IOPscience iopscience.iop.org

This content has been downloaded from IOPscience. Please scroll down to see the full text.

Download details:

IP Address: 91.62.162.143
This content was downloaded on 29/03/2024 at 12:10

Please note that terms and conditions apply.

You may also like:

Functional Carbon Materials

Scientific Journals on the Net:
Colin Finney, PUS Research Group, Science Museum, London, UK

Special issue on applied neurodynamics: from neural dynamics to neural engineering
Hillel J Chiel and Peter J Thomas

The activities and funding of IRPA: an overview
Geoffrey Webb

Part II

Physics developments to tackling major issues affecting the lives of citizens

IOP Publishing

EPS Grand Challenges

Physics for Society in the Horizon 2050

Mairi Sakellariadou, Claudia-Elisabeth Wulz, Kees van Der Beek, Felix Ritort, Bart van Tiggelen, Ralph Assmann, Giulio Cerullo, Luisa Cifarelli, Carlos Hidalgo, Felicia Barbato, Christian Beck, Christophe Rossel and Luc van Dyck

Original content from this work may be used under the terms of the Creative Commons Attribution NonCommercial 4.0 International license. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, publisher and DOI and you may not use the material for commercial purposes.

Chapter 5

Physics for health

Ralph Assmann, Darwin Caldwell, Giulio Cerullo, Henry Chapman, Dana Cialla-May, Marco Durante, Angeles Faus-Golfe, Jens Hellwage, Susanne Hellwage, Christoph Krafft, Timo Mappes, Thomas G Mayerhöfer, Thierry Mora, Andreas Peters, Chiara Poletto, Juergen Popp, Felix Ritort, Petra Rösch, Lucio Rossi, Marta Sales-Pardo, Iwan Schie, Michael Schmitt, Friedrich Simmel and Aleksandra Walczak

5.1 Introduction

Ralph W Assmann[1,4], Giulio Cerullo[2] and Felix Ritort[3]
[1]Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany
[2]Politecnico di Milano, Milan, Italy
[3]Universitat de Barcelona, Barcelona, Spain
[4]Present affiliation: GSI Helmholtzzentrum für Schwerionenforschung, Darmstadt, Germany

The fundamental research on the physics of elementary particles and nature's fundamental forces led to numerous spin-offs and has tremendously helped human well-being and health. Prime examples include the electron-based generation of x-rays for medical imaging, the use of electrical shocks for treatment of heart arrhythmia, the exploitation of particle's spin momenta for spin tomography (NMR) of patients, and the application of particle beams for cancer treatment. Tens of thousands of lives are saved every year from the use of those and other physical principles. A strong industry has developed in many countries, employing hundreds of thousands of physicists, engineers, and technicians. Industry is designing, producing, and deploying the technology that is based on advances in fundamental physics.

Major research centers have been established and provide cutting-edge beams of particles and photons for medical and biological research, enabling major advances in

© 2024 The Authors.
doi:10.1088/978-0-7503-6342-6ch5

Published under license by IOP Publishing Ltd

the understanding of structural biology, medical processes, viruses, bacteria, and possible therapies. Those research infrastructures serve tens of thousands of users every year and help them in their research. Modern hospitals are equipped with a large range of high technology machines that employ physics principles for performing high-resolution medical imaging and powerful patient treatment. Professors and students at universities use even more powerful machines for conducting basic research in increasingly interdisciplinary fields like biophysics and robotics. New professions have developed involving physicists and reaching out to other domains. We mention the rapidly growing professions of radiologists, health physicists, and biophysicists.

While physics spin-offs for health are being heavily exploited, physicists in fundamental research keep advancing their knowledge and insights on the bio-chemical mechanisms at the origin of diseases. New possibilities and ideas keep constantly emerging, creating unique added value for society from fundamental physics research. This chapter does not aim to provide a full overview of the benefits of physics for health. Instead, the authors concentrate on some of the hot topics in physics- and health-related research. The focus is put on new developments, possible new opportunities, and the path to new applications in health.

Angeles Faus-Golfe and Andreas Peters describe the role of particle accelerators and the use of their beams for irradiating and destroying cancer cells. State-of-the-art machines and possibilities for new irradiation principles (i.e., the FLASH effect) are introduced. As physics knowledge and technology advance, tumors can be irradiated more and more precisely, damage to neighboring tissue can be reduced, and irradiation times can be shortened.

Darwin Caldwell looks at the promise and physics-based development of robotic systems in the macroscopic world, where they are complementing human activities in a number of tasks from diagnosis to therapy. Friedrich Simmel looks at the molecular and cell-scale world and explains how nanorobotics, biomolecular robotics, and synthetic biology are emerging as additional tools for human health (e.g., as nano-carriers of medication that is delivered precisely).

Henry Chapman and Jürgen Popp describe the benefits of light for health. Jürgen Popp is considering the use of lasers that have advanced tremendously in recent years in terms of power stability and wavelength tunability. Modern lasers are used in several crucial roles in cell imaging, disease diagnosis, and precision surgery. Henry Chapman considers the use of free-electron