ROYAL SOCIETY OF CHEMISTRY

DISCOVERY

Digital futures

A new frontier for science exploration and invention

Introducing our new perspectives series

In a world where global challenges and advances in technology bring both uncertainty and new possibilities, the chemical sciences have a critical role to play. But what will that role be? How can we maximise the impact we make across academia, industry, government and education? And what actions should we take to create a stronger, more vibrant culture for research that helps enable new discoveries?

Our perspectives series addresses these questions through four lenses: talent, discovery, sustainability and knowledge. Drawing together insights and sharp opinion, our goal is to increase understanding and inform debate – putting the chemical sciences at the heart of the big issues the world is facing.

Discovery

Chemistry is core to advances across every facet of human life. But where do the greatest opportunities lie? How will technology shape the science we create? And what steps should we take to ensure that curiosity-driven research continues to unlock new opportunities in unexpected ways?

Sustainability

Our planet faces critical challenges – from plastics polluting the oceans, to the urgent need to find more sustainable resources. But where will new solutions come from? How can we achieve global collaboration to address the big issues? And where can the chemical sciences deliver the biggest impacts?

Talent

Talent is the lifeblood of the chemical sciences. But how do we inspire, nurture, promote and protect it? Where will we find the chemical scientists of the future? And what action is required to ensure we give everyone the greatest opportunity to make a positive difference?

Knowledge

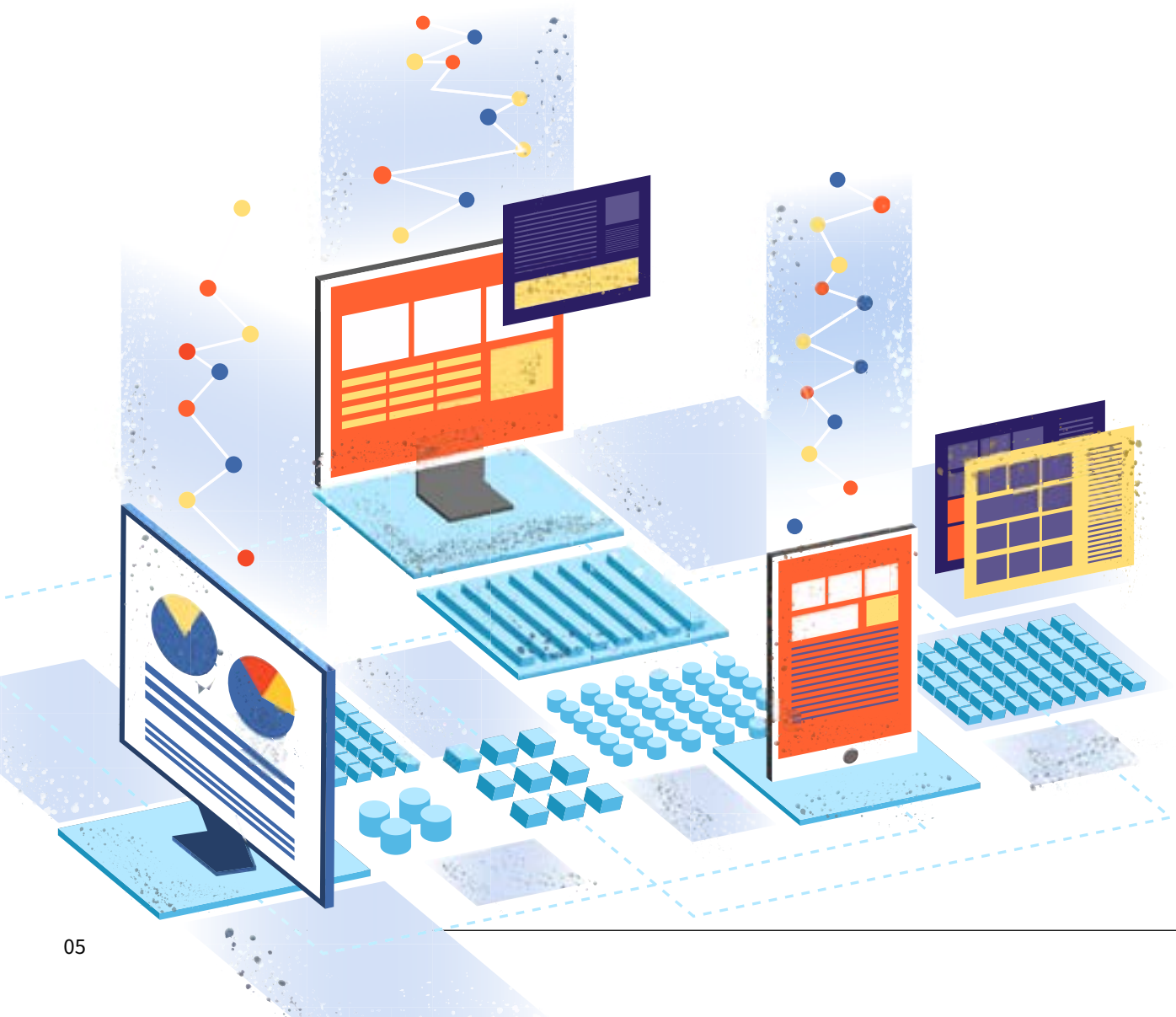
Around the world research fuels scientific progress but the way we are sharing new knowledge is changing. What are the big challenges of the digital era? How can open access become a global endeavour? And what do chemical science researchers really think about the constantly evolving landscape?

Find out more at **www.rsc.org/new-perspectives**

Contents

Key findings

1. Digitisation is a global ‘megatrend’ impacting society at every level from individuals and households through to large companies and national systems.
2. ‘Digital’ encompasses many types of data and of technologies that generate, interpret and act on it.
3. Digital technologies have huge potential in chemistry-using industry sectors. They will increase efficiency and sustainability across the chain from sourcing raw materials, to product development and manufacturing, to distribution, consumption and end of product life.
4. Scientists in universities, research institutes and companies are harnessing multiple digital technologies in R&D. The pace of development and adoption of digital tools is increasing and includes:
    - Computational modelling and simulation
    - Imaging & visualisation
    - Machine learning (ML) and artificial intelligence (AI)
    - Advanced measurement and sensing
    - Robots and automated systems.



5. Harnessing digital technologies for science R&D will enable scientists to deliver new benefits for society and the economy faster. Using digital technologies in R&D will enable:
    - Faster, cheaper, safer, more efficient innovation
    - More effective diagnosis, prevention and treatment of disease
    - New sustainable technologies, from better batteries and solar cells to next generation plastics and resource efficient industrial processes
    - Environmental decision-making and regulation informed by high quality data and analysis from multiple sources
    - Breakthroughs and new knowledge in physical, life and digital sciences.
6. Digital technologies will enable and challenge human scientists to go faster and to think at a higher level. They will extend human ambition and creativity, enabling multidisciplinary teams to solve bigger problems.
7. For the foreseeable future human input and supervision will be essential in harnessing data and digital tools for scientific discovery in a way that is efficient, effective and ethical.
8. The fundamental elements of the scientific method will not change, but digital technologies will transform each step in it and, crucially, the links between steps like:
    - Developing a question or goal
    - Generating a hypothesis and making predictions
    - Experimentation, observation and measurement
    - Interpreting data and drawing conclusions
    - Identifying avenues for further investigation and application.
9. Rather than ‘using digital for digital’s sake’, the optimal combination of digital techniques to develop and use depends on the research question, target application and wider societal and economic context.
10. The volume of scientific data and the sophistication of techniques to collect and interpret it will continue to increase. There are huge opportunities to harness this data to bring new insights, make discoveries and inform decisions. There are also significant challenges in sharing and in rigorously interpreting data.
11. To harness the chemistry-digital interface we need new digital skills, roles and careers in science discovery. New multidisciplinary collaborations, communities and capabilities will also be crucial.

12. Leadership and strategic vision, combined with insights from active researchers, will be key to ensuring we seize the opportunities at the chemistry-digital frontier, considering dimensions like:

    - *Breakthroughs and disruption:* What are areas that need urgent disruption and where digital capabilities will accelerate breakthroughs?
    - *Complexity:* What are areas where conventional techniques simply cannot handle the levels of complexity involved?
    - *Transdisciplinary problems:* What are the challenges that are too big for one lab, company or discipline to solve?
    - *Key enabling facilities:* What technologies, platforms and capabilities could have a transformative impact across multiple science discovery and application areas?

Examples of possible focus areas are:

- Sustainable energy e.g. next generation batteries or breakthrough catalysts to enable low carbon fuel production.
- New medicines & diagnostics to enable prevention, early detection and treatment of everything from bacterial infections, tropical and emerging diseases to cancer, dementia and obesity.
- Predicting and reacting to environmental impacts using multiple, distributed, real-time sensing systems and models.
- Tackling the 'plastics problem' will involve science and technology innovation challenges including new ways of designing and making polymer building blocks and plastic recycling behaviours and technologies.
- Key enabling platforms combining infrastructure and expertise in areas like automation of synthesis or formulation, high performance computing, modelling, data-sharing and advanced data analysis or measurement. These platforms can be connected to enable transfer of data and samples from one to another, and will underpin advances in multiple challenge and discovery areas.

13. There are many opportunities for everyone to push forward the chemistry-digital interface for the benefit of society: for individuals; for the chemistry community in partnership with other communities in the physical, life and digital sciences; as well as for research and teaching institutions, companies, funders and governments.

Key areas for action include:

- Lifelong training in digital skills.
- Roles and career progression for digital experts in research outside digital industries.
- Fostering multidisciplinary collaborations and communities.
- Supporting and enabling data sharing.
- Leadership and advocacy for the digital futures of science R&D.

1

Introduction

Digitisation – a global 'megatrend'

'Digitisation' or 'digitalisation' is a major political, economic and public focus, with significant investments and media attention around the world. It is a strategic priority for governments, industry and institutions globally. Discussions often concentrate on AI and robotics, including opportunities for growth and prosperity as well as concerns about ethical issues and threats to skills and employability.

Digitisation in fact involves many different types of technologies and concepts, which overlap and are interdependent. It includes robotics, sensors, high performance computing, 3D printing and wireless communications systems, along with data mining and supervised or unsupervised machine learning (roughly speaking what is often referred to as AI).

There are cross-cutting advances in algorithms, optimisation, modelling, simulation, visualisation, cryptography, as well as protocols and standards enabling interoperability of systems. There are concepts that leverage, integrate and enrich technologies and methods e.g. blockchain, data streaming, digital twins, cloud computing, and the Internet of Things.

Digitisation permeates many layers across society and the economy, in addition to the digital or IT sector itself. Digital technologies have implications and bring opportunities for industries from finance, law and logistics to pharmaceuticals, materials and energy, as well as for private and public systems across healthcare, security, regulation, transport and education. The paradigm characterised by this fusion between physical, biological and digital 'megatrends' has been called the *Fourth Industrial Revolution* or *Industry 4.0*[1].

More specifically in chemistry-using industry sectors, from chemicals, materials, pharmaceuticals, biotech and agri-tech to energy, automotive, aerospace, IT and consumer products, there are opportunities under the 'digitisation' umbrella across the spectrum from raw materials sourcing, supply chains, product development, R&D, manufacturing, logistics and distribution, sales and consumer testing.[2] There are also opportunities to link these different dimensions in end-to-end or closed loop systems.

[1] *The Fourth Industrial Revolution*, Klaus Schwab, Penguin (2016)

[2] *Strategic Innovation and Research Agenda*, SusChem (2019)
www.suschem.org/press-area/press-releases/suschem-publishes-new-strategic-innovation-and-research-agenda

Strategic Advisory Forum & participants

In our Science Horizons project, we engaged with over 700 academic researchers globally to seek views on key trends and emerging research areas in the chemical sciences and its interfaces.[3]

Scientists have always developed and adopted new techniques, but we heard from researchers a new sense of excitement about the range of techniques that have come online in recent decades, the pace at which they are evolving and converging, and the quantity of data and potential insight they bring. These techniques range from advanced measurement and computational modelling to AI and robotics.

We also heard a degree of scepticism about the extent to which big data, AI and robotics will be truly transformative in scientific discovery. This is set against a backdrop of high levels of public and private sector investment in digitisation globally as well as significant public concern about jobs and ethics.

In our first *Strategic Advisory Forum* we set out to gain a more in-depth understanding of the long-term promise of and concerns about the use of data and digital technologies for scientific discovery. We invited experts from different scientific fields and sectors to discuss and set out a *Digital Futures* vision.

[3] *Science Horizons*, Royal Society of Chemistry (2019) www.rsc.org/new-perspectives/discovery/science-horizons

The participants in the Forum were:

- **Prof Varinder Aggarwal** Professor of Synthetic Chemistry, University of Bristol
- **Dr Niklas Blomberg** Director, ELIXIR
- **Prof Muffy Calder** Professor of Formal Methods, University of Glasgow
- **Prof Andy Cooper** Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory
- **Prof Charlotte Deane** Professor of Structural Bioinformatics, University of Oxford
- **Dr Martin Jones** Deputy Head of Microscopy Prototyping, The Francis Crick Institute
- **Prof Jacqueline McGlade** Professor of Sustainable Development and Resilience, University College London, and Professor of Public Policy and Governance, Strathmore University
- **Prof Kristin Persson** Professor in Materials Science and Engineering, UC Berkeley
- **Dr Edward Pyzer-Knapp** Research Lead, Machine Learning and Artificial Intelligence, IBM Research UK
- **Dr Elizabeth Rowsell** Corporate R&D Director, Johnson Matthey
- **Dr James Weatherall** Vice President, Data Science & AI, AstraZeneca
- **Dr Horst Weiss** Vice President, Knowledge Innovation, BASF SE
- **Dr Chris White** President, NEC Labs America
- **Prof Sophia Yaliraki** Professor of Theoretical Chemistry, Imperial College London

This vision set out in this white paper will inform long-term thinking by the Royal Society of Chemistry (RSC) Leadership Team and Board of Trustees, and we hope will be useful for other individuals and organisations considering the potential of data and digital technologies in the context of scientific discovery and application.

We held the Strategic Advisory Forum in Burlington House, London on 16 September 2019, moderated by Greg Foot, Science Presenter and Producer, and held under the Chatham House Rule. This white paper is based on discussions at the Forum as well as pre-interviews with participants, but does not necessarily reflect the views of individual participants. The RSC is also grateful to Prof Lee Cronin, University of Glasgow and Dr Stefan Platz, AstraZeneca for interview insights. This white paper written by Dr Deirdre Black and Dr Wendy Niu at the RSC, which takes responsibility for any errors or omissions.

The images used in this white paper are taken from the accompanying video, which can accessed at www.rsc.org/new-perspectives/discovery/digital-futures

2

Societal & economic benefits of digital technologies in science R&D

This white paper focusses on digitisation in science R&D, identifying opportunities across the spectrum from fundamental discovery research to science targeting real-world applications. Digitisation in this context is relevant for universities, research institutes and R&D-intensive SMEs or large multinational companies, as well as for science-informed policy.

