculations are needed to predict amplitudes for hadronic processes in BSM scenarios that might be seen in experiments (**C:1**, **C:5** and **C:9**). BSM theories give rise to effective interactions that couple quarks inside hadrons to new particles, resulting in novel hadron decay processes or modified rates for decays already observed. In neutral meson mixing, new physics can produce 4-quark operators with different spin-colour structure to those in the SM. RBC/UKQCD is currently computing the corresponding matrix elements in the kaon sector (relevant to NA62 and the future Project-X at Fermilab). The domain-wall formalism avoids artificial mixing of left- and right-handed amplitudes in discretization artefacts and so simplifies the approach to BSM amplitudes with non-SM handedness. Using new renormalisation schemes a precision of 5% was achieved for the relevant BSM matrix elements. Further improvements using physical *u/d* quark masses and finer lattices, together with step-scaling to higher energy scales will constrain the error to ~1%.

Another strand of BSM activities deals with the theory of gravity at the very early stage of the Universe and requires a proper understanding of quantum gravity (**A:1**, **A:3** and **C:3**); the Cosmic Microwave Background (CMB) provides an observational window to this epoch. Quantum gravity is generally expected to have a holographic nature, i.e. it can be described by a dual quantum field theory (DQFT) in 3-dimensional flat space. Through this conjecture the CMB spectrum can be computed from the unknown DQFT. The predicted power spectrum has recently been tested against WMAP (Wilkinson Microwave Anisotropy Probe) and Planck data. Two important facts emerged: firstly, from an observational point of view the holographic description of the CMB is competitive with the standard model of cosmology, the ΛCDM model. Secondly, Planck data strongly suggests that the DQFT is a Yang-Mills theory coupled to non-minimal scalars with quartic interactions. The main caveat is that the CMB spectrum is currently reconstructed from a perturbative calculation in the DQFT, which suffers from strong infrared (IR) divergences; however due to the presence of an IR fixed point this is not thought to be an issue in the full theory. Moreover, the first comparison with Planck data indicates that the DQFT is strongly coupled in the low-multipole region of the CMB spectrum, so perturbative approaches cannot succeed. Both issues can be solved using lattice simulations of the DQFT. The models favoured by Planck can be unambiguously established or rejected through lattice.

**Computing demands:** By 2021, mixed representation composite Higgs calculations including Majorana fermions and requiring of order 10 PetaFlop-years (PFy) per representation will need to be routine. Lack of computing power currently limits the physical realism of dark matter calculations:

an exhaustive study of a single dark matter model requires ~1 PFy. The inclusion of physical *u/*d quark masses and finer lattices in BSM hadronic process amplitudes requires a factor ten increase in available computing resources, equivalent to a dedicated ~40PF system, to perform the proposed calculations.

**Track Record**


[1] *Bennett et al., JHEP 1803 (2018) 185*
[2] *Cossu et al., arXiv:1904.08885 [hep-lat].*
[3] *Boyle et al. [RBC and UKQCD Collaborations], JHEP 1710 (2017) 054*
[4] *Drach et al., EPJ Web Conf. 175 (2018) 08020*


## 3.10 LHC Phenomenology

*Contributors: Banfi, Caola, Englert, Forshaw, Klein, Maitre, Seymour, Spannowsky*

**Vision** – The Large Hadron Collider (LHC) is the highest energy particle experiment devoted to tackle the science challenge "What are the basic constituents of matter and how do they interact?" Large-scale perturbative QCD calculations will probe physics beyond the standard model using LHC data.

**Key research challenges:** Since the Higgs discovery in 2012, the LHC community has reinforced its efforts to find new particles via direct and indirect measurements. Beyond new resonant peaks in data, signals of new physics beyond the Standard Model of elementary particles are accessible only through deviations between state-of-the-art predictions and observed data. The extraordinary precision of many measurements challenges the accuracy of available theoretical calculations, in particular QCD predictions in the perturbative, high-energy regime.

So far, data have shown very good agreement with the Standard Model. However, with more luminosity, more differential measurements will be available, which could probe interactions that have not been accessible to date. Interpretation of future data requires accurate predictions for processes with more particles in the final state, and/or the modelling of exclusive experimental cuts. At the same time, analysing new data in the light of effective field theories will make it possible to constrain new physics scenarios and to gain insight on the typical scale of new physics. This quest for precision will be at the core of the LHC phenomenology program until the end of its high-luminosity phase (~2040).

An important research direction is perturbative QCD calculations at the highest possible accuracy, currently next-to-next-to-leading order (NNLO). Low-multiplicity perturbative calculations (Higgs, vector bosons and tops) will still play a decisive role in the interpretation of LHC data. On the one hand, theoretical efforts will concentrate on novel techniques to improve their speed and accuracy (Caola [1]). On the other hand, comparison of NNLO calculations with data will allow extractions of the parameters of the Standard Model with significantly higher precision, as well as an improved understanding of the structure of the proton (Klein). At the same time, extending NNLO calculations to higher multiplicities, e.g. vector bosons with additional jets, requires strategies to optimise the use of CPU time by storing event files (Maitre [2]). More exclusive measurements push perturbative QCD to the edge of its applicability, and either analytical resummations (Banfi) or Monte-Carlo simulations with parton-shower event generators (Forshaw, Seymour) are required to obtain meaningful predictions. Parton-shower event generators are being pushed at the amplitude level to incorporate more quantum interference effects that are crucial at the sub-10% level [3]. Last, development of effective field theories and their implementation in parton-shower event generators will be crucial to exploit future LHC data to constrain new physics scenarios (Englert, Spannowsky [4]).

**Computing demands:** Theoretical predictions for LHC phenomenology mainly consist of serial Monte-Carlo integrations with modest RAM and storage requirements. The key factor to ensure

timely predictions is the CPU speed. In particular, NNLO calculations require the largest amount of supercomputing resources. Describing the tails of differential distributions, or multi-differential distributions, where signals of new physics might hide, require theoretical control over higher multiplicity final states. The resources required increase significantly with the multiplicity of the final state: for example a 2→1 process (e.g. Higgs production) requires ~1k core hours, while a 2→2 process (e.g. Higgs plus one jet/dijet production), the current state of the art and a key goal of precision phenomenology over the next five years, require from 0.1M to 1M CPU hours per process being studied. Therefore, an increase of a factor of 10-20 in computing resources is needed, not only to increase the number of predictions available, but also to tackle difficult integration regions that constitute the bottleneck of many QCD calculations and can easily increase the CPU requirements by a factor of 10. In the longer term, 2→3 processes (e.g. Higgs plus two jets) represent the next computational frontier. Without any breakthrough in computational efficiency, these would require a similar relative uplift in computing power as that seen in moving from 2→1 to 2→2 processes.

**Track record**

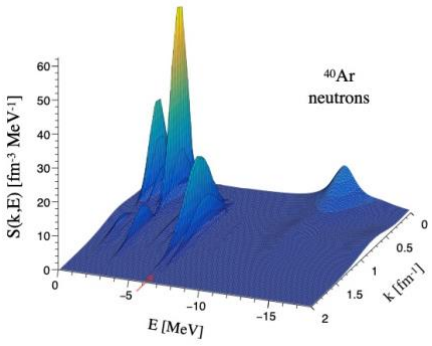

[1] *F. Caola, K. Melnikov, R. Röntsch, arXiv:1902.02081 [hep-ph]*
[2] *D. Maître, J.Phys.Conf.Ser. 1085 (2018) no.5, 052017*
[3] *R. A. Martínez, M. De Angelis, J. R. Forshaw, S. Plätzer, M. H. Seymour, JHEP 1805 (2018) 044*
[4] *C. Englert, P. Galler, A. Pilkington, M. Spannowsky, arXiv:1901.05982 [hep-ph]*


## 3.11 Nuclear physics

*Contributors: Barbieri, Pastore, Dobaczewski, Rios, Stevenson, Walet*

**Vision** – Perform parameter-free, ab initio calculations of the properties of nuclei up to A~140 and beyond. Development of a predictive and accurate description of nuclear fission, rooted in quantum many-body theory.

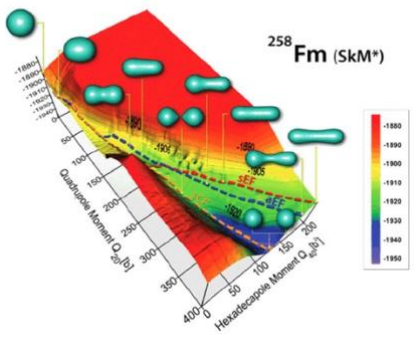
**Key research challenges: 1) *Ab initio* studies of nuclei and stellar matter at the limits of stability.** *Ab initio* nuclear theory aims at parameter-free predictions of nuclei, based on the interweaved advances of (1) theories of the nuclear force, constrained by QCD, and (2) computational many-body theory. *Ab initio* methods have recently achieved significantly higher levels of accuracy and are addressing several scientific questions: the systematics of nuclear radii, nuclear driplines and optical potentials will be crucial to the science programs of future facilities like the Facility for Antiproton and Ion Research (FAIR; Germany) and the Facility for Rare Isotope Beams (FRIB; USA); information on hyperon-nucleon systems from the Japan Proton Accelerator Research Complex (J-PARC; Japan) and from Lattice QCD simulations can shed light on neutron star matter (and on the largest possible star mass); the *ab initio* computed spectral function (see figure for the $^{40}$Ar isotope) is a crucial ingredient to neutrino-nucleus scattering. Predictions of the latter will benefit oscillation experiments to learn the neutrino mass hierarchy (e.g., at DUNE, USA) and will help clarifying the dynamics of supernova and neutron star mergers seen by multi-messenger astrophysical observations.




Supercomputing resources are essential to address the above questions. *Self-consistent Green's function* theory (SCGF) [1] will be used to study neutrino reactions on $^{40}$Ar by the end of 2019 and will push the limits of SCGF to masses of A~140 in 2020. The FRIB facility (USA) will start operations in 2021 to study several proton-rich isotopes and to discover the driplines in the *pf*-shell. The long-term challenge (2022-2030) will be introducing innovative approaches, based on SCGF and

*diagrammatic Monte Carlo* to predict nuclear scattering up to 100 MeV/A, hopefully eliminating phenomenological inputs in the interpretation of experiments with radioactive beams.

**2) Theoretical support for experimental nuclear physics: a novel strategy for the theory of nuclear fission.** Nuclear fission is a fascinating phenomenon in which the atomic nucleus, a compact self-bound mesoscopic system, undergoes a quantum transition into two or more fragments. A predictive, and accurate description needs to be rooted in quantum many-body theory and remains one of the biggest challenges in science. Current models assume adiabatic changes of internal degrees of freedom at thermal equilibrium, but if the fission occurs at sufficiently high energies and/or short times the process will be non-adiabatic and non-thermal. These effects will impact our understanding of fundamental astrophysical process. UK theory groups exploit nuclear density functional theory (DFT), novel nonlocal functionals and supercomputing techniques to go beyond adiabatic approximations and to obtain a unified description of fission at varying excitation energies.




The UK has recognised expertise in the domain of nuclear DFT and plans to perform extensive calculations of potential energy surfaces (PES) of all fissile nuclei using state of the art functionals [3]. The figure illustrates a typical PES (quadrupole and hexadecapole). The lines located at the bottom of the valley indicate the most probable fission paths. The project will create a complete database of fission properties of all nuclei and provide the astrophysical community with reliable data for their simulations. Fission plays a crucial role in determining the end point of the r-process nucleosynthesis—recently observed in neutron star mergers via the production of extra free neutrons and by the fission recycling mechanism [4].

**Computing demands:** Ab initio SCGF calculations of neutrino reactions on isotopes with A~140 will require ~1M CPUh per isotope and we expect to use 20M CPUh to investigate this region of the nuclear chart (where the first electron-ion collision experiments are planned). The *auxiliary field diffusion Monte Carlo* (AFDMC) method [2] has greater potential to understand nuclear shapes but is more computationally demanding: $^{40}$Ca currently requires >4M CUPh alone. We are improving the method and expect accurate results to be possible with 1.0-1.5M CPUh by 2021. The new AFDMC will require ~80M CPUh to support the interpretation of FRIB data over the 2021-2024 period. Calculation of a PES for a fissile nucleus costs 0.4M CPUh, however, a truly quantitative description can only be achieved with several more degrees of freedom. Adding octupole deformations—crucial to understand asymmetric fission—would require 80M CPUh for each fissile nucleus, while exascale simulations will be necessary to predict triaxiality and pairing correlation effects.

**Track Record**

[1] *Raimondi, F. and Barbieri, C., Physics Review* ***97****, 054308 (2018)*
[2] *Lonardoni, D. et al., Physics Review C* ***97****, 044318 (2018).*
[3] *Raimondi, F., Bennaceur, K. & Dobaczewski, J., J. Phys. G: Nuclear & Particle Physics,* ***41****(5), 055112 (2014).*
[4] *Goriely, S et al., Physical Review Letters,* ***111****(24), 242502 (2013)*

## 3.12 Lattice QCD and Flavour

*Contributors: Boyle, Davies, Jüttner, Bouchard, Del Debbio, Flynn, Horgan, Horsley, Kenway, McNeile, Portelli, Rakow, Sachrajda, Wingate*

**Vision** – To uncover and exploit the cracks in the Standard Model (SM) of particle physics that may lead to new physics requires both accurate experimental results from a range of large facilities and accurate results from theory. The ultimate goal here are answers to: 'What are the fundamental particles?', 'Is there a unified framework?' and 'What is the origin of matter-antimatter asymmetry?' by making stringent tests of the SM until it breaks.

**Key research challenges:** The success of Lattice QCD in recent years in providing validated, fully-controlled and model-independent numbers for a range of hadronic masses and matrix elements has led to a host of tests for new physics and the extraction of SM parameters. The UK has played a strong role in this through calculations on DiRAC and has the world's most accurate results and most advanced techniques in strange, charm and bottom physics. We have hosted Lattice 2016 (Southampton), received the Ken Wilson Award for work on hadronic K decays and play a major role in the Flavour Lattice Averaging Group. Key players in this area using DiRAC facilities have been the collaborations HPQCD (Glasgow/Cambridge/Plymouth), QCDSF/UKQCD (Edinburgh) and RBC/UKQCD (Edinburgh/ Southampton/Liverpool).

The energy- and intensity-frontier experiments High Luminosity LHC, BelleII@KEK, NA62@CERN, Muon-g-2@Fermilab and J-PARC will provide the increased precision that is needed over the next decade; improved theory accuracy from lattice QCD requires a pre-exascale system in the short term (2021) and an exascale system in the medium (2024) to long term (beyond 2024). On the same timescale experiments aimed at understanding neutrino physics, DUNE, HyperK and T2K will need accurate lattice QCD input to capitalise fully on their results.

A pre-exascale facility is required in the **short** term to allow high statistics and very fine space-time lattices (with spacing of 0.03 fm) reducing systematic errors for heavy b quarks and allowing access to the important high-momentum region for the light hadrons produced in B-meson decay. It will also allow calculations of 'long-distance' effects in rare-K decays for NA62@CERN/KOTO@J-PARC and a decades-long puzzle around decay amplitudes of the neutral K mesons should be resolved; QED effects in leptonic and semileptonic meson decay matrix elements will be included and the 0.5% uncertainty needed for QCD effects in the muon's magnetic moment will be achieved through numerically-intensive studies of multi-hadron effects.

In the **medium term** a further improvement in precision will be achieved by increasing the simulation volume and further decreasing the lattice spacing. This will allow sub-1% precision in results for b- and c-meson leptonic, semileptonic and mixing amplitudes required to interpret BelleII and LHCb results. Accurate calculations of a range of baryon form factors (numerically more challenging) will allow new flavour-physics tests. Nucleon form factors accurate to a few percent for neutrino physics should become possible. The study of long-distance physics in rare K decays and neutral kaon mixing will enter a precision era and constitute reliable precision probes for new physics. Improved determination of quark masses will be used to test the nature of the Higgs boson. Inclusion of QED effects in more complicated decay channels like hadronic rare K decays and neutral meson mixing will become computationally feasible.

In the **long term (beyond 2024)** a step change in Lattice QCD+QED simulations will require substantially larger volumes to reduce residual finite volume effects and very fine lattice spacings to be able to accommodate all relevant scales from the pion mass up to charm and bottom quarks in full QCD+QED simulations of a large class of hadronic matrix elements. This will require simulating L/a~256 or even 512 ensembles. Continued and new theoretical and algorithmic developments (e.g. multilevel Markov Chain Monte Carlo with impact likely also on statistical inference of big data) over

a five- to ten-year time scale together with exascale resources might also allow attacks on much harder problems like hadronic D-decay or D-mixing and even inclusive B-decays, which the flavour community is waiting for.

**Computing demands:** Achieving the short-term (2021) goals above requires lattices with spatial extent L/a~128 and deflation techniques are required to achieve these goals. Based on current performance on DiRAC resources (Extreme Scaling), scaling computing requirements up from L/a~64 leads us to conclude that generating configurations and producing physics results will require a dedicated resource of the order of 20-30PFlop/s as currently only available to our competitors in the US. Similarly, the increased complexity of the computations required to deliver the medium term (2021-24) goals (large number of contractions, noise-to-signal problems, multi-channel hadronic states analyses together with larger volumes (L/a~256 and larger), smaller lattice spacing) will require 50-100PF of dedicated computing resources. To deliver the step change in simulation capabilities required beyond 2024, simple scaling of current algorithms, network fabric and computer architecture suggests that generating the required ensembles and producing physics output will require exascale supercomputers.

**Track record**


[1] *Chakraborty et al., 2017, HPQCD, PRD96:034516*
[2] *Bai et al., 2015, RBC/UKQCD, PRL115*
[3] *Flynn et al., 2015, PRD91 no.7, 074510*
[4] *Blum et al., 2018, PRL121 no2. 022003*
[5] *Chakraborty et al., 2015, PRD91:054508*

# 4 Climate, weather and earth sciences

*Editor: Bryan Lawrence (National Centre for Atmospheric Science [NCAS], University of Reading)*

Supercomputing is used in three main ways in environmental science: to support the creation of large and complex observational datasets; to simulate the environment; and to analyse the data resulting from observation and/or simulation. Often these are interlinked in complex workflows which exploit supercomputing throughout.

In this section, we present seven specific areas where next generation supercomputing is expected to enhance UK environmental science from advancing numerical weather prediction for science and society to computational mineral physics, geodynamics and seismology. In each case the use of exascale computing will deliver both enhanced understanding of the physical system and the capacity to use that knowledge in the service of society. Many share the same modelling system or systems (derivatives of the existing UK seamless modelling suite). All will also depend on international collaboration, in terms of shared approaches to techniques, analysis, and exploiting model diversity to understand the model driven uncertainty.

Where these areas of science include simulation, they also share the same underlying computing constraints which arise from the way that the earth is simulated. Most modelling systems utilise a blend of two very different ways of representing phenomena: resolved equations and sub-grid scale parameterisations. Resolved equations describe large-scale phenomena and are solved numerically on a grid covering the domain of interest (e.g. the entire globe for climate science and some numerical weather prediction) with unresolved phenomena within each of the grid cells represented by forcing terms arising from parameterisations of their bulk properties.

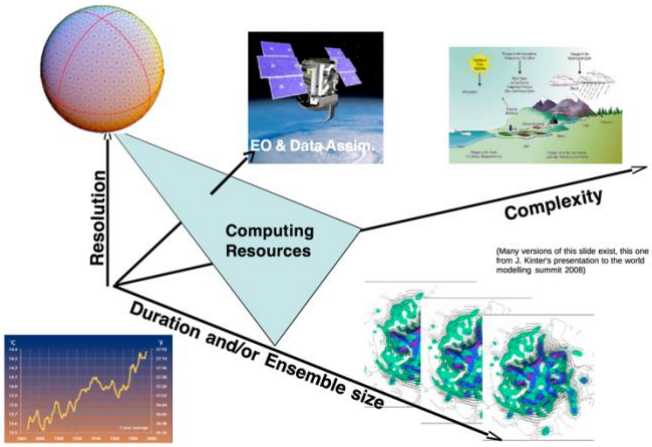



*Figure 3 Different ways of exploiting the available computing. More computing is needed to advance down these axes (see text).*

At any given time the available computing power can be used in one of five different ways: to increase the resolution (use smaller grid cells), to run more simulations which can be used to estimate uncertainty via sampling the available parameter space (bigger ensembles), to make better use of data assimilation to initialise, guide or evaluate simulations, to represent phenomena more accurately via more equations and/or improved parameterisations (increase complexity), or to run simulations for longer (duration). Some of these options are depicted in Figure 3. The bulk of the cases presented in this section will exploit increased supercomputing by advancing down one or more of these axes, utilising some combination of increasing resolution, complexity, ensemble size and the use of data. In terms of resolution, key processes of interest in the atmosphere include storms and convection, and in the ocean, eddies, boundary currents and regions of strong upwelling, and so the ability of models to resolve or permit these features is of great importance – but only if doing so does not preclude other important aspects of their use (such as the ability to run for long enough for development and/or production purposes). In terms of complexity, modelling will improve support for biogeochemistry, hydrology, and land-surface modelling related to agriculture, forestry, and ecosystems.

The problems presented include capability grand challenge experiments which could require the use of an entire T0 machine (e.g. the highest resolution experiments) as well as large ensembles that would need to be analysed in-situ, again using an entire T0 machine, to preclude the production of

unmanageable amounts of data. These capability experiments would be supplemented by many capacity experiments, running on the T0 platform or elsewhere as load and resources permit.

In most cases, running for substantially longer than is possible with current models and architectures will not be possible without advances in mathematics and computational science, as discussed in Section 0. In particular, both software and algorithms will need to evolve to address architectural diversity, and to provide additional strong scaling to speed up lower resolution runs. The use of a common modelling system across application areas will facilitate the pull through of these advances into production code. The scientific problems discussed here will all depend on these advances, particularly for the longer-term ambitions – without which we will not be able to provide the scientific results required to find solutions to important issues such as extreme events, energy infrastructure and food-supply risks, and fully understanding our changing regional climate and weather patterns.

Future supercomputing is also integral to the exploitation of data, both from real world observations and from simulation. In both cases, the volumes of data are increasing rapidly and have done so for decades –instruments going into space exploit more and more on-board computing to increase their data rate, and all the directions of exploiting computing in simulation result in larger datasets for analysis. As a consequence, where scientific data analysis could once be done on relatively small servers, much now also requires supercomputing to handle the necessary volumes and data rates necessary to sustain scientific workflows – scientific and industrial competitiveness could not be met if UK scientists spend substantially longer to produce knowledge from these datasets than their international peers. Such supercomputing also needs to support shared fast access to large datasets, since multiple copies of the largest datasets would be prohibitively expensive. All but one of the cases in this section require customised supercomputing for environmental data handling, which we might expect to be delivered by an additional T1 system complementary to a T0 simulation platform. Such a T1 data analysis system would itself be a “data analysis” capability platform, supporting communities of data users, as well as analysis by the modelling teams. All are also likely to make significant use of artificial intelligence and machine learning techniques via the introduction of AI/ML to replace some parameterisations, to emulate complex systems where speed is necessary (e.g. in spinning up oceans), and particularly, in data analysis and the production of downstream products, such as those necessary to advance hydrology and agricultural science.

## 4.1 Advancing Numerical Weather Prediction (NWP) for Science and Society

*Contributors: S. Vosper (Met Office)*

**Vision:** Development of next-generation NWP systems, based on ensembles of “convection permitting” coupled global atmosphere-ocean-land-sea-ice models and convection-permitting coupled ensemble data assimilation, operating at grid lengths of a few kilometres (1-4 km).

“As a computational problem, NWP is highly complex - comparable to the simulation of the human brain and of the evolution of the early Universe”, [1]. Recent decades have seen significant progress in our ability to predict the weather. This is a direct result of improvements in the numerical models, observations (e.g. from satellites), data assimilation techniques for model initialisation and ensemble approaches to describe forecast uncertainty. These achievements have been facilitated by advances in massively-parallel supercomputers and software.

**Key Research Challenge*:*** State-of-the-art global NWP models currently operate with grid spacings of order 10 km. These are unable to explicitly resolve important small-scale phenomena such as convective clouds and gravity waves, which are instead approximated through parametrizations. Imperfections in certain parametrizations (e.g. convection) inhibit our ability to predict high-impact weather phenomena and limits predictability in NWP systems, partly because the nonlinear upscale energy cascade of these phenomena is poorly represented. The simplicity of some parametrization

schemes means we are also unable to explore the importance of complex interactions e.g. between cloud microphysical processes and aerosols. Improvements in accuracy also require advances in data assimilation, utilising greater numbers and a wider variety of observations. In future these are likely to be based on ensemble techniques, requiring very large ensembles of short-range forecasts. For forecast ranges beyond a few days the coupling between the atmosphere, ocean and sea-ice is also known to be important.

Significant enhancements to model resolution, model complexity, ensemble size and data assimilation are all constrained by compute capacity. Yet the ability to explore these is vital to enable the research required for further advances in NWP. The vision of a next generation coupled modelling systems operating at atmospheric grid-spacings of a few km would lead to improvements in our fundamental understanding of the processes which influence predictability and increased skill in operational NWP on forecast timescales of hours to months ahead. The global predictions would also provide boundary information for even finer-scale (grid lengths of order 100m) limited-area NWP, for very detailed (cloud resolving) high-impact weather.

**Goals:** *Medium term (2022-2025):* A global km-scale coupled NWP capability as a research tool to explore the behaviour of the global atmosphere in unprecedented detail, develop convection-permitting data assimilation and new "scale aware" parametrizations and to evaluate the relative benefits of resolution, model complexity and ensemble size.