It is important to cultivate and harness the interfaces between physical, life and digital sciences because of the potential to bring all three to a new level of discovery and impact. The *Digital Futures* discussion explored a future for scientific discovery and application in which natural scientists have harnessed diverse digital technologies to enable, accelerate and extend what they do today. This will build into a virtuous circle in which the natural and digital domains propel one another forward, generating important new research directions in both.

Data and digital technologies will enable scientists to answer questions and find solutions faster and more efficiently. They will enable scientists to tackle bigger problems, to think at a higher level and to uncover possibilities that humans alone would not have found. Key benefits include:

- Faster, more efficient science discovery and innovation
- New molecules and materials for energy, environment and health
- Better diagnostics and decision-making for environment and health
- Smart and resource efficient manufacturing

In the context of science R&D, digital technologies include sensors and other measurement instruments, modelling and simulation techniques, data sharing and knowledge management systems, machine learning, data mining, visualisation, robots and automated systems. There are further underpinning areas like algorithms, optimisation methods, abstractions and representations, mathematical modelling and theory, digitally enabled closed-loop systems, hardware like high performance computers and storage, software and graphical user interfaces (GUIs).

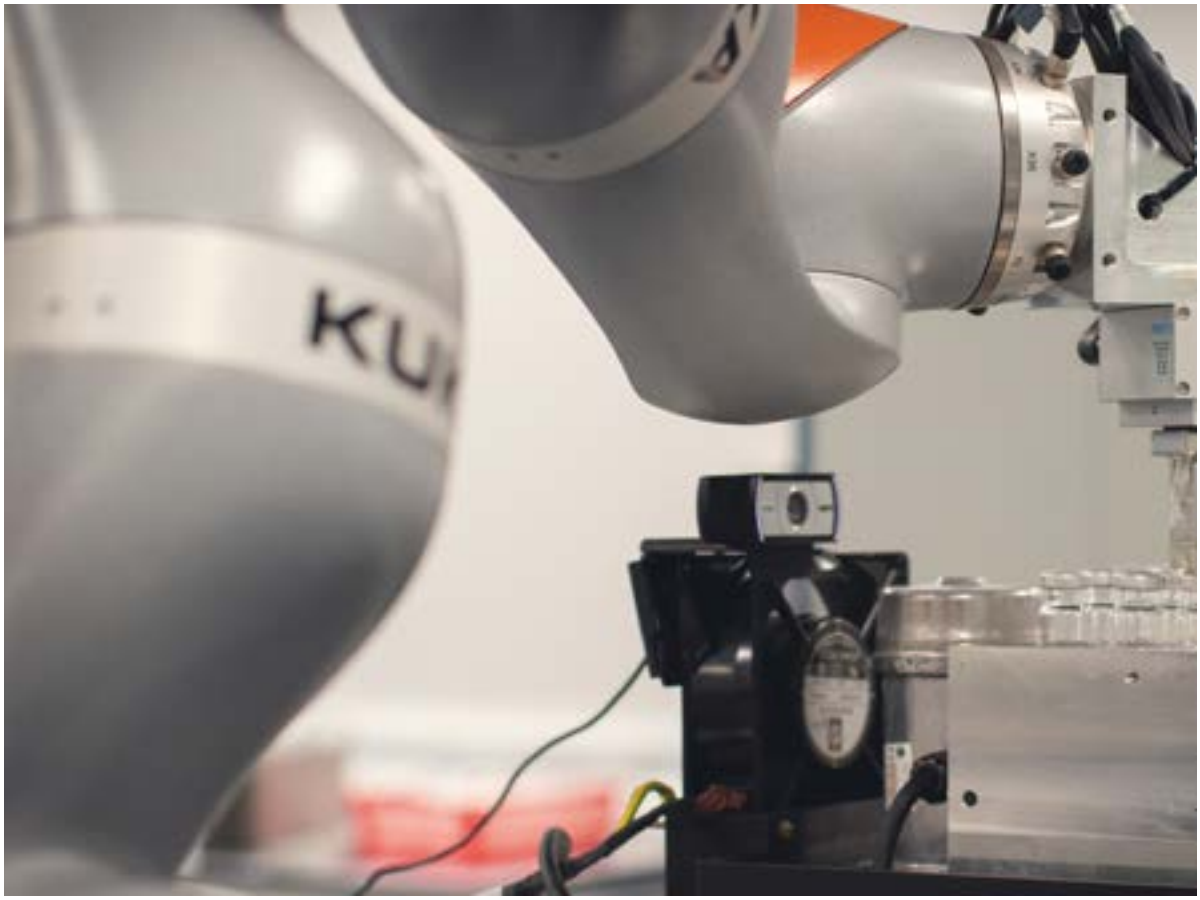
What to prioritise and where to focus effort, so that 'digitisation' of scientific discovery will have the most impact, is context dependent. It depends on the scientific question and the target application as well as on capabilities and priorities in research and innovation systems on local, national and international scales. We cover this in Sections 5 and 6.

Faster, more efficient science discovery and invention

Digital technologies will enable scientists to discover and innovate faster, making discovery and invention more efficient, reproducible and safer. This will bring benefits for multiple chemistry-using industry sectors – from chemicals, materials, pharmaceuticals and biotechnology to energy, automotive and aerospace – as well as policy, government and the digital technology sector.

For example:

- **Increasing the odds of finding new technologies quickly:** Especially in areas like energy, environment and health where there are urgent challenges and a need for breakthroughs and disruptive technologies.
- **Reducing cost and risk:** Saving time and money that would have been spent pursuing options that look promising but ultimately are not safe, practical or commercially viable. This will shave years off the time spent ruling out leads that are ultimately dead-ends in everything from new medicines to new batteries.
- **Reducing duplication** by ensuring scientists have access to all relevant knowledge generated in the past and being generated now.
- **Freeing up human time and capability** to work on higher-level creative thinking by using digital technologies to do repetitive tasks. Robots are already able to do certain kinds of physical experiments thousands of times faster than humans, and computers can hold and process much more information than a human.
- **Expanding exploration:** Finding patterns and structures that human beings alone might not see or have time to reach – whether that is investigating all of the data generated by modern imaging instruments or exploring more of the possible molecular or materials structures that could physically exist.
- **New digital innovations:** As multidisciplinary groups tackle big challenges involving chemistry this will also push the frontiers of R&D in digital areas like machine learning, robotics, modelling and computer science.



New molecules and materials for energy, environment and health

New molecules and materials are essential for everything from next generation medicines to sustainable energy materials to biodegradable or recyclable plastics. Digitalisation will transform chemists' ability to design, make and modify materials and molecules to meet societal challenges.

For example:

- **Health:** Next generation therapeutics including new medicines and vaccines; new materials and formulations for drug delivery, wound healing, bone and tissue regeneration. These new therapeutics will be developed within the broader paradigms of personalised and systems medicine (see below).
- **Energy:** Materials for technologies like next generation solar panels, batteries and thermoelectrics. Targeting cheaper and more efficient technologies that have longer lifetimes in real-world conditions, use fewer rare earth or toxic elements and are more easily recycled.
- **Environment:** Materials for environmental applications from water and air purification to recyclable or biodegradable polymers. Non-toxic catalysts made from abundant elements to enable energy efficient transport and industrial processes.
- **Disruptive technologies:** Targeting breakthroughs in fuel and raw material production (solar fuels, hydrogen production, CO_2 capture, storage and use), water harvesting, nitrogen production, new data storage and display technologies.

Combining tools like computational modelling, data mining, AI, robotics and advanced measurement, scientists will predict, screen, discover and test new molecules and materials faster and more efficiently than ever before. They will also find molecules and materials that would never have been discovered without these tools.

Enhanced diagnostics and decision-making for environment & health

Advanced sensors and sensor networks, combined with data streaming, modelling and visualisation, will enable monitoring and decision-making in multiple contexts, for example:

Environment

- **Precision environmental chemistry:** Understanding how any combination of pollutants will interact with the local environment and how they will travel around the planet in air and ocean systems. Using a suite of digital technologies including modelling to understand complex molecular interactions that take place in the atmosphere, soil and oceans as well as sensors that can detect and monitor with high specificity in real-world conditions.
- **Modelling and visualisation to enable environmental policy:** Enabling science-informed policy by using tools to represent predictions of climate models, flows of waste or the spread of disease. This can include simulating scenarios to predict impacts and risks associated with policy interventions, in order to inform decisions and the development of new policy and regulation.
- **Keeping and sharing records to ensure regulatory compliance:** Increasing the effectiveness of monitoring and regulatory compliance, using data and digital tools enables companies, governments and agencies to track and share data in standardised formats across the chain from suppliers and manufacturing to consumers and end of life.
- **Maximising agricultural productivity and minimising food waste:** Using sensors to detect the ripeness of a crop and therefore inform decisions about when to harvest a crop, deliver or consume a food product, and whether to apply pesticides in a particular section of agricultural land.

Health

- **Precision and systems medicine:** Using a range of diagnostic and measurement tools like imaging, spectroscopy and genomics, researchers will harness insights across multiple scales – from understanding the molecular origins of disease or of the interactions within cells to global patterns in patient populations – to develop prevention and treatment strategies with minimal side-effects for individual patients.
- **Enabling clinical trials for rare diseases:** Remote sensors and diagnostic tools that patients use at home make it possible to trial new therapeutics for patients who are widely geographically distributed or where it is not possible for them to come regularly to a clinic for monitoring.
- **Monitoring and prompting adherence to medical treatments:** For conditions such as asthma, monitoring the frequency with which a patient is taking medicine, reminding a person to take medicine and also picking up signs that someone may need to take more medication.
- **Personal well-being and preventative medicine:** Enabling people to make informed decisions based on data about everything from local levels of pollutants and content of their food to personal exercise levels or local weather conditions.
- **Worker safety:** Monitoring conditions in factories and plants across multiple sectors, including for corrosion, leaks and faults as well as for predictive maintenance.

Smart and resource efficient manufacturing

Digital technologies enable cost-saving and efficiency in R&D and scale-up of technologies from laboratory to plant. Some examples are:

Optimising industrial processes: Industrial processes involve a complex interplay between many different physical and chemical factors. They also depend on local conditions and equipment. It is challenging to optimise processes so that they yield as much product as possible in a way that is safe for workers, has minimal impact on the local environment and uses minimal energy. There are also overarching decisions about increasing the efficiency, safety and lifespan of existing assets as well as designing and building new plants.

Comprehensive data collection across different elements of a process or plant can provide a real-time view of a system. Combined with data analysis and modelling techniques, chemists in industry will be able to better predict the impacts of potential changes to any system, enabling more systematic and agile optimisation decisions.

Resource efficient manufacturing and products: Digital technologies enable companies to track energy and materials use across the whole pipeline from raw materials extraction to manufacturing and distribution through to the end of a product's life, for example:

- **Life cycle thinking:** Acquiring the data needed to do product life cycle assessments, and using this data to make decisions about suppliers, manufacturing processes and product design.
- **Advanced measurement and sensing techniques:** From sensors that detect waste molecules or side products with high sensitivity to devices that monitor energy use in a plant or in a product.
- **Products and processes with lower environmental footprint:** For example, new processes that run at lower temperature or use more sustainable catalysts or solvents, products made from abundant and non-toxic raw materials, or products that enable more efficient use of energy and materials with lower environmental impact.
- **Centralisation and decentralisation:** Everything from having a smaller number of specialised testing facilities and optimising which samples or processes are tested where, to having decentralised or on-demand manufacturing to minimise waste and distribution footprints.

3

People, machines and scientific discovery

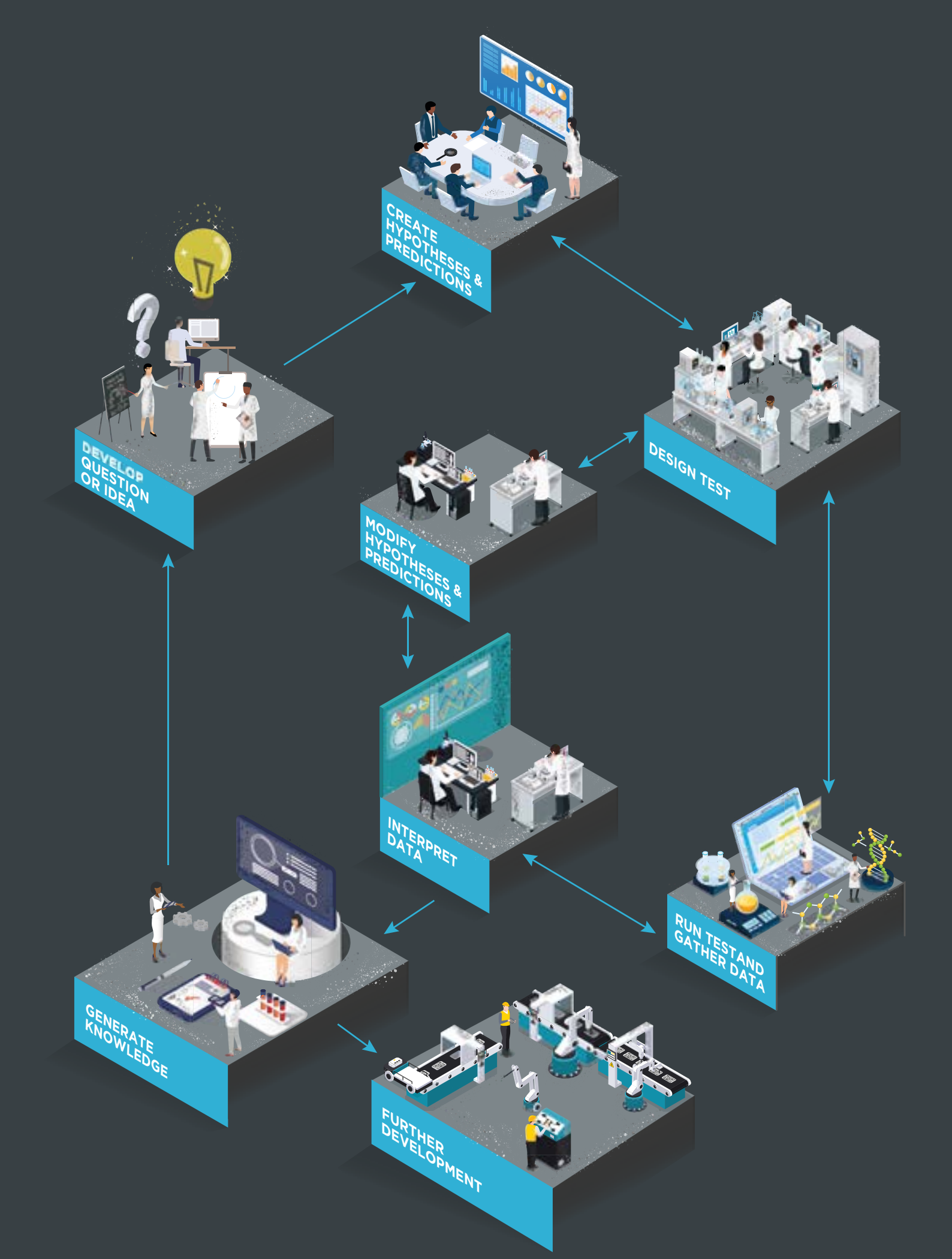
Digitally extending and augmenting science discovery

Digital technologies have the potential to enhance every aspect of the scientific process. The fundamental elements of the scientific method – developing a question or goal, generating a hypothesis and making predictions, experimentation and observation, measurement and interpreting data, and drawing conclusions or identifying avenues for further investigation – will not change, but digital technologies will transform each of these steps and, crucially, the links between them.[4]

> **What will be really disruptive is how digital tools will push us to develop new or different ideas and hypotheses, they will amplify human creativity, identifying new possibilities or opportunities.**
>
> **Dr Elizabeth Rowsell,** Corporate R&D Director, Johnson Matthey

> **The key thing that digital will enhance in science is the connectivity between steps, making best use of all existing knowledge so that you go maximally informed into each stage in the discovery, development and innovation process, storing and sharing knowledge each time you do new experiments.**
>
> **Dr Horst Weiss,** Vice President, Knowledge Innovation, BASF SE

[4] See also Ezer D & Whitaker K, eLIFE (2019) https://elifesciences.org/articles/43979


CREATE HYPOTHESES & PREDICTIONS
DEVELOP QUESTION OR IDEA
DESIGN TEST
MODIFY HYPOTHESES & PREDICTIONS
INTERPRET DATA
RUN TESTAND GATHER DATA
GENERATE KNOWLEDGE
FURTHER DEVELOPMENT

Many scientists already use a suite of digital tools in research, including everything from computational modelling, advanced measurement and simulations to visualisation, machine learning and automation.