*Longer term (2025-2030):* Further development of km-scale NWP and significant testing, including hero simulations (1 km grid). The experiments will lead to pull through of research into operational NWP (Met Office, the European Centre for Medium-Range Weather Forecasts (ECMWF) and other centres), delivering significant improvements in predictability and socio-economic benefits to the UK. Just increasing resolution to 2km from 10km would require at least $5^3$=125 times as much compute!

The UK has a well-deserved strong reputation in this field, as a pioneer in seamless weather and climate modelling [2]. Research would be coordinated as a Joint Weather & Climate Research Programme (JWCRP) activity, involving Met Office and NERC centres (National Centre for Atmospheric Science (NCAS), Centre for Ecology and Hydrology (CEH), National Oceanography Centre (NOC) and National Centre for Earth Observation (NCEO)) plus leading groups at UK universities.

**Computing demand***: Medium-Term:* Compute capacity requirements for research simulations are 10-20x that currently available for global NWP/climate research (140K CPU cores). Data assimilation requires massive observational databases with long-term archives of historical observations.

*Long-Term:* The compute capacity required is 50-100x current capacity.

## 4.2 Exploring the Frontiers of Climate Modelling

*Contributors: Vidale, P-L[1], Roberts, M.[2] ([1]NCAS, University of Reading, [2]Met Office)*

**Vision:** Next-generation modelling system capable of integrating at least ten-member global coupled simulations at sub-10 km resolution at speeds of 1-10 simulated years per day.

**Key Research Challenge:** Our main focus is on exploiting the new class of Global Weather Resolving Models (GWRMs [3]), with mesh sizes at sub-10km in the atmosphere ("convection permitting") and eddy-resolving models in the ocean (1/12 and beyond). There are three synergistic activities: i) develop General Circulation Models (GCMs) so that they can perform coordinated experiments at high-resolution, including incorporating new physical parameterisations; ii) design, carry out and assess core experiments (hindcasts and predictions) at the process level, as well as targeted sensitivity experiments, to identify the benefits of increased resolution in the ocean and atmosphere; iii) explore the frontiers of climate modelling through experiments at ultra-high

resolution, increasingly focussing on the multi-scale interactions between cloud forcing and circulation.

**Goals:**

*Medium term (2022-2025):* exploit the new class of GWRMs, with mesh sizes at sub-10km in the atmosphere, and eddy-resolving models in the ocean (1/12 and beyond). Process focus is on storms, circulation and impacts, including a robust assessment of changing societal risks by phenomenon.

*For the atmosphere*, current performance, with HadGEM-GC3.1 at 10km (N1280) delivers 0.14 simulated years per day (SYPDs) using 7,236 cores; at 5km (N2560), the current configuration delivers 6 simulated days per day (SDPD), using 12,400 cores, but new optimisations promise to deliver up to 20 SDPDs. It is expected that LFRIC, which we will start to experiment with in 2020, should deliver similar turnaround, very likely using more cores (~20K at 5km). We aim for ensembles (size 10) of 5km seasonal simulations in the next 1-2 years and for an AMIP run (30 years duration) by 2022.

*For the ocean*, the focus is on developing methods to quickly spin-up eddy-resolving simulations, as well as exploiting coupling to the atmosphere. The 1/12 ocean model (run on its own) can currently deliver 720 SDPD (2 years) on 10K cores, and 1/36 is estimated at 360 SDPD on 80K cores.

The ideal coupled configuration will be about 10km in the atmosphere and eddy-resolving in the ocean at centennial scales, to exploit improved air-sea coupling and consequent heat uptake processes. We also plan simulations with 5km atmosphere and eddy-resolving ocean to enable investigation of improved diurnal cycle processes on atmosphere-ocean coupling.

*Longer term (2025-2030):* the community goal is to achieve a 1km global climate simulation in the next decade. This will enable several advances in our understanding of climate sensitivity, the simulation of extremes, quantification of environmental risks. Robust assessment of risk will depend on internationally coordinated exploitation of model outputs.

The UK has a well-deserved strong reputation in this field, as a pioneer in seamless weather and climate modelling [4]. Research would be coordinated as a JWCRP programme, involving Met Office and NERC centres (NCAS, CEH, NOC and NCEO) plus leading groups at UK universities.

**Computing demand***: Medium-Term*: we will need access to O(100K) cores for months at a time in order to run our ensembles. Storage and analysis on JASMIN[9] are crucial to scientific output.

*Long-Term*: we will need to develop the capability to turn on hybrid architectures, e.g. exploiting GPUs. For the 1km capability we will need sustained access to order of 1 million cores.

## 4.3 Seasonal to Decadal Prediction with Digital Oceans

*Contributors: Coward, A (NOC)*

**Vision:** Delivering next generation models to support both the UK Marine sector and coastal communities in making sense of global- and regional-scale change and variability – to protect lives and livelihoods.

[9] The Joint Analysis System Meeting Infrastructure Needs (JASMIN) is a "super-data-cluster" which delivers infrastructure for data analysis. JASMIN is funded by the Natural Environment Research Council (NERC) and the UK Space Agency (UKSA) and delivered by the Science and Technology Facilities Council (STFC).

**Key Research Challenge:** Delivering robust statistics of the potential changes in ocean circulation on seasonal to decadal timescales with large, high-resolution, global coupled atmosphere ocean models with multiple nested regional coastal seas.

Many coastal communities depend on the delicate interplay between ocean circulation and local ecosystems. Even minor changes in ocean circulation (position of a current, salt intrusion etc.) can significantly alter local conditions on which livelihoods depend. While coastal communities are in the frontline for adverse effects related to climate change, the economic and societal benefits of robust seasonal to decadal prediction will be felt by all.

Recent UK research [5] has shown that key processes which set and maintain sub-surface ocean conditions over winter for re-emergence the following season are significantly better represented in coupled models with high resolution components. Increasing resolution is not the only challenge since the physical conditions experienced by the local ecosystems are not the only drivers for change. Concerns over other factors such as the distribution and fate of plastics within the ocean will be addressed by ever-more complex biogeochemical models. Such models will, for example, track plastics not merely as passive tracers but as active components in the food chain with potential consequences for the long-term health of our oceans [6,7]. The correct representation of the export of quantities from the shelf regions to the deep ocean is a key challenge for both climate change mitigation and marine environment preservation. Whether investigating off-shelf carbon fluxes or the fate of pollutants from coastal waters, a seamless interface between intricate shelf processes and open ocean counterparts is an active research goal.

The twin drivers of increasing resolution and complexity will demand codes and algorithms adapted to the future architectures delivering exascale computing. The UK is not alone in this effort – it is a full participant in international collaborations with clear development strategies (see https://www.nemo.ocean.eu) and projects underway to deliver improved codes to the Copernicus Marine Environment Monitoring Service (CMEMS) within the 5 years of the IMMERSE EU-project.

**Goals:** *Medium term (2022-2025):* The challenge for the next decade will be to run high resolution coupled models routinely in 50+ member ensembles in order to provide robust statistics for seasonal prediction. This will require a robust and efficient nested modelling capability designed to enable multiple coastal regional nests within a global model. Such a facility is likely to see significant uptake in 5-10 years with more regional modelling groups using global models with regional nests to address local issues. *Longer term (2025-2030):* A nested model approach can never be truly seamless since processes existing within the coastal models may not exist in the outer model. A longer-term goal will be to have a reference configuration of a 1/60$^{o}$ global model supporting processes traditionally reserved for regional studies (e.g. internal waves, wetting and drying, tides etc.)

**Computing demand:** *Medium-Term*: A single such member currently utilises O(500) nodes (8%) of ARCHER for a third of a year; we can anticipate a need for a 4 x increase in computing resource in the short term for this task alone rising to 20-25 x current capacity within 5-10 years *Long-Term*: : A 1/60$^{o}$ global model itself represents a 125x (being the resolution change cubed) increase in resource requirement. The introduction of more complex processes increases the requirement further both because of more complex calculations and because the greater internal variability associated with the processes will require more ensemble members to address. An increase of 300x current capacity is probably required to achieve this aim.

Under-pinning both mid and long-term aims is a need for curation and processing services to keep pace with the expanding volumes of output that need to be assessed and mined for scientific outcomes. These services allow many more scientists beyond the core supercomputing experts to access model results and provide vital scientific return on the supercomputing investment (for example to enable a large community to exploit multi-PB output from a 1/60$^{o}$ global model).

## 4.4 Earth system modelling: Developing safe futures for the full Earth System

*Contributors: Colin Jones (NCAS, University of Leeds)*

**Vision:** An enhanced understanding of the coupled Earth system and its response to future human forcing through development of UK community Earth system models that are: *storm-resolving* (in the atmosphere), *eddy-resolving* (in the ocean), and include accurate representation of processes key in determining the response of the coupled Earth system to future forcing.

**Key Research Challenge:** Earth system models (ESMs) constitute our best tools for assessing the response of the coupled global environment to human influence. They are the tool of choice for developing future socio-economic and mitigation pathways aimed at avoiding dangerous global change and provide the primary link between global change science and global change policy. ESMs build on physical climate models and interactively include key biogeochemical cycles (e.g. carbon), allowing urgent questions to be addressed, such as; what amount of anthropogenic $CO_2$ emissions are available to stay below a given global warming level? ESMs also include atmospheric chemistry and aerosols, supporting analysis of the near-term mitigation potential from reducing non-$CO_2$ greenhouse gases, as well as allowing future climate and air quality risks to be studied in a single model system. ESMs are used to assess the risk of abrupt, potentially irreversible, changes in the coupled Earth system. For example, the risk of rapid Antarctic ice loss and impacts on global sea level, permafrost thaw, methane release and the risk of amplified Arctic warming or future wildfire risk and impacts on terrestrial ecosystems.

The 1st UK Earth system model (UKESM1) was recently completed and simulations with this model are now supporting the 6th Intergovernmental Panel on Climate Change (IPCC) Assessment Report (AR6) [8]. By necessity, these simulations are multi-centennial in length and require large ensembles to sample the range of possible future emission pathways. Combined with the need for both process completeness and realism and the availability of supercomputing, UKESM1 has a spatial resolution of ~1°in the ocean and atmosphere. To increase the reliability of knowledge around future Earth system change we urgently need to further develop UKESM1 to be capable of accurately simulating atmospheric weather phenomena and key ocean processes, such as boundary currents and ocean eddies, coupled with advanced parameterizations for key climate, biogeochemical and cryosphere components of the Earth system.

**Goals:** *Medium term (2022-2025):* Further develop a hybrid-resolution UKESM2, where physical model components are simulated at a higher resolution than biogeochemical ones, with regular coupling between components [9]. We aim to be *synoptic resolving* (~0.5°) for the physical atmosphere and *eddy permitting* (~0.25°) for the physical ocean, interactively coupled to a set of advanced ~1° biogeochemical models. This model will support a future IPCC AR7 and UK research into the coupled Earth system.

*Longer term (2025-2030):* By 2030 we aim to develop a new UKESM3 model, with significantly increased resolution in both the physical and biogeochemical model components. We aim to be *storm-resolving* (in the atmosphere) and *eddy-resolving* (in the ocean), implying ~0.1° resolution, or better, in both components. We will further include a suite of new and improved process descriptions critical for determining the future evolution of the coupled Earth system. These include more advanced cloud-aerosol processes, marine biology, soil-vegetation processes, including interactive permafrost and wildfires, as well as coupled models of the Antarctic and Greenland ice sheets. Such a model will support a major advance in our understanding of the coupled Earth system, including its sensitivity to human influence, delivering unprecedented scientific support to UK climate policy.

The UK has world-leading standing in Earth system modelling, with UKESM1 the most advanced ESM in the word today. UKESM1 resulted from a 5-year JWCRP programme involving tight collaboration between the Met Office and 7 NERC centres (NCAS, NOC, CEH, NCEO, Plymouth

Marine Laboratory (PML), British Antarctic Survey (BAS) and the Centre for Polar Observation and Modelling (CPOM)), as well as important contributions from 14 UK universities. We envisage future work will continue this tight collaboration, supporting both national capability and academic research.

**Computing demand***: Medium-Term*: For UKESM2, based on the hybrid resolution approach, we envisage requiring of order 6x the compute power of today (i.e. sustained access to 60K cores) and 4x the storage capacity (i.e. 50+ PB). *Longer-term*: To realise a coupled Earth system model (UKESM3) at 0.1° resolution a compute capacity of order 200x that of today is required, combined with a factor 50x increase in storage capacity.

## 4.5 Understanding the Earth System from Space

*Contributors: Martyn Chipperfield, Amos Lawless, John Remedios (National Centre for Earth Observation)*

**Vision:** High-quality environmental datasets to enhance our understanding of the Earth system, combining multiple sources of satellite data with state-of-the-art forward models, easily accessible to the wider scientific community.

We are now entering a new satellite era in which we will be able to observe the Earth system like never before. Novel space-based instruments, such as the planned Earth Explorer missions, will operate alongside highly capable operational systems in the Metop/ Meteosat and Sentinel series of satellites. Their combined power offers unprecedented views into the workings of the climate system and how its different components interact. The collection, processing and synergistic use of data from these satellites will permit great leaps in our understanding of how the Earth system is changing and consequential improved levels of prediction on time scales ranging from hours to years.

**Key Research Challenge:** Mathematically precise retrievals of datasets from satellite data have already contributed to our understanding of the climate system with, for example, sea-surface temperature retrievals being fundamental for IPCC reporting [10]. Ultimately, leading research will drive forensic and predictive environmental modelling and high quality, global data sets which, in combination will also enable huge impact in key policy areas of government activity and in growing markets (e.g. [11]). Achieving this goal requires a step change in our capacity to store and process Earth Observation (EO) data. Over the next 5 to 10 years, the amount of satellite data available will at least increase by an order of magnitude in size with the coming online of Sentinel-4 and -5, ESA Earth Explorer missions and the concurrent running of geostationary platforms, each containing a number of relevant EO instruments. The current ESA EO data archive contains around 20 Petabytes of data, but this is expected to increase to 100 Petabytes by around 2025. A step-change in computing resource and storage capacity would significantly benefit both long time series of environmental data documenting change but also near-real time applications. Faster processing would allow a greater breadth of Earth system variables to be retrieved from satellite datasets with robust uncertainties on the data.

A specific limitation on current methods comes from the forward models used in satellite retrievals, which can be very complex and operate at the limit of current computer resources. Increased computing resources would allow increases in the spatial resolution to be used, the complexity of known processes to be included (e.g. in chemistry schemes), the length of model simulations to be used, and the number of sensitivity simulations (or ensemble members) to be performed.

**Goals:** *Medium term (2022-2025):* Capability to perform retrievals of new satellite data as it comes online, producing climate-relevant datasets, stored and available to the wider community, incorporating new forward models such as atmospheric chemistry-aerosol models at ‘air quality’ resolution (0.1 degree).

*Longer term (2025-2030):* Detailed, increasingly long-timescale, coupled model runs suitable for a robust evaluation of the UK Earth System Model, both in terms of its individual components and when coupled, by objective comparison with a variety of EO data. Development of real-time monitoring systems to enable rapid response to ongoing events, such as mapping of wildfires or monitoring of other significant air pollution events which impact health.

**Computing demand***: Medium-Term*: The computing capacity for retrievals requires an increase which exceeds that in processing power as storage-bandwidth is not following computability. Analysis systems to make the data available to the wider EO community requires an increase in user storage capacity of the order many TB for individual experiments, along with petascale active archive storage, and large-scale systems and networks that allow fast access to these data (including from archive).

*Long-Term*: Of the order of several thousand cores for processing and storage capacity of the order of many Petabytes. Systems that allow users to concurrently access such data seamlessly.

## 4.6 Data Assimilation for Earth System Models

*Contributors: Martyn Chipperfield, Amos Lawless, John Remedios ([1]NCEO)*

**Vision:** Development of next-generation data assimilation systems, capable of bringing high volumes of observational data into complex Earth system models to provide robust state and parameter estimates with associated uncertainties.

Data assimilation (DA) is the process of combining observational data with numerical models in a way that extracts the most information from both, by allowing for their respective uncertainties. It allows us to (i) build a complete picture of an environmental system from disparate data; (ii) initialise forecasts for environmental prediction; and (iii) validate numerical models by a robust comparison with observational data. Besides being used in operational weather forecasting, the assimilation of Earth Observation (EO) data has allowed us to understand more about the different components of the Earth system, for example to predict ocean biogeochemistry through the assimilation of chlorophyll data [12].

**Key Research Challenge:** Many current DA systems are built on simplified assumptions, such as the approximate linearity of the systems, the ability to neglect certain space and time scales and a lack of correlation in the errors between different measurements. The latter assumption requires over 95% of satellite data to be ignored in numerical weather prediction systems. With emerging complex Earth system models, often with coupled components including processes at different scales, and the huge increase in the quantity of satellite data coming online, the traditional assumptions no longer hold and new DA methods are required that go way beyond our current techniques. Modern DA systems estimate the model uncertainty through ensembles of model simulations, while some current research is investigating the boundaries between DA and machine learning techniques. In order to increase our capacity in this area it is fundamental that we have increased facilities for storage of the observational data, fast I/O for bringing the data into the modelling systems and increased processing power, to deal with the processing of an increased number of observations and performing large ensembles of numerical simulations.

**Goals:** *Medium term (2022-2025):* Large-scale scientific studies using DA on component models, including (i) a global 1000-member ensemble DA scheme run on a daily basis, with the Joint UK Land Environment Simulator (JULES) land surface model, to assimilate a diverse range of satellite data, for example soil moisture, vegetation indices, carbon fluxes, and land surface temperatures; (ii) the first high resolution pan-tropical carbon cycle reanalysis, identifying sources and sinks of carbon linked to climate and land use change [13], by combining data from the ESA Biomass mission (launch 2021) with disturbance, biophysical and meteorological data.

*Longer term (2025-2030):* A community DA framework using new DA methods to produce state and uncertainty estimates for very high dimensional Earth system models, such as coupled atmosphere-ocean-sea-ice models, including nonlinear ensemble methods and incorporating ideas from machine learning. Ability to assimilate very high resolution and high frequency data from satellites such as Himawari-9 into numerical weather prediction and hydrology models in order to improve nowcasting of hazardous weather, such as flash flooding.

**Computing demand***: Medium-Term*: Sustained access to the equivalent of several thousand CPU cores for processing. Very fast I/O possibilities for bringing large data streams into models.

*Long-Term*: A further 5 to 10x increase in compute for processing. Large data stores, of the order of Petabytes, storing observational data, with fast I/O and accessible by the wider EO community.

## 4.7 Solid Earth Science (SES)

*Contributors: J. Brodholt[1], T. Nissen-Meyer[2], J. Van Hunen[3] ([1]UCL, [2]Oxford, [3]Durham)*

**Vision:** To transform our understanding of the dynamic behaviour of the solid Earth and core and the impact on surface volatiles, life and strategic mineral resources, to enhance our understanding of volcanic and earthquake risk, and to understand the evolution of the Earth and other terrestrial planets from accretion to now.

**Key research challenges:** Solid Earth Science incorporates near surface geology through to the inner core. It is necessarily multidisciplinary and involves geodynamical modelling, seismology and computational mineral physics. Key challenges are to quantify fluxes and chemical interactions between the surface and interior, how the near-surface and deep Earth interact, improved forecasting and mediation of geological hazards, and long-term carbon sequestration.

***Geodynamical Modelling*** is required from the very local to global scales. Key challenges include incorporating vastly different scales of influence, from meters to 1000s of kilometres and from seconds to billions of years, as well as huge variation in rheological scales between melts and fluids to solid silicates. Particularly challenging are the many applications where those vastly different scales appear in one system, for example the intrinsic coupling between rapid earthquakes and slow tectonic evolution, or thin sediment layers, faults, or porous media fluid flow affecting the deformation and decoupling of entire lithospheric plates on a global scale [14]. Modelling increasingly forms the essential tool to link a range of Earth Science observations into a physically viable model. For example, modelling will be used to interpret seismological and geochemical signals at subduction zones. This will transform our understanding of fluxes of volatiles and strategic mineral resources between surface and interior, as well as supporting volcanic and earthquake risk probabilities. Finally, very large geodetic and seismological datasets make joint Bayesian inversion modelling possible, in which a large set of forward models are performed (e.g. in a Monte-Carlo approach) to fit complex physical processes directly to available observables. All these approaches require a significant increase of supercomputing resources.

***Seismology*** is our is our primary data-driven “sight” of both the near surface and deep Earth. Seismic instrumentation has been surging in recent years, accumulating huge amounts of data requiring advanced methods on data mining, data analysis, seismic wave modelling and inverse techniques. Seismological applications are amongst the most computationally demanding scientific endeavours, regularly providing finalists for the Gordon Bell Prize (winner, 2017). Enhanced computational resource is needed to resolve problems ranging from to real-time seismic hazard analysis, understanding earthquake rupture, nuclear monitoring, seismic exploration and carbon sequestration. Ever-evolving supercomputing hardware requires seismic methods to be adapted and updated quickly; and seismic methods have been at the forefront of being tested on hybrid infrastructures, partly due to their excellent parallel scalability. Machine learning applications will

become central and inevitable tools in seismology, requiring significantly more supercomputing resources in the near future [15].

***Computational Mineral Physics*** underpins solid Earth challenges by providing key data on rheological and thermodynamic properties. Future challenges will be to model grain boundaries and provide chemical accuracy results. Grain boundaries control grain size and mobilities, as well as chemical pathways throughout the solid Earth. In the next 5 to 10 years our ambition will be to combine crystal prediction methods on complex, realistic grain boundaries using density-functional and hybrid-functional methods, underpinned by Quantum Monte Carlo and other higher-order methods.

**Computing demand***:* Solid Earth Science is already a major user of supercomputers and currently accounts for about 5% of the UK's ARCHER national facility, as well as using international facilities such as Titan (US Department of Energy (DOE)) and PRACE (EU) computers. Our ambitions require an order of magnitude increase in CPU resources every five years.

**References**


[1] Bauer, P et al. 2015. Nature 525, 47–55 https://doi.org/10.1038/nature14956.

[2] Brown, A et al. 2012. BAMS, 1865-1877 https://doi.org/10.1175/BAMS-D-12-00018.1

[3] Stevens et al. (2019), DYAMOND: The DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains. Progress in Earth and Planetary Science

[4] Roberts, M. J., Vidale, P. L., Senior, C., Hewitt, H. T., Bates, C., Berthou, S., Chang, P., Christensen, H.M., Danilov, S., Demory, M.-E., Griffies, S. M., Haarsma, R., Jung, T., Martin, G., Minobe, S., Ringler, T., Satoh, M., Schiemann, R., Scoccimarro, E., Stephens, G. and Wehner, M. F. (2018) The benefits of global high-resolution for climate simulation: process-understanding and the enabling of stakeholder decisions at the regional scale. https://doi.org/10.1175/BAMS-D-15-00320.1

[5] Jeremy P. Grist, Bablu Sinha, Helene. T. Hewitt, Aurélie Duchez, Craig MacLachlan, Patrick Hyder, Simon A. Josey, Joël J.-M. Hirschi, Adam T. Blaker, Adrian. L. New, Adam A. Scaife, Chris D. Roberts. 2019 Re-emergence of North Atlantic subsurface ocean temperature anomalies in a seasonal forecast system. Climate Dynamics (in revision)

[6] Holt, Jason; Hyder, Pat; Ashworth, Mike; Harle, James; Hewitt, Helene T.; Liu, Hedong; New, Adrian L.; Pickles, Stephen; Porter, Andrew; Popova, Ekaterina; Allen, J. Icarus; Siddorn, John; Wood, Richard. 2017 Prospects for improving the representation of coastal and shelf seas in global ocean models. Geoscientific Model Development, 10. 499-523.10.5194/gmd-10-499-2017

[7] Pecl, G. T., et al., (2017). "Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being." Science 355(6332). 10.1126/science.aai9214

[8] Kuhlbrodt, T., Jones, C.G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., Hill, R., Graham, T., Ridley, J., Blaker, A. and Calvert, D., 2018. The Low-Resolution Version of HadGEM3 GC3. 1: Development and Evaluation for Global Climate. *Journal of Advances in Modelling Earth Systems*, *10*(11), pp.2865-2888.

[9] Stringer, M., Jones, C., Hill, R., Dalvi, M., Johnson, C. and Walton, J., 2018, October. A Hybrid-Resolution Earth System Model. https://doi.org/10.1109/eScience.2018.00042.

[10] Merchant, C.J., O. Embury, N.A. Rayner, D.I. Berry, G.K. Corlett, K. Lean, K.L. Veal, E.C. Kent, D.T. Llewellyn-Jones, J.J. Remedios, R.A. Saunders (2012), A 20-year independent record of sea surface temperature for climate from Along-Track Scanning Radiometers. Journal of Geophysical Research: Oceans, 117, C12013, https://doi.org/10.1029/2012jc008400.

[11] Ganesan, A.L., M. Rigby, M. F. Lunt, R. J. Parker, H. Boesch, N. Goulding, T. Umezawa, A. Zahn, A. Chatterjee, R. G. Prinn, Y. K. Tiwari, M. van der Schoot and P. B. Krummel (2017), Atmospheric observations show accurate reporting and little growth in India's methane emissions, Nature Communications 8, 836, https://doi.org/10.1038/s41467-017-00994-7.

[12] Skákala, J; Ford, D; Brewin, RJW; McEwan, R; Kay, S; Taylor, BH; de Mora, L; Ciavatta, S. (2018), The Assimilation of Phytoplankton Functional Types for Operational Forecasting in the Northwest European Shelf. Journal of Geophysical Research-Oceans. https://doi.org/10.1029/2018JC014153.

[13] Bloom, A., J-F Exbrayat, I.R. van der Velde, L. Feng, and M. Williams (2016), The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times, PNAS, 113 (5) 1285-1290, doi: 10.1073/pnas.1515160113

[14] Riel, N., Bouilhol, P., van Hunen, J., Cornet, J., Magni, V., Grigorova, V., & Velic, M. (2019). Interaction between mantle-derived magma and lower arc crust: quantitative reactive melt flow modelling using STyx. *Geological Society, London, Special Publications*, *478*(1), 65-87. https://doi.org/10.1144/SP478.

[15] Moseley, Benjamin, Andrew Markham, and Tarje Nissen-Meyer. "Fast approximate simulation of seismic waves with deep learning." *arXiv preprint arXiv:1807.06873* (2018).

# 5 Computational Biology

*Editors: Ewan Birney (EBI), David Ford (Swansea)*

Health informatics and the broader bioscience, including structural biology, cell biology, genetics and genomics, as well as more traditional clinical research such as epidemiology and clinical trials have been revolutionised with the dropping cost in data generation on living systems. An exemplar has been genome sequencing which has dropped over a million fold in cost over the decade, but similar large gains have been achieved in other omics technology and in imaging. This drop in cost has had a profound effect how this science is performed, with this becoming a first class data science. This in turn has a profound effect on the change in data and compute required for this science. Much of the impact is around data - both volumes and heterogeneity of data, with the added complication for the need for careful authorisation of data use as much of the health data is personal data. The data-driven nature of this science means that any supercomputing strategy in this area must be well coordinated with the data intensive compute strategy of UKRI.

As well as compute coupled to the data, as required for the data intensive science in this area, there are also pockets of work which require large scale compute resource for either complex simulations or machine learning methods which are compute intensive. These simulations span biophysics (eg, protein structure simulation) through to effective exploration of genetic architectures' underlying traits. A key distinction to make is whether the compute is inherently data intensive, where data input/output (IO) rates might limit the collective CPU horsepower, or cases where compute is limiting. In those cases where compute is limiting, the more simulation-centric compute solutions are ideal for these biological use cases.

## 5.1 Biomolecular Simulations: From Molecules to Subcellular Assemblies

*Contributors: Khalid S[1], Essex J.W[1], Biggin P.C[2], Sansom M.S.P[2] ([1]Southampton, [2]Oxford)*

**Vision:** To combine exascale computing and machine learning with multiscale physics to integrate spatially and temporally resolved data from high-throughput experimental methods. The resulting quantitively predictive biology will maximize the impact of basic molecular and cellular biology on the bioscience economy.

**Key research challenges:** We are already leading the world in biomolecular simulation, however, additional resource is required to extend the scope of our research such that (i) societal and healthcare benefits can be maximally realised and (ii) the UK can maintain its position as a global leader. Development of novel drugs, exploitation of biological molecules for gene sequencing devices, and designing biocatalysts require an understanding of the chemistry of key molecules, and their interaction with their biological surroundings. Molecular simulations already offer insights into these areas, and the scope, efficiency and accuracy of these insights will be vastly extended in the medium to longer-terms. In the **short term,** our focus is on characterisation of the dynamics of proteins implicated in disease as well as improvement of methodologies for both more detailed and larger-scale calculations. In the **medium-term**, our goals are to address the following areas:

1. **Affinity and selectivity of drugs:** development of more accurate force fields and progress towards full quantum-level representations. Characterising the selectivity of drugs will be tackled by exploring drug binding to off-target proteins, through large-scale screening and integration with machine learning approaches. Biologics, such as engineered antibodies, fusion proteins and oligonucleotides are larger and more complex than small-molecule drugs, making identification of target sites, improvement of bioavailability and stability *in vivo*, even more difficult for these molecules. In the medium-term we will apply methods that are working for small molecules to biologics to characterise areas for focus with better computational resources in the longer term.

Computation of the binding affinity and selectivity of molecules is central to patient-specific drug treatment and design, one of the key research challenges of Computational Biomedicine (section 6). Leading edge aspects are already taking place on emerging exascale architectures such as Summit.

2. **Bioavailability**: We will simulate drug-exporter recognition processes involved in drug-resistance mechanisms such that drugs likely to exported and therefore to be ineffective *in vivo* can be readily predicted. Prediction of the pathways *via* which drugs enter cells and modifications for improved cellular entry will be possible by combining machine learning and molecular modelling.
3. **Subcellular level characterisation of mammalian and bacterial biology:** We will perform simulations of realistic models of portions of cells, for example the bacterial cell envelope. This will provide unprecedented insights into the *in vivo* behaviour of cellular compartments and will produce training datasets for machine learning methods to enable, in the longer-term, prediction of subcellular response to new drugs, environmental hazards and other stimuli. The data from our medium-term goals will lead to more efficient development of drugs for a range of diseases including cancers, Central Nervous System diseases and bacterial infections, in collaboration with our industrial partners. In the **longer-term**, given the necessary exascale computing and beyond, we aim to bridge the gap between molecular-level events and whole-organism biology. This will enable us to achieve: (i) simulations with direct and real-time impact on medicine. For example, in combination with next generation sequencing methods and machine learning, by 2030 we aim to be able to run simulations to predict any new protein mutations and viability of drugs, while the patient waits. (ii) implement on-the-fly simulations of subcellular structures as they are being determined by methods such as cryogenic electron microscopy (cryo-EM) or cryo-electron tomography (CryoET), thus the dynamics will be added automatically as the structures are resolved. This will provide insights into biological processes that can be utilised by medicine, biotechnology and synthetic biology applications.

**Computing demand***:* Currently, a typical molecular dynamics (MD) simulation of multiple proteins within a membrane consists of approximately ~1 million particles, produces ~1TB of data and requires 10,000 KAUs on ARCHER to generate a trajectory that is state-of-the-art in terms of length. For a single study, many such trajectories are needed. Our future research aims necessitate much larger and longer simulations for which resources at least x20 the current national facility would be required.

**References**

[1] *Basak S et al, Nature. 2018*
[2] *Chorev DS et al, Science. 2018*
[3] *Zanetti-Domingues LC et al, Nature Comms. 2018*
[4] *Aldeghi et al, JACS, 2017*

## 5.2 Large Scale Genomics - for human health and worldwide diversity

*Contributors: Tim Cutts, Julia Wilson (The Wellcome Sanger Institute)*

**Vision:** Genomes underpin most biological research and are integrating rapidly into medical practice. Genomes constitute enormous quantities of data. A continuing and dramatic expansion of computing infrastructure is needed to support their handling and interpretation.

**Key research challenges:** A sample of the decade-long, Grand Challenges in genomics that the Wellcome Sanger Institute is tackling, usually in the context of global collaborations include:

- The **Human Cell Atlas** [1], an international project to create comprehensive genomic reference maps of all cells in the human body (37 trillion) as a basis for understanding human

health and diagnosing, monitoring, and treating disease. This requires the integration of sequencing and high-resolution microscopy data.

- **Human Genetics**: sequencing genomes from 50,000 UK Biobank [2] participants, 50,000 interval cohort participants, patient disease cohorts, and 13,000 children with rare developmental disease. A further complexity is integration of genomic data with related unstructured health care data, which has privacy implications.
- **Infectious disease**: sequencing genomes to understand the emergence and spread of diseases and drug resistance by global surveillance of pathogens using genomics as well as developing tools to analyse, visualise and deliver information. Understanding the microbiome in health and disease, and to treat diseases that are associated with unwanted imbalances in humans' micro-organism communities.
- The **Darwin Tree of Life** [4] project which will sequence all 66,000 eukaryotic species across the British Isles to address fundamental questions in biology and disease and accelerate medicinal drug discovery.

In 2018, the Sanger Institute had read more than 5 petabases of DNA. Each analysis requires at least one sweep over the data, usually implemented using a High Throughput Computing scheme, but also coupled to specialist hardware and analytics. The resulting condensed data file is then used in a variety of complex statistical approaches, including large factor analysis / mixed models, clustering and machine learning. This component of the compute has more challenging CPU and memory requirements depending on the task. Improved technologies and the increase in speed and scale suggests this will at least double in two years. Sequencing technology has been exceeding Moore's law in price-performance for more than a decade, which multiplies the impact on data analysis costs.

The above scientific projects, while diverse in their aims, have a number of common themes which require national scale computing infrastructures:

- **Machine learning**: Genomic, imaging and other biodata sets are growing too large for manual analysis by expert humans to be feasible. Machine learning approaches are essential to the creation of models from the data. These models can then be applied as inference tools to individual samples, essential for future real-time tools the healthcare professional can use in the clinic. Both genetic and imaging data are increasingly seeing the application of this technology.
- **Complex assembly problems**: *De novo* genome assembly, as needed by Darwin Tree of Life, and meta-genome assembly as required for microbiome studies, are complex large memory problems which can advantageously be tackled on supercomputing machines.
- **Distributed cloud computing techniques**: Sharing of the data, and allowing other scientists to compute against that data, is critical to delivering large collaborative science. This lends itself to a hybrid of high-performance computing and web-scale cloud computing approaches.

**Computing demand:** Our data and compute requirements have been increasing exponentially by 15-20% per annum for more than 10 years, a trend which shows no sign of slowing. Current capacity is 30,000 cores and 55PB of usable storage.

**References**


[1] *The Human Cell Atlas. Regev A, Teichmann SA et al. Elife 2017;6*
[2] *UK BioBank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. Cathie Sudlow et al. PLoS Med 12(3): e1001779.*
[3] *Microbial Genomics and Infectious Disease. D. A. Relman. N Engl J Med 2011; 365:347*
[4] *https://www.sanger.ac.uk/news/view/genetic-code-66000-uk-species-be-sequenced*

## 5.3 Biomedical image analysis - accelerating discovery of disease mechanisms and drug discovery

*Contributor: Declan O'Eegan (MRC-LMS & Imperial College London)*

**Vision:** To integrate simulation and measured organ data for complete understanding of cardiac function in health and disease.

**Key Research Challenge**. Quantitative analysis of medical imaging has enormous potential for precision risk prediction, understanding of disease mechanisms, and accelerated drug discovery but until recently has been impeded by the difficulty and expense of acquiring and analysing large population data sets with linkages to health outcomes and multiomics [1]. Transnational biorepositories are now addressing this problem directly by curating high-quality, imaging data with corresponding health outcomes and genomic sequencing.

For instance, deep learning imaging analysis of UK Biobank data (100,000 participants by 2020) is already enabling rapid progress in efficient image phenotyping at scale revealing discoveries about the genetic architecture of the brain, human development and ageing processes [2]. High-resolution motion tracking of the heart using MRI is also now demonstrating the potential role for imaging in automated risk prediction and classification in both health and disease [3,4]. In the next few years we would expect further substantial progress to continue in using image-derived phenotypes for both genetic association studies and causal inference – bringing rapid benefits to understanding the molecular basis of disease and targeting of potential therapeutic pathways. Coupled to this data driven approach, orthogonal haemodynamic simulations provide a model first view of the heart. In the next 5 to 10 years it is very likely that imaging data volume and complexity will continue to grow, for example exploiting multi-dimensional and super-resolved datasets, as well computational demands for image interpretation, quantification, and modelling driven by the rapid evolution of algorithms, data storage, and cloud computing technologies [5].

Work in this domain is being actively supported by the establishment of the UK Health Data Research Alliance, supported by Health Data Research UK, who will develop and co-ordinate the adoption of tools, techniques, conventions, technologies, and designs that enable the use of health data in a trustworthy and ethical way for research and innovation. The UK's position as a world leader in biomedical AI is further strengthened by Industrial Strategy Challenge Fund (ISCF) investment in a national network of centres of excellence, across the UK, using digital systems and machine learning to improve diagnosis and deliver precision treatments.

**Computing demand:** To fully realise the healthcare potential of these investments, and to leverage the increasing availability of low-cost sequencing for genotype-phenotype modelling, requires a globally-competitive supercomputing platform with scalable provision of GPU-intensive resources coupled with high-throughput storage for efficient deep learning workflows. This will accelerate research discoveries and clinical applications that exploit machine learning for phenotyping and biological inference using high-dimensional human imaging inputs offering powerful and scalable data-driven paradigms with applications in genomics, pharmacodynamics and risk-prediction.

**References**


[1] *Miller et al.* ***Nat Neurosci****. 2016.*
[2] *Elliott et al.* ***Nature****. 2018.*
[3] *Bello et al.* ***Nat Mach Intell****. 2019.*
[4] *Schafer et al.* ***Nat Genet****. 2017.*
[5] *Ellenberg et al.* ***Nat Methods****. 2018.*

# 6 Computational Biomedicine

*Contributors: Marzo A[1], Narracott AN[1], Dall'Ara E[1], Li X[1], Bhattacharya P[1], McCormack K[1], Cantwell C[2], Doorly D[2], Sherwin S[2], Vincent P[2], Weinberg P[2],, Alastruey J[3], De Vecchi A[3], Niederer SA[3], Nithiarasu P[4], van Loon R [4], Mengoni M[5], Rodriguez B[6], Coveney P[7], Richardson R[7], Patronis A[7], Diaz V [7], Bernabeu MO[8], Sudlow C[8], Trucco E[9] ([1]University of Sheffield, Insigneo Institute for in silico Medicine; [2]Imperial College London; [3]Kings College London; [4]Swansea University; [5]University of Leeds; [6]University of Oxford; [7]University College London; [8]University of Edinburgh; [9]University of Dundee)*

**Vision:** The 'grand plan' for *in silico* medicine is the modelling of biomedical processes across all relevant length and time scales, from genome to whole human and beyond. The objective is to improve the scientific understanding of medicine in order to simulate any combination of physiological and pathological processes, for the purposes of developing healthcare solutions and improving clinical practice through stratification and personalised care. Such a capability would be of considerable value to industry and regulatory bodies. Though the initiative is worldwide, the UK has a leading role in this endeavour, not least through its widespread presence in most international *in silico* research activities.

**Key research challenges:** Research in all areas of computational biomedicine has benefited enormously from technological advances over the past few decades. This includes innovation in imaging technology (e.g. Magnetic Resonance Imaging), *ex vivo* and microfluidic assays (cardiovascular) and increased access to omics data. Grand challenges that remain are the robust quantification of anatomy and function from medical images, patient-specific molecular drug design and treatment, the computation of non-observable variables (such as mechanical forces, mass transport, material properties, molecular binding poses and dynamics), the integration of information into biophysical and/or statistical models capable of predicting the onset/progression of conditions and the translation of these models into clinical applications.

In general, *in silico* computational activities fall into six broad categories ('C'), each with major computational requirements:

C1 - Fundamental research, to develop and populate a framework of interconnectable and quantitatively accurate biomedical models.

C2 - Intermediate research, in which combinations of such models are used to explore integrated physiology and pathologies related to ageing, comorbidities, drug discovery, or simply understand the limitations of current healthcare solutions.

C3 - Industrial research, developing and optimising healthcare solutions for clinical deployment (including pharmaceuticals and other treatments based on diet and/or physical activity, or where physiological manipulation is impaired for ethical or technical reasons).

C4 - Specific applications in reduced-order model (ROM) technologies, most notably in Computational Fluid Dynamics, Electrophysiology, Musculoskeletal mechanics, cellular, fluid-solid-thermal interactions and multi-scale models

C5 - Creation of in-clinic, patient-level healthcare systems requiring the construction of ROMs (e.g. HeartFlow).

C6 - Direct usage of full computational modelling in batch-mode clinical practice to support clinical decisions on diagnosis, prognosis, or treatment.

The current comparative immaturity of technologies and models requires a decade's investment by each strand in model development and application. As modelling becomes universal, demand will rise exponentially.

**Key Deliverables:** Major objectives for each clinical domain are identified below, together with those members[10] of the UK chapter of the Virtual Physiological Human (VPH) Institute[11] who are most closely involved:

| Domain | Leaders | 2021 | 2024 | 2030 |
|---|---|---|---|---|
| Neuro-Musculoskeletal | Sheffield, Imperial, Leeds, Cardiff | Accurate individual-specific prediction, uncertainty | Multi-scale/ multi-physics | Strong coupling between all multiscale parts |
| Cardiac/Cardiovascular | Sheffield, Oxford, UCL, KCL, Imperial, Swansea | ROMs, uncertainty, accurate individual-specific prediction, treatment prediction | Uncertainty, Multi-scale/ multi-physics | Strong coupling between all multiscale parts |
| Respiratory | Sheffield, Imperial | ROMs, uncertainty, accurate individual-specific prediction | Multi-scale/ multi-physics | Strong coupling between all multiscale parts |

**Computing demand**: The table below provides an estimation of the key parameters that characterise the *in silico* computational landscape in the UK over the short (2021), medium (2024) and long term (2030).

| | 2021 | | | | | | 2024 | | | | | | 2030 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Challenges | C1 | C2 | C3 | C4 | C5 | C6 | C1 | C2 | C3 | C4 | C5 | C6 | C1 | C2 | C3 | C4 | C5 | C6 |
| Number of Institutions | 30 | 20 | 25 | 15 | 10 | 5 | 50 | 30 | 45 | 30 | 25 | 25 | 150 | 90 | 120 | 150 | 250 | 55 |
| Applications per Institution | 3 | 2 | 2 | 2 | 2 | 1 | 10 | 5 | 5 | 5 | 5 | 5 | 20 | 10 | 20 | 15 | 20 | 10 |
| **Total Applications per category** | **90** | **40** | **50** | **30** | **20** | **5** | **500** | **150** | **225** | **150** | **125** | **125** | **3k** | **900** | **2.4k** | **2.25k** | **5k** | **550** |
| **Total Applications in UK** | **235** | | | | | | **1.3k** | | | | | | **14k** | | | | | |

To support this, the community requires an immediate **2x** increase in computing relative to the anticipated 25PF scale of the Archer2 service, **10x** over 2 years and **120x** by 2025, with projects such as CompBioMed and Verified Exascale Computing for Multiscale Applications (VECMA), led by the Centre for Computational Science at UCL, being poised to exploit exascale services by the end of 2021.

## 6.1 Representative examples

**Cardiac Electrophysiology (Challenge focus: C1, C2)**: Cardiac arrhythmias are a leading cause of morbidity and mortality in the developed world and their global prevalence is increasing. Numerous

[10] Participating Universities: Bedford, Cambridge, Cardiff, Durham, Edinburgh, Exeter, Glasgow, Imperial, KCL, Leeds, London, Loughborough, Nottingham, Oxford, Sheffield, Sheffield Hallam, Southampton, Surrey, UCL

[11] VPH Institute: International non-profit organisation, mission to realise the ambitions of the Virtual Physiological Human initiative.

mechanisms for treating them have been proposed, particularly for atrial fibrillation (AF), but there is no clear consensus on the process driving fibrillatory activity due to the complexity and seemingly chaotic nature of activation patterns and inherent variability between patients. Computer modelling has already demonstrated its ability to both improve our understanding of arrhythmia mechanisms and develop comprehensive diagnosis and personalised therapeutic strategies which achieve termination and require minimum intervention. However, the broad range of timescales involved in the physical processes being modelled leads to long runtimes on current petascale systems – at least 10-1000 times slower than real time, depending on the sophistication of the models used. To be able to solve the inverse problem and associated uncertainty quantification in personalising model parameters and assess the reliability of the model prediction requires a significant number of forward executions of complex models in order to identify parameter values from a large range of multi-modal spatio-temporal imaging and electro-anatomic mapping data. In addition, to utilise electrical measurements, these simulations must be undertaken during the procedure, requiring close to real-time execution. Even the personalised model must be further executed multiple times for each scenario being tested, to define confidence bounds in the predictions and inform the clinical decision maker. At the lowest length and time scales, the electrophysiological response is controlled by the transport of ions through transmembrane proteins, which is amenable to study using molecular modelling and simulation methods leading to the design of new drugs to address heart disease. The sheer scale of computation required to effectively utilise modelling in the electrophysiology domain (as well as cardiovascular, respiratory, etc) means that exascale computing is a necessary ingredient.

**Musculoskeletal modelling (Challenge focus: C1, C2, C6):** Osteoarthritis, rheumatoid arthritis, bone metastases, osteoporosis and sarcopenia are the most common musculoskeletal diseases that are associated with a high and growing socioeconomic burden worldwide [1] (in the UK the combined morbidity is nearly 12 million people). A common theme across these diseases is the disparity in scale between the cause of the disease and the clinical end-point. Arthritis and osteoporosis are caused by changes in cellular activity that are precipitated, for example, by systemic inflammation or hormonal changes. The clinical endpoints – unbearable joint pain or bone fracture – only appear after disease progression that can last decades. Pharmacological interventions attempt to modify cellular activity, but understanding how this effects clinical endpoints in a population of individuals remains a highly sought-after goal of disease prevention and treatment planning.

Musculoskeletal diseases possess features from individual cells to whole bones/muscles in space and from microseconds-long bone fracture events to years-long bone/muscle loss processes in time [2-4]. The problem is massively multiscale, and spans from molecule to cell, tissue, organ and body levels. The future demand in musculoskeletal modelling lies in the effective coupling of multiphysics algorithms across these scales. From the bottom, agent-based modelling is used to predict single cell response to various molecule concentrations. Ordinary Differential Equations (ODEs) use the combined action of multiple cells as input to predict tissue remodelling over space and time. This information can then be coupled with tissue-level micro- Finite Element (FE) models based on high resolution micro- computed tomography scans (***microFE***), and thereafter organ-level homogenised FE models (***CT2S***[12] ) to predict whole organ (bone/muscle) behaviour, as well as body-level dynamics models for movements. At the tissue, organ and body scales, the variability of loading and boundary conditions over time is represented by stochastic distributions. These distributions can either be used directly as input to large (~ billions of DOFs) data-parallel spectral stochastic FE methods, or sampled from and used to execute many simulations in a task-parallel Monte-Carlo (MC) setting (***ARF0 [4,5]*** *&* ***ARF10 [4]***) at the population level. In addition, there is also variability within populations, both of cells (bottom level) and individual subjects (top level), where MC-like approaches are again commonly employed. The next step, i.e. the estimation of the effectiveness of

[12] https://ct2s.insigneo.org/ct2s/

new therapies or of multifactorial interventions comprising existing therapies, requires these multiscale models to be analysed for large patient data-sets that represent the true variation in the target population. Regulatory bodies e.g. the U.S. Food and Drug Administration (FDA) have underlined the massive expectations from such *in silico* clinical trial methodologies by preparing guidelines for such activity [6]. Yet, a leading front in this endeavour can be maintained only by an excellent exascale computing support.

Simulations at each of these scales also carry unique challenges. A few common themes across the scales including the highly nonlinear hierarchical nature of biological tissues, with the need to account for heterogeneity and anisotropy. In addition, tissue failure can be manifested under different mechanical and disease conditions, making it critical to simulate a full range of scenarios in order to identify the critical conditions. In order to provide strong coupling across different temporal and spatial scales, we also need to consider the multi-scale nature of contact/interface problems (between different tissues and organs, or with an implant), which can be critical in determining mechanical competence or tissue health. Multiphysics models should also be considered when solid and fluid dynamics models should be coupled to predict tissue remodelling over time. To ensure sufficient convergence and VVUQ (verification, validation and uncertainty quantification) of such models requires high-end supercomputing resources and support.

Advances in the last three decades have led to sufficient maturity in models at each scale. Yet, multiscale models that combine approaches across scales are only beginning to appear and are already saturating existing computational limits. The largest tissue scale models that we are currently running in ShARC (Sheffield, T3) have approximately 500M DOFs (microFE models of vertebral body using microCT images as input). The storage of multiscale models will reach 100s to 1000s of TB in the next 5 years, and the transfer of data (input/output across different scales) will be in the region of 10s of TB between institutions and HPC facilities. With the increase in image resolution (e.g. Synchrotron images are typically in 100s GB), models will progressively become much larger and require supercomputers to solve. However, the move towards HPC also poses other challenges, in terms of scalability of workflows, porting of codes to HPC, development of alternative codes to commercial software, and the maintenance of HPC-suitable codes, which require expert support and abundant computational power.

**Human *in silico* clinical trials in post-myocardial infarction (Challenge focus: C2, C3, C6):** Cardiovascular disease stands as the major cause of death worldwide, primarily dominated by ischaemic heart disease. Ischaemic heart disease results in complex electro-mechanical abnormalities in the human heart, which may lead to lethal cardiac arrhythmias and/or heart failure. Limitations in experimental and clinical techniques hamper their investigation, and there is an urgent need for novel approaches such as multiscale modelling and simulation to yield effective improvements in diagnosis and treatment to decrease the mortality of such deadly conditions.

Supercomputing enables systematic computational investigations into the complex nonlinear processes underlying human cardiac activity and the variability in the response to treatment. Clinical imaging and electrophysiological datasets are being integrated to construct patient-specific anatomically-based electro-mechanical models of ischaemic human hearts, requiring tens of millions of nodes due to numerical constraints. The large disparity in scales of the cardiac electro-mechanics problem further exacerbates convergence requirements, and the need for highly efficient and multi-physics solvers and supercomputing. In the short to medium term, the aim will be the intensive validation of personalised electro-mechanical models based on clinical electrophysiological recordings and image-based deformation maps. In the long term, systematic human *in silico* trials based on supercomputing simulations will be conducted to evaluate the safety and efficacy of pharmacological therapy building on established collaborations with clinical and pharmaceutical partners.

**Cardiovascular modelling (Challenge focus: C1, C2, C3, C4, C5):** Clinical simulations of blood flow have great potential to aid medical decision-making processes through the provision of additional information and interpretation using patient-specific information that is currently available from various scans that are already performed on patients with vascular diseases. Such clinically-oriented simulations have typically focussed on relatively small regions of the cerebrovasculature or used reduced-order modelling to encompass the entirety of the cardiovascular system [7-9].