Digitisation in chemistry will be a continuation of the advances we already see in areas like computational chemistry, synthesis and analytical science and holds the promise of opening really new fields, perhaps analogous to the genomics revolution.

While in 10 years' time we will not have fully 'interoperable' digital tools used universally across research and innovation, digital technologies will have accelerated progress in science discovery and application, automating some areas and expanding the possibilities and patterns humans explore.

Digital technologies will not replace people, but they will extend and augment human capabilities. Some tasks that are done by humans as part of science today will be automated. However, new tools will not replace scientists or solve big, difficult challenges by themselves. We will need knowledgeable, skilled people to develop, maintain and upgrade robots, algorithms and code as well as to design and supervise digital systems.

Moreover, digital technologies will overall free up time and both enable and push scientists to ask higher-level questions and go after bigger more complex challenges. As digital technologies find patterns and possibilities that a human brain cannot see, this will expand the options scientists consider in everything from deciding which experiments to carry out to what diagnostic signs to follow up on for medical or environmental interventions.

By using digital technologies in many different and context-dependent ways scientists will be able to:

- think at a higher and more abstract level, with more time for creativity;
- take on bigger and more complex challenges, linking end-to-end across scientific discovery and application;
- assemble a toolkit of techniques that will accelerate and enhance their research; and
- work in new ways, collaborating across disciplines and countries to harness all relevant information.

> Digitisation is important across the whole spectrum in the pharmaceutical and biotech sector – from early stage R&D to late development to commercialisation. For instance at the very early stage of identifying promising leads for new medicines, computers can handle databases of millions of molecules, and even suggest novel molecular designs. Getting computers to search these large spaces means that not only do we consider a wider range of possibilities, but we free up our scientists to spend more time on more complex higher-level creative thinking and other scientific tasks.

Dr James Weatherall Vice President, Data Science & AI, AstraZeneca

> Data and digital technologies have great potential to improve the quality of work in science. Like having a higher resolution microscope or a faster computer, they open new paths to help generate and digest information, to create knowledge and then to manage knowledge. Ultimately, digital techniques are tools to help our smart people. An algorithm can deal with ten parameters better than a human brain, but knowing which ten and how to combine them is where you need a person.

Dr Horst Weiss Vice President, Knowledge Innovation, BASF SE

As researchers work more closely with digital technologies, interactions between people and machines need to be as seamless as possible. Researchers who are not digital experts will need to convey what they want to do, and to interpret and use results and recommendations from digital tools. The following are important factors when designing future human-machine interactions:

- **Explanation:** When recommendations from digital technologies are fully explained researchers can better understand the decisions they make based on them. This is particularly important when a model or algorithm is suggesting an unexpected route of action, say for experimental design. Potential benefits of using the digital technology are negated if researchers ignore the suggestion.
- **Trust:** Researchers may not trust 'black box' systems about which they do not understand the detailed workings. Developers of digital tools can build trust, particularly with non-expert users, by documenting data sources and explaining the methods and assumptions used.
- **Intent:** Currently, chemistry researchers typically interact with digital technologies using direct instructions. An intent-driven paradigm would be more akin to the way humans communicate, but requires a level of interpretation on the part of technology.

With more widespread adoption, digital technologies continually become 'commoditised' for new applications in natural sciences research. As for any new scientific tools, it is important to have domain experts from the chemistry and digital spheres working together to develop and apply new techniques. Once they are commoditised, users who are chemistry domain experts will still need enough knowledge and skill to critically evaluate the outputs or guidance from any technique.

We must ensure that adoption of digital technologies is within robust ethical frameworks. This includes ensuring that there are secure protocols for handling data about people and that there is appropriate supervision and review of decision-making that depends on digital tools like machine learning, predictive modelling and sensing. Predictive models and machine learning may be based on incorrect assumptions, incomplete information, unreliable literature and biased or limited data. Sensors may be faulty and an algorithm may not do what a user assumes it does. It will be essential to have experts from the digital and chemistry domains co-developing and testing techniques and their applications in scientific discovery.

> A faulty sensor in a network is more dangerous than a sensor that does not work at all. You've got to factor this into any design and especially decision-making based on sensors or networks of sensors.

Prof Muffy Calder Professor of Formal Methods, University of Glasgow

> Trust depends on language and education. You're still an expert in something but need to have the ability to communicate with others. A big barrier is that the use of words is very different in different disciplines. What I mean by "proved", "certain" or "complete" may be very different from someone in another discipline.

Prof Charlotte Deane Professor of Structural Bioinformatics, University of Oxford

> Trust in a system comes down to integrity. We need to be much more strict about documenting what system components, like sensors, are doing, and about the provenance of data.

Prof Jacqueline McGlade Professor of Sustainable Development and Resilience, University College London and Professor of Public Policy and Governance, Strathmore University

> Ethics is really important. We should not rely blindly on some interpretation because there is a lot of data or some clever analysis of it. Ethics is more important in certain circumstances than others. If there is a mistake in materials discovery you may lose time and money but no one gets hurt, but in areas like precision healthcare or the environment we should not simply accept the first outputs or options generated from a digital tool.

Prof Sophia Yaliraki Professor of Theoretical Chemistry, Imperial College London

It is important not to have 'digital for digital's sake'. Researchers need to understand the benefits and limitations of existing digital tools to determine the optimal combination of digital techniques for their particular research question or target application. For example:

- Options for generating hypotheses and designing experiments will depend on different types of modelling across multiple scales, the availability of underpinning theoretical insights, knowledge about application contexts, existing experimental data and intuition.
- One measurement technique may be easier to automate than another.
- Structures like molecules may be easier to represent in a standardised way than formulations or colloids.
- We are likely to see major improvements in our ability to predict reactions and synthesis routes in some areas of synthetic organic chemistry much sooner than for heterogeneous catalysts, cathode materials or alloys.
- There will be a spectrum from situations where you can design a material on a computer to those where you get a rough guide as to where to start, with everything in between.

Questions about the prioritisation of wider benefits of digital technologies and of where and how to foster the interface between the digital, physical and life sciences are discussed in Section 5.

> "Most science and technology challenges are multi-length scale and complex, so you need to use a variety of techniques that give different insights on a problem and answer different questions."

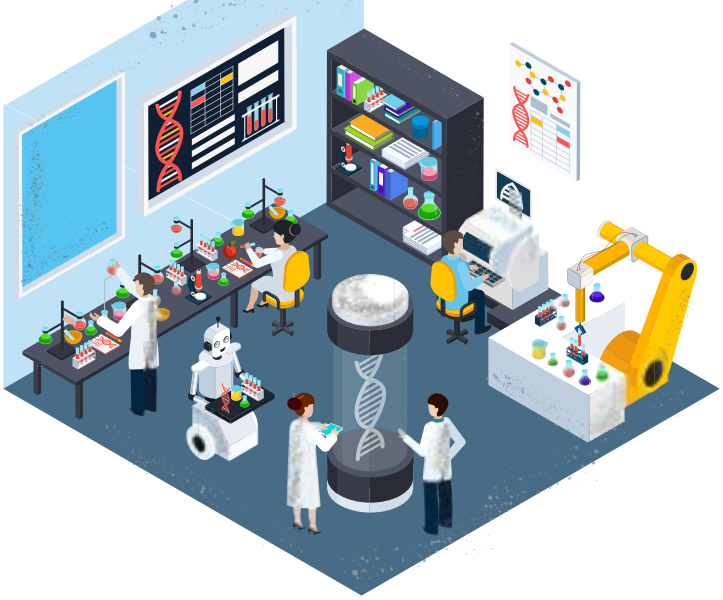
Prof Andy Cooper Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory

Digital technologies will make science more reproducible. This is important in order to validate results and to reduce duplication of both successful and unsuccessful avenues of inquiry, meaning that when something is done in one lab others can build directly on it, rather than 'reinventing the wheel'.

This applies to building on science that has been done in past and, using data streaming, should become possible in real-time for people working in different locations on the same challenge. As an example, organic synthesis experiments can be incredibly complex, with small changes in experimental parameters significantly altering outcomes. Highly accurate automated systems can give chemists more precise control over experimental parameters, automatically recording data throughout an experiment. This makes it possible to repeat the experiment and to track errors or deliberate changes as researchers optimise a process or property.

> "Imagine a world where we have a chemistry lab with a wide range of cameras and sensors in all the equipment. By sensing everything a student researcher is doing as they go through a synthesis process, we're capturing all that knowledge and information without adding to the workload."

Dr Chris White President, NEC Labs America



Enabling higher-level problem solving and abstract thinking

Using a blend of diverse digital tools scientists will be able to work in new ways, at a higher level of abstraction and on bigger problems. Being able to do things orders of magnitude faster opens new possibilities in terms of the challenges or questions people pursue. Computers can also hold more data than a human brain, and digital tools can systematically explore exponentially more possibilities, searching for structures and optimising functions. People will work with digital technologies to associate meaning with this information, using it to draw conclusions or make decisions.

Multiscale and multi-modal insights

Researchers in many fields are integrating different types of measurement, modelling and theory to gain insights into physical and biological systems. They are also making connections between our understanding of structure, properties and interactions of systems from nano to micro to macro scale. For example:

- **Whole-cell understanding:** Researchers combine many different measurements and insights to gain deeper understanding of cells – these include imaging, protein metabolomics, genome sequencing, small molecule probes a nd modelling. This enables a system-level approach to designing new drugs and treatments, as researchers can predict and control the impact a treatment will have on multiple components and pathways within a cell.
- **Monitoring and understanding pollution:** Researchers use a range of measurements from ground and ocean-based stations as well as airborne and satellite instruments. Using modelling and visualisation they work to predict and understand everything from molecular level questions about impacts on human health and local environment to questions about how pollutants spread through the soil, oceans and atmosphere globally.

> “We’re increasingly seeing an interplay between different modalities as scientists bring together techniques like proteomics, metabalomics, sub-cellular sampling and imaging techniques, along with genome sequencing and chemical interventions like genome regulation and small molecule probes.”

Dr Niklas Blomberg Director, ELIXIR

> “People wanted to probe the entire cell before, but couldn’t do it. Now we have the tools and techniques to answer the bigger question. Eventually we might probe the entire human – not in the next 10–20 years, but it will happen one day.”

Prof Charlotte Deane Professor of Structural Bioinformatics, University of Oxford

Building application insights and constraints into design and discovery phases

Digital technologies enable more 'end-to-end thinking'. Building in multiple higher-level insights and constraints from target applications right from the beginning saves time and money. For example:

- **Materials:** There are many criteria that determine the suitability of a material for applications: typically the material should be relatively cheap and easy to process, non-hazardous, efficient and stable under real-world operating conditions, and easy to recycle. Depending on the application, R&D may be targeting robustness in environments that have high temperature, moisture or alkalinity. It may seek a range of opto-electronic, electrical, catalytic or mechanical properties. There may also be goals to use minimal amounts of elements like mercury, indium or platinum, or to ensure materials that satisfy regulations for applications in health. There is no one-size-fits all approach, but using a combination of computational modelling, high throughput experimental screening and advanced characterisation can eliminate leads that would fail at application stage. This both accelerates discovery of useful materials and reduces the cost of discovery and of applications testing or clinical trials.
- **New drugs and therapies:** Researchers are building in knowledge about variations in efficacy and side effects of medicines for different individuals and populations. Taken together with understanding across multiple scales – starting with molecular level insights on how a drug molecule binds to a target or of protein-protein interactions, and moving up through to probing a cell, tissue or organ – research has moved away from the paradigm of designing treatments that target a single component or pathway with little understanding of potential side effects. Instead, researchers take a system-level approach to predict and control the precise impact a treatment will have throughout the human body.
- **Scale-up:** Often the way that you first make a new molecule or material in the lab is not the best way to make it on large scales for application. There are therefore iterations between early discovery and scale-up from both a production and a product point of view. There are opportunities for interactions between humans and digital tools in optimising scale-up, for example using iterative intelligent modelling and automation to explore around a starting set of properties and experimental or process set-ups.

> "In the UN, we see the results of science hitting the road for decision-making. People don't just think about tackling one risk, followed by another, followed by another... now we have the capability to think about a multi-hazard complex system.

Prof Jacqueline McGlade Professor of Sustainable Development and Resilience, University College London, and Professor of Public Policy and Governance, Strathmore University

> "Materials properties are a chain of links and a material is only as good as the weakest link in the chain. You need to consider the whole cohort of properties that matter to a commercial product. Digital technologies won't just give you the answer, but say you screen 100,000 formulations for a battery cathode material, looking at capacity, voltage, stability etc. You might find 100 promising avenues. Then you can think about the critical things you've got to have, the most difficult properties to obtain and what you can rule out. This is a guide as to where to focus, and where not to focus, next.

Prof Kristin Persson Professor in Materials Science and Engineering, UC Berkeley

> "Catalysts for use in transport and industry often contain a metal. If we can understand at an atomistic scale what the metal is doing and where it needs to be while in use we can figure out how to use less of it and to recover and reuse it. Multiscale modelling and in-operando measurement are key for resource efficient invention.

Dr Elizabeth Rowsell Corporate R&D Director, Johnson Matthey

Finding and using new patterns in large datasets

Digital tools like machine learning, data mining, modelling and automation can enable scientists to discover useful patterns in data and make decisions informed by those patterns, even without a complete understanding of why the pattern exists. The pattern itself often generates interesting fundamental questions. Some examples are:

- **Health:** Statistical techniques, including machine learning, can identify parameters that are a 'surrogate' for patient benefit. Drawing on multiple types and sources of data, they can provide researchers and clinicians with an understanding about treatments that are likely to be suitable for a patient. Alternatively, an individual may get a cue to consult their doctor based on comparisons of a set diagnostic measurements with their equivalent when the person was previously ill. These techniques generate many new fundamental questions about why certain parameters or patterns relate to the presence or likely onset of a medical condition.
- **Formulation:** Scientists may be targeting a property for a consumer product like a biodegradable polymer or a formulation giving a clear shampoo or detergent. Using a combination of machine learning and automation, they might find a correlation between a recipe and a property, or a set of heuristics that link to a property. The initial dataset and analysis will give ideas for designing new experiments, both to optimise formulations and to pursue fundamental questions about, for example, the molecular properties of dispersions, or why formulations look or behave as they do.
- **Empirical learning:** Tools which use machine learning to extract patterns from existing empirical data are especially powerful in areas that are hard to fully describe by theory but where large volumes of data could be available, such as drug discovery and formulation development. An example is polymeric foams, a very complex area where there is no rigorous theory or complete underpinning understanding. However, even without a detailed first principles understanding of how a foam builds, scientists can use all data about performance in real-world environments, processing steps and characterisation, and recipes or composition, to see useful relationships between starting conditions, processes and foam properties.
- **Exploring higher dimensional data:** Scientists usually look for trends in data represented in two or three dimensions, sometimes with time sequences, but with AI can discover and investigate patterns in many more dimensions. For example, an imaging mass cytometer can have as many as 40 different channels.

> "Computers now can hold far more information and complexity, and explore more possibilities than an individual human brain. We need to harness this to help us see patterns and options that a human being would not find.

Prof Charlotte Deane Professor of Structural Bioinformatics, University of Oxford

> "The idea that we break complex systems down to sub-components and find the connections is only needed if we have a limited amount of sensing and data. Now there's an opportunity to learn every time we gather data, and the speed at which we figure things out will dramatically increase.

Dr Chris White President, NEC Labs America

Increased speed enabling higher-level thinking

In addition to saving time and money, being able to do things faster can make it possible for researchers to develop new higher-level and more complex questions. For example:

- **Computational modelling:** With increased computing performance, calculations that used to take months or years can now be computed in hours or minutes. An example is materials science, where rather than focusing on understanding structure-property relationships for one or two specific materials, researchers can now seek insights across families of materials and can identify the underlying principles and mechanisms of action.
- **Faster experimentation:** If a synthesis step that previously might have taken a year to complete can be done in a few hours, that means researchers can target more and more complex syntheses, focussing human effort on the most challenging steps that cannot be automated. Researchers can explore questions about general classes of and connections between properties, structures and processes.
- **Data mining:** The ability to search through all available historical and current data in a way that an individual cannot creates possibilities to frame new problems and look for connections and correlations that would not have been possible before.

> “With automation, we can do some materials synthesis experiments 1000 times faster than an individual person can. This kind of step-change in the number of experiments we can do means we can afford to tackle bigger questions. We can be more ambitious in the challenges we take on rather than incrementally exploring around what we already know.”

Prof Andy Cooper Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory

> “If you were a student in crystallography 20 years ago, you solved the structure for a single interesting protein in your PhD. Now computers and robots routinely solve 10s or 100s a month. This means you can ask totally different questions. In my field that allows you to think in terms of multi-component systems - you move from designing something that binds or interacts at one site in a particular way to designing something that impacts in exactly the way I would like it to with several different components of a system.”

Prof Charlotte Deane Professor of Structural Bioinformatics, University of Oxford

> “My students work in a field that has been revolutionised in the last 20 years. From computing one property of one material in a year, they can now perform hundreds or thousands of computations in a week. This allows you to answer different types of questions – higher-level questions. Not just how an ion moves in one crystal structure, but what kinds of structures enable fast ionic movement. This level of abstraction was not available 20 years ago in the field.”

Prof Kristin Persson Professor in Materials Science and Engineering, UC Berkeley

Revealing unforeseen research directions and solutions

As digital technologies expand human capability, there is the exciting prospect of transformational discoveries and opening completely new research directions. Digital technologies will offer scientists hypotheses, observations and options that they may never have seen or considered to be optimal or interesting. In the same way that it would have been difficult to predict the wealth of insight and application heralded by the genomics revolution, it is impossible to predict what these new directions will be, but some examples could be:

- **Discovering new reactions as well as new molecules and materials** with structures or properties that we do not yet know are able to exist
- **Generating new research questions in the digital sphere**, for example new challenges in robotic science and engineering, in machine learning and algorithm development, or questions about the mathematical and computational abstractions that describe chemical structures and processes.
- **Revising our fundamental assumptions about systems**. Examples of assumptions that we have already seen, at least partially, overturned include that a particular small molecule drug interacts only with its target; that just one protein, or gene, is implicated in a particular disease; and, that a particular G-protein receptor has just two states, 'on' and 'off', when in fact there is a partial or influencing state depending on a neighbour.
- **Designing systems for decision-making and applications.** Scientists can be part of framing transdisciplinary problems at a higher level and designing systems in the optimal way to answer a question or inform a decision. An example is big picture thinking about questions related to chemicals in the environment, starting with articulating the decisions that a local or national government, regulator or company needs to make. This may involve inter-related technical choices about the kinds of sensors that are needed, how they will be networked, how the data will be used and by who, how the data will be represented and interpreted, all framed within wider considerations such as risk, cost, and acceptable levels of uncertainty or thresholds for action.

“ ”

As a computer scientist, what interests me is not only how ‘digital technologies’ might transform chemistry, but the kinds of new computer science research questions this transformation will generate. Do we need new abstractions and representations to talk about chemical processes and structures in a cyber-physical-chemical world? How do you develop a hypothesis, what processes do we want to understand and what do we want to model or measure?

Prof Muffy Calder Professor of Formal Methods, University of Glasgow

“ ”

Using a blend of robotics guided by computation means we can do things faster, but what’s really exciting is the prospect of finding things we wouldn’t discover otherwise. For example we’ve never discovered a heterogeneous catalyst mixture with 20 components, but I can imagine that kind of scenario evolving from intelligent robotic discovery. This would then generate a whole new scientific challenge for humans, which is to figure out why that catalyst works.

Prof Andy Cooper Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory

“ ”

Chemists discover and invent new properties. What I’m excited to see is digital technologies making it possible to discover some really novel molecular and materials structures with extraordinary properties.

Prof Charlotte Deane Professor of Structural Bioinformatics, University of Oxford

“ ”

We can be sure that digitisation in chemistry will lead to avenues that we can’t even predict. For example, when the Human Genome Project finished we went after mapping the proteome and the brain, and more will come.

Prof Sophia Yaliraki Professor of Theoretical Chemistry, Imperial College London

Better ways of working

Automated closed-loop systems

Scientific discovery and application involves sequences of steps, generally with iterations between individual steps as well as around the whole cycle from developing an idea or hypothesis, to carrying out experiments in a lab or in the field, to gathering and interpreting data, and then making decisions and taking action.

Automated closed-loop systems remove manual connections between steps in a cycle. Autonomous systems are connected by feedback loops and can also be guided by artificial intelligence at each step from prediction to experimentation to analysis so that the system can iterate and self-optimise without intervention from scientists.

Scientists will need to work closely with automated closed-loop systems, with a significant role in the design, direction and supervision of such systems. There are opportunities to use digital augmentation in combination with human insight, intuition and creativity, at each step as well as around the whole loop.

An example is closed-loop molecular or materials discovery which aims to accelerate fundamental discovery research as well as innovation in areas from new drugs and therapies to energy technologies.[5] This will include discovering and making compounds or materials that human scientists would never have found.

The initial steps of developing a hypothesis and designing an experiment are based on all relevant existing data about molecular/materials structures, properties and target applications along with computational modelling and simulation based on known theories, models and previous data. Experiments use intelligent robotic and automated systems to make and characterise new molecules and materials, systematically exploring and iteratively optimising their properties. The resulting structures and properties are added to the body of input data to inform a new hypothesis and sets of experiments.

General closed-loop systems are a long way off. For example for molecular or materials discovery, the optimal combination of in silico screening, automated intelligent experimentation and high throughput physical screening will vary depending on the question or application.

[5] See for example Stein S and Gregoire J, Chemical Science (2019) pubs.rsc.org/en/content/articlehtml/2019/sc/c9sc03766g

> I think there will be examples in the next 10 years where it's entirely closed loop – automated synthesis and measurement guided by computation. It won't happen across the board, the question is whether it will get beyond proof of principle in a university lab and how widely adopted it will be.

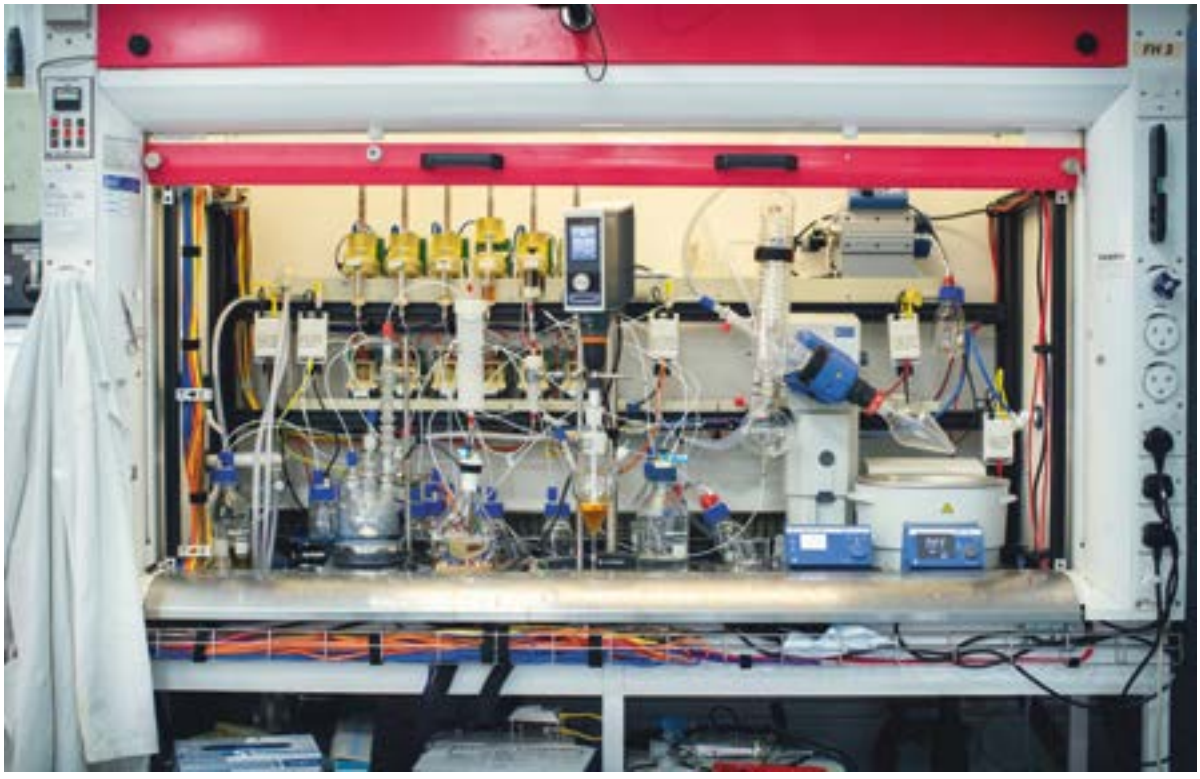
Prof Andy Cooper Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory

> We finally have the beginnings of a revolution in materials science. We can use quantum mechanics and first-principles simulations to predict novel materials, but we need robotics and automation and to use a closed-loop system. Otherwise we won't fail fast enough to find new battery cathode materials for the future.

Prof Kristin Persson Professor in Materials Science and Engineering, UC Berkeley

> There's a huge opportunity to reduce the time and cost from hit to lead in drug discovery. Traditionally we need to do the design-make-test-analyse cycle many times on promising molecules before we are confident enough to proceed. Through AI and automation we are looking at potentially reducing the timescales to do this, from months or weeks, to days or hours.

Dr James Weatherall Vice President, Data Science & AI, AstraZeneca



Computationally guided experimental design

Digital technologies can both support and push scientists to think more broadly about what hypotheses to test, as well as how to test them.

In developing hypotheses and designing experiments, computers can explore high volumes and complexity of previous empirical data and uncertainties around it. They can also simultaneously consider inputs and outputs from theory and computational modelling as well as constraints related to target applications and experimental configurations. Machine learning and automation can also enable iteration between hypothesis generation and experimentation.

It is important for researchers to consider the recommendations of these tools seriously or else any efficiency gains or potential new insights will be lost.

As an example, in order to design more potent drug molecules, chemists may start with a known molecule and make structural changes that incrementally increase potency. A computer may suggest instead to experiment on less effective molecules if their 'failures' might reveal key insights about the underlying mechanism of action of the drug.

"

It can take 4-5 steps to make even a simple molecule, with overall hundreds of ways of completing all the steps. For more complex molecules there can be thousands of routes to making the molecule. Tools that predict how well each route will work – including predicted yield of each step – will be a great asset to scientists in evaluating which route to pursue.

Programmes for predicting reactions are getting better, but still based on limited data. They use retrosynthesis engines to suggest ways of making molecules and combine with assessment for what is most likely to be the optimum way. In ten years' time, with more complete data including failed reactions, they should be really good.

Prof Varinder Aggarwal Professor of Synthetic Chemistry, University of Bristol

"

Optimisation is one thing that computers do well. If you can accurately list your constraints, they can tell you the possible experiments that will give you the largest amount of information. The question is whether we are prepared to list our constraints truthfully, and then honestly follow what the computer says.

Prof Charlotte Deane Professor of Structural Bioinformatics, University of Oxford

"

The experiments you can do are not always the ones you want to do. We often do experiments that give you a nice and strong effect, but they are not always the most scientifically instructive.