Using HemeLB, a lattice-Boltzmann based solver developed at UCL in studies of flow in and around aneurysms, allowed the simulation of their treatment with flow-diverting stents, and magnetically steered drug delivery to target sites (for example tumours) [10-13]. However, a true virtual model of the human vasculature must take into account the full circulatory system and its interaction with the relevant systems and organs. Work is currently underway to connect a patient scan of the arterial tree to the corresponding venous tree, and to a full electro-cardiovascular model of the heart (using the ALYA[13] code). Despite the high computational costs necessitated to do this (most likely 100k+ processors for the high resolution cases necessary for high Reynolds number flow in some regions of the body) this is merely a start. In the short term, the aim is to integrate further relevant organ models over time to produce a virtual representation of the human body. Test runs with HemeLB of the full human arterial tree, resolved in 3D, at very high core counts (in excess of 300,000 processors) have exposed many of the challenges associated with extreme-scale simulation. As the ratio of FLOPs to memory capacity (and bandwidth) is increasing with the emerging exascale, the development of memory-conscious applications is becoming essential. HemeLB has undergone considerable optimisation to operate at the multi-PF scale. Very careful programming is required to utilise memory effectively. It is also important that one attempts to ensure data locality, since the performance of IO at all levels cannot keep pace with computations. *In situ* visualisation is being explored in HemeLB as a strategy to minimise data movement, a very strict requirement of future applications. Hybrid programming must be adopted to exploit the heterogeneous architectures of future machines, e.g. the inclusion MPI+X, where X is usually OpenMP. To combine MPI and threads in a sensible manner is a considerable task that maximises parallelism for efficient operation of the application. The use of MPI alone is seemingly adequate to O(100,000) when only traditional multi-core processors are in use, but some issues relating to support for 64-bit count and internal buffer allocation exist at scale.

Exascale computing will discover new mechanisms of cardiovascular disease and improve patient management at scale. Over the past years, ARCHER has enabled the simulation of a patient's heart on 16,000 CPU cores, allowing a single heart beat to be solved in 4 minutes. This step change in simulation speed has already led to a clinical trial of model and image guidance of pacemaker implantation (NCT03495505). Furthermore, the need for characterising the inherent variability across populations has led to the development of virtual cohort approaches [14]. This will lead to moving from simulations of 10's of patients to simulating 100's to 1000's of cases in the 2019-2024 period. This volume of computation cannot be currently accommodated in ARCHER. Furthermore, close collaboration between UK academic and industrial partners have demonstrated the use of up to 50,000 CPU cores of the ARCHER supercomputer to simulate blood flow in the major arteries of the brain in a patient-specific fashion [15] and to enquire into mechanisms of vascular development often compromised in cancer. Work is ongoing to achieve full machine utilisation (120,000 CPU cores), which demonstrates that codes capable of taking advantage of Exascale computing resources exist. Alongside biophysical simulations (BS), Artificial Intelligence (AI) approaches for image quantification and clinical decision-making support have emerged in recent years. A £7M project funded by the National Institute for Health Research (NIHR) is leveraging AI methods

[13] https://www.bsc.es/research-development/research-areas/engineering-simulations/alya-high-performance-computational

developed in the UK to deliver precision medicine for diabetes in India, showcasing the global outreach of UK cardiovascular research.

Both BS and AI approaches rely on having access to abundant and cheap computational power. Relying on private providers of computational time will inevitably raise concerns regarding patient-data security and the potential for misutilisation. A steady stream of governmental investment into supercomputing facilities, including flagship exascale computing resources, is of paramount importance for the UK to keep at the forefront of the worldwide pace of discovery in the cardiovascular research domain.

**References**


[1] *Briggs AM, Woolf AD, Dreinhöfer K, et al. Reducing the global burden of musculoskeletal conditions. Bull World Health Organ https://doi.org/10.2471/BLT.17.204891.*
[2] *Bekas C, Curioni A, Arbenz P, et al. Extreme scalability challenges in micro-finite element simulations of human bone. Concurr Comput Pract Exp https://doi.org/10.1002/cpe.*
[3] *Taylor M, Perilli E, Martelli S. Development of a surrogate model based on patient weight, bone mass and geometry to predict femoral neck strains and fracture loads. J Biomech https://doi.org/10.1016/j.jbiomech.2017.02.022.*
[4] *Bhattacharya P, Altai Z, et al. A multiscale model to predict current absolute risk of femoral fracture in a postmenopausal population. Biomech Model Mechanobiol.* ***18****, 301-318 (2019), https://doi.org/10.1007/s10237-018-1081-0*
[5] *Li Q et al. Prediction of Age-specific Hip Fracture Incidence in Elderly British Women based on a Virtual Population Model, CBMC 2019, https://www.compbiomed-conference.org/wp-content/uploads/2019/07/CBMC19_paper_99.pdf*
[6] *Viceconti M, Cobelli C, Haddad T, et al. In silico assessment of biomedical products: The conundrum of rare but not so rare events in two case studies. Proc Inst Mech Eng Part H J Eng Med https://doi.org/10.1177/0954411917702931.*
[7] *Narata AP, De Moura FS, Larrabide I, et al. The Role of Hemodynamics in Intracranial Bifurcation Arteries after Aneurysm Treatment with Flow-Diverter Stents. Am J Neuroradiol 2018;39.*
[8] *Mandaltsi A, Grytsan A, Odudu A, et al. Non-invasive stenotic renal artery haemodynamics by in silico Medicine. Front Physiol https://doi.org/10.3389/fphys.2018.01106.*
[9] *Melis A, Clayton RH, Marzo A. Bayesian sensitivity analysis of a 1D vascular model with Gaussian process emulators. Int j numer method biomed eng 2017;33.*
[10] *Patronis A, Richardson RA, Schmieschek S, et al. Modeling patient-specific magnetic drug targeting within the intracranial vasculature. Front Physiol https://doi.org/10.3389/fphys.2018.00331.*
[11] *Groen D, Richardson RA, Coy R, et al. Validation of patient-specific cerebral blood flow simulation using transcranial doppler measurements. Front Physiol https://doi.org/10.3389/fphys.2018.00721.*
[12] *Franco CA, Jones ML, Bernabeu MO, et al. Dynamic Endothelial Cell Rearrangements Drive Developmental Vessel Regression. PLoS Biol https://doi.org/10.1371/journal.pbio.1002125.*
[13] *Groen D, Hetherington J, Carver HB, et al. Analysing and modelling the performance of the HemeLB lattice-Boltzmann simulation environment. J Comput Sci 2013;4:412–22.*
[14] *Roney CH, Williams SE, Cochet H, et al. Patient-specific simulations predict efficacy of ablation of interatrial connections for treatment of persistent atrial fibrillation. Europace https://doi.org/10.1093/europace/euy232.*
*[15] Coveney P, Groen D, Schmeischek S et al. Archer: Modelling blood flow around the vessels of the brain. https://www.archer.ac.uk/casestudies/ARCHER_casestudy_HemeLB.pdf.*

# 7 Engineering and materials

*Editor: Nilanjan Chakraborty (Newcastle University)*

Supercomputing is used in three main ways in the broad area of engineering and materials: (i) to create large simulation databases by numerically solving mathematically complex equations (e.g. Navier-Stokes equation) governing multi-scale multi-dimensional problems; (ii) to carry out controlled numerical experiments to isolate the effects of different physical mechanisms to aid fundamental understanding; (iii) to analyse and process large databases created by expensive simulations to extract physical insights to aid development of high-fidelity predictive models. These three areas are closely interrelated and depend on supercomputing and its advancements. Improvements in terms of capability and accessibility of supercomputing play key roles in reducing simplifying assumptions and the level of empiricism. Therefore, advancements in the supercomputing capability are pivotal to future technological progress in engineering and material science.

In this section, ten specific areas are presented where investment in next-generation supercomputing can offer step changes in the advancement of a range of different areas including aerodynamics, turbulent reacting and non-reacting flows, plasma physics and material science in the UK. In all these disciplines, the computational cost depends on the computation grid size, which in turn is dictated by the scientific challenges and the level of detailed information to be extracted from simulation data.

For example, due to the advancement in computational power, it has now become possible to resolve all the relevant length and time scales of turbulence without any physical approximation using Direct Numerical Simulations (DNS). DNS data can, in principle, be considered equivalent to time-resolved experimental data with resolution up to the smallest length scale of turbulence. However, strong assumptions are often invoked in DNS for the purpose of simplification of chemistry, droplet atomisation, flow-structure interactions and particle-fluid interactions. The strategic investment in supercomputing and e-infrastructure will offer the opportunity to eliminate some of the aforementioned limitations and tackle some of the outstanding challenges in the analysis of turbulent flows by carrying out DNS for larger values of Reynolds numbers and obtaining fundamental physical insights by postprocessing simulation results. However, DNS is computationally the most expensive approach, and it is still not possible to simulate engineering systems of practical interest, which involve complex geometries. Engineering simulations, therefore, rely upon techniques where a suitably averaged/filtered flow field is obtained. The averaging/filtering procedure gives rise to unclosed terms, which are approximated using turbulence and combustion models. These simulations are computationally cheaper than DNS but the fidelity of the simulation predictions are strongly dependent on the accuracy of the turbulence and combustion models. One of these methodologies is Reynolds Averaged Navier-Stokes (RANS) simulations where all the governing equations are solved in an averaged sense and all the pertinent turbulent processes take place at the sub-grid level and thus need to be modelled. As a result, the performance of RANS simulations is strongly influenced by the accuracy of the models. A further technique known as Large Eddy Simulation (LES) solves the filtered governing equations and resolves the turbulent processes which are associated with length scales greater than the computational grid spacing. However, the physical processes occurring at the unresolved sub-grid scale still need to be modelled. Since part of the turbulence is resolved in LES, the implications of the modelling inaccuracy are less serious in LES than in RANS, although LES is more computationally expensive than RANS. For the foreseeable future, RANS and LES will continue to be used for engineering simulations of practical problems (e.g. design-cycle of next generation tidal and marine turbines, simultaneously fuel-efficient and low-pollution cars, aeroplanes and gas turbines, etc.), with LES likely to play an increasingly important role in the future with the increased affordability and deployment of supercomputing.

Similarly, advancements in supercomputing will enable more detailed plasma physics and material science calculations than currently possible, which will enhance the predictive capabilities of the

simulations and accelerate future innovations. In laser-plasma physics, UK-based researchers are currently unable to run ensembles of 3D simulations, although these are common in the US, where researchers have access to national laboratory facilities. The UK, through the development of novel numerical techniques (e.g. EPOCH particle-in-cell), is also ideally positioned to take a lead in the study of laser-plasma interactions for fusion but is limited to 2D simulations of single speckles while the international standard is rapidly becoming multi-speckle or 3D.

The aforementioned discussion indicates that supercomputing is an indispensable tool for future research in engineering and material science and future scientific breakthroughs will increasingly come from simulations enabled by supercomputing [1]. Thus, investments in supercomputing and improving the UK's capability in this respect will be instrumental in terms of maintaining and reinforcing its world-leading status in technology developments and innovation. All of these will help the UK to address the grand challenges faced by current society (e.g. low emission power generation and propulsion, fuel efficiency, alternative energy generation), which will drive wealth creation, a better quality of life and the development of highly-skilled personnel in the future.

**References**

[1] *Panel report on recent significant advances in computational science, http://sc.doe.gov/ascr/ProgramDocuments/Docs/Breakthroughs2008.pdf*

## 7.1 Combustion simulations: towards the delivery of a safe energy economy

*Contributors: Nilanjan Chakraborty[1], David R Emerson[2] ([1]Newcastle University; [2]STFC Daresbury Laboratory)*

**Vision:** Energy is a key national infrastructure that is recognised as a major global challenge. Today's energy mix includes traditional supplies of combustion energy alongside emerging low-carbon solutions from renewables and nuclear energy. Combustion is a dominant and vital component in delivering reliable and cost-effective energy for both power generation and transportation and will remain so for many decades. Its safe and efficient use, and understanding its impact on health, transportation, and climate change, pose major societal and economic challenges. The UK Consortium on Turbulent Reacting Flows (UKCTRF) will make optimum use of supercomputing, combined with an ambitious software development programme, to enhance current modelling of turbulent reacting flows and provide improved fundamental understanding to ensure environmental friendliness.

**Key research challenges:** The UKCTRF brings together 44 experts across 20 UK institutions, experienced in using high-performance computing to enable concerted collaborative Computational Fluid Dynamics (CFD)-related fundamental and applied research on turbulent reacting flows to reduce duplication, and tackle challenges grander than individual attempts. Since its inception in 2014, the UKCTRF has achieved significant scientific and industrial impact with hundreds of high-impact journal and conference papers[14] which utilised cutting-edge computational simulations by exploiting the recent advancements in supercomputing and software developments. High-fidelity simulation is at the core of the UKCTRF and this will enable researchers to investigate many important phenomena, including those that are currently beyond today's capability. Examples that are just becoming feasible include (i) simulation and modelling of multiphase reacting flows (e.g. droplet and pulverised coal/biomass combustion); (ii) combustion analysis of biogas and low calorific fuels derived from coal gasification; (iii) flame-wall interaction, including sprays with associated chemical kinetics; and (iv) combustion at elevated pressures; many of these topics have only recently become accessible due to the advancement of supercomputing [1, 2]. Continued progress in

[14] http://www.ukctrf.com/index.php/sample-publications/

supercomputing will enable challenges related to (i) and (ii) to be tackled in the next three to five years and as we advance towards exascale computing it will be possible to make significant advances in understanding in (iii) and (iv) in the 2024-2030 timeframe. However, it will only be possible to tackle these challenges if we have corresponding progress in software development that will enable significant advances in algorithms to deliver the required scalability and performance on modern hardware with its complex memory hierarchy and interconnect. It is also worth stating that the challenges outlined go beyond timely intellectual problems; they are firmly rooted in the need for improved industrial understanding of gas turbines and engine cylinders under elevated pressure conditions.

**Computing demand***:* The problems outlined involve turbulent reacting flows, with complex interactions occurring between chemical reactions, turbulence, heat transfer, radiation, fuel atomisation, and vaporisation, and further influenced by the separation of length and time scales. These problems exhibit a variety of multi-physics phenomena that are at the forefront and often beyond experimental capability. The UKCTRF community can perform Direct Numerical Simulations (DNS) that can resolve all length and time scales, without physical approximations, to produce the equivalent of a time-resolved experiment. As the length scale separation in turbulence is a strong function of turbulent Reynolds number $Re_t$, the grid spacing for DNS needs to decrease with increasing $Re_t$. This makes DNS of reacting flows extremely computationally expensive, and the computational cost increases steeply with turbulent Reynolds number (~$Re_t^3$). Thus, DNS of reacting flows is often limited to moderate values of $Re_t$, making it impossible to conduct DNS of combustion systems of practical interest involving complex geometries. Thus, engineering simulations rely on techniques involving suitable averaging/filtering of the flow field in the context of Reynolds Averaged Navier-Stokes (RANS) and Large Eddy Simulations (LES), respectively, but these methodologies lead to unclosed terms, which are approximated using turbulence and combustion models. However, the fundamental insights obtained from DNS data are already used by UKCTRF to develop high-fidelity models for RANS and LES, which in turn play pivotal roles in developing future generation energy efficient and environmentally friendly combustors. However, to realise this capability and maintain UK's world-leading status in turbulent reacting flow research requires access to substantial amounts of computational time, far beyond what is provided through the existing grant-based allocations. This is summarised in Table 1 where the challenges to be tackled by the UKCTRF community to remain globally competitive are listed. It is worth noting that the advancements in computational power will enable reactive flow simulations to achieve $Re_t$ values which are typical of

| | **Short term (~2021)** | **Medium term (~2024)** | **Long term (~2030)** |
|---|---|---|---|
| **Sample problems to tackle** | Simulation & modelling of multiphase reacting flows (e.g. droplet and pulverised biomass/coal combustion) and turbulent premixed/non-premixed/partially-premixed flames for high values of turbulent Reynolds number and thermodynamic pressure. | The research themes in the left hand column but in laboratory scale geometries with full details of primary and secondary atomisations + combustion analysis of biogas and low calorific fuels derived from coal gasification; flame-wall interaction, including sprays with associated chemical kinetics. | Systems level high-fidelity simulations (e.g. DNS and LES) for high pressure alternative fuel combustion in the presence of wall in internal combustion engines and gas turbines. |
| **Grid points needed** | $10^{10}$ | $10^{11\sim12}$ | $10^{13\sim15}$ |
| **Supercomputing requirements** | 10 PF | 100 PF | 1 EF |

engineering applications and laboratory-scale experiments. Thus, one-off simulations in short term (or medium term) are expected to be routine in medium term (or long term). Furthermore, it will enable UK researchers to aim for multi-scale multi-physics simulations in the future, which are currently inaccessible because of computational limitations.

A strategic approach is needed to meet the challenges faced by the UKCTRF community to develop efficient and clean industrial combustion technologies. This strategy should meet the computational demand outlined in the table above and also facilitate data storage, data analysis, and visualisation of the simulation results to ensure the wealth of information is available for the benefit of the UK.

**References**


[1] *Combustion-Pele: Transforming Combustion Science and Technology with Exascale Simulations* https://www.exascaleproject.org/project/combustion-pele-transforming-combustion-science-technology-exascale-simulations/

[2] *HPC4: High Performance Computing for Energy Whitepaper* https://hpc4e.eu/sites/default/files/hpc4e_project_files/HPC4E_Whitepaper.pdf


## 7.2 Materials Chemistry

*Lead Contributors: Scott M. Woodley, Alexey A. Sokol, C. Richard A. Catlow (UCL)*

**Vision:** The major challenge in material simulations is the transition from modelling systems in equilibrium to describing complex quantum phenomena, as well as open systems under kinetic control both in nature, industrial environments (e.g. construction materials for chemical and nuclear storage /reactors) and devices (e.g. batteries, solar cells, computers, lighting). We solve the atomic and electronic structures of materials with increasing system size and discover mechanisms that underpin longer scale processes with the focus moving from the single scale regime to a multi-scale level and to the interaction between electrons, vibrations and light. To enable these advances, we will continue to lead in the international development of methods and software including interatomic potentials, density functional theory (DFT) and higher-level theories for not just qualitative but quantitative predictions.

**Key research challenges:** Current computational capabilities mean that the implementation of electronic structure theories beyond DFT is essentially impossible to perform in the UK on all but the simplest systems; even a single project would need several times the processing power of ARCHER, since sophisticated electronic structure theories such as quantum Monte Carlo (MC) and the random phase approximation are essential to describe accurately the properties of many materials. As exotic quantum properties of materials are increasingly being revealed in experiments, it is essential that computational facilities are available to describe, predict, and understand these materials with accurate electronic structure techniques. Examples of key challenges include:

• The search for novel smart materials with self-cleaning and self-healing surfaces and interfaces (addressing the problems of corrosion, aging, surface and interface reactivity, growth, diffusion and nucleation), where by 2021 we will need to run DFT dynamics for 1000 atoms for 1ps; by 2024 for 5000 and 2ps; and by 2030 for at least 10000 and 2ps. These simulations will rely on being able to model atoms and electrons moving through interfaces between materials under electric field.

• The rational design of target (metastable) materials, e.g. biomaterials, pharmaceuticals and catalysts; predictive modelling of and simulation driven synthetic conditions; materials screening (for tuning a desired property); and structure prediction (global optimisation). Targeting specific materials, selective nano-particles, -tubes, etc. (sizes, shapes and compositions) and nucleation conditions by 2021; dial-a-polymorph, material/particle with specific properties after 2024.

• The development of an integrated, in operando representation of active sites in heterogeneous catalytic applications, including non-equilibrium conditions, for accurate quantitative description (and reaction rates) and comparison with experiment. Wider use of non-periodic, beyond-DFT approaches by 2021 along with multiple site sampling, making possible routine investigation of site interaction and synergy in catalysis by 2024. Connection between microscopic studies through kinetic MC to macro-reactor simulations including the flow and links to concurrent experiments will be pursued. Essential soft- and hardware support for studies of complex catalytic systems in (DC/AC) electric fields, interaction with light and under radioactive irradiation beyond 2024.

**Track Record:** Founded in 1994, the Materials Chemistry Consortium (MCC) has received continuous EPSRC support. The current scientific programme embraces eight related themes: Reactivity and Catalysis; Materials for Energy Generation, Transport and Storage; Environmental and Smart Materials; Soft Matter and Biomaterials; Materials Discovery based on screening and global optimisation; Fundamentals in Bulk Properties; Surfaces and Interfaces; Low Dimensional Materials. The MCC develops and optimises key materials software, and supports over 500 registered users on ARCHER from 80 research teams based in 29 UK institutions. The leading international standing of our members is evident from the range of the international collaborations, the frequent invitations to speak at major international conferences and the volume and quality of MCC outputs (over 120 papers pa, including [1-5]). Our representative applications include atomistic molecular dynamics simulations of high energy particles (radiation) in a material containing 0.5 billion atoms using 60,000 parallel processors in blocks of 12 hours and routine static electronic structure calculations on thousands of atoms.

**Computing demand***:* Currently, we have on average access to 18% of ARCHER (TIER-1) and 15% of THOMAS (TIER-2) High End Computing (HEC) resources. Even after suppressing expectations, our members request over twice the available HEC resources available to us. Our work can only continue to be world leading, cutting edge, with the opportunity of revolutionising the use of computational design in materials, with a 50-100 fold increase in current computing capabilities and, crucially, investment in software development. Catalysis, for example where such advances will make an impact, depends on phenomena at different time and length scales, from electronic transitions in the fs and ps regime to reagent transport on the ms to 10s regime. The integration of these scales is currently not possible and phenomenological models are used. However, with a 50-100 fold increase in power it will be possible, with suitable software advances, to encompass these length/time scales in a single calculation. In the longer term, more accurate methods will need to be employed, which scale less favourably with the number of particles of interest. Even at the current "standard" level of accuracy, to increase the system size by just one order of magnitude requires a 1000-fold increase in available resources, indicating the scale of resources required later in the decade.

**Track Record**


[1] *Nature, **546**, 280-284 (2017)*
[2] *Nature, **543**, 657-664 (2017)*
[3] *Nature Chemistry, **10**, 1112-1118 (2018)*
[4] *PNAS, **115**, 5353-5358 (2018)*
[5] *JACS, **140**, 7301-7312 (2018)*


## 7.3 High-power lasers and Quantum Electro-Dynamics (QED) science


*Contributors:* Remi Capdessus[1], Chris Ridgers[2] and Paul McKenna[3] ([1]Strathclyde, [2]York, [3]Strathclyde)


**Vision:** Exploit the development of high-power lasers in the QED-Plasma regime for basic science and as a driver for compact particle and radiation sources.

**Key research challenges:** Use the new high-power laser-plasma interaction regime to study relativistic plasma dynamics and quantum electrodynamics (QED) processes. The latter includes the generation of high-energy radiation by accelerated electrons and the production of electron-positron pairs. This so-called *QED-plasma* regime is postulated to exist in extreme astrophysical environments such as pulsar magnetospheres, but is yet to be explored in the laboratory. In the QED-plasma regime, the electrons radiate a significant fraction of their energy as an extremely bright flash of g-ray (>MeV energy) photons by nonlinear (inverse) Compton scattering (NLCS). These extreme multi-MeV to GeV g-ray flashes are extremely bright and potentially spectrally tuneable, enabling applications such as nuclear resonance fluorescence, inspection of cargo, deep penetrative imaging; softer x-ray sources will allow for phase-contrast imaging at unprecedented resolution and with negligible dose delivery. UK researchers have first explored the QED-plasma regime numerically [1] and in 2018 published the first experimental results demonstrating the high-field phenomenon of radiation reaction [2,3]. Maintaining the UK at the forefront of this new research needs new approaches to numerical modelling of laser-plasma interactions that include QED processes.

There are large active research groups working on both QED and accelerator science in Strathclyde, York, Imperial, Belfast and RAL and research on this subject is a major theme in the UK's annual high-power laser meeting with over 60 attendees. These groups have international links with international groups and the ELI laser programme. Addressing the physics of this new state requires us to: (i) develop our fundamental understanding and description of strong-field QED processes (where, unlike standard QED, we must use heuristic rather than *ab-initio* descriptions of the processes due to the large number of photons involved). For example, the role of electron spin dynamics in these processes is currently poorly understood; (ii) include the resulting descriptions in our large scale plasma modelling tools, mainly particle-in-cell (PIC) codes (the required numerical algorithms are in presently in their infancy); (iii) use these new codes to plan and interpret experiments investigating strong-field QED-plasma physics. These experiments will be performed on current and upcoming, international-scale, high-intensity laser facilities. Without an understanding of QED-plasma behaviour we will be unable to realise the applications of next generation multi-PW lasers, such as drivers of compact particle and radiation sources. Supercomputing capabilities are essential to perform QED-plasma simulations and demonstrate new scalable algorithms.

***In the short term (up to 2021)***: Develop new approaches to numerical modelling of QED processes in high temperature plasmas. Specifically, in order to be able to properly interpret and explain the future experimental results it will be necessary to include new QED processes– such as the electron spin effect– in the EPOCH PIC code. This will guide the design of experimental tests to benchmark numerical simulations. This will require at least similar computing resources to those deployed previously, i.e. simulations up to 30,000 core hours. As an order of magnitude estimate we expect the four very active groups in this area (Strathclyde, Imperial College London, York and Queen's University Belfast) to require on the order of 5-10 production runs of this type per year resulting in a requirement of ~1M core hours per year.

***In the medium term (up to 2024)****:* Investigate new interaction regimes in which different high energy areas of physics are connected, such as plasma physics, particle physics and astrophysics. This is likely to lead to the design of experiments in which some aspects of extreme astrophysical environments are replicated in the laboratory using up-coming multi-petawatt laser facilities. The design of such experiments requires accurate (and therefore costly) numerical simulations. New, multi-£100M multi-PW laser are expected to become operational in this timeframe, requiring more realistic simulations of experiments, in particular the interaction of lasers with dense plasmas. This requires at least one order of magnitude increase in particles per cell (to >500) to reduce simulation noise, resulting in a requirement of >10M core hours per year. Improving spatial resolution is desirable and an additional 10M core hours would enable several highly resolved simulations (~5 per year in total across the community) with double the spatial resolution in each direction.