Dr Martin Jones Deputy Head of Microscopy Prototyping, The Francis Crick Institute

"

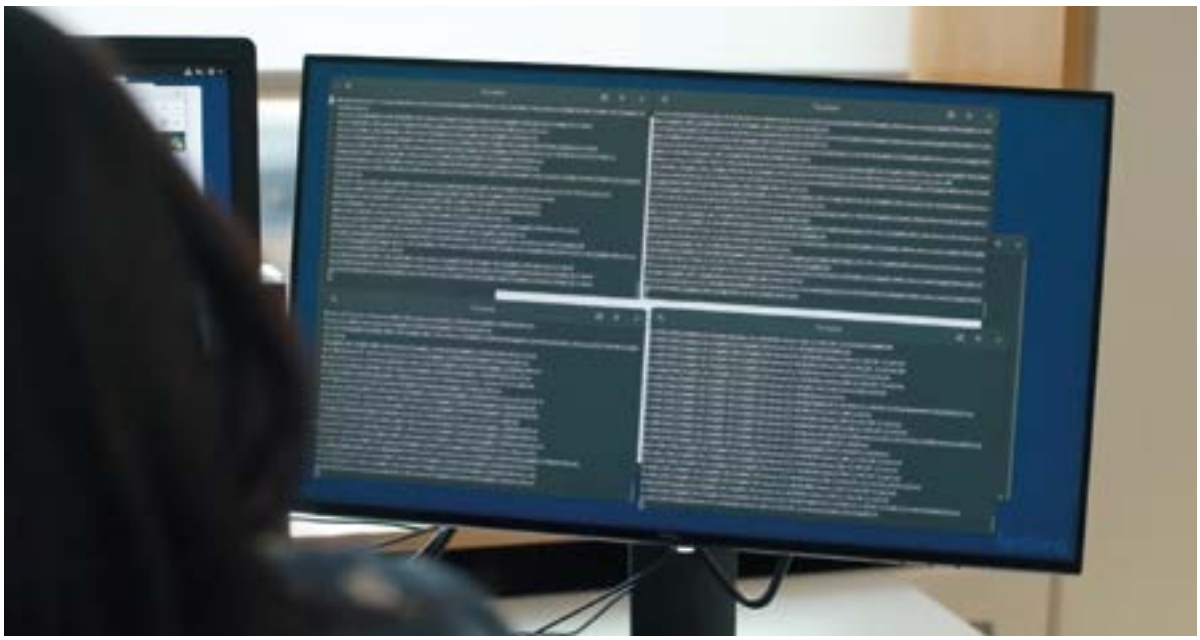
You don't need to understand the system to drive optimisation, but a judicious use of digital tools in experimental design can lead to better products and better scientific understanding.

Dr Edward Pyzer-Knapp Research Lead, Machine Learning and Artificial Intelligence, IBM Research UK

"

Cognitive algorithms reasoning over a large dataset can suggest hypotheses in a more objective way – not necessarily the experiment I might want to do or intuitively do, but the experiment that I should do.

Dr James Weatherall Vice President, Data Science & AI, AstraZeneca



Harnessing all relevant data and knowledge

Digital technologies can enable collaboration by making it easier to share information among individuals and groups in different locations. Digitalisation can more broadly enable scientists, citizens, companies and governments to more effectively identify, share, build on and benefit from accumulated data and knowledge. This includes:

- **Knowledge management systems enhanced by AI engines and visualisation tools** which highlight the most relevant and valuable information, allowing researchers to efficiently navigate and harness existing knowledge and spend less time digesting large volumes of information or duplicating effort. Machine learning algorithms combined with data mining can also uncover connections that people cannot see. For example, machine learning applied to data mined from historical materials science research publications recommended new thermoelectric materials several years before their actual discovery.[6] Commercial systems such as IBM Cognitive Discovery can also automate the extraction of knowledge from data accumulated inside organisations, and use this to make predictions or recommendations.[7]
- **Data streaming and real-time data sharing:** 5G networks enable real-time data sharing between devices and labs as a result of improvements in bandwidth, latency, and intelligence. This will potentially enable real-time optimisation of experiments or computations based on data or new knowledge being generated now, rather than publications that appear several months after data was recorded.
- **Distributed learning and data brokers:** In some situations data cannot be shared, for example personal data in fields like healthcare or private data between commercial competitors. Distributed learning algorithms take a federated approach, allowing partners to learn from each other's data without direct access to the original information. An example is MELLODDY, a project to develop a distributed machine learning platform for drug discovery.[8] Another way to share private or proprietary data involves third party data brokers via which anonymised data is made available to multiple parties. An example is Lhasa Limited, which facilitates data sharing in areas like toxicology.[9]

[6] Tshitoyan et al, Nature (2019) www.nature.com/articles/s41586-019-1335-8

[7] IBM Cognitive Discovery www.zurich.ibm.com/cognitivediscovery

[8] Machine learning ledger orchestration for drug discovery (MELLODDY) www.imi.europa.eu/projects-results/project-factsheets/melloddy

[9] Lhasa Ltd www.lhasalimited.org

- **Data standards, formats and interoperability** underpin collective use of data. The FAIR initiative, a framework for structuring data to be Findable, Accessible, Interoperable and Reusable, is being widely applied to data management, and this encourages domain-specific standards within wider data discovery best practice. In chemistry, community efforts to develop and maintain standards and formats exist in pockets, as well as widely adopted commercial standards, for example for molecular information (e.g.SMILES, InChI, MOLfiles) or the Crystallographic Information File (CIF) as the de-facto standard for representing crystallography data. There are numerous standards and formats related to other aspects of chemistry, as well as proprietary standards and formats such as those developed by instrument manufacturers.
- **Data reduction and storage:** In scenarios where there are large data volumes – like microscopy or with streamed data – it becomes inefficient and prohibitively costly from a financial and computational point of view to put all data on disc. Data reduction tools reduce the amount of primary or raw data stored. AI can be used to filter and identify relevant data but it is important in data reduction to consider how data might be used not only for an immediate research project but for future research also.

As research becomes more multidisciplinary we need to broaden what we consider to be 'chemical data', for example making decisions about how to handle real-time sensing data or whether to include descriptors capturing environmental footprint in data about a molecule, material or process.

> "Many of the problems chemistry will attack are big science problems, like the circular economy, biofuels or healthcare. One of the biggest challenges in collaborative projects is managing information – open sharing and accessibility is the cheapest way to manage large volumes of information in distributed and disconnected teams. Looking back at the Human Genome Project, the BERMUDA declaration of open science was crucial as it enabled researchers to work independently but on a core of open data. The 'download and analyse' paradigm will also have to go away because of data volumes."
>
> **Dr Niklas Blomberg** Director, ELIXIR

> "We already use specialised computing infrastructure for cryo-electron microscopy and volume electron microscopy, where the instruments produce data faster than it can be written to standard discs. Next generation microscopes will be a hundred times faster. There are opportunities to learn from fields like particle physics which already have machines producing petabytes of data per second."
>
> **Dr Martin Jones** Deputy Head of Microscopy Prototyping, The Francis Crick Institute

> "Knowledge management systems augment scientists, enabling them to be more creative and giving them time for complex thinking. Researchers don't need to spend time copying literature, or duplicating things that have already been done."
>
> **Dr Edward Pyzer-Knapp** Research Lead, Machine Learning and Artificial Intelligence, IBM Research UK

> "Our company is 200 years old and has a huge amount of data. I'd like to be able to use this to make things more intelligently with less footprint and to understand the implications of our waste."
>
> **Dr Elizabeth Rowsell** Corporate R&D Director, Johnson Matthey

> "Some questions are so complicated that a single company cannot solve them, and there is a realisation that we need think carefully about how to share information. In a multinational and multi-trust environment, we need to take this seriously along the value chain, which means we need mechanisms for exchanging information, having measures to tell us if the information is somehow broken."
>
> **Dr Horst Weiss** Vice President, Knowledge Innovation, BASF SE

4

Using a digital toolkit in scientific research

Natural scientists increasingly use a range of digital techniques, with the specific combination of techniques depending on the kind of research they are doing. Examples are:

- Computational modelling and simulations
- Automation of physical experiments
- Advanced measurement and sensing
- Imaging and visualisation
- AI and machine learning
- Computer hardware and architectures

Computational modelling and simulations

Computational modelling and simulations, increasingly integrating machine learning, are powerful tools in making predictions, exploring options to guide thinking and in interpreting empirical data. Advances in any underlying theories and/or having larger and more complete training datasets are critical for continually improving models. Examples of the use of modelling are:

- **Predicting materials structures and properties:** First principles calculations, combined with existing experimental data, can provide a starting point for further modelling or for experiments. For example, if scientists find a material with particular structure and properties, they may look for materials with related structures and similar properties. Modelling can provide promising candidates that may be unexpected as well as indications of what is unlikely to work, all before lengthy or costly physical experiments and screening.

- **Environmental modelling:** Predictive modelling and simulations are particularly important in environmental chemistry and science, as it is often not possible to do the equivalent of controlled clinical trials, especially for large-scale impacts.
- **Simulations:** Advances in theory and computational modelling will lead to more accurate simulations in the next ten years. The predictive power of simulations increases as they get closer to representing the complexity of real-world systems, meaning they can be used to predict the performance of new materials and processes without lengthy lab experimentation. An example is seeking to understand the underlying electron conductance process in new types of metal-air battery electrodes by comparing simulations with experiment.[10]

“ ”

For complex multiscale problems it is unlikely we will ever be able to do complete “in silico” predictions, but theoretical and computational techniques will be increasingly powerful tools in guiding us where to look.

Prof Andy Cooper Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory

“ ”

There is a huge role for models to improve our preparedness and response on environmental issues. You can do a lot of experimentation in the model to better inform policymakers about environmental issues, both understanding what we observe now and exploring what-if scenarios based on predictive models.

Prof Jacqueline McGlade Professor of Sustainable Development and Resilience, University College London, and Professor of Public Policy and Governance, Strathmore University

“ ”

Our vision at the Materials Project was to democratise the use of first principles calculations – so everyone doesn’t need to have a detailed understanding of how to do Density Functional Theory (DFT) calculations, but can use the result.

Prof Kristin Persson Professor in Materials Science and Engineering, UC Berkeley

“ ”

Chemistry has lots of messy, complex, multiscale data. You’re dealing with multiple kinds of structures, reactions and measurement techniques, often at the same time. So it’s important to bring in as much theory, computational modelling and data-driven approaches as we can to guide us in navigating and drawing conclusions from this data.

Prof Sophia Yaliraki Professor of Theoretical Chemistry, Imperial College London

[10] See for example the Mat4Bat project http://nano-bio.ehu.es/project/silico-design-efficient-materials-next-generation-batteries-mat4bat

Automation of physical experiments

Benefits of partially or fully automating experiments using robots and automated instruments include:

- **Freeing up researcher time:** So people can do fewer repetitive tasks and can spend more time on complex and creative thinking.
- **Speed:** Increasing throughput and productivity.
- **Accuracy:** Automated systems are more precisely controlled. Changes and errors are also easier to trace because data is captured automatically in a systematic, standard way.
- **Reproducibility:** Automated experiments are highly repeatable within the same setup. Automation facilitates reproduction of experiments by others as experimental and analytical protocols can be published and shared, sometimes even directly as executable code.
- **Safety:** Researchers can be isolated from potentially hazardous substances as parts of an experiment are carried out by machines.

Examples of areas where physical automation already exists and is developing to varying degrees are:

Synthesis: Automated systems are already used, especially in industry, for some more straightforward steps in solution-based chemistry and synthesis. These include flow systems and batch reactors. There are significant challenges in bridging to the next level of automation, including automating the connections between different steps in a synthesis, automating more difficult steps, and synthesising complex molecules.[11]

For inorganic materials there are many different synthesis methods and also challenges in solid materials handling. There are opportunities for modular approaches, breaking synthesis and characterisation into individual automated steps. There are also examples of mobile robotic systems which can carry out and connect different steps.

Measurement: Techniques like mass spectrometry and gas chromatography are automated for routine applications in industry. There are also large elements of automation in techniques from fluorescence spectroscopy to NMR to calorimetry. Large scale X-ray screening experiments are also being automated, for example the X-Chem partnership between the Diamond Light Source and Structural Genomics Consortium covers steps including aspects of sample preparation, automatic data collection, and data analysis, implemented as a streamlined process, allowing up to 1000 compounds to be screened individually in less than a week.[12]

[11] Peplow M, *Chemical & Engineering News* (2019) cen.acs.org/synthesis/Automation-people-Training-new-generation/97/i42

[12] XChem: X-ray structure-accelerated, synthesis-aligned fragment medicinal chemistry www.diamond.ac.uk/Instruments/Mx/Fragment-Screening.html

“ ”

A challenge in chemical synthesis is that it is can be very hard to fully understand the sensitivity of a particular reaction to all possible conditions, for example, to moisture or oxygen. Automated systems and computers can keep track of and control more variables than a human, systematically exploring the full range of the experimental set-up. People may not record every detail when they describe a synthesis, meaning there can be issues with reproducibility. Capturing experimental protocols in a systematic way will be important for machine learning algorithms and should also help with reproducibility.

It could be really transformative for chemistry if we could use automation to make really complex molecules like anti-microbial and anti-cancer compounds. This is really difficult because these structures have more than 10 carbons, multiple heteroatoms, and multiple stereogenic centres. It would be especially attractive if such complex molecules could be made by iterative automated chemistry, but that is really difficult and beyond the capability of current methods.

Prof Varinder Aggarwal Professor of Synthetic Chemistry, University of Bristol

“ ”

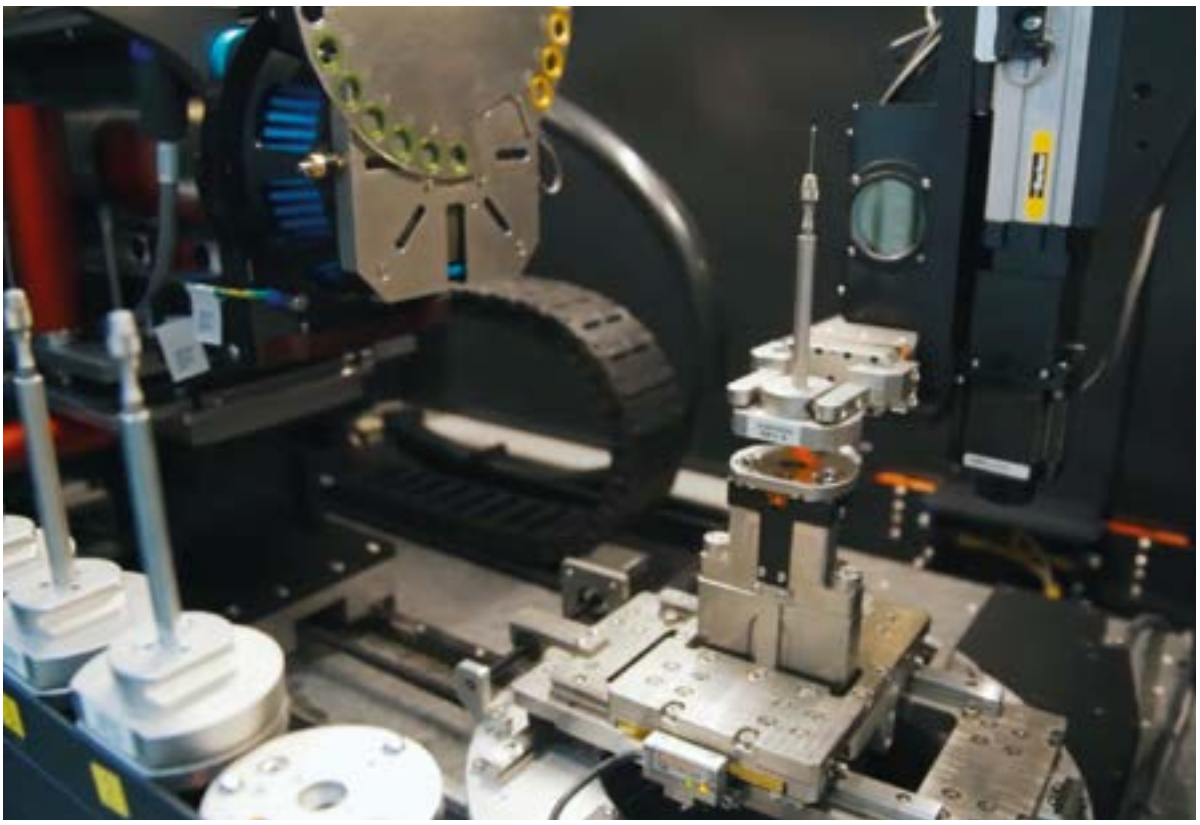
An awful lot of things in materials research are just not automated yet. The techniques are a bit more advanced in organic synthesis. It’s not just about stringing things together – some of the computational and automated tools don’t exist at all yet, but they certainly could on a 10-year horizon.

Prof Andy Cooper Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory

“ ”

I think robotics and automation that help us with chemistry are plausible at a certain scale. But it’s not robotics for robotics’ sake. One thing doesn’t fit all, and I struggle to see how they can do everything.

Dr Elizabeth Rowsell Corporate R&D Director, Johnson Matthey



Advanced measurement and sensing

In our Science Horizons project we heard from researchers that modern measurement techniques – from different kinds of spectroscopy, spectrometry and microscopy to robust miniaturised distributed sensor networks – are giving unprecedented insights about the structure, properties and interactions of systems from atoms and molecules through to materials and surfaces through to organisms and ecosystems.

Some examples of how researchers are bringing these techniques together and taking them to the next level in terms of capability and application are:

- **Smart microscopy:** Integrating light and electron microscopy by using automation and machine learning to analyse a sample using fluorescence microscopy and then guiding an electron microscope about where to look.
- **Multi-modal approaches:** Scientists use multiple techniques to study the same system or answer a question. An example is advancing our understanding of protein structure and function using a combination of X-ray diffraction, cryogenic electron microscopy (cryo-EM) and mass spectrometry.

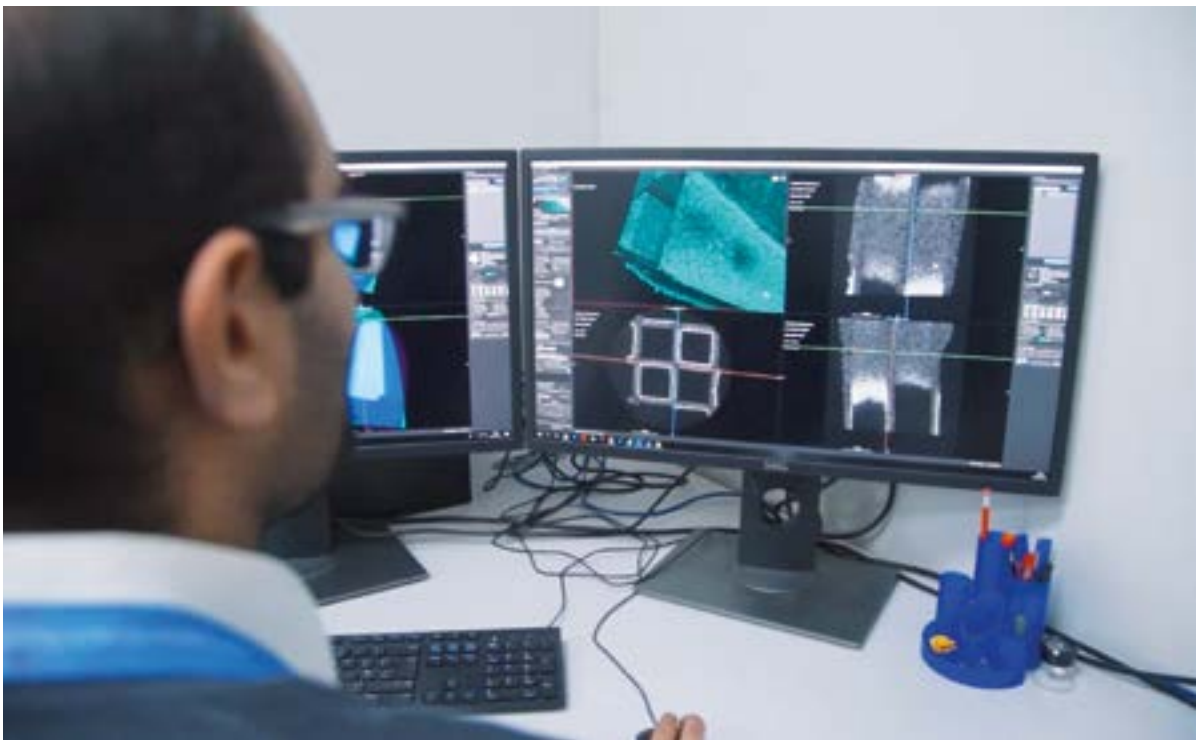
- **In-operando measurement:** Sensors and measurement instruments that enable researchers to study the performance of components or products while in use. These insights can be used to monitor safety and to improve future designs. Examples are monitoring a battery while in use to see when the electrolyte is about the break down, monitoring the interactions of a drug with its target in a cell, using smart or active coatings on a medical implant or drug delivery system to get read-outs about its interactions in the patient's body.
- **Distributed real-time and in-situ sensors:** Often combined with AI, networked sensors are important for everything from environmental monitoring and large-scale agriculture to applications in industrial health and safety such as monitoring gas release or corrosion, or knowing when equipment needs to be maintained or replaced.

> **We're seeing huge tech changes for environmental monitoring. We already deploy wearables, satellites that use radar pulses to tell you about wind speed, and affordable sensors for volatile organic compounds and heavy metals. But we don't have very good sensors or models yet for microplastics.**

Prof Jacqueline McGlade Professor of Sustainable Development and Resilience, University College London, and Professor of Public Policy and Governance, Strathmore University

> **I'd really like to see sensing or diagnostic systems built into products that feed back in real time, e.g. seeing a battery in service. Then you can understand crunch points and design your product better. All of this is enabled by digital technologies, but we're not there yet.**

Dr Elizabeth Rowsell Corporate R&D Director, Johnson Matthey



Imaging and visualisation

Scientists use imaging and visualisation techniques widely, in some cases using a combination of automation and statistical methods such as machine learning to accelerate complex and time-consuming data analysis and calculations.

- **Image analysis:** Machine learning algorithms can extract relevant data from images with high accuracy and throughput. Many researchers have developed algorithms for biological imaging, for example to classify images, or identify and track objects.
- **Medical diagnosis:** Machine learning can be used to aid healthcare professionals in interpreting medical images to support diagnosis and decision-making.[13]
- **From big data to relevant data:** Machine learning can be used to filter data and decide what can be thrown away so that scientists have more data relevant to the question they are trying to answer or problem they want to solve.
- **Visualisation** is useful for summarising data and representing it in an accessible way. For example, visualisations of predictions of climate models, of observed waste flows, or the spread of disease, can be useful tools in a policy context to support development of regulation or decision-making.

> **We need analysis tools that can filter down large datasets to highlight what's useful, for example pinpoint specific features in images. Many groups around the world are looking at what machine learning can do on this, both for research and for medical diagnostics.**

Dr Martin Jones Deputy Head of Microscopy Prototyping, The Francis Crick Institute

[13] For a perspective on machine learning in clinical translation see Saria et al, PLOS Medicine (2018) blogs.plos.org/speakingofmedicine/2018/11/28/better-medicine-through-machine-learning-whats-real-and-whats-artificial

“ ”

In the next 10 years we will have increasing use of visualisation techniques, both to request what we want to do and to understand the outputs at the end.

Dr Edward Pyzer-Knapp Research Lead, Machine Learning and Artificial Intelligence, IBM Research UK

“ ”

Visualisation can be a very powerful tool for interpreting data, but you have to understand what exactly you are looking at because once you visualise you've already skewed the data, you're looking at a transformation of it. For example you may only be looking at a low-dimensional slice of data about a system with many degrees of freedom and you may be looking at a linear, non-linear or geometric projection of this data which may lead to different conclusions.

Prof Sophia Yaliraki Professor of Theoretical Chemistry, Imperial College London

AI and machine learning

Researchers are using machine learning in many different ways, including in all of the previous areas discussed in this section.[14] For example:

- **Combining machine learning with physical automation:** Researchers can create an automated experimental set-up that intelligently explores a range of parameters as it, for example, identifies promising properties or patterns and rules out less promising avenues.
- **Enhancing imaging and diagnostics:** Researchers are using machine learning techniques to identify patterns in images for scientific research and for medical applications. This can accelerate diagnosis and identify patterns that a person may not see.
- **Data reduction:** Machine learning algorithms are used to filter large volumes of data produced from some instruments or through streaming, in order to identify and save relevant data.
- **Enhancing measurement and sensing:** Machine learning is being used in everything from optimising measurement systems to informing decision-making based on data from sensors.
- **Quantum chemistry:** Using existing calculations as training data, machine learning algorithms can in some cases accurately predict properties such as electronic structure and potential energy surfaces faster than first principles quantum mechanical calculations.[15]

“ ”

Machine learning and artificial intelligence is good at exploration and to highlight directions for innovation, and it's under-utilised in science in this way. But it still needs a person to interpret and take it to the next level.

You need intelligence to turn data into something relevant and useful. For example machine learning methods and toy models can be used to give insights in areas where we have limited theoretical understanding or don't know the initial conditions. Machine learning can vary around the toy model, exploring it better even if a human does not fully understand the changes. This way we can get a more rigorous model with well-defined failure modes and limits of applicability.

[14] For more information and AI and machine learning generally see https://royalsociety.org/topics-policy/projects/machine-learning/

[15] For example Smith et al, Chemical Science (2017) pubs.rsc.org/en/content/articlelanding/2017/SC/C6SC05720A#!divAbstract

Dr Chris White President, NEC Labs America

Computer hardware and architectures

The continual advances in computer hardware will enable researchers to go faster and to handle more complexity in everything from modelling and simulations to robotics, data analysis and visualisations.

It is important for chemical science researchers to engage with digital technology researchers and developers at the cutting edge in order to understand emerging innovations and current and future technologies that could significantly accelerate and augment their research.

Examples of evolving computer hardware and architectures that will be important for chemical scientists to be aware of are:

- **Edge computing:** Distributed architectures where data is processed near the point of collection, including having an AI engine on the sensor. Enabled by smart devices, this leads to faster, more reliable responses and uses less energy. Edge computing also has benefits for private and sensitive data, for example in healthcare, as data is not transferred to a centralised processing point.
- **Stream computing:** Designed for continuous processing of large datasets, stream computing analyses high velocity data flows and offers insights in real-time. For techniques like microscopy, stream computing may be an approach to overcome the transfer, storage and processing challenges associated with high data volumes.
- **Quantum computing:** A computing paradigm that uses quantum superposition and entanglement to manipulate information. It can be much faster than classical computers for tasks such as search but is only applicable for processes that are described by quantum operations. It has clear potential applications in quantum chemistry.
- **Specialised processors:** Optimised for specific task(s), specialised processors can greatly improve performance for computationally-demanding processes and can become mainstream. For example Graphics Processing Units (GPUs) have enabled the development and application of computationally-expensive deep learning algorithms. In the future, we may see chemists use custom-made heterogeneous systems that combine specialised processors.

5

Barriers & enablers

Data opportunities and challenges

The volume of scientific data and the sophistication of techniques to collect and interpret it will continue to increase. This encompasses data from individual techniques like cryo-EM, phage displays and next generation sequencing as well as from distributed and real-time sensing. Multi-modal approaches to research also create larger and more complex datasets as scientists draw on multiple types and sources of data relevant to a research challenge.

Digital technologies create many opportunities to harness data to make new discoveries and innovate faster as discussed in Section 3. However there are also significant pitfalls and challenges.

Recording and accessing data

Data provenance and integrity: This is key in drawing conclusions from or making decisions informed by data.

> **Data provenance and integrity are crucial. Which equipment did the data come from? What was the test set-up? How do I verify that my sensing set-up is delivering information I can trust? What was the data collected for? Who is maintaining it? What representation of the data is being used? What standards are being used? What models are assumed in the representation of the data and what are those models based on?**
>
> **Prof Muffy Calder** Professor of Formal Methods, University of Glasgow

Standards and formats are key for data sharing and interoperability. As digital technologies and data sharing become more prevalent in R&D, reaching an agreed understanding about what data to record, how to represent it, and how to format it, enables scientists to compare, combine and analyse data from different sources in an efficient and reliable way.

Standards for metadata are essential, for example describing the origin of and assumptions about data – especially when integrating data from different sources. Higher level ontologies and classifications such as SNOMED that sit above databases used in different groups can also facilitate interoperability without requiring that data from different sources is formatted in exactly the same way.[16] Registration processes or identifiers for data can also facilitate combination of different datasets.

> **Algorithms for reaction prediction would be better served if published chemistries could be curated and placed in a database. Another issue is that most of the chemistry generated isn't reported in a form that is easily accessible to software.**
>
> **Prof Varinder Aggarwal** Professor of Synthetic Chemistry, University of Bristol

> **Standards are critical for distributed learning. Can you really come up with sensible results if you're comparing data between hospitals in New York and Paris? They have different healthcare systems and may collect information about patients in different ways.**
>
> **Dr James Weatherall** Vice President, Data Science & AI, AstraZeneca

[16] SNOMED, US National Institutes of Health, National Library of Medicine www.nlm.nih.gov/healthit/snomedct/index.html

Future-proofing beyond a specific research context: To ensure longevity, decisions about how to generate, format, store and share data should ideally factor in thinking about how data collection techniques might evolve, as well as potential future uses of data beyond the immediate research question or current instruments and technologies. An option is to consider standards and formats that are extensible so that they can evolve at the same pace as the techniques for gathering data. It is also important to consider how data acquired to answer a particular scientific question may be valuable for answering other questions also.

Sharing all data including failures and negative results: Sharing data enables scientists, citizens, companies and governments to identify, build on and benefit from accumulated data and knowledge.

Digital tools enable exploration of vastly larger volumes of data and it is important to publish all data, rather than one 'slice' of it like a graph in a scientific journal article. It is also important to track data and analyses that were published and looked promising at one time but later were identified as having mistakes or problems.