***In the long term (up to 2030)**:* Develop a sustainable program of numerical simulations which supports the experimental campaigns/programs using multi-petawatt laser facilities and develop an ambitious and cross-disciplinary research program on QED science. As multi-PW laser systems become more ubiquitous the UK community should play a leading role in first experiments, given its historical strength in this field. In order to understand these experimental results a full programme of 3D simulations would be highly desirable contingent on at least a continuation of >10M core hours per year. As multi-PW lasers become the norm in the field we do expect significant growth (in the number of large-scale, well resolved 3D simulations required) however and an increase in computational resources to >100M core hours per year would help accommodate this.

**Computing demand**: PIC codes are the core simulation tool for high intensity laser-plasma interactions. Currently 3D PIC simulations require on the order of 30,000 core hours for one simulation producing approximately 1TB of data. Doing 3D simulations without significant approximations and assumptions is currently not possible. Simply doubling the resolution of such simulations will require 16 times the core hours to complete. For accurate modelling of near-solid densities the resolution will need to increase by at least a factor of 10 requiring in excess of 10,000 times the core hours. Therefore, there is significant long-term future demand for additional computational resources. As an example, a minimum of 10M core-hours would be needed for the next 5 years. This is assuming no growth in the community working on high-intensity laser interactions in the QED regime (~5-10 full time researchers across 5 institutions) although significant growth is expected as new multi-PW lasers begin to reach their expected operating intensities in the next few years. 10M core hours corresponds to around 5-10 small scale (30k core hours) 3D simulations per researcher (1-2 larger scale 3D simulations at ~100k core hours) or approximately 100 times as many 2D simulations per year. This number of 2D simulations is required to perform large parameter scans. These are essential as the underlying strong-field QED and ultra-relativistic plasma processes are not well understood – a situation that must be addressed in the multi-PW era. These 2D simulations scans will enable carefully selected 3D simulations, essential to plan the first experiments to explore this regime.

**Track Record**


[1] *C. P. Ridgers et al. Phys. Rev. Lett.* ***108*** *165006 (2012)*
[2] *J. Cole et al. Phys. Rev X* ***8****, 011020 (2018)*
[3] *K. Poder et al. Phys. Rev. X* ***8****, 031004 (2018).*


## 7.4 Plasma accelerators and light sources


*Contributors: Roman Walczak[1], Zhengming Sheng[2], Martin King[2], Paul McKenna[2] ([1]Oxford, [3]Strathclyde)*


**VISION**: Drive UK development of compact laser-driven or particle-driven particle and radiation sources towards potential applications in science, medicine, industry and defence.

**Key research challenges**: The interaction of intense lasers (or particle beams in some cases) with matter provides the capability to accelerate electrons and ions to high energies, over micron- to metre-scale distances and to generate ultrabright sources of X-rays and gamma radiation as well as secondary particles such as positrons and neutrons. These sources have the potential to be applied across a wide range of areas, including fundamental science, medicine, industry and security. Supercomputing provides an important capability for numerical modelling of the laser-plasma interaction physics underpinning the development of these sources.

The UK has several internationally leading groups working on wakefield acceleration and applications, both laser-driven (LWFA – Laser WakeField Acceleration) and beam-driven (PWFA –

Plasma WakeField Acceleration)), which contributed to breakthroughs in this field [1-4]. Since 2016, the UK research on wakefield acceleration has been nationally coordinated by the Plasma Wakefield Accelerator Steering Committee (PWASC, see http://pwasc.org.uk/ ) which worked out a roadmap for the years 2018-2050 (see http://pwasc.org.uk/ ) with 12 recommendations which form the basis for requests for resources such as supercomputing. In the short and medium terms up to 2024, the stable generation of high quality electron beams from LWFA and PWFA and 10 GeV single stage acceleration of electrons from LWFA will be among the key interests in this area, which also form the foundation for applications.

The UK is also at the forefront of laser-driven ion acceleration. Maximum proton energies of close to 100 MeV have been demonstrated and simulations on ARCHER were vital in understanding the ion-acceleration mechanism giving rise to that [5] and underpinning physics such as relativistic self-induced transparency physics in ultrathin foils [6]. Further supercomputing resource will allow for improved resolution along with the inclusion of computationally expensive additional physics (e.g. particle collisions, ionisation) which will enable further optimisation of the acceleration. This, along with identifying and exploiting new physical processes, will keep the UK at the international forefront of the field as next generation laser facilities, operating at higher intensities, come online in 2019-20. In addition to the UK programmes, there are a few big projects in the Europe such as ELI (Extreme Light Infrastructure), AWAKE (Advanced WAKEfield Experiment) and EuPRAXIA (European Plasma Research Accelerator with eXcellence In Applications), with many UK groups providing theory and simulation support.

**Computing demand**: PIC codes are again the core simulation tool requiring similar computational demands as those outlined in the high-power lasers and QED case. For accurate modelling of beam-driven or multi-pulse laser-driven plasma acceleration and radiation processes (incoherent and coherent from terahertz to even gamma-rays), high resolution (grid cells required increase with density and simulation scale) and low noise simulations will be needed. The future computational requirements for the largest simulations are indicated below requiring over 100 runs per year.

| | **Short (~2021) term** | **Medium (~2024) term** | **Long (~2030) term** |
|---|---|---|---|
| **Complexity** | Near-critical density plasma, attosecond coherent X-ray radiation, multi-pulses, QED pair production | Solid density plasma, particle collision and ionisation, large scale hybrid-PIC simulation, coherent hard X-ray radiation, multi-pulses | Full scale integrated kinetic simulation with low noise, accelerator start to end simulations including diagnostics |
| **Grid cells needed** | $10^9$ | $10^{10}$ | $10^{12}$ |
| **Performance** | 5 PF | 20 PF | 100 PF |

**Track Record**


[1] *S.P.D. Mangles et al. Nature **431**, 535-538 (2004)*
[2] *W.P. Leemans et al. Nature Physics **2**, 696-699 (2006)*
[3] *H.P. Schlenvoigt et al. Nature Physics **4**, 2 (2008)*
[4] *E. Adli et al. Nature **561**, 363-367 (2018)*
[5] *A. Higginson et al., Nature Communications **9**, 724 (2018)*
[6] *B. Gonzalez-Izquierdo et al., Nature Physics **12**, 505-512 (2016)*


## 7.5 Magnetic Confinement Fusion (MCF)


*Contributors: RJ Akers[1], M Barnes[2], B Dudson[3], B* McMillan[4]*, FI Parra[1], CM Roach[1], HR Wilson[3] ([1]CCFE, [2]Oxford, [3]York, [4]Warwick)*


**Vision:** Growing a leading-edge supercomputing capability is essential for the UK to reap the benefits from its investment in fusion.

**Key research challenges:** Success in harnessing nuclear fusion for terrestrial energy production will result in a safe, carbon neutral energy source. The UK is a world leader in Magnetic Confinement Fusion, where supercomputing simulations are transforming our understanding and maximising scientific output from experiments including: the European flagship, JET (Joint European Torus), which will soon operate in reactor relevant deuterium-tritium (D-T) fuel mixtures; the UK's, MAST-U (Mega Amp Spherical Tokamak Upgrade), which starts operation imminently and will test novel solutions to handle plasma "exhaust"; and the international ITER experiment under construction in France designed to demonstrate high fusion power gain. Research to optimise fusion energy reactor designs spans a wide range of problems that are extremely demanding computationally: optimising the plasma to achieve fusion conditions in the heart of the device; designing plasma facing materials to tolerate high radiation and thermal loads over sufficiently long times for the reactor to be economically viable; and developing complex components needed to operate the reactor. These challenges require the integration of models to resolve disparate scales in space (, velocity space for the plasma) and time: e.g. from the macroscopic response of the material structure over the reactor lifetime, to the nanosecond timescale associated with the immediate local impact of neutron bombardment; and plasma dynamical processes from microsecond timescales to the operating period of the reactor. Designing a fusion power plant requires the capability to do such calculations reliably and iteratively, and so the following are essential: (i) substantial access to the most advanced supercomputing hardware resources available, and (ii) development of new mathematical and computational algorithms to make previously "inaccessible" challenges tractable. MCF fusion researchers are strongly engaged in such activities both internationally and in the UK.

Here we focus on UK MCF research challenges requiring supercomputing in three areas that are directly related to the plasma:

***Plasma Turbulence*** causes loss of heat and particles from the plasma core, and determines fusion performance. Modelling requires supercomputing as the turbulent structures span broad ranges of scale in space and time. Research questions include: (1) Interactions between plasma flows and turbulence [1] with flows generated by external inputs or the plasma itself [2]; (2) The interaction between turbulence at disparate electron and ion space-time scales, so computationally demanding that few calculations have been performed; (3) A lower computational cost theory [3] to assist in understanding the impacts of cross-scale interactions; (4) High performance plasmas where turbulence acquires large magnetic perturbations [4], and burning plasmas where the turbulence is influenced by the fusion products; (5) Advances in gyrokinetic (GK) simulations will be tested against results from JET and MAST-U, make predictions for ITER and inform longer term development of predictive models. First principles transport calculations to predict performance using GK turbulence models [7] typically require $10^5$ cores running for ~$10^2$ hrs even with basic models. Such calculations will be needed more routinely in the future.

***Edge Plasma*** influences global performance and determines the wall loads from plasma exhaust. Novel exhaust solutions are needed and will be explored in MAST-U. Challenges include: Modelling the edge turbulence responsible for the spreading of the plasma exhaust. In the short term fluid plasma modelling will be used to interpret/guide MAST-U experiments and assess advanced divertor configurations. This will transform our understanding of the interactions between high-power turbulent plasmas, neutral gases and solid surfaces. In the medium term, simulations capturing both plasma physics and detailed engineering design, will support ITER operation and power plant design. In larger, hotter devices such as ITER, fluid models will need augmenting with more expensive GK models, requiring theoretical [5] and computational advances.

***Macroscopic Instabilities*** limit plasma confinement and cause rapid losses that damage the device. Eruptions known as edge localised modes must be controlled to avoid serious erosion in future tokamaks. Another example is neoclassical tearing mode (NTM), which degrades confinement and can cause disruptions that rapidly deposit plasma energy onto the walls. Disruptions cannot be

tolerated in ITER. NTMs can be controlled using microwave heating and current drive, but this needs demanding kinetic modelling to resolve the roles of electrons and ions on disparate length- and time-scales [6]. Supercomputing is needed to model such instabilities in future tokamaks.

**Computing demand***:* MCF is already a major user of the world's largest supercomputers, e.g. the XGC1 PIC code used 90% of Titan (27 petaflop) for 3 days to simulate ~1ms of ITER real time. It is our ambition that such scales of simulation should become routine in the next 5-10 years. UK MCF computational plasma researchers have made discoveries [1,3,4,6] and pioneered new algorithms [3,5,7], and would benefit considerably from having access to national computing resources that compete with those available to researchers in US, Germany and Japan.

In the table below, the increased numbers of lattice points in the medium and longer term are based on our estimates of the increased resolutions that will be needed for higher fidelity simulations to model: (i) cross-scale interactions between turbulence driven at the disparate perpendicular length scales associated with the electron and ion Larmor radii (requiring larger grids in both directions perpendicular to the magnetic field), (ii) electromagnetic turbulence (which requires higher resolutions in both real space and velocity space), and (iii) the impact of turbulence-driven transport on equilibrium profiles (where multiple local simulations are required to model turbulence across the device). In the short term, our largest simulations will require access to a 10PF machine, and we anticipate running ~O(10-20) such simulations per year. The demand for such simulations is expected to grow so they can be fully exploited to help guide high profile international fusion experiments that will be coming on line.

| | **Short (~2021) term** | **Medium (~2024) term** | **Long (~2030) term** |
|---|---|---|---|
| **Sample problems to tackle** | Impact of plasma flow on turbulence.<br>Model cross-scale turbulence interactions. | Self-organised flows.<br>DNS of cross-scale interactions.<br>Core profile predictions. | Full-device transport predictions. |
| **Lattice points needed** | $10^{12}$ | $10^{13\sim14}$ | $10^{15}$ |
| **Supercomputing requirements** | 10 PF | 100 PF | 1 EF |

**References**


[1] *CM Roach et al, PPCF, 51, 124020 (2009)*
[2] *Barnes et al, PRL, 111, 055005 (2013)*
[3] *Hardman et al, PPCF, 61, 065025 (2019)*
[4] *Dickinson et al, PRL, 108, 135002, (2012)*
[5] *Geraldini et al, PPCF 59, 025015 (2017)*
[6] *K Imada et al PRL 121, 175001 (2018),*
[7] *Barnes et al, Physics of Plasmas, 17, 056109 (2010)*


## 7.6 Engineering Design and Optimisation Enabled by Mesoscopic Simulation of Multiphase Flows

*Contributors: Professor Kai Luo (UCL)*

**VISION**: To enable engineering design and optimization through scale-bridging mesoscopic simulation of multiphase flows.

**Key research challenges**: Multiphase flows are ubiquitous in nature, sciences, biomedicine and a variety of engineering disciplines. In various energy and propulsion devices, liquid fuels are injected

into a gaseous environment, which undergo atomization, evaporation and mixing. In electricity generation using fossil fuels, nuclear and solar technologies, liquid water is converted into vapour of supercritical conditions. Gas/oil extraction involves multiphase flow from nano-pore to reservoir scale. In the process and pharmaceutical industries, foams, emulsions and powders are often transported by a multi-component flow system. In manufacturing, liquid materials flow into moulds to make components or deposit on surfaces to create desired patterns. Finally, in biomedicine, "smart" drugs are being developed, which can swim in complex fluids crowded by red blood cells and proteins.

Multiphase flows are characterised by a large number of properties such as density ratio, surface tension, wettability, hydrophobicity. Correspondingly, there are many nondimensional numbers that define multiphase flows, like Atwood number (A), Capillary number (Ca), Eötvös number (Eo), Marangoni number (Mg), Morton number (Mo), Ohnesorge number (Oh), Reynolds number (Re), Richardson number (Ri), Stokes number (St), Weber number (We). Moreover, there are a wide range of geometric parameters associated with multiphase flows, such as droplet/power diameter and shape. For practical applications, for example, the droplet diameter ranges from 1 nm to 10 mm. In one litre of water, there are more than $10^{12}$ nonmodisperse droplets of 10 μm in diameter. To resolve droplets of 10-micron droplets in 1 $cm^3$ cube would require $10^{15}$ grid points. Much more grid points are needed to capture topological features of droplets in collision, merging, separation. Due to the Eulerian nature of the carrier phase (flow) and the Lagrangian nature of the dispersed phases (droplets, particles), conventional CFD based on the Navier-Stokes solvers is not the best method. In contrast, the lattice Boltzmann method (LBM) has several advantages: automatic interface capturing, ease to handle complex geometries (non-spherical droplets/particles, geometries of engineering systems), and extremely high parallel efficiency on massively parallel CPUs and/or GPUs. This is supplemented by dissipative particle dynamics (DPD) and molecular dynamics (MD) to provide molecular level insights.

**Computing demand**:

| | **Short (~2021) term** | **Medium (~2024) term** | **Long (~2030) term** |
|---|---|---|---|
| **Sample problems to tackle** | Laboratory experiments: droplet collisions & surface deposition; channel multiphase flow. | Devices: inkjet printing; drug production; computer cooling. | Systems: spray in gas turbines; fuel cells & batteries; human blood circulation. |
| **Lattice points needed** | $10^{12}$ | $10^{13\sim14}$ | $10^{15}$ |
| **Supercomputing requirements** | 10 PF | 100 PF | 1 EF |

**Track Record**


[1] *Q. Li, et al. Prog. Energy Combust. Sci. 52, 62-105, doi:10.1016/j.pecs.2015.10.001 (2016).*
[2] *M.O. Bernabeu, et al. J. R. Soc. Interface 11, 17, doi:10.1098/rsif.2014.0543 (2014).*
[3] *G.G. Wells, et al. Nat. Commun. 9, 7, doi:10.1038/s41467-018-03840-6 (2018).*
[4] *H.H. Liu, et al. J. Fluid Mech. 837, 381-412, doi:10.1017/jfm.2017.859 (2018).*
[5] *R. S. Qin. Sci Rep 7, 7, doi:10.1038/s41598-017-09180-7 (2017).*

## 7.7 Computational Aerodynamics

*Contributors: Mark Savill[1], Dave Emerson[2] ([1]Cranfield University; [2]Daresbury Laboratory)*

**Vision:** Maintaining and building UK competitive international position for Aerospace requires continual advances in computational engineering simulation & design capability.

**Key research challenges:** The UK has the *number one aerospace industry in Europe*, and is globally *second only to the US*, with good potential for growth [1]. It is already a major contributor to the UK economy supporting over 100,000 jobs; generating revenue in excess of £24 billion annually. The research and development (R&D) challenge is to maintain and build on this strong position.

Key needs are to replace expensive full-scale experimental testing with Virtual-Test-Bed analyses; to address the Civil Aviation Authority aspiration of Digital Certification by 2050; to advance Computational Engineering Design methods for the environmentally friendly aircraft needed in the 2030 timeframe (ultimately to meet 2050 emissions targets [2,3]); and at the same time to impact goals for wind turbine power generation, high speed train and performance car developments.

The UK R&D base is in a good position to address these major, global challenges [4] via the UK Applied Aerodynamics High End Computing (HEC) Consortium, which brings together nationally and internationally recognised experts from over 40 UK academic and industrial organisations experienced in latest scientific computing methods on the most advanced supercomputers for Computational Fluid Dynamics (CFD) simulations, as part of the EPSRC/STFC Collaborative Computational Project for Engineering Science to advance necessary cross-disciplinary and multi-scale simulation and modelling for identified EPSRC Grand Challenges. In particular a key need is to address a range of transient and unsteady phenomena including gust, flutter responses for aircraft wings; turbomachinery primary-secondary flow path interactions and off-design instabilities; turbine and wake interference effects in wind farms; particulate track-bed and terrain interactions with high speed trains and vehicles; and overall Multi-disciplinary Design Optimisation (MDO).

The Aerospace Technology Institute Strategy provides a succinct road map for impact, but sufficient supercomputing resources are needed to allow high fidelity virtual propulsion systems and other Virtual Test Beds to be established within the medium term; as well as to ensure full MDO and multi-point operational design trade-offs are achievable in the medium to long term.

Digital Certification for whole aircraft is extremely challenging, but given increased resources it should also then be possible to use advanced simulations to replace at least some of the current certification testing within the longer-term timeframe.

Computational engineering is already playing an increasingly important role in understanding and developing high speed ground transportation, as identified by the Royal Academy of Engineering assessment of Eurostar re-design; while Formula 1 companies in particular seek more interactive advanced simulation to drive rapid prototyping; and future energy supply will rely in part on renewable resources that require computational optimisation.

**Computing demand***:* Current ARCHER Leadership Project work has been producing flagship, component-level, high-fidelity simulations, multi-fidelity aerodynamic optimisation or multi-disciplinary analysis for: Airbus transonic wing buffet; Bombardier thrust reverser; international high lift wing, Rolls-Royce gas turbine engine fuel injectors aero-thermomechanics; DNV-GL offshore wind turbines; MBDA hypersonic vehicle test case; high speed train (HST) in cross flow and gust; Bloodhound air-brake and surface interactions. In the short term (~2021) 1-10PF compute resources are required for a wider range of such component-level, one-off, high-fidelity (~1Bn cell) virtual-test-bed solutions to address fundamental engineering challenges, industrial proof-of-concept, and for multi-objective design optimisation. In the medium term (~2024) similar dedicated 10-100PFlops resources will be needed to progress to handling whole products: airframes, engines, wind-turbine

farms, re-entry vehicles, HST and other ground transports (at ~1T cell fidelity). In the longer term (~2030) the need to address new aircraft configurations, with fully integrated blended fuselages and turbo-electric engine systems, as well as uncertainty quantification and management to assess design sensitivities generally, will require 100PFlop to Exascale resources in order to handle the necessary shift to largely time-elapsing one-off and stochastic analyses, and to enable sufficiently closely-coupled multi-physics and multi-scale analyses for routinely deploying full MDO across flight envelope operating conditions.

**References**


[1] *Lifting Off: Implementing the Strategic Vision for UK Aerospace, UK Government Report (2013)*
[2] *Flightpath 2050: Europe's Vision for Aviation*
[3] *ACARE: Realising Europe's vision for aviation; Strategic Research and Innovation agenda*
[4] *CFD Vision 2030 Study: A Path to Revolutionary Computational Aerosciences, NASA/CR-2014-218178.*


## 7.8 Quantum mechanics-based materials modelling


*Contributors: Matt Probert[1], Carla Molteni[2], Chris Pickard[3] ([1]University of York; [2]King's College London, [3]University of Cambridge); UK Car-Parrinello (UKCP) Consortium [1]*


**Vision:** To be world-leading in the first principles simulation of materials, by continually improving the codes we develop and the calculations we perform on cutting-edge hardware.

**Key research challenges:** Developments in materials science underpin many new technologies and promise to deliver solutions to many of the major global challenges. There is an ever-growing demand for new molecules / materials to meet the challenges of the future, but developing them by the traditional 'trial and error' experimental route is getting exceedingly expensive. Our solution is 'materials by design' – to use the computer programs we have developed to do 'in silico' design and test of materials – which is now possible due to the accuracy of the methods and the computing power available to execute these programs on sufficiently complex systems.

As an example, the group of Prof Carla Molteni (Kings College London Physics) is currently exploring the electronic structure effects of mutations affecting the binding of neuro-transmitters to ligand-gated ion channels (LGIC) [2], which mediate fast synaptic transmission and are involved in several neurological disorders. With current computing power, it takes several months to explore each mutation, which is the goal in the short-term (up to 2021). Whilst this is a very useful step in the "Understanding the Physics of Life" grand challenge, it will require a step-change in computing power to screen mutations effectively, which is the goal for the medium-term (up to 2024). To go beyond this to "Developing Future Therapies" grand challenge by developing a new biomolecule that could target a particular receptor in a biomolecule and hence be a new candidate drug for medical advances, will require exascale computing which is the long-term aim (up to 2030).

As another example, the group of Prof Matt Probert (York Physics) is developing the widely-used CASTEP [3] program to study potential new thermoelectric materials that can efficiently harvest waste heat from industrial processes and convert it into useful electrical energy as part of the "Energy" grand challenge. In the short term, we can do calculations to test if a given material / stoichiometry / structure is a good thermoelectric, but we are limited as the computational search space is large and the number of calculations required per candidate is daunting. With access to exascale computing power, we could automatically screen a large set of possible materials / compositions / structures using the AIRSS approach (Ab initio Random Structure Searching; [4]) developed by Prof Chris Pickard (Cambridge Materials) and rapidly focus attention onto a few promising candidates, dramatically reducing the time-to-discovery. The same methodology can be applied in many different fields of science and technology, for example the work of Dr Paul Bristowe

(Cambridge Materials) on developing lead-free perovskite photovoltaic materials [5], and the work of Prof Mike Finnis (Imperial College Materials) on developing ultra-high temperature ceramics [6].

**Computing demand***:*

| Example study | Short (~2021) term | Medium (~2024) term | Long (~2030) term |
|---|---|---|---|
| Mutations and LGIC | Single mutation = 3 months | Single mutation = 1 week | Develop new targeted biomolecule |
| Thermoelectric material screening | Single material = 3 months | Single material = 1 week | Comprehensive search for new materials |
| Supercomputing requirements | System size: 10 PF using 1% total / project | System size: 100 PF using 1% total / project | System size: 10 EF using 1% total / project |

**Track Record**


[1] http://www.ukcp.ac.uk/
[2] https://pubs.acs.org/doi/10.1021/acs.jpclett.8b03431
[3] https://doi.org/10.1524/zkri.220.5.567.65075
[4] https://doi.org/10.1088/0953-8984/23/5/053201
[5] https://doi.org/10.1021/acs.chemmater.6b03944
[6] https://doi.org/10.1103/PhysRevB.91.214311


## 7.9 High fidelity simulations of turbulent flows


*Contributors: Laizet S[1], Sandham N[2] ([1]Imperial College London, [2]Southampton); UK Turbulence Consortium (UKTC)*


**Vision:** Using large-scale high-fidelity simulations to improve our understanding of turbulent flows and to learn how to manipulate them for practical applications.

**Key research challenges:** Our daily life is surrounded and sustained by the flow of fluids. Blood moves through the vessels in our bodies, and air flows into our lungs. Fluid flows disperse particulate air pollution in the turbulent urban as well as indoor environments. Fluid flows play a crucial role for our transportation and in our industries. Many of the environmental and energy-related issues we face today cannot possibly be tackled without a better understanding of the dynamics of fluids. From a practical point of view, fluid flows relevant to scientists and engineers are turbulent ones; turbulence is the rule, not the exception. In the last 24 years, the UK Turbulence Consortium (UTKC), with over 50 members across 22 institutions, has provided vital information required to make many advances in our understanding of turbulence by simulating these flows using supercomputers. Examples include: identification of the structures involved in the turbulence cascade process; discovery of the mechanisms that sustain near-wall turbulence and of new dissipation laws; design of turbulence models redefining industry standards; innovative turbulence control techniques. Hundreds of manuscripts[15], generating thousands of citations, have been published in high-impact factor journals, demonstrating the impact and quality of the UKTC research. The outstanding work from the UKTC has also helped its members to secure multi-million £ grants from governmental funding bodies and industries.

[15] https://www.ukturbulence.co.uk/publications.html

While the Navier-Stokes equations constitute a broadly accepted mathematical model to describe the motions of a turbulent flow, their solutions can be extremely challenging to obtain due to the chaotic and inherently multi-scale nature of turbulence. The smallest scales impact the largest scales, and small changes to boundary conditions, initial conditions, or grid resolution, for example, can have a dramatic impact on the solution. A particular challenge is the non-linear cascade of turbulent energy from large eddy scales to the small-scale eddies, at which turbulent energy is converted to heat and dissipated. The turbulent scales are typically separated by many orders of magnitude. A simulation that captures/resolves all of these scales is called direct numerical simulation (DNS). Except for the simplest of problems, simulating turbulent flows require supercomputing resources. However, even with today's state-of-the-art algorithms and petascale systems, DNS is only feasible for a small class of problems, namely those at moderate Reynolds numbers (defined as the ratio of inertial forces to viscous forces) and in simple geometries.

**Computing demand***:* An exascale machine would allow an incredible three- to five-fold increase in Reynolds number by comparison to current systems, with the possibility to include more complicated physical models (e.g., combustion or aero-elasticity). DNS of a complete aerospace system (airplane with full engine simulation, aircraft in manoeuvring flight, etc.) or a full scale offshore wind farm (with hundreds of large-scale turbines) during operation will become possible in the 2030 timeframe. Such DNS, based on ~100-1000 billion grid points (depending on the range of turbulent scales of the problem), and performed on 10-100 million cores for 100-1000 hours (depending on the complexity of the flow), will be used to validate and improve our theoretical understanding of turbulence, and to develop and evaluate turbulence control techniques. Such simulations will become the norm, not the exception. They will also be important for the design of turbulence models, crucially needed when cost constraints for industrial designs or short turnaround times (such as in the Formula one sector) means simulating all the scales of the flow is impossible. DNS at exascale may require significant reformulation of existing flow solvers, implementation of new physics, and development of a more nuanced problem formulation, but with the potential to produce significant advances in our quest towards a greener future. An exascale supercomputer will generate substantial industrial impacts in the transportation, energy supply/generation, biomedical and process sectors [1]. These sectors have an enormous impact on the environment and the global energy budget and any improvements in aero/hydrodynamic, mixing and heat transfer efficiencies as well as reducing emissions, noise and drag rely in large parts on a better understanding of the overarching subject of turbulence.

**References**

[1] https://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/turb_flow_exascale.pdf

## 7.10 Atomic, Molecular and Optical R-matrix calculations

*Contributors: Jonathan Tennyson (University College London); UK Atomic Molecular and Optical R-matrix (UK AMOR) consortium*

**Vision:** Atomic, Molecular and Optical (AMO) physics addresses a wealth of key fundamental and technological problems concerning physical systems. The range of these problems has grown hugely with the development of ultrafast lasers and ultracold physics as well as the interest in quantum information and computing. The quantum mechanical equations which govern these problems are well known but in almost all cases defy solution. The calculable R-matrix method is a UK development which provides the dominant methodology for solving many AMO problems. and is being increasingly adopted in nuclear physics and ultracold chemistry. UK AMOR exploits the R-matrix methodology to address leading edge problems in AMO physics. Our code base is

**Key research challenges:** *Short/Medium term* (1) Providing a full ab initio quantum description of the complex processes which occur in intense laser fields as probed by modern experimental facilities such as XFEL (X-Ray Free-Electron Laser Facility): such calculations have traditionally proved to be beyond the reaches of ab-initio theory; (2) Fusion occurs in stars and offers the hope almost limitless cheap electricity on Earth: AMOR aims to provided the detailed atomic and molecular data, unobtainable experimentally, to aid design of modern fusion experiments and the interpretation of stellar spectra. *Medium/Long term*: (3) Low energy electron collisions with living tissue are now known to be the major cause of radiation damage via DNA strand breaks but the processes need to be understood at the molecular level require detail elucidation allow molecular level radiation damage models to be built. *Long term*: (4) development of R-matrix based methodology for treating ultracold collisions is underway and will in due course become a significant user of computer time.

**Computing demand***:* ARCHER is addressing some of the issues in (1) and (2) above. However, for example the (atomic) intense laser field project (1) above needs to be extended to consider new physics in the form of molecular processes, double ionisation and spin orbit coupling effects. On current account this work would take about 1.5 Million kaus per year. This would only be deliverable with a significantly faster / bigger machine. Project (2) will probably continue at approximately the current usage; however the *Medium/Long term* projects (3) and (4) are currently using no ARCHER time and can be anticipated to use substantial quantities (maybe up 1 Million kaus per year each).

**Track Record**


[1] *O. Hassouneh, N. B. Tyndall, J. Wragg, H. W. van der Hart, and A. C. Brown, Phys. Rev. A 98, 043419 (2018)*

[2] *D. D. A. Clarke, G. S. J. Armstrong, A. C. Brown, and H. W. van der Hart, Phys. Rev. A 98, 053442 (2018)*

[3] *S.P. Preval, N.R. Badnell, M.G O'Mullane, J. Phys. B: At. Mol. Opt. Phys. 52, 025201 (2019)*

[4] *J.D. Gorfinkiel, S. Ptasinska, J. Phys. B: At. Mol. Opt. Phys. 50, 182001 (2017).*

# 8 Digital humanities and social sciences

*Editors: David de Roure (University of Oxford), Andrew Prescott (University of Glasgow)*

An exponential growth is occurring in the quantity and range of data recording human life, behaviour and society, as a result of developments such as the Internet of Things and growth of 'smart' devices in home, health, transport and urban life. When linked to developments in AI and data analytics, this growth in social data has the potential to deliver transformational benefits across society. However, achieving such transformations is dependent on innovation in computational methods and access to much greater computing power than at present. The way in which evidence-based social and governmental decisions are made could be made considerably more efficient and effective if supercomputing power capable of analysing and modelling vast amounts of data could be made available. Greater computing power and the new supply of data will enable us to better understand society, and ensure that the UK is well placed to exploit future developments in this rapidly changing arena.

Access of humanities researchers to supercomputing facilities has hitherto been patchy due to lack of supercomputing capacity. However, as historians, literary researchers, archaeologists and others increasingly work with large digital resources, there is a pressing need for more supercomputing capacity. Historians need to interrogate large corpora of e-mail and web archives. Humanities scholars in the UK wishing to use born-digital archives are already falling behind other European countries such as Denmark. Language is central for future developments in fields such as robotics and supercomputing facilities, and supercomputing will be essential to exploit the large linguistic corpora now being developed. Work in computational musicology and virtual environments require supercomputing infrastructure to create spin-offs for the creative industries. Supercomputing can also be valuable in exploring paper archives, facilitating for example the tracing of information about individuals across widely dispersed archives of the Holocaust.

The following case studies illustrate how supercomputing will have a transformative impact on the nature of research in the arts and humanities and facilitate innovative work in key areas such as:

- retrieval of information from born-digital data and archives
- improved modelling of language, supporting developments in areas such as robotics and cognition
- use of AI to retrieve widely dispersed historical data
- documentation of fragile built heritage, supporting tourism
- new computational techniques for creative industries such as music
- improved corporate and public policy decision-making.

Through work in areas such as these, supercomputing in the humanities and social sciences has the potential to be a driver of major social and economic growth and development.

## 8.1 Humanities research in born-digital archives

*Contributors: J. Winters (University of London)*

**Vision:** Enabling historians, literary scholars and other humanities researchers to easily explore massive born-digital archives such as web archives, corporate e-mail archives and social media.

**Key research challenges:** Born-digital archives are presenting researchers in the humanities with unprecedented challenges of scale and complexity: in the fourth quarter of 2018, Twitter reported 321 million ‘monthly active users’ (126 million ‘monetisable daily active users’) [1]; in 2015, a single government department was found to have an email server with half a billion emails [2]; a single hashtag - #metoo – was used 1.5 million times on Instagram in 2018 [3]. As internet traffic increases

exponentially, so will the size of web archives and the computing power required to interrogate them for research purposes.

The challenges for humanities researchers investigating born-digital material are illustrated by web archiving. Web archiving dates back to the early activities of the Internet Archive in late 1996, and in the UK the British Library has been archiving the .uk country code Top Level Domain (ccTLD) since April 2013, following the extension of legal deposit legislation. The legal-deposit UK Web Archive is an enormously rich resource for researchers in a range of humanities disciplines, including history, literature and languages, area studies, history of science and technology, and digital humanities, but quantitative analysis in the UK remains rare. Lack of access to computational infrastructure is currently determining the kind of research that is being undertaken with this data, and that is small-scale and qualitative.

In 2017, the UK Web Archive held approximately 500TB of data, and 60-70TB is added every year. Even the derived data made available by the British Library is dauntingly large for a researcher based in a typical humanities department. To take one example, the list of URLs crawled between 1996 and 2013 (data purchased from the Internet Archive to supplement the legal-deposit collection) is too large to be hosted on GitHub, and the data for 2013 alone is 73GB [4]. At the time of writing, the Internet Archive holds more than 351 billion archived web pages.

Humanities researchers working with the archived web are already being constrained in what they can do by the unavailability of supercomputing or super-computing facilities. This is a problem that will only worsen as the archives of the web increase, and become more impenetrable, each year. Our aim is for the UK to become, by 2030, the leading international exponent of the retrieval and analysis of large born-digital archives and to have pioneered key pipelines and tools for such forms of research.

**Computing demand***:* Although the UK has been world-leading in using web archives for humanities research, the continued growth in born-digital data requires a step change in provision of computational infrastructure. In Denmark, researchers have been able to take advantage of access to supercomputing through the Cultural Heritage cluster at the University of Aarhus [5]. This enables the high-level analysis of the Danish Web Archive as a whole, and is not limited to the study of websites selected more or less manually or to analysis of derived datasets which omit multimedia content. Greater computing power is required to extend this appoach and to deal with video and sound content.

**References**


[1] *Lauren Feiner, 'Twitter crashes 10% after it says expenses will rise sharply', CNBC Tech https://www.cnbc.com/2019/02/07/twitter-q4-2018-earnings.html (7 February 2019)*

[2] *The digital landscape in government: 2014-15: business intelligence review (London: The National Archives, 2016) http://www.nationalarchives.gov.uk/documents/digital-landscape-in-government-2014-15.pdf*

[3] *'Instagram year in review 2018', Instagram Info Center (12 December 2018) https://instagram-press.com/blog/2018/12/12/instagram-year-in-review-2018/*

[4] *British Library, 'Crawled URL index: Jisc UK web domain dataset (1996-2013) http://data.webarchive.org.uk/opendata/ukwa.ds.2/cdx/*

[5] *Cultural Heritage Cluster https://dighumlab.org/cultural-heritage-cluster*

## 8.2 Corpus Linguistics

*Contributors: Martin Wynne[1], David de Roure[2] ([1]Bodleian Libraries, University of Oxford, [2]University of Oxford)*

**Vision:** A deeper understanding of society and the human cultural record through large-scale processing of digital records of spoken and written interaction.

**Key research challenges:** The analysis of our vast and growing text corpora has led the way in developing data-driven methods both in the humanities and social sciences. These have revolutionised our understanding of language from historical periods through to the born digital social record of today, and enabled studies ranging from the genealogy of knowledge to the social media analysis of elections. Mining texts for information is an increasingly important process across all disciplines, as well as in business and public life.

The field of corpus linguistics is therefore a key enabler as well as a valuable area of research in its own right. With applications from lexicography and translation to sentiment analysis and question answering systems, it offers the knowledge and tools necessary for sophisticated exploration of the textual record. It also makes a major contribution to commercial activities involving language processing. AI techniques have long been explored, and language models are valuable outputs of computational linguistics research as other domains adopt these methods.

The availability of corpora across multiple languages underpins this work, and it demands an infrastructure for full text datasets at huge scale — including today's social media archives — accompanied by processing power, interfaces and APIs for exploration and analysis. Records of speech are also part of this infrastructure, hugely important in terms of language use, accents, dialects, and cultural content; they may come from oral history projects, broadcast media and home recordings, alongside the growing capture of voice at scale by modern smart devices.

**1. Data-driven linguistics:** Over the past fifty years corpus linguistics has led the way in developing data-driven methods in the humanities and social sciences. This has not only enabled researchers to test assertions about language against real evidence of usage, but it also led to new types of research question involving larger scales of data and time-frames, new theories of grammar, and it has revolutionized numerous areas of applied linguistics such as language learning, translation studies, and lexicography, as well as making a major contribution to a number of commercial activities involving language processing. Closely related to these developments is the increasing prominence of stylometry in literary studies which is facilitating understanding of authorial characteristics and enabling authors of anonymous texts to be identified. Language continues to evolve and change as societies change, and data-driven research into trends in language usage is therefore a constant, ongoing challenge.

**2. Text mining:** Large datasets of text and speech are not just for linguists but for everyone, although extracting reliable information from texts relies on linguistic knowledge, methods, tools and datasets. Mining texts for information is an increasingly important process in almost all disciplines, and in many areas of business and public life. Corpus linguistics is therefore a key enabler as well as a valuable area of research in its own right, offering the knowledge and tools necessary for sophisticated search based on linguistic knowledge, and methods for interpreting the results.

**3. Exploiting historical text collections:** Many libraries across the world, ranging from the National Library of Wales with its comprehensive online archives of newspapers and periodicals published in Wales, to the National Library of Norway, which is digitising its entire collection, are making their collections available in digital form. If the outputs of the mass digitization projects which are underway were consistently made available not just as page images in digital libraries, but as full text datasets accompanied by processing power and interfaces for exploration and analysis, then many more researchers could participate in new forms of digital research. A student with a desktop

computer can nowadays access more texts than a senior researcher could track down in their lifetime a few years ago, and there is huge potential for democratizing research and applying the wisdom of crowds to the understanding of history.

**4. Freeing the speech archives:** Historical records of speech are sparse, but hugely important as holders of information about language use, accents and dialects, and of cultural content. Recordings of speech in oral history projects, broadcast media, home recordings, business archives contain information about people's lives and experiences and can provide unique insights, but are mostly held on analogue media, with very few are available digitally to researchers. Computing facilities to digitize, store, transcribe, annotate, make available and preserve these archives could create a renaissance in oral history and transform our understanding of the recent past.

**5. 'Social climate change':** Nowadays digital text and speech are being produced everywhere, potentially searchable, downloadable and analysable in real time. New facilities and instruments are necessary to handle this data deluge, which presents opportunities not only to understand language change in real time, but to understand how society might be changing, by capturing information about the 'social climate' through discourse, in the same way that millions of digital sensors allow us to track the weather and understand it better.

In order to achieve these aims, we need to create, by 2030, very large-scale linguistic corpora recording human use of language in a wide range of contexts.

**Computing demand:** In additional to increased computing power, there are specific requirements for the following:

1. **Long-term data storage**, since virtually all original data is of high value and should be preserved in perpetuity. Recordings of speech or writing are not reproducible datasets. They are unique records of human behaviour, and a part of our cultural record.
2. **Secure access and authorization in a wide domain of trust**, involving millions of users, and commercial data providers, since many of the potentially most valuable datasets for research are also commercially valuable and not freely available at zero cost. In order to exploit the possibilities of text mining more widely and thoroughly, the domain of trust for secure access to protected resources needs to include data producers and owners, such as publishers and social media enterprises, allowing a much wider range of datasets to be made available for data mining and other research purposes, with users numbering in the millions worldwide. Access to data and interfaces in the humanities and social sciences cannot be restricted to a core team of more or less fixed size or duration, as might happen in the experimental sciences - it is likely to be of potential interest to large and growing numbers researchers, teachers, other professionals, students and members of the public. This requires access and authorization mechanisms which are sustainable in the long-term and which are scalable to millions of data items and users.

## 8.3 Develop Cultural Analytics Capacities for Archives

*Contributors: Tobias Blanke, Michael Bryant, Reto Speck, Mark Hedges (King's College London)*

**Vision**: Advance UK capacities in the analysis of large-scale historical collections in archives and content-holding institutions using advanced machine learning and AI technologies.

**Key research challenges**: To move to the next generation of research in history and enrich a data-driven understanding of the past we need supercomputing to process data across large-scale collections we have built up in recent years.

In 2010, Nature covered the topic of big cultural data and compared typical data sets used in cultural research with those in the sciences that can be considered big [1]. While data sets from the Large Hadron Collider are still by far the largest around, cultural data sets can easily compare to other

examples of big sciences. The Holocaust Survivor Testimonials Collections by the Shoah Foundation contained 200 TB of data in 2010. The present project is based on these kinds of Holocaust collections brought together in the large-scale European collaboration European Holocaust Research Infrastructure (EHRI) [2]. Here, we cite particular examples of the supercomputing-relevant work in EHRI, but archives and other memory institutions will become increasingly important in this context, with the new discipline of cultural analytics catching up fast with other more advanced disciplines. For archives to realise analytical benefits, it is essential that they fully understand their roles and responsibilities in a big data environment with supercomputing capacities.

Our example from EHRI shows how advanced computational modeling can address the specific challenges posed by dispersed and complex archives with poor metadata and catalogue information [3]. It starts from dictionary-based approaches to develop an initial evaluation of the Holocaust memories in survivor testimonies using unsupervised learning. Using generative Recurrent Neural Networks (RNNs), we then proceeded to generate a larger training corpus of positive and negative memories using 'distant supervision' [4]. With this corpus, we were able to train a highly accurate neural network that qualitatively and quantitatively improved the baseline dictionary model. Based on the accuracy of the advanced model we are confident that we can analyse the Holocaust memories to current research standards. To this end, we employed three advanced methods of computational semantics. These helped us decipher the decisions by the neural network and understand, for instance, the complex sentiments around family memories in the testimonies. While we generally succeeded with our objectives, several limitations remained mainly with regards to the testing and training infrastructures. Because the processing of neural networks requires advanced computational infrastructure, which we lacked, we were not able to apply appropriate testing of various hyper-parameters. These include neural network hyper-parameters such as sequence length or network architectures, which we should have tested further, but also the sentiment lexicon or the length of the testimony parts. For the lexicon, e.g., we relied on a single example, which though commonly used might not be the best choice in our circumstances. Further experiments will improve our performance further.

The second example from EHRI works through some of the data management challenges and uses the recent AI innovation of graph databases and triple stores. The automatic extraction of names, locations, etc. in the textual resources is an important task for Holocaust digital research [5]. Being able to automatically identify names allows for automatic metadata enrichment of the available resources and thus for better semantic indexing and search; improved document retrieval for specific people; better research related to groups of people affected by the Holocaust; automated authority list creation, etc. EHRI has developed a high-performance person extraction service, which is based on a named entity recognition pipeline tuned to the specific needs of EHRI. Especially challenging are historical geographies, which dynamically change over time. We developed sophisticated disambiguation mechanisms, which selects the appropriate location from our knowledge base when more than one candidate exists and dynamically updates the knowledge base with new sources in the EHRI data store. We developed high-performing synonymisation techniques for multi-labelling of place names; dynamic detection of place hierarchies for disambiguation; and co-occurrence statistics. However, the graph database we are using for this data modelling does much of its processing entirely in-memory, which limits our data ingestion capacity.

We wish now to develop these experiments to use AI across very large dispersed archives. This will result in new methods and tools which will represent an innovative use of AI and facilitate investigation of dispersed archives.

**Computing demand**: For the EHRI machine learning experiments, we currently used P2.xlarge EC2 instance with GPU support, on which it took on the order of 10 hours to train the generative RNN at 50 epochs. This does not allow for any kind of cross-validation or other advanced evaluation. On a NVIDIA Quadro P4000 with 8GB (GPU) memory, CUDA 10.1, running Tensorflow r1.13 GPU, the same experiment still lasted 6 hours for the full 50 epochs of the text generation using RNNs. While

this was approx. 60% faster than the AWS (Amazon Web Services) P2.large GPU instances it is still quite slow for serious analytical work with the testimonies and 'distant reading' experiments. In the future, VMs optimised for Deep Learning workloads that could reduce training time by e.g. a factor of three or more could have a transformational effect on how we could develop and iterate models and test different hyperparameters. Regarding our second use case and our data management challenges, we need much better access to in-memory processing. High-memory instances are needed for running triple stores or other so-called no-sql databases that either use entirely in-memory models or greatly benefit from fitting entirely into RAM. This is particularly useful when working with datasets such as DBPedia or Geonames, or graph-like models which require wide or deep traversals to perform semantic enrichments.

**References**


[1] *Hand, Eric. "Culturomics: word play." Nature 474, no. 7352, pp. 436-440, 2011.*

[2] *https://ehri-project.eu/*

[3] *Blanke, Tobias, Michael Bryant, and Mark Hedges. "Understanding memories of the Holocaust—A new approach to neural networks in the digital humanities." Digital Scholarship in the Humanities, OuP, Oxford, 2019.*

[4] *Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1, no. 12 (2009): 2009.*

[5] *De Leeuw, Daan, Mike Bryant, Michal Frankl, Ivelina Nikolova, and Vladimir Alexiev. "Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure." 2018 IEEE 14th International Conference on e-Science (e-Science), pp. 58-66. IEEE, 2018.*


## 8.4 The Fragile Heritage Hub: providing a digital record to support reconstruction and to creating resilience for our global cultural legacy


*Contributors: Professor A. Wilson, Dr C. Gaffney and Professor V. Gaffney (University of Bradford)*


**Vision:** To provide a challenge-led, transformative approach to the assessment and anticipation of global heritage risks through a scaleable, digital infrastructure supporting 3D object creation through web scraping and data fusion.

**Key research challenges:** The recent catastrophic fire in Notre Dame follows a number of similar, high profile events including fires at Windsor Castle, York Minister, the Glasgow School of Art and the National Museum in Brazil, and demonstrates that widespread and irrevocable destruction of global heritage is amongst the world's most intractable of problems. Aside from cultural value these events represent lost opportunities for communities across the world to benefit economically, or take comfort socially, from heritage assets which may be damaged from conflict, neglect, looting, vandalism, natural disaster, environmental change, and unregulated development

The UK has a leading role in heritage protection and creative sector development internationally and can assist in addressing multi-facetted and complex heritage threats through tested digital responses from the global to community levels. However, the Arts lack the capacity to provide or give access to supercomputing or storage to provide a national capacity for this work. Relevant funding streams often have narrow geographic/ single disciplinary focus, and yet heritage threats have global reach and arise from complex combinations of factors relating to development, poor governance, societal division, environmental risks and the lack of technical expertise.

The Fragile Heritage Hub derives from the Arts and Humanities Research Council (AHRC)-funded pilot project "Curious Travellers". The results from this project demonstrate the need for a significant

supercomputing infrastructure and the requirement for a challenge-led, approach to heritage risk[16]. Curious Travellers infrastructure anticipates potential damage or loss through automated web scraping of vast amounts of existing digital heritage imagery, integration of these with donated images through a live Web and App interface and merging such data with existing HD digital records including laser scanning. The infrastructure could support digital heritage recording nationally and provides a heritage disaster response system with global potential. The capacity of the existing infrastructure to scale up to international requirements has been proven through fieldwork on historic buildings damaged in the 2015 earthquake in Nepal, on remote recording of Fountains Abbey in the UK and in reconstruction of the Temple of Bel in Palmyra, destroyed by ISIS in 2015. A single scrape of data for the Temple of Bel involved automated filtering, calibration and rectification of more than 187,000 images. Overall, the pilot project has processed more than 6 million images, mainly from social media, but also through public donation. If supported such an infrastructure would underpin national strategies to use heritage relating to development, entrepreneurship and sustainability. This is not possible through conventional funding routes or with existing supercomputing capacity.

**Supporting Documentation**

- Digital heritage recording contributes to the protection of heritage assets and to a range of UN Sustainable Development Goals. **Evidence**: DCMS Heritage Statement (2017). https://www.gov.uk/government/publications/the-heritage-statement-2017
- Digital heritage recording promotes partnerships for sharing and leads to the generation of digital resources with societal legacy, academic and economic value. **Evidence:** V&A 2017 ReACH Declaration https://www.vam.ac.uk/research/projects/reach-reproduction-of-art-and-cultural-heritage
- Digital heritage recording preserves heritage at risk by creating a new proactive response to natural and human disaster. **Evidence:** UNESCO 2010 managing disaster risks https://whc.unesco.org/en/managing-disaster-risks/; British Council 2017 – In Harm's Way. https://www.britishcouncil.org/sites/default/files/in_harms_way_-_second_edition_online_version.pdf
- Cascading digital products supports the broad networks and inclusive sets of stakeholder groups required both to preserve heritage andalso create research outputs that directly impact the field of heritage protection. **Evidence**: UNESCO 2016 report - Culture Urban Future. https://unesdoc.unesco.org/ark:/48223/pf0000246291
- Digital heritage engagement at local levels enhances the understanding, value and importance of tangible and intangible heritage assets. **Evidence:** Historic England 2017 report Heritage Counts https://historicengland.org.uk/research/heritage-counts/; Culture White Paper 2016 https://www.gov.uk/government/publications/culture-white-paper.
- Digital heritage recording of tangible and intangible heritage supports maintenance and celebration of cultural identity which can be leveraged in peace keeping activity. **Evidence**: British Council 2017-20 Corporate Plan. https://www.britishcouncil.org/sites/default/files/corporate-plan-2017-20.pdf
- The outputs of digital heritage recording have an economic value that can be leveraged and contribute to an increase in sustainable tourism, social cohesion and community well-being. **Evidence:** Historic England 2017 report – Heritage Counts https://historicengland.org.uk/research/heritage-counts/; British Council 2016 – Arts Connect Us https://www.britishcouncil.org/arts/about.

[16] http://www.visualisingheritage.org/news.php

## 8.5 Computational Musicology

*Contributors: David De Roure[1], Tim Crawford[2], Mark Sandler[3], Kevin Page[1] ([1]Oxford, [2]Goldsmiths, [3]QMUL)*

**Vision:** Continue innovation in computer analysis of music and its application in the creative industries, building on the opportunity of new computational methods over increasingly large scale audio data.