Sharing both positive and negative results is particularly important for training machine learning systems.

> **For machine learning it is really important to record both positive and negative results. Having more information about what works and what doesn't means that the algorithms are much better able to judge which pathways to use in retrosynthesis or in the forward direction.**
>
> **Prof Varinder Aggarwal** Professor of Synthetic Chemistry, University of Bristol

It is also important where possible to share algorithms and code although, as with data, there will be constraints and limitations around this.

Combining data from different places is important as a way of bringing together different silos in order to draw higher-level insights, but not if this is simply creating an unstructured 'data lake'. For example in the pharmaceutical sector there will be many types and sources of data from scientific literature, imaging, pharmacology, toxicology and clinical trials.

Data-sharing culture and behaviours

People have different motivations, limitations and constraints when it comes to data sharing and it is important to have a realistic understanding of these in order to create the conditions that will be most likely to encourage, enable or compel individuals, companies or institutions to collaborate and share.[17] Being aware of these factors can also prevent unintended consequences, such as reduced participation of corporate researchers in scientific fora if associated requirements around sharing data or code are incompatible with company policy.

Examples of different kinds of constraints and disincentives are:

- Commercial competition will constrain the extent to which companies may share data, methods or results.
- Competition between individuals, especially in academia, may make them reluctant to share complete datasets, algorithms or code.
- Sharing data within and between countries needs to be thought through, especially in the case of personal data about individuals or data related to national security.
- National governments and citizens may have concerns about data that has been generated by publicly funded research being shared with and delivering benefit to other countries.
- Proprietary formats and software which require maintenance and upgrading may be part of the business model for companies that make scientific instruments.
- Cost implications of systems to enable data sharing on a long term basis need to be considered, and clarity about who will bear these costs.
- For individual researchers or groups in universities or companies, generating data in a format that can be shared inter-operably may be unappealing as it can be tedious, especially if done manually.

There are many possible incentives and enablers for data sharing ranging from making it technically easier and less time-consuming, to identifying situations where there is a personal or collective benefit.

Technical enablers

- Private and decentralised algorithms in which no one sees each other's data but can interact with it. This allows everyone to benefit, for example by creating better models because datasets are bigger, more diverse and less biased.
- Using anonymised or encrypted data.
- Collectively annotating public datasets, which can inform what annotations to use on private datasets, making it easier to import public data in a structured way using machine learning.
- Repositories for data or code which are easily searchable can save a lot of time. They can give researchers access to quality datasets to build on rather than having to generate similar data themselves. A PhD student can use code that another student may have spent a sizeable part of their PhD developing, reducing duplication and moving the research area forward faster.
- Institutional databases, automated data backup, centralised data storage and warehousing all enable data sharing.
- Machine learning to automate aspects of data classification, tagging and formatting to support standardisation and sharing.
- Data streaming and other solutions discussed in Section 3 on Harnessing all relevant data and knowledge.

[17] For another perspective see blogs.lse.ac.uk/impactofsocialsciences/2018/11/14/the-main-obstacles-to-better-research-data-management-and-sharing-are-cultural-but-change-is-in-our-hands

Commercial and national interest incentives

- Pre-competitive or non-competitive situations, for example sharing toxicology data that is important across many sectors in a supply chain. Having larger datasets enables all companies to ensure employee and customer safety as well as regulatory compliance.
- Reproducibility and efficiency: Sharing data both inside and outside companies can lead to greater efficiency and reduce duplication.
- Consortia like Public-Private Partnerships can consider contributing data as a benefit-in-kind, which means that sharing internal data is monetised by funding agencies.
- Designing local or national scientific initiatives so that they both generate and capture value from research data in a particular field, including attracting highly skilled workers and inward industrial investment.
- Customer demand for standardised inputs and outputs from scientific instruments.

Individual and academic incentives

- Defining minimum reporting standards to collectively drive a scientific area forward by addressing issues with reproducibility or duplication.
- Collective vision from leaders, for example community papers articulating how senior researchers propose to approach minimum information standards and community conventions, how data ought to be recorded and published in their field.
- Publisher requirements around deposition of data and code. For example in chemistry if published work includes a crystal structure then publishers generally require that it is deposited with the Cambridge Crystallographic Data Centre.[18]
- Collective benefit of sharing negative results combined with publisher and institutional systems to enable and reward sharing of negative results.
- Shifts in academic research culture to incentivise and reward collaboration and publication of data.
- Creation of professionalised career paths for people with knowledge and skills in data sharing and knowledge management.
- Data Champion roles to advocate for data sharing and enable peer-to-peer support.

[18] Cambridge Crystallographic Data Centre www.ccdc.cam.ac.uk

> In synthetic chemistry, we need much more data about what works and what doesn't so we can develop better algorithms for the prediction of reaction outcomes. We need to record every reaction, both positive and negative. Publishers can play a big role by specifying the type and format of data that needs to be made available. Getting academics and journals to change their practices is hard!

Prof Varinder Aggarwal Professor of Synthetic Chemistry, University of Bristol

> You need to have a lot of senior people within a field coming together to think about minimal information standards. If you look at biology, there have been a number of community-led conventions for how data ought to be recorded on the day of publication. These have been put out as community papers, and that is what enables journals or editors to hold people accountable.

Dr Niklas Blomberg Director, ELIXIR

> It's always difficult to agree standards or create umbrellas enabling interoperability. Someone has to go first, put things forward and people will gravitate towards that. Once you have a standard then you can reinforce it, for example by making it a requirement for funding applications or publications, or by incentivising SMEs to develop data interoperabiltiy solutions.

Dr James Weatherall Vice President, Data Science & AI, AstraZeneca

> It's important to manage change in digitalisation. In order to reach agreement on new technologies, you need to have an idea of what it means for your business or research area. It's hard to get people "thinking digital", for example people may not use electronic lab books because it doesn't reflect the way they work, doesn't offer enough flexibility. It's a culture change – you need to show researchers and the organisation that something comes out of it afterwards, what the net benefits are.
>
> The ideal is to have good design of experiment and systematic data so you can draw conclusions and reproduce things instead of doing things by careful trial and error. This requires a culture change because getting clean data with the right descriptors is not so straightforward. For example how do you capture the meaning of "heat up" or "wet" or "pale", which might be recorded in a lab notebook, in a way that can be recorded unambiguously later? People need to see the benefits for themselves and for their company of setting things up so as to get all the data in a way that is easy to interpret afterwards.

Dr Horst Weiss Vice President, Knowledge Innovation, BASF SE

New skills and roles

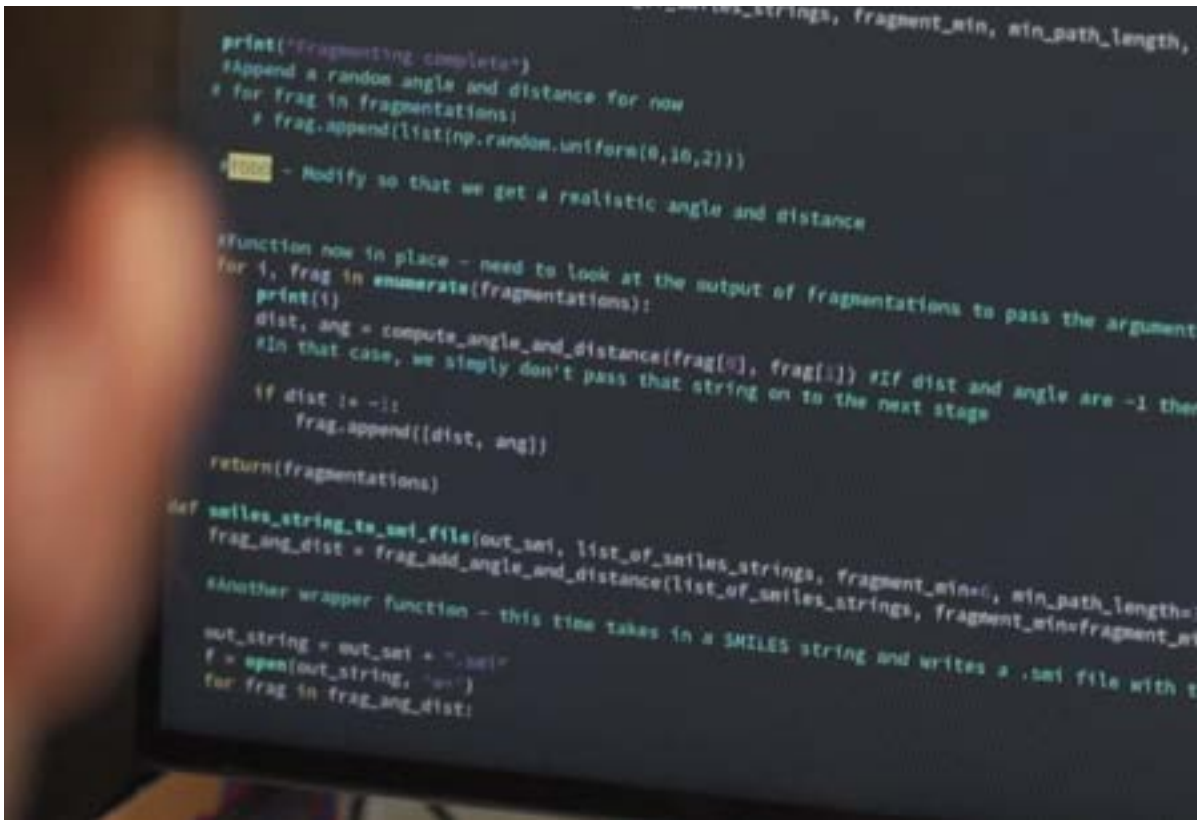
Skills

It will become increasingly important for researchers across career stages in the chemical sciences to develop and update their digital skills. The type and level of knowledge and skill will vary from baseline computational, mathematical and statistical competence for all chemical scientists to deeper understanding and knowledge for people working in a multidisciplinary way to adopt or develop new digital tools for chemical sciences discovery and application.

Concrete examples include broad skills in areas like design of experiment, algorithms and coding, use of specialised software and tools for making predictions, setting up an automated computational or wet lab experiment, and visualising or drawing inferences from data.

As datasets get bigger it will be even more essential to use computational and statistical techniques. Depending on the individual, this ranges from basic things like recording data in searchable formats and using hypothesis testing and linear regression, to established techniques in machine learning, to more novel aspects and applications of deep learning or Bayesian statistics.

In current chemical science research there is a spectrum from areas and groups where the use of tools like machine learning is lagging behind state-of-the-art algorithm development, to multidisciplinary groups that have been working for some time to push new research frontiers in chemistry, mathematics, computer science and engineering.



While not all chemists need to be involved in developing and 'commoditising' new digital tools it is important that all have enough knowledge to decide which tools to use, understand their strengths and limitations, and to critically evaluate outputs or suggestions generated by them.

Chemists will need to be 'T-shaped', with deep knowledge and expertise in chemistry, and a broad base of supporting digital skills. This combination will enable chemists to benefit from digital technologies in their own research and to 'speak the language' of digital experts, leading also to more fruitful collaborations in multidisciplinary teams.

As with multidisciplinarity and interdisciplinarity more broadly, what is important for science and for the economy overall is to have a range of people from single domain experts to interdisciplinary experts.

As the chemistry-digital interface develops, it will be crucial to maintain deep expertise in core chemistry. Rather than replacing chemistry content in undergraduate degrees with large numbers of courses in mathematics, statistics, computer science or coding, it will be important to tailor undergraduate course content to include opportunities to develop those wider skills in ways that are relevant to chemistry. It will be important to draw on perspectives from employers in articulating digital skills requirements and advising on curricula.

As data and digital technologies play an increasing role across the wider economy and society, chemists who have digital skills will be increasingly in demand to use their expertise to benefit other disciplines and sectors.

> **Digital technologies cut across many fields, and this is a way to train the workforce of the future. How can you get Masters and PhD graduates with skills in this area and make sure they are employable by industry? In Sweden, the Wallenberg Foundation is funding the Wallenberg AI programme, getting PhD students with the right skillset for Swedish industries.**
>
> **Dr Niklas Blomberg** Director, ELIXIR

> **Today you shouldn't train a chemist who can't programme. They won't necessarily do the programming but they need to understand it. As datasets become larger you won't be able to do anything without computational or statistical methods. There are basic things like knowing how to format data sensibly that people should know as well.**
>
> **It is important to prepare people for working in a cross-disciplinary environment in which you may be representing your discipline and be the only person with deep disciplinary knowledge in your area.**
>
> **Prof Charlotte Deane** Professor of Structural Bioinformatics, University of Oxford

“ ”

The power materials scientists of the future will need to use data generated from both computations and experiments. We're going to need people with the data skills to access, understand and interpret data themselves. This includes statistics as well as coding and using modern web languages. Combining data properly is especially important for machine learning because if you use machine learning as a black box, without understanding the original datasets, the answers are not well-informed.

Prof Kristin Persson Professor in Materials Science and Engineering, UC Berkeley

“ ”

Skills and training are very important. We don't want people who use commodity data science toolkits without understanding what assumptions they are making and why they are making them. Training in high school is key. You don't need to go to university to learn how to use Python.

Not all chemists need to become deep experts in AI, modelling or robotics to benefit from the value they bring, but everyone does need to understand where these systems have come from and what they give back.

Dr Edward Pyzer-Knapp Research Lead, Machine Learning and Artificial Intelligence, IBM Research UK

“ ”

It's more important than ever to develop critical thinking, problem solving, statistical and teamwork skills as chemists work on increasingly multidisciplinary problems and need to be aware of the social implications and assess the risks of what they do. They need to be able to collaboratively tackle problems at a higher level, critically evaluating results and drawing on expertise from multiple fields and subfields.

Dr Elizabeth Rowsell Corporate R&D Director, Johnson Matthey

“ ”

In Bell Labs we say you never have an electrical engineer solve an electrical engineering problem, because then you get the same solution other electrical engineers would create. You want to send a chemist to solve an electrical engineering problem to get truly creative and innovative solutions. What you need are students with enough depth so they can project the problem in front of them into the space of problems they've already seen, and then re-project that out in a way that might give an innovative and disruptive solution.

Dr Chris White President, NEC Labs America

Roles, career paths and recognition

In order to harness data and digital technologies to their full potential in life and physical sciences research, it will to crucial to attract and retain more people with digital expertise into these areas. This includes data scientists, knowledge management experts, roboticists and research software engineers whose skills will also be much in demand in the digital tech sector.