**Key research challenges**: Research challenges include conversion of recorded audio to a musical score, audio source separation, audio to score alignment, and structural analysis of music. These approaches have applications throughout the music production pipeline, from composition through to discovery and reuse, and in the study of musical works.

The importance of the creative industries to the UK economy is widely recognised and recently articulated in terms of cultural value [1]. The music industry represents significant Gross Value Added (GVA; over £3Bn) and is one of the most investment-intensive industries in the economy and a significant export industry.

The UK is a leading player in a very active international R&D community which continuously innovates in the algorithms for music information retrieval. These are applied to the growing corpus of digital recorded audio, with increasing computational demand. For example, an international "Digging into Data" project on "Structural Analysis of Large Amounts of Music Information" utilised 1000s of compute hours to perform algorithmic structural analysis of a corpus of western music recordings, with a subset analysed by experts to obtain a ground truth [2]. All these approaches have seen a step change in effectiveness in the last 5 years with the adoption of Deep Learning approaches.

These research outcomes have applicability in musicology and in the music industry. The AHRC Transforming Musicology project, in the Digital Transformations Programme, explored how emerging technologies for working with music as sound and score can transform musicology both as an academic discipline and as a practice outside of universities [3]. The EPSRC FAST Programme Grant (Fusing Audio and Semantic Technologies) has brought the latest techniques to bear on the entire recorded music industry, end-to-end, producer to consumer and is an early adopter of Data Science techniques in music analysis (e.g. [4]). AI techniques are now being used to assist music composition and production.

While many of the algorithms are performing complex signal processing tasks, there is also a significant research activity around the surrounding workflows, metadata, annotation, storage, discovery and retrieval of musical content [5][6]. There is significant interest in the use of web-scale quantities of Linked Data and in the automatic generation of ontologies from audio.

The UK has several international leading groups in this field, with teams at QMUL, Goldsmiths, Oxford, City, and the Alan Turing Institute.

**Computing demand**: The computational demands are increasing with the increasing availability of digital audio data from new recordings, live performances, and studio production. We have already moved from using legacy monaural and stereo recordings to multitrack, but there is a clear industry trend towards high resolution spatial audio, use of microphone arrays, audio capture for augmented reality (AR) and virtual reality (VR). In addition to this step change in audio data supply, these techniques can be combined with video processing. The physical modelling of instruments (virtual acoustics) already makes significant supercomputing demand. A pressing computational challenge arises when we process data in real time for live applications, which also drives algorithmic and computational innovation.

**Track Record**


[1] *Geoffrey Crossick and Patrycja Kaszynska, Understanding the value of arts & culture: The AHRC Cultural Value Project. 2016. https://ahrc.ukri.org/documents/publications/co-creation-report-2015/*

[2] *AF Ehmann, M Bay, JS Downie, I Fujinaga, D De Roure. Music Structure Segmentation Algorithm Evaluation: Expanding on MIREX 2010 Analyses and Datasets. ISMIR, 561-566*

[3] *Lewis, Richard; Crawford, Tim and Lewis, David. 2015. Exploring information retrieval, semantic technologies and workflows for music scholarship: The Transforming Musicology project. Early Music, 43(4), pp. 635-647.*

[4] *K. Choi, G. Fazekas, K. Cho and M. Sandler, "The Effects of Noisy Labels on Deep Convolutional Neural Networks for Music Tagging," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 2, pp. 139-149, April 2018. doi: 10.1109/TETCI.2017.2771298*

[5] *K. R. Page, S. Bechhofer, G. Fazekas, D. M. Weigl and T. Wilmering, "Realising a Layered Digital Library: Exploration and Analysis of the Live Music Archive through Linked Data," 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, 2017, pp. 1-10. doi: 10.1109/JCDL.2017.7991563*

[6] *D De Roure, KR Page, B Fields, T Crawford, JS Downie, I Fujinaga. An e-Research approach to web-scale music analysis. Philosophical Transactions of the Royal Society A. 369(1949), pp 3300-3317. 2011. https://doi.org/10.1098/rsta.2011.0171*


## 8.6 Computational modelling for decision-making


*Contributors: David De Roure[1], Nigel Gilbert[2] ([1]Oxford, [2]Surrey); also see [1]*


**Vision:** In order to deal with an increasingly complex world, we need ever more sophisticated computational models that can help us make decisions wisely and understand the potential consequences of choices.

**Key research challenges**: Examples across the sector include: enhancing the quality of decision-making and policy design in public policy; deploying new data sources in urban modelling in the retail and transport sectors; identifying and managing risk and forecasting how economies will evolve in the finance sector; enabling innovative high-quality design and manufacturing, more efficient supply chains and greater productivity; guiding policy and business decisions in environmental modelling.

The 2018 Blackett Review “Computational modelling: technological futures” [1] provides a review of UK computational modelling capabilities, summarised in [2]. It anticipates that large-scale availability of data about individuals (e.g. from ubiquitous sensors) will transform modelling as we go forward, with modelling spanning many scales. Uptake is set to increase, with modelling used for strategic and policy-level issues and increasing engagement of senior decision-makers. Looking ahead, more models will be built by computer and help to train computers.

Increasing computing power and greater availability of data have enabled the development of new kinds of computational model that represent greater detail and enable virtual experiments before trying things out for real. Modelling approaches include discrete/continuous, dynamic/static, stochastic, Markovian, individuals/population, logics, automata and algebraic models, complex and emergent systems, game theory, machine learning and ensemble modelling. To focus on one with immediate exascale demand, agent-based modelling codes have been developed for supercomputers and clusters and used in multiple disciplines. These contain large sets of agents that each represent individuals or groups, with differing levels of autonomy, different characteristics and different behaviours. Large scale simulations enable modellers to explore emergent properties, or predict when tipping points will be reached.

To demonstrate the scale of a simulation: modelling the passage of children through the UK school system (e.g. to test approaches to schools admission policies, the effect of school exclusions, and teacher recruitment and retention policies), using data from the Pupil Level Annual school Census (annual since 1996), would require simulation through time of around 8.5M pupils, 440K teachers and 25K schools per year; each would be represented by an agent, and model the interactions between pupils in the same class, teachers in the same school, and schools in the same area.

**Computing demand**: The clear exascale requirement occurs as simulations scale up, e.g. to the scale of the population. A comprehensive 2017 review [3] identifies codes for multiple platforms; some codes (e.g. REPAST HPC) are designed for cross-platform use, with scalability testing on Top500 resources. The Blackett review notes that such explorations are often very computationally intensive, but recent advances in computational infrastructure have the potential to make them viable, and that ensuring appropriate national facilities are available to both academia and industry will be essential. Social simulation increasingly appears in the project portfolios of major supercomputing centres, for example the 440 TF facility in ETH, Zurich, is used to support "social supercomputing" applications.

**References**


[1] *Walport, M., Calder, M., Craig, C., Culley, D., de Cani, R., Donnelly, C., Douglas, R., Edmonds, B., Gascoigne, J., Gilbert, N., Hargrove, C., Hinds, D., Lane, D. C., Mitchell, D., Robertson, D., Rosewell, B., Sherwin, S. and Wilson, A., (2018) Computational modelling: technological futures. Report. Government Office for Science, London.*

[2] *Calder M et al. 2018 Computational modelling for decision-making: where, why, what, who and how. R. Soc. open sci.5: 172096. http://dx.doi.org/10.1098/rsos.172096*

[3] *Sameera Abar, Georgios K. Theodoropoulos, Pierre Lemarinier, Gregory M.P. O'Hare. Agent Based Modelling and Simulation tools: A review of the state-of-art software, Computer Science Review, Volume 24, 2017, Pages 13-33. https://doi.org/10.1016/j.cosrev.2017.03.001.*


## 8.7 Large Scale Network Analytics


*Contributors: Pete Burnap[1], Rob Procter[2] ([1]Cardiff, [2]Warwick)*


**Vision:** Establish techniques and capability for the analysis of interconnected networks at scale, including social (e.g. Web, social networks), technical (e.g. telecommunications, Internet of Things), and economic (e.g. financial transactions).

**Key research challenges**: Understanding and explaining the evolution of the Internet and Web as large-scale, socio-technical systems, including social media analytics (e.g. in the context of elections). Detection of anomalies in financial transactions. Cybersecurity of future digital ecosystems such as smart cities, homes and infrastructure; and detection of malevolent activity.

Today we live life in digital networks of all kinds, generating massive longitudinal data sets of millions of people, including location, financial transactions, communications, and minute-by-minute interactions [1]. This data has unprecedented breadth, depth and scale. Analysis of this data occurs in the fields of Network Science, Internet Science, Web Science and more broadly Computational Social Science [2], as well as in cybersecurity.

Analytics of this large scale network or graph data has a range of applications, including anomaly detection in the context of financial transactions, analysis of healthcare data from millions of smart homes for precision medicine, cybersecurity (e.g. network intrusion and insider threat detection) national security (e.g., tracking emergent terrorist threats) and 'information operations' (e.g., fake news and trolling). These have immediate – and in some cases urgent – societal and economic benefits, and underpin the success of the UK industrial strategy in AI and the Data Economy.

Web Science and Internet Science take a socio-technical systems approach in studying the ways in which society and technology co-constitute one another. This helps understand the evolution of the Web and the Internet, respectively, and helps us design the future systems for continuing sustainability and social benefit [3]. The notion of "social machines", due to Tim Berners-Lee, has become established as an analytical lens onto these large-scale, socio-technical systems [4].

These are interdisciplines which engage the analytical power of a diverse range of researchers including mathematics, physics, sociology, economics, psychology, law and computer science.

**Computing demand**: Large graph analytics is one of the most computationally intensive methods in computational social science. There are multiple implementations of Breadth-First Search and Single Source Shortest Path algorithms for supercomputers: code development and optimisation is challenging and these algorithms are also used for benchmarking (for example, Graph500 which is based on a breadth-first search in a large undirected graph). As problem sizes scale, best practice is becoming established in the tradeoffs between algorithms, data structures, systems and storage. With increasing scale of data supply, these research methods are set to make full use of exascale capability.

**References**


[1] *David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational Social Science, Science 6 February 2009: 323 (5915), 721-723. DOI:10.1126/science.1167742*

[2] *Thanassis Tiropanis, Wendy Hall, Jon Crowcroft, Noshir Contractor, and Leandros Tassiulas. 2015. Network Science, Web science, and Internet Science. Commun. ACM 58, 8 (July 2015), 76-82. DOI: https://doi.org/10.1145/2699416*

[3] *James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, Daniel Weitzner. Web Science: An Interdisciplinary Approach to Understanding the Web, Communications of the ACM, July 2008, Vol. 51 No. 7, Pages 60-69. DOI: 10.1145/1364782.1364798*

[4] *Nigel Shadbolt, Kieron O'Hara, David De Roure, Wendy Hall. The Theory and Practice of Social Machines. Springer 2019. DOI: https://doi.org/10.1007/978-3-030-10889-2*


## 8.8 New and Emerging Forms of Data


*Contributors: Mark Birkin[1], David De Roure[2] ([1]Leeds, [2]Oxford)*


**Vision:** Establish innovative computational methods in social data analytics to realise the value of the massively increasing number and growing diversity of continuously generated and interconnected data sources.

**Key research challenges**: *Analysis*: How can we derive inferences, often in real-time, from billions of data points relating to patterns of movement, social interaction or economic behaviour? *Design:* How do we design complex systems which combine the interactions between billions of agents (both human and physical) to deliver next generation technologies e.g. mobility as a service, combining vehicles, sensors, people and infrastructure to deliver socially beneficial outcomes? *Policy*: How can we manage interventions to deliver economic benefits, quality of life, healthy or sustainable environments, using New and Emerging Forms of Data (NEFD) in combination with e.g. deep learning, massive ensembles or data assimilation to validate policy conclusions? These challenges require innovation in streaming data analytics, privacy-enhancing technologies, and data linkage at scale.

The increasing pervasiveness of technology and connectivity is enabling new ways of living and working, changing how social, economic, political, and cultural processes are created, and the

resultant variety of new data rapidly emerging is important to supplement, augment, and in some cases replace datasets collected by traditional means [1]. The benefits of these new sources are significant for social science and evidence based policy; e.g. in informing policy around active travel or pollution for health, or links to housing policy for economic development and sustainability.

However, this brings a significant computational challenge because new sources, such as the 'cyberphysical' devices of the Internet of Things, provide continuously generated (streamed) data, with an anticipated scale of billions of devices in coming years through deployments such as Consumer IoT, Smart Cities and connected vehicles. The computational challenge is deepened by the growing scale and diversity of these sources, together with the requirement to link sources while protecting privacy. Innovating in these new algorithmic methods requires significant computational capability and itself informs the requirements for future infrastructures.

Two influential reports have scoped this area of innovation and investment in new methods. The OECD report on New Data for Understanding the Human Condition [2] describes the new data sources in the social science infrastructure context, while the Blackett Review of Internet of Things [3] recommends that the UK will be a world leader in the development and implementation of the Internet of Things, to enable goods to be produced more imaginatively, services to be provided more effectively and scarce resources to be used more sparingly.

UK investments in this area include the Economic and Social Research Council (ESRC) Urban Big Data Centre, the EPSRC PETRAS IoT research hub [4], the Urban Analytics programme at the Alan Turing Institute, and the ESRC "Data Analytics and Society" Centre for Doctoral Training. This research and development community develops social data science and AI methods alongside spatial analysis, geostatistics and a wide variety of disciplinary perspectives, engaging with new sources including IoT and video, in order to understand process, structure and interactions across spatial and temporal scales.

**Computing demand**: Infrastructures for NEFD are rapidly emerging and gaining uptake. Underpinning data has been supplied through investment in the Admin Data Research Network, Urban Big Data Centre and Consumer Data Research Centre. A new blend of AI and Data Science is emerging which will promote deployment of reinforcement learning, data assimilation, machine learning etc for insight into social process and human behaviour, and from there to policies and interventions. This research calls upon a breadth of infrastructures and algorithmic approaches, and we are seeing rapid change in programming practice. Scenarios with exascale demands include massive linkage of anonymised data, and the combination of modelling and analytic approaches needed to make sense of the data. With the increasing scale of data supply we anticipate significant computational experiments as part of this programme of work.

**References**


[1] *Mark Birkin. Big Data for Social Science Research, Ubiquity, Volume 2018, January 2018*

[2] *New Data for Understanding the Human Condition – International Perspectives. OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences. OECD February 2013.*

[3] *Internet of Things: making the most of the second digital revolution. Government Office for Science. December 2014.*

[4] *Taylor, P et al. Internet of Things: realising the potential of a trusted smart world. Royal Academy of Engineering: London. 2018.*

# 9 Mathematics and Science of Computation

*Editors: Beth Wingate[1], Mark Parsons[2], Jeremy Yates[3] ([1]University of Exeter; [2]University of Edinburgh, [3]University College London)*

The skill-sets required to successfully use supercomputers for applications in the sciences, the humanities, and industry and commerce also drive the development of the technology. Because computer architectures themselves are undergoing a living evolution, it is essential that the skill sets that create the algorithms, technology and software, be encouraged and cultivated, allowing the use and development of advanced computing technology to flourish in the UK.

This section highlights the mathematics and science of computation through six challenges targeted to evolve the way we research and work together as the age of advanced computing technology, and forthcoming new architectures, unfolds. These challenges were discussed at the Royal Society Meeting on Numerical Algorithms for high performance computing[17] (Higham, Grigori and Dongarra), and at a UKRI meeting on Software for Exascale Computing (both in 2019). To date the majority of funding for UK universities and companies in this area has come from the European Commission's Horizon 2020 programme, specifically the Future and Emerging Technologies in HPC initiative (FETHPC).

The challenges are:

- **Developing the new area of *Mathematics at scale***, which highlights the notion of mathematics in computation;
- **Performance modelling and next generation benchmarking,** which highlights the key role of performance prediction and measurement for exascale systems;
- **Composable languages and tools across supercomputing applications,** which highlights how mathematics, languages and computational tools work together;
- **Working with industry to design the next generation of supercomputing systems,** which highlights the role of working with companies that design and develop new architectures that are candidates for future supercomputers;
- **Next generation development cycle**, which highlights how we must find new ways of working together in the digital era;
- **New REF Unit of Assessment for Computational Science**. The Research Excellence Framework (REF) is the UK Government's system for assessing the quality of research in UK Higher Education Institutions. This challenge proposes a way to nurture and encourage the development of these skill sets in the UK.

The title of this section has been chosen carefully. It reflects the authors' contention that the future of modelling and simulation requires new mathematical thinking in the context of the computing technologies that are emerging as we approach the exascale era. In the past, users of computer systems focussed on science ***by*** computation. To develop new software technologies and algorithms to support the next generation of research results we must focus on mathematics and the science ***of*** computation.

[17] https://royalsociety.org/science-events-and-lectures/2019/04/high-performance-computing/

## 9.1 Mathematics at Scale

*Contributors: Beth Wingate[1], Colin Cotter[2] ([1]University of Exeter; [2]Imperial College)*

**Vision:** Develop new mathematical paradigms to advance and accelerate the entire pathway from application, algorithm, through to software. In taking advantage of universal mathematical structure, this will also advance co-creation of next generation computer architectures.

**Key research challenges:** Traditional uses of supercomputing include analysis and explanation of science, but it is only the starting point for the utility of models [1]. Today's computational models, especially those required to meet the goals of the Industrial Strategy, go well beyond traditional purposes. ***We now expect to be able to blend models with data in order to predict, optimise and control complex systems not only for science, but for social and economic phenomena, and to do so on machines that are very different than in the past***. Fortunately, many of our most important applications have shared mathematical or statistical infrastructure we can take advantage of to create new methods and mathematical algorithms in order to accelerate the performance of simulations, but also to participate in the co-creation of the new computer architectures themselves. In addition, the UK has invested in many years of high quality mathematical sciences that are ripe to be applied to these problems. One can think of computational models as different types of racing cars; they all share the need for a chassis, but depending on their use, the chassis will be built in different ways. For supercomputing, the chassis is the application's inherent mathematical structure of the entire pathway from application → algorithm → to software required to run efficiently on new computers. The key challenges are:

*The Outer Loop* is a phrase that acknowledges the idea that we now need to develop new mathematics, algorithms, and methods, to use our models for understanding how to control, predict, and optimise complex systems. These subjects include Data Assimilation, Uncertainty Quantification, Optimisation, and some applications of Data Analytics. Understanding and exploiting these are mathematical research challenges to create new mathematics and algorithms that can be shared by many applications, whether in the humanities or science, and could significantly accelerate our use of supercomputers.

*The development of a new branch of mathematics (Mathematics at Scale)* that can create new mathematics and algorithms to provide computational blueprints for new architectures and, once understood, provide guidance about how new architectures could transform the underpinning equations and processes for better application performance. New mathematical ideas need to be developed, such as time-parallelism, simulation reconstruction, and probabilistic computing, and shared with all applications that can benefit.

*Connecting mathematicians with computational science domain experts and supercomputing researchers* is a major challenge because in the past, computational applications have evolved independently of each other. Developing new ways of working together to connect applications that have shared mathematical infrastructure and developing new mathematical and algorithmic ideas to accelerate their progress is essential to advancing the economic benefits of supercomputing research. The importance of this activity was noted in the supercomputing white paper which recommended the establishment of "a Collaborative Computational Project, including computational scientists, computational mathematicians and ICT researchers to enable a joint technology development programme that develops algorithms, hardware and software." A specific example of such a project was the EPSRC-funded NAIS (Numerical Analysis and Intelligent Software) project which helped bridge the gap between numerical analysts, computer scientists and supercomputing software developers. The proposed collaborative computational programme, consisting of multiple projects, each working on specific challenges, should be the main co-ordinator of research and

technical development activities and disseminator of results such that all communities benefit from developments which may have broad applicability.

**Computing demand:** Algorithms are widely recognized as a key aspect of the oncoming exascale era of computing. An example of this in the UK is the joint NERC/STFC/Met Office Gung Ho project which transformed the mathematical and computational infrastructure of Met Office next generation model. Another example of recent UK-led work is the CRESTA (Collaborative Research into Exascale Systemware) project which employed a software co-design approach to plan how six major codes could be modified to run on an Exascale system [5]. The importance of mathematics at scale as a key step in supercomputing research has been documented internationally. One example is from the 2019 SIAM CS&E (Society for Industrial and Applied Mathematics – Computational Science and Engineering) quote of Jack Dongarra, "*A steady stream of software hits arises from an artful coupling of the right mathematical formulation with algorithms and software shaped to exploit evolving computational architectures*", which is detailed in [4]. For additional examples see the US DOE's Applied Mathematics Research for Exascale Computing [2], which presents the scientific case for mathematics research in the US Exascale Computing program. In Europe [3], the original European ETP4HPC Strategic Research Agenda (2013) omitted Mathematics and Algorithms and their impact on system software and programming mathematical development. This was quickly seen as too limiting and, in the second draft (ETP4HPC Strategic Research Agenda 2016), an entirely new section focused on mathematics and algorithms was added. In the current draft (2017) of the 3rd edition, Mathematics and Algorithms appear throughout.

**References**


[1] *Blackett Review: "Computational Modelling: Technological Futures",* https://www.gov.uk/government/publications/computational-modelling-blackett-review

[2] *Applied Mathematics Research for Exascale Computing, US DOE,* https://www.osti.gov/biblio/1149042-applied-mathematics-research-exascale-computing, *doi* 10.2172/1149042

[3] *ETP4HPC Strategic Research Agenda versions 2013, 2016, 2017*

[4] *Dongarra et al, SIAM Review, Vol 60, No. 4, pp. 808–865, 2018*

[5] *CRESTA Whitepaper: "Architectural Developments Towards Exascale", Parsons et al, The University of Edinburgh, 2014, doi 10.5281/zenodo.3234480*


## 9.2 Performance Modelling and Next Generation Benchmarking


*Contributors: Mark Parsons[1], Michèle Weiland[1], Stephen Jarvis[2], Simon McIntosh-Smith[3], Jeremy Yates[4] ([1]University of Edinburgh; [2]University of Warwick, [3]University of Bristol, [4]University College London)*


**Vision:** Supercomputing continuously advances. Effective investment focussed on scientific need requires accurate benchmarking of new technologies and modelling of performance at scale.

**Key research challenges:** Benchmarking of supercomputing systems has been a key component of system development and procurement for decades. For many years, benchmarks that were not representative of scientific application patterns were employed with the aim to extract the best possible performance from the hardware. In recent years, however, benchmarking has become broader with benchmarks now focussing on real-world scenarios and specific system characteristics for a variety of application types. The purpose of this benchmarking is to get a clear picture of a new system's likely performance at scale, and to understand the behaviour of an existing system, both under a realistic workload. Because developments happen globally, by their very nature, such

activities involved international collaboration. The scale of next generation systems poses many challenges for performance modelling and benchmarking as follows:

*Performance prediction at scale:* new hardware technologies may only be available in small quantities at first. Robust, repeatable methods of predicting performance are therefore needed in order to predict behaviour of new technologies at scale (these can be new processors, accelerators, memory or interconnect components, as well as combinations thereof and full system simulation). This is a very active area of research at present with the largest systems available today being only a fraction of the size of forthcoming exascale systems.

*Supporting development of new algorithms:* benchmarking of new algorithms and comparison with existing techniques is a key tool when honing new ideas. Of particular importance is studying the effect on algorithm performance on real-world problems rather than artificial datasets. This is an area of research which is often neglected – leading to poor algorithm design. Developing targeted synthetic input datasets for benchmarking, for example by exploiting Machine Learning techniques, that are representative of real-world data can provide broader testing coverage and therefore more robust algorithm design.

*Efficiency focussed benchmarking:* supercomputing cannot ignore the climate change agenda and the power and energy usage of systems is becoming a key component of many procurements. The scale of forthcoming systems and their efficiency, particularly with regard to their specific application load, should be a key discriminator in purchasing decisions in future. Benchmarks must take into account a system's ability to efficiently complete complex and varied workloads on an extreme scale while minimising the waste of resources (including power and energy) [1].

*Fast prototyping of new systems:* making decisions with regard to the development of new hardware and software technologies can be an expensive challenge. New models of in-silico prototyping are required so that decisions on where to invest limited R&D funds can be made quickly and effectively.

*Understanding how applications perform on systems:* all the above will rely upon a detailed understanding of how applications actually perform on systems from the CPU core, cache, memory and local disk hierarchies, across interconnects to other nodes and a cluster IO and storage system. This will involve the characterisation of both applications and system components, including system software performance. It is this activity that will provide the data that will support the above activities.

All of these areas of research can benefit from a mini-app [2] approach, but research is needed to ensure that the use of such mini-apps accurately reflects the behaviour, and thus the likely performance, of full applications once they are ported to new systems and architectures.

**Computing demand:** Contributing to the TOP500 [3] list requires long HPL [4] runs on new systems, however the advent of more representative benchmarks such as HPCG [5] has lowered the full system time required for benchmarking. In percentage terms of total system availability, the requirements of performance modelling and benchmarking will be low – however, full system runs will be needed periodically.

**References**


[1] *M. Weiland & N. Johnson. Benchmarking for power consumption monitoring. Computer Science - Research and Development (2015). https://doi.org/10.1007/s00450-014-0260-1*

[2] *UK Mini-App Consortium. http://uk-mac.github.io*

[3] *TOP500. http://www.top500.org*

[4] *J. Dongarra. Performance of Various Computers Using Standard Linear Equations Software", University of Tennessee, 37996, Computer Science Technical Report Number CS - 89 – 85.*

[5] *J. Dongarra, et al. High-performance conjugate-gradient benchmark: A new metric for ranking high-performance computing systems. International Journal of High Performance Computing Applications, 2015*

## 9.3 Composable Languages and Tools Across supercomputing Applications

*Contributors: David A. Ham[1], Lawrence Mitchell[2] ([1]Imperial College; [2]Durham University)*

**Vision:** Advances in computational science will only be possible if core computational components and mathematical algorithms can be directly combined without constant reimplementation.

**Key research challenges:** Monolithic reimplementation of simulation capabilities by individual application communities is unsustainable and will impede the development of the new capabilities required to address grand challenge problems. The experience of being unable to update methods or implementations chosen many years ago is common; advancing capabilities by adding one feature per PhD is a widespread experience. In an era of rapidly increasing parallelism and algorithmic sophistication, the result will be that the UK's computational science capabilities will stagnate. Conversely, by developing composable tools which enable application communities to directly exploit mathematical and algorithmic advances without recoding, the UK will establish itself as a leader in the development of simulation science capabilities with global impact.

At a low level, composable, reusable components are standard practice: tools such as MPI, OpenMP, and OpenACC provide standardised, higher-level models of parallelism than offered by "bare metal" programming. There are isolated examples of excellence in numerical software, such as the PETSc system (Portable, Extensible Toolkit for Scientific Computation; [1]) which provides model developers and users with *programmable* access to many of the world's most advanced solvers and preconditioners for sparse linear and nonlinear problems. Similarly, the Kokkos [2] library offers a programming model that hides details of multicore or GPU parallelism. At an application level, domain specific languages like the FEniCS [3] and Firedrake [4] systems show how, for mesh-based discretisations of partial differential equations, modellers can write a high-level, mathematical description of their problem, and automatically obtain an efficient, parallel implementation. Tools such as Dolfin-adjoint [5] and DefCon [6] in turn exploit this high-level interface to enable users to solve inverse problems and systems with multiple stable states. Closer to the application domain, the Cactus [7] system facilitates code sharing across computational astrophysics. Cactus, FEniCS, and Firedrake all rely on PETSc, which in turn relies on MPI; an illustration of the composability and cross-community impact of this approach.

These isolated examples, in several of which UK researchers play leading roles, demonstrate the potential of the composable approach. They enable bidirectional, interdisciplinary work between mathematics, application areas, and computer science, by allowing each area to focus on their strengths, while exploiting the strengths of others. A key advantage over a monolithic development strategy is that composable software tools *automate, and scale up, the delivery of expertise.* A single supercomputing expert can provide optimal implementations to entire communities, rather than just a single research group.

**Implications for supercomputing research:** The UK needs a productive supercomputing application environment in which increasingly complex combinations of simulated systems, algorithms and hardware are routinely combined to achieve previously impossible grand challenges. For this to be achieved, participation in the ecosystem of composable tools needs to be seen as a necessary part of all supercomputing research. Numerical methods development is not complete when there is a proof of concept: the methods must be delivered in a widely usable tool; algorithmic advances must be delivered in the composable toolchain, not just in a demonstrator; and application development must build on the appropriate composable tools, instead of coding from scratch. Proposals for new simulation science must be judged on how they enrich and exploit this broader ecosystem, and stand-alone efforts should be rejected as poor value for money. This approach further depends on the abstractions, languages and libraries which comprise the composable

toolchains expanding in scope to span the full gamut of supercomputing algorithms and applications. Computational science research directed at achieving this universal coverage and delivering the requisite tools is therefore critical.

**References**


[1] http://www.mcs.anl.gov/petsc

[2] *Carter Edwards et al. Kokkos: Enabling manycore performance portability through polymorphic memory access patterns. J. Parallel and Distributed Computing (74) 3202-3216, 2014* doi:10.1016/j.jpdc.2014.07.003

[3] *Logg et al. Automated Solution of Differential Equations by the Finite Element Method. Springer, 2012* https://fenicsproject.org

[4] *Rathgeber et al. Firedrake: automating the finite element method by composing abstractions. ACM Trans. Math. Softw., 43(3):24:1–24:27, 2016.* doi:10.1145/2998441 https://firedrakeproject.org

[5] *Farrell et. al* Automated derivation of the adjoint of high-level transient finite element programs*, SIAM Journal on Scientific Computing 35.4, pp. C369-C393.* doi:10.1137/120873558 http://www.dolfin-adjoint.org

[6] *Farrell et. al The computation of disconnected bifurcation diagrams* arXiv:1603.00809

[7] *Goodale et. al The Cactus Framework and Toolkit: Design and Applications VECPAR'2002* http://edoc.mpg.de/3341 https://cactuscode.org


## 9.4 Working with Industry to Design the Next Generation of Supercomputing Systems


*Contributors: Mark Wilkinson[1], Peter Boyle[2], Mark Parsons[2], Jeremy Yates[3] ([1]University of Leicester; [2]University of Edinburgh, [3]University College London)*


**Vision:** *A technology foresight programme co-funded by UKRI and industry to provide UK researchers with early access to new technologies, facilitate strategic planning for software development and enhance the impact of UKRI research through sharing knowledge and skills.*

**Key research challenges:** Imminent developments in supercomputing hardware offer significant opportunities for new research breakthroughs as calculations of previously unachievable scale and complexity become routine. However, realising these opportunities presents major challenges in terms of software and algorithm development: the increased power of next generation processors will only be accessible to those codes and algorithms which can fully exploit all available levels of parallelism within the hardware. Supercomputers are scientific research instruments and a greater understanding of the underlying hardware by the research communities which exploit them drives increased research productivity. A co-design programme is an ideal mechanism to ensure that UKRI researchers can take advantage of new processor technologies to maintain their world leadership, while also increasing the numbers of UKRI researchers with the skills needed to participate in, and drive, technology transformations. However, such a programme for the UK will require international collaboration because the UK, unlike Japan, the USA or China, does not have an indigenous processor industry[18] except for niche devices which are largely in the small embedded space.

The increasingly competitive processor market also presents an opportunity to develop heterogeneous systems combining multiple architectures on a common fabric with shared storage in order to support each stage of complex workflows with the most appropriate hardware. The entrance of processors based on Arm designs into the supercomputing market offers the potential to lower the bar for chip-level co-design around specific research computing workflows, as multiple

[18] While the UK does host Arm's global headquarters, Arm is a processor design company which produces designs for manufacturers to develop into processors.

manufacturers develop their own tailored products based on the Arm instruction set. A proliferation of options using Graphics Processor Units (GPUs) is appearing, including an established eco-system based around the NVIDIA GPU, and planned systems based on AMD and Intel GPUs. For some of these technologies, the programming environments are challenging and diverse, requiring both software and algorithm development but presenting substantial opportunities for applications that are able to exploit them. The potential applications of Quantum Computing must also be considered.

In such a rapidly evolving environment, experience has shown that the returns on supercomputing investments can be maximised through co-design projects with the supercomputing industry. The White Paper *"UKRI National Supercomputing Roadmap 2019-30"* notes the value of "a rolling co-design/technology foresight programme" and that public funding to support this will "unlock co-funding from industry partners to establish proof-of-concept systems for benchmarking of novel hardware". However, as noted below, recent work has focussed on exploring new hardware technologies after they have been developed rather than co-designing them with industry as has happened in the past. In order to re-invigorate this activity, the UK will have to collaborate internationally.

Due to the long timelines involved (the Japanese co-design project which resulted in the A64FX processor, which is now used in the Fugaku system, was a decade long project) we emphasise that co-design programmes must span or go beyond the life-cycle of supercomputing infrastructures to prepare for the replacement of hardware and algorithms. Knowledge and skills will be shared in both directions: insights into new hardware developments from the supercomputing industry will drive software development and porting within the UKRI community, while next generation algorithms can precipitate important course corrections in hardware architecture roadmaps.

The UK already has a strong track record in the area of supercomputing co-design encompassing silicon-level, software-level and system-level activities, although at the silicon design level this is largely historical and in recent times the UK has focussed on software and system-level co-design. In addition to the direct research benefits from the highly productive supercomputers which now underpin the scientific outputs of UKRI researchers, this continued cycle of co-design by a number of UK facilities (e.g. DiRAC, EPCC, Hartree) has also generated direct economic benefits through intellectual property creation, commercial product development and inward investment into the UK. Examples include:

- Silicon-level chip co-design: Peter Boyle (Edinburgh/DiRAC) was a full member of the design team for the IBM BlueGene/Q (the fastest computer in the world in 2012) and 8% of the BG/Q chip was designed by UK academics, leading to 6 joint US patents with IBM in supercomputer design, the 2012 Gauss Award paper and a joint IEEE invited paper with IBM. Edinburgh currently has a joint project with the Intel Pathfinding and Architecture Group on the codesign of future Intel supercomputing systems with QCD simulation codes, a unique activity around accelerators. The focus includes the exascale design planned for installation in the US in 2021 (DOE Aurora 2021) and Edinburgh participates in the DOE Exascale Computing Project Pathforward workshops with Intel. This work has already led to a joint Intel-Boyle patent [1] on floating point representations for machine learning being filed with the US patent office.
- Acccelerating IO: Production by DiRAC, in collaboration with DELL EMC and the Universities of Cambridge and Durham, of functioning 300+TB SSD array filesystems for checkpointing (2018, Durham) and general IO access (2019, Cambridge). These have brought significant performance enhancements to large scale CFD simulations and the Cambridge deployment topped the June 2019 io500, a factor of two above the USA's Summit supercomputer.
- Next generation I/O technologies: EPCC has led the Horizon 2020 NEXTGenIO project which has designed and built a new server motherboard with Fujitsu and Intel to make use of Intel's 3D XPoint™ memory for supercomputing systems.

- Novel network topologies: DiRAC influenced multithreaded extensions to Intel's MPI, driving improved concurrency in their MPI and Omnipath software which has become part of the shipping product in Intel MPI 2019, and the subject of a joint paper, as a result of DiRAC stress testing.
- The HPE/Arm/Suse Catalyst UK [2] initiative placed three, 4000-core Arm-based clusters in UK universities (Bristol, Edinburgh, Leicester/DiRAC) to explore the potential of Arm processors for large-scale supercomputing services and support the development of the Arm software stack.

However, a lack of quick to access funding for similar activities has resulted in other opportunities being missed, such as the Hewlett Packard Enterprise (HPE) Comanche programme where UK sites were invited to take part but could not access the required $250k at short notice, in contrast to their US competitors. Recently, the ExCALIBUR project has introduced a limited programme to explore new hardware technologies and it is hoped that this programme will be expanded to give the computational science community greater access to forthcoming technologies in order to explore application portability and performance.

In the context of research outputs, examples of the benefits which can be derived from the careful interplay between algorithms and the engineering of the computer hardware include:

- RSE work on the Grid software for Lattice QCD (Boyle, Edinburgh) led to the development of a new "eigenvector"-based algorithm which both reduces statistical errors (variance reduction) and improves algorithmic efficiency of first principles calculations of the properties of hadrons by a factor of around fifty. A factor 6 speed-up in the single-node performance of the MILC code was achieved using block solvers (Yamaguchi, Edinburgh). Combined with system-level co-design of the new DiRAC Extreme Scaling system at Edinburgh, these enhancements ensure that every communication link is simultaneously operated at wire speed.
- DiRAC, Intel and the Institute for Computational Cosmology at the University of Durham co-funded the development of the open-source SWIFT code for large-scale cosmological simulations of galaxy formation. SWIFT uses task-based parallelism to achieve much greater code efficiency than comparable codes and will support UK leadership in the performance of more physically realistic calculations in many areas of cosmology and astrophysics.
- As part of the NEXTGenIO project, ECMWF with EPCC have explored the use of 3D XPoint™ memory in their IFS (Integrated Forecasting System) application's I/O. The use of this memory has transformed the performance of their complex weather forecasting dataflow.
- Increased flexibility to procure more cost-effective and energy-efficient hardware without jeopardising research productivity, enabling the deployment of more powerful services.

We therefore propose that the UK should develop or join one or more international collaborations (for example with Japanese or American partners) to co-design future hardware and software technologies. Particularly with regard to hardware co-design, it is understood this will be a long-term activity and must be a broad community effort such that the needs of the UK's modelling and simulation communities are properly reflected in the hardware advancements that are developed.

**Computing demand:** The computing requirements for this activity are modest. On average, for close-to-market technologies one or two production-scale test systems would be deployed per annum, along with several smaller proof-of-concept systems for less developed technologies. For longer-term challenges, for example the development of a new processor, more significant funding would potentially be required. This model of long-term and short-term activities will provide a natural route into co-design for PhD students and other early career researchers, who can experience the research benefits from the concurrent evolution of hardware, algorithms and software and potentially become involved more directly in technology development in the future.

Researchers are often understandably reluctant to invest time in code development to take advantage of new hardware which may not be available at scale for several years or may even prove not to have longevity: scientific leadership demands that research funding is used to deliver science

results now. By offering novel technologies at production scale, including Tier-2 scale provision when appropriate, and funded RSE effort to support the development work, the potential for immediate science benefits will provide an incentive for more communities to plan for software enhancements in line with technology roadmaps and will ensure that when new technology is incorporated into the UKRI supercomputing infrastructure, there are workflows ready to exploit it. Across UKRI, this will drive software and algorithm innovation and aid with the identification of those technologies which are most appropriate and cost effective for particular science workflows, providing opportunities for international leadership.

**References**


[1] *See USPTO.gov: Application No.: 20190042544; Title: "FP16-S7E8 MIXED PRECISION FOR DEEP LEARNING AND OTHER ALGORITHMS".*

[2] *https://www.top500.org/news/hpe-will-supply-arm-powered-hpc-systems-to-three-uk-universities*


## 9.5 Next Generation Development Cycle


*Contributors: Beth Wingate[1], Mark Parsons[2], Michèle Weiland[2], Jeremy Yates[3] ([1]University of Exeter; [2]University of Edinburgh, [3]University College London)*


**Vision:** Co-creating a flexible, community-driven development cycle to support the evolution of new algorithms and methods into advanced scientific software for the purpose of 1) accelerating scientific discovery and 2) participating in the design of new efficient computing algorithms and technologies, as well as new technological skill sets to attract industry to the UK.

**Key research challenges:** Success in developing new, agile ways of working together will propel the country forward in scientific computing in this important time in history. As preparations for the exascale era are continuing, the concept of co-design, which seeks to bring together hardware and system software developers, computational scientists, applications experts, research software engineers and visualisation experts to jointly develop Exascale hardware and software, is becoming increasingly important. Building on lessons learned from early examples of co-design we have identified new ***co-design cycles***, taking place on different time scales and at different levels of disruption of the status quo, that can benefit from national cross-cutting communities. These cycles, which at their heart are an infrastructure of people, include (but are not limited to) activities as follows:

1. Short time cycles (1-2 years)
    - Systematic profiling of codes to identify potential areas of incremental change.
    - Community workshops and Special Interest Groups to discuss algorithms which underpin key science codes/areas and explore potential alternatives.
    - Hackathon programmes to accelerate code porting projects bringing users, code owners and optimisation experts together.
    - Proofs of concept of disruptive change, with identified challenges that evolve as we advance.
    - Systematic analysis of common mathematical concepts and structures that can be advanced by novel algorithms and which are shared by many numerical methods and applications.
2. Medium time scales (2-5 years)
    - Mathematical and algorithmic advances identified earlier in the development cycle, but requiring significant and potentially disruptive design and implementation efforts.
    - Prototyping key cross-domain software libraries and domain specific languages.

- Designing novel hardware and software architectures driven by requirements from application challenges.

3. Longer term evolution (5- years)
    - Interaction and interoperability of libraries and APIs with different computing architectures.
    - Building robust key cross-domain software libraries and domain specific languages.
    - Co-designing disruptive novel architectures as a basis for new hardware platforms.

All of this work will require close alignment between academic leaders and industry partners. Given the power requirements at the exascale, a key component in all of this work will be co-design for efficiency. At any one time short, medium and long-term cycles of development will be progressing concurrently.

A key challenge is to develop the community of supercomputing researchers, Research Software Engineers, applied mathematicians and algorithm developers to focus on working on these challenges in a sustained way.

To this end the creation of a UKRI Collaborative Computational Project, including computational scientists, computational mathematicians and ICT researchers to enable a joint technology development programme that develops algorithms, hardware and software, was recommended by the UKRI Supercomputing White Paper. This programme, consisting of multiple projects, each working on specific challenges, should be the main co-ordinator of research and technical development activities and disseminator of results such that all communities benefit from developments which may have broad applicability. Different scales and timelines must considered within the portfolio of individual projects.

**Computing demand:** The UK community is already participating in such activities, for example the EPSRC-funded ASiMoV (Advanced Simulation and Modelling of Virtual Systems) Prosperity Partnership project [1]. This project, which will develop the world's first high fidelity "digital twin" model of a Rolls Royce gas turbine, will need exascale levels of performance to deliver the necessary detailed accuracy for virtual certification of new gas turbine engine designs. A full simulation is expected to need a significant part of an exascale system for up to a month. It can only succeed by bringing together a co-design community to work on the problem.

However, much co-design work will require small run times but potentially on large proportions of the largest supercomputing systems available.

**References**

[1] *See https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/S005072/1.*

## 9.6 New Research Excellence Framework (REF) Unit of Assessment for Computational Science

*Contributors: Beth Wingate[1], Neil Chue Hong[2], Jeremy Yates ([1]University of Exeter; [2]Software Sustainability Institute / University of Edinburgh, [3]University College London)*

**Vision:** A REF Unit of Assessment in Computational Science which includes supercomputing research, Research Software Engineering (RSE) developments, computational mathematics, co-design of new computational platforms, and computational science across its many domains, will grow today's group of highly skilled researchers to ensure the UK remains at the forefront of computational science aligned with the Government's Industrial Strategy.

**Key research challenges:** The Research Excellence Framework (REF) is the UK Government's system for assessing the quality of research in UK Higher Education Institutions. It reports once every 5 years. Computational science challenges drive the development of supercomputing and AI hardware and software technologies worldwide. The UK has a strong history of leadership in computer science and computational science. However, this latter category has been falling behind in recent years because there have been limited opportunities to publish research software engineering papers. This position is now improving through journals such as the Journal of Open Research Software, and astronomy journals such as Monthly Notices of the Royal Astronomical Society are becoming open to the publication of papers on the design of supercomputing systems, analogous to the instrumentation papers associated with large-scale observational projects. In order to recognise this key area of research creativity, this challenge seeks to grow the domain by including a new REF Unit of Assessment for Computational Science in future REF submissions. The expectation is that this new Unit of Assessment will sit in Main Panel B but be separate and clearly distinguished from the existing Computer Science and Informatics UoA.
This will enable the UK to build on its history of innovation and excellence in this area, by recognising the people that drive it. It will encourage computational scientists and research software engineers to tackle Grand Challenge problems and ensure the UK reaffirms its strength in this important area which spans all scientific and industrial research disciplines.

However, for the current REF cycle, efforts must be made to ensure that this kind of work is recognised via existing criteria, such as innovation and impact, which are of increased importance compared to previous REF cycles. In addition, efforts could be made to identify those academic journals which do now give credit to this type of research, e.g. the Supercomputing Frontiers and Innovations Journal – an international open access journal focussing on "research and development findings and results at the leading edge of supercomputing systems, highly parallel methods, and extreme scaled applications".

# 10 Conclusions

The United Kingdom has always been a world leader in computational science which is predominately enabled today by supercomputing. Over the past 40 years, supercomputing has become so fundamental to scientific understanding that, in many fields of research, the delivery of world class insight and innovation is entirely dependent on it. The world is now entering the 'era of exascale computing' where computer technologies are evolving quickly and new uses of these technologies could enable dramatic new scientific discoveries, provide understanding for social change and economic growth, and even allow the potential of foreknowledge of our changing environment. *But it is not only these new research discoveries alone that are important.* The skill sets developed by the activity of participating in next-generation supercomputing, and the choices we make about future computer architectures and their applications, will themselves feed into the direction the technology itself takes. Our active participation in supercomputing in this rapidly evolving era of exascale computing will ensure that we maximise the scientific, societal and economic benefits delivered by supercomputing over the next decade. All of the four initial Grand Challenges[19] in the UK government Industrial Strategy require supercomputing resources for their success. However, as Table 1 shows, the UK is lagging behind its competitors in the provision of large-scale supercomputing services.

This document has presented a cross-section of the research goals of the UKRI research community across seven broad themes. In summary, these themes have shown:

1. **Expanding the frontiers of fundamental sciences:** increased supercomputing infrastructure will enable the UK community to deliver the first simulations of the observable Universe capable of fully resolving Milky Way-like galaxies and to uncover new physics from experimental data collected by facilities such as the Large Hadron Collider.
2. **Climate, weather and earth sciences:** next generation supercomputing will increase our understanding of climate processes, support industry and society in adapting to the changing climate, and improve forecasting of geological hazards such as earthquakes.
3. **Computational biology:** the data driven nature of this science means that any supercomputing strategy in this area must be well coordinated with the data intensive compute strategy of UKRI. As well as compute coupled to the data, as required to exploit genomic data at scale for insights into health, there are also pockets of work which require large scale compute resource for either complex simulations of proteins and other molecules involved in living processes or applications of compute intensive machine learning methods to understand biological systems.
4. **Computational biomedicine:** the 'grand plan' for in silico medicine is the creation of a complete mathematical representation of human physiology that encompasses the entirety of the anatomy from genome to organism and permits the simulation of any combination of physiological and pathological processes. As modelling becomes universal, demand for computing resources will rise exponentially in order to support the development and optimisation of personalised healthcare solutions for clinical deployment.

   **Engineering and materials:** supercomputing is an indispensable tool for future research in engineering and material science and future scientific breakthroughs will increasingly come from simulations enabled by it. Simulations of new thermoelectric and photovoltaic materials are key to the development of alternative energy generation methods, along with the design of fusion reactors. The advent of digital twins for products from cars to gas turbine engines will pave the

[19] The Industrial Strategy White Paper identified four initial grant challenges: Artificial Intelligence and data; Ageing society; Clean growth; Future of mobility (see https://www.gov.uk/government/publications/industrial-strategy-the-grand-challenges/industrial-strategy-the-grand-challenges for more details).

*Table 1: Comparison between the largest supercomputers available to UK researchers with those of their international competitors based on data from the June 2020 Top 500 list[20].*

| Country | Max PF | System Name | No of systems above 5 PF | No of systems above 10 PF |
|---|---|---|---|---|
| Japan | 416 | Fugaku | 7 | 3 |
| USA | 149 | DOE ORNL Summit | 15 | 7 |
| China | 93 | Sunway TaihuLight | 2 | 2 |
| Italy | 22 | Marconi | 2 | 2 |
| Switzerland | 21 | Piz Daint | 1 | 1 |
| Germany | 19 | SuperMUC-NG | 5 | 1 |
| S. Korea | 14 | Nurion | 1 | 1 |
| France | 12 | Tera-1000-2 | 3 | 1 |
| Australia | 9 | Gadi | 1 | 0 |
| Taiwan | 9 | Taiwania 2 | 1 | 0 |
| UK | 7 | Met Office | 1 | 0 |
| Spain | 6.5 | Mare Nostrum | 1 | 0 |
| Saudi Arabia | 5.5 | Shaheen II | 1 | 0 |
| Finland | 5 | MAHTI | 1 | 0 |
| Norway | 4.4 | Betzy | 0 | 0 |
| India | 3.8 | Pratyush | 0 | 0 |
| Canada | 3.6 | Niagara | 0 | 0 |
| Sweden | 2.9 | Tetralith | 0 | 0 |
| Austria | 2.7 | VSC-4 | 0 | 0 |
| Russia | 2.4 | Lomonosov 2 | 0 | 0 |
| UK | 2.3 | Cumulus (STFC DiRAC/EPSRC Tier 2/ Cambridge) | 1 | 0 |
| Brazil | 1.8 | Santos Dumont | 0 | 0 |
| UK | 1.8 | Scafell Pike (STFC Hartree) | 1 | 0 |
| Poland | 1.6 | Prometheus | 0 | 0 |
| UK | 1.6 | Archer (EPSRC EPCC) | 1 | 0 |

way for virtual design and certification, reducing both costs of innovation and lead times to production.

5. **Digital humanities and social sciences:** exponential growth is occurring in the quantity and range of data recording human life, behaviour and society. Greater supercomputing power and

[20] https://www.top500.org/lists/top500/list/2020/06/

the new supply of data will deliver step changes in the computational handling of music, language processing for robotics and decision-making in public policy.

6. **Mathematics and science of computation:** the future of modelling and simulation requires new mathematical thinking in the context of the computing technologies that are emerging as we approach the exascale era. In the past, users of computer systems focussed on science ***by*** computation. To develop new software technologies and algorithms to support the next generation of research results we must also focus on mathematics and the science ***of*** computation, in collaboration with international partners.

Each of these domains has presented world-class research challenges that require supercomputing and associated data science technologies to solve them. Detailed assessments of future needs have been presented, each showing how incremental improvements (of around 10X today's resource) could have an immediate effect on the domain and how previously unattainable results will be possible as we enter the exascale age (around 100X today's resource).

By 2023, the USA, Japan, China and all European Union Member States and Associated Countries will have access to exascale supercomputing resources. This will not simply allow these countries' research communities (from academia and industry) to do more of the same more quickly, it will allow these researchers to tackle research challenges using supercomputers that have previously been impossible. During this period of great changes in computing technologies, these countries will be the ones that have the most potential for prosperity. They will emerge from the next decade as the leaders of these technologies and the leaders in these research areas.

The UK Research & Innovation Science Case for Supercomputing has sought to show how UK researchers both in academia and industry would benefit from long-term, increased investment in supercomputing, and e-Infrastructure in general, in order to maintain our position as a global leader in science and innovation. The combined scientific goals of the UKRI computational research communities require a significant uplift in UK supercomputing resources and such investment would realise significant scientific and societal benefits. The case is clear.