Key factors for funders, universities, research institutes and chemistry-using companies to bear in mind are:

- Considering the 'value proposition' that natural sciences research offers people with digital expertise whose skills are in high demand and highly rewarded in other sectors.
- Creating and supporting long-term, secure technical and research roles with a clear framework for career progression.
- Viewing these areas as key enabling rather than service roles. People in the roles are at the cutting edge and pushing the frontiers of their techniques.
- Ensuring opportunities for continual professional development, for example training in advances in a person's area of technical expertise and opportunities to evolve or develop new techniques as part of their contributions to research.
- Ensuring that contributions to research and its applications are recognised.
- Addressing the fact that, especially for universities and in the context of research funding structures, there are many short term roles which results in loss of knowledge and lack of continuity. Related is the fact that development of algorithms, code or software is often highly inefficient and sub-optimal when carried our by non-experts as a side project.
- Ensuring that there are people with appropriate understanding of knowledge management and awareness to ensure compliance with ethical standards.

> "New technologies will never stick and make long-term changes unless we start to professionalise some roles. The idea of long-term technical staff has been eroded in universities and the idea that you train every chemistry PhD student to be a roboticist or software engineer every four years is totally unsustainable."

Prof Andy Cooper Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory

> "In academia developing, maintaining and upgrading software is often a kind of side activity, and a dead end on a research career path because it is not recognised as a contribution to research. Even if a piece of software developed by a postdoc or PhD student is used by ten thousand people it's not always citable. I also see a lot of duplication and the issue of 'abandonware' because when a person writes specialised software and then leaves there's a lack of continuity and no documentation."

Dr Martin Jones Deputy Head of Microscopy Prototyping, The Francis Crick Institute

> "If you buy a new Scanning Tunnel Microscope or any advanced instrument, it's clear you need skilled technicians to operate, maintain and upgrade it. It's the same for software, and funding models need to reflect this. At the Materials Project, we have people who permanently work on the API, analysis and recipe-based workflow codes. If you don't maintain and upgrade software it's inefficient as you can't get the same level of output long term."

Prof Kristin Persson Professor in Materials Science and Engineering, UC Berkeley

Multidisciplinary collaborations, communities and infrastructure

Collaborations and communities

New types and levels of collaboration will be needed for everything from developing optimal solutions and approaches on specific technical challenges to co-design of major research programmes by researchers in the chemical and digital spheres.

Creating new communities and collaborations takes time as people develop a sufficient common language, with an understanding and appreciation of one another's expertise. It is important to have long-term structures and projects to support these communities and collaborations.

Joint challenges and projects that are important and interesting for everyone are key to developing interdisciplinary collaboration. This can be at different scales such as:

- Collaboration on specific workflows along the pipeline from upstream data acquisition through to processing and interpretation. This may bring together developers who care deeply about how to make Fast Fourier Transforms work efficiently, wet lab scientists who are not at all expert in coding and downstream data management experts.
- Joint supervision of PhD students on interdisciplinary research projects.
- Internships, secondments and consulting to work on projects within and between academia, SMEs and large companies.
- Cross-disciplinary curriculum development projects in universities, for example between chemistry, computer science, mathematics and statistics departments.

- Co-development and application of new techniques. For example, a connection between a person or group that is expert in machine learning or robotics and a person or group that has a chemistry research question. Ideally this will push both areas forward, creating new algorithms or robotic systems and new chemistry insights.
- Big high-level challenges that will require multidisciplinary solutions. Researchers from across different areas of science and engineering will need to co-design research programmes, identifying which approaches are most likely to succeed. This will generate novel challenges and insights in multiple arenas as different science and engineering areas propel each other forward, with each opening new and potentially transformative research directions for the other.

While 'interdisciplinarity for interdisciplinarity's sake' tends not to work, creating shared spaces and fostering new communities is important in enabling collaboration. For example:

- Networks and events, focussed on a digital technology with applications in multiple areas of chemistry, biology, materials science and physics or in companies that compete in different areas. An example is the UK AI^{3}SD Network.[19]
- Hack spaces and hack hours which can be hosted in universities or between research institutes and local companies.
- Online fora. For example in image analysis, open source software like Matlab and ImageJ are important. Fora focussed around different software have merged so that if a person wants to know how to do a particular operation for an image analysis problem using one tool someone who has solved that problem using another tool comes forward. Ten years' worth of questions about both types of software have been merged on one searchable forum.
- Shared facilities and resources also naturally bring together user groups and facilitate collaboration.

[19] Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery www.ai3sd.org

“ ”

For effective multidisciplinary working, you need to both be able to pose a problem and understand an answer in the language of other disciplines, while having networks of experts within your own discipline to cross-check your thinking.

Prof Muffy Calder Professor of Formal Methods, University of Glasgow

“ ”

Now we really need to combine chemistry, computer science and mathematical knowledge. We need a common language – this is really important as there has to be enough understanding to identify what's really important, relevant, possible or useful.

Prof Andy Cooper Professor of Chemistry, University of Liverpool, and Director of the Materials Innovation Factory

“ ”

Ultimately you usually have some horizontal scaling across domains, so you only need domain knowledge and the technique or tool is commoditised. However, as for all techniques and applications, especially in the early stages you need people working together from both the technique and the domain ends.

Maximising the potential of digital technologies in physical sciences R&D is going to require a fundamental shift in the academic system away from 'heroing it on your own' to really multidisciplinary teams bringing together experts from many several different disciplines.

Dr Edward Pyzer-Knapp Research Lead, Machine Learning and Artificial Intelligence, IBM Research UK

“ ”

It's crucial to define problems that require multiple disciplines, so that there is something exciting in them and everyone is motivated by them. Having this co-design culture will create a virtuous circle where talented people from multiple disciplines are invested in solving big, important problems and in turn their chances of success are higher because the problem has been framed as a multidisciplinary one from the get-go, identifying the best approaches to apply to that challenge.

Prof Sophia Yaliraki Professor of Theoretical Chemistry, Imperial College London

Capability & infrastructure

New capabilities, functions and physical infrastructures are an important dimensions in harnessing data and digital technologies for discovery and innovation in the physical and life sciences.

The best approach depends on the scale of a problem and the number of users. Large companies and national labs may bring in and customise existing technologies. They may have centralised teams and facilities with expertise in diverse areas, from high performance computing, modelling, data analytics and image analysis to machine learning, data mining or automation. These functions work with and support multiple business units across the company.

On the other hand, SMEs and smaller university research groups will generally not have the critical mass in terms of volume of work to warrant full-time experts or infrastructure in many areas. This can lead to inefficiency as a chemistry researcher may need to learn to use many different digital tools but use them in a sub-optimal way.

Shared facilities and centres can enable smaller labs and SMEs to have access to expertise or techniques when needed without having an in-house expert on everything. An example is the Alan Turing Institute, the UK's national institute for data science and AI.[20] The institute consolidates the expertise of people doing cutting-edge machine learning and big data research, and is open to collaborations so that other disciplines can benefit also.

Distributed systems that connect and build on existing initiatives can be easier to set up than large centralised infrastructure, although this can bring challenges related to interoperability and sharing. An example is ELIXIR, an intergovernmental organisation that integrates and sustains bioinformatics resources across its member states in Europe.[21]

There are also different options related to the scale of facilities and types of research question. As an example, an automated synthesis facility that optimises reactions on a scale of 20-50 experiments may not be worth travel or set-up time for researchers from another location. However, for some research questions access to a higher throughput facility enabling optimisation over thousands of reactions may be appealing.

[20] The Alan Turing Institute www.turing.ac.uk

[21] ELIXIR https://elixir-europe.org/

Leadership & vision

Leadership and a long-term strategic vision will be key to ensuring science and society benefit from the opportunities at the chemistry-digital frontier. Bottom-up initiatives, creativity and communities that develop around exciting scientific areas will also be important.

In order to frame and co-design high-level challenges, experts from different areas will need to come together to identify where digital technologies will have the greatest impact, for example in accelerating innovation, increasing efficiency or safety, addressing a societal challenge or opening new frontiers in discovery.

Some perspectives to consider in guiding decision-making about how to prioritise or structure research and innovation initiatives are dimensions like the following, as well as overlaps between them:

- **Breakthroughs and disruption:** What are areas that need urgent disruption and where will digital capabilities accelerate breakthroughs?
- **Complexity:** What are areas where conventional techniques simply cannot handle the levels of complexity involved?
- **Transdisciplinary problems:** What are the challenges that are too big for one lab, company or discipline to solve?
- **Key enabling facilities:** What technologies, platforms and capabilities could have a transformative impact across multiple science discovery and application areas, and for multiple users?

Examples that straddle one or more of these dimensions are:

Sustainable energy including next generation batteries to support increasing global demand for energy storage due to electrification of transport, portable electronics and renewable energy generation from intermittent sources. Also, catalysts to enable water-splitting and CO_2 conversion to produce fuels or materials with low carbon footprint.

New medicines & diagnostics to enable prevention, early detection and treatment of everything from bacterial infections, tropical and emerging diseases to cancer, dementia and obesity.

Predicting and reacting to environmental impacts using multiple, distributed, real-time sensing systems and models. This will involve inter-related technical choices about the kinds of sensors that are needed, how they will be networked, how the data will be used and by who, and how the data from them will be represented and interpreted. All of this will need to be underpinned by modelling, and framed within wider considerations, such as risk, cost, and acceptable levels of uncertainty or thresholds for action.

Tackling the 'plastics problem' in a realistic, long-term way will involve science and technology innovation challenges, including:

- Detailed understanding of how plastics degrade in soil and water, and the impacts of the degradation product on living organisms.
- Design of plastics that biodegrade or are recyclable.
- Models for circulation of mirco and macro scale plastics in global ecosystems.
- New sensors, sensor networks, and data analytics.
- Life cycle assessments and comparisons of products containing plastics and potential alternatives.

This is in addition to business, policy, social and psychology challenges as diverse as design and economics of municipal recycling systems, consumer behaviour and regulatory frameworks.

Key enabling platforms combining infrastructure and expertise in areas like automation of synthesis or formulation, modelling, data-sharing and advanced data analysis or measurement. These can combine centralised and distributed facilities, networked to enable transfer of data and samples from one to another. These platforms would underpin advances across challenge areas as well as in fundamental discovery across multiple fields and sub-fields.

> **One challenge for chemistry is to set up community collaboration and consensus which is required to secure large-scale and long-term investment in important research areas.**

Dr Niklas Blomberg Director, ELIXIR

> **It's important to have a balance between top-down structures and creating communities that draw talent into them because they are working on a problem that is interesting and exciting.**

Prof Charlotte Deane Professor of Structural Bioinformatics, University of Oxford

> **Today there are problems too big for one university, company or even country to solve. The direction of travel is consolidation, identifying commonalities around big questions or platforms that will underpin progress in multiple areas. It's really important for chemistry to do this and that way you create a virtuous circle by attracting and retaining really good people. The message that 'we can solve this problem which is internationally vital and we can collectively do it' is important.**

Prof Jacqueline McGlade Professor of Sustainable Development and Resilience, University College London, and Professor of Public Policy and Governance, Strathmore University

6

Conclusions & what needs to happen

There are many opportunities to nurture and push forward the interfaces between the chemical sciences and digital sciences and technologies in order to deliver impact for society. These include opportunities for individuals, for the chemistry community in partnership with other communities in the physical, life and digital sciences, as well as for research and teaching institutions, companies and funders.

Lifelong training in digital skills

- Continued integration of digital skills into the school curriculum and undergraduate chemistry courses, e.g. maths, statistics, programming, computer science. *[Government, curriculum developers, schools, colleges, universities and research institutes]*
- Lifelong training in digital skills for researchers across career stages. *[Colleges, universities and research institutes, companies, individuals]*
- Identification of future digital skills needs for R&D, and co-creation of teaching and training with educators. *[Companies, chemistry community, universities and research institutes, scientific societies, curriculum developers]*
- Training for researchers on ethical issues in the context of digital technologies. *[Chemistry community, universities and research institutes, companies, funders]*

Roles and career progression for digital experts in research outside digital industries

- Long-term technical and research roles for digital experts in academia, including career progression and development opportunities e.g. data scientists, research software and robotics engineers. *[Universities and research institutes, funders]*
- Mindset and culture change regarding the role of digital professionals in research, including recognition e.g. attributions and references in journal articles, new prizes and awards. *[Chemistry community, universities and research institutes, companies, publishers, scientific societies]*

Fostering multidisciplinary collaborations and communities

- Two-way engagement between academia and industry, and between chemistry and other disciplines, for knowledge and skills transfer e.g. training and secondments. *[Funders, universities and research institutes, companies, individuals]*
- Opportunities to foster multidisciplinary exchange and communities e.g. shared spaces, networks, fora and events. *[Universities and research institutes, funders, scientific societies, individuals]*
- Creating communities for digital experts and users within the chemical sciences and across its interfaces in the physical and life sciences. *[Chemistry community, universities and research institutes, scientific societies]*

Supporting and enabling data sharing

- Proactive sharing of research data, including positive and negative results. *[Individuals, universities and research institutes, companies, publishers and data service providers, funders]*
- Collaborations with digital experts to develop tools and platforms that facilitate data sharing, including making it easy for users and understanding IP, privacy and security concerns. *[Chemistry community, funders, companies, data service providers]*
- Agreed data standards and formats for data. *[Chemistry community, scientific societies, publishers, companies, data service providers, funders]*
- Culture, incentives and mandates that ensure data sharing is the default insofar as is possible. *[Chemistry community, scientific societies, companies, funders, governments, publishers, universities and research institutes]*

Leadership and advocacy for the digital futures of science R&D

- Horizon-scanning and identification of research areas and capabilities where digital will have transformative impacts, in partnership with experts from other science and technology fields. *[Chemistry community, funders, scientific societies]*
- Engagement between digital and chemistry domain experts, and policymakers, to enable better use of digital tools for data-informed policy making. *[Government and policy makers, individual researchers, scientific societies]*
- Horizon-scanning to identify and prioritise future needs for facilities and resources, including business models with access for smaller labs and SMEs. *[Chemistry community, funders, scientific societies, companies]*
- Ethics framework for responsible use of data and digital technologies in science R&D. *[Government and policy makers, companies, universities and research institutes, chemistry community, scientific societies]*
- Articulate benefits and pitfalls of digital technologies in R&D, facilitating debate to respectfully engage with public concerns around the use of digital technologies e.g. ethics, privacy, transparency. *[Chemistry community, scientific societies, universities and research institutes, companies, individuals]*
- Reach out to inspire the next generation of scientists with examples of digitally-augmented and enhanced research. *[Chemistry community, scientific societies, universities and research institutes, companies, individuals]*

ROYAL SOCIETY OF CHEMISTRY

Thomas Graham House
Science Park, Milton Road
Cambridge CB4 0WF, UK
T +44 (0)1223 420066

Burlington House
Piccadilly, London
W1J 0BA, UK
T +44 (0)20 7437 8656

International offices
Beijing, China
Shanghai, China
Berlin, Germany
Bangalore, India
Tokyo, Japan
Philadelphia, USA
Washington, USA

www.rsc.org/new-perspectives

@RoyalSocietyofChemistry
@RoySocChem
@roysocchem
@wwwRSCorg
linkedin.com/company/roysocchem

Registered charity number: 207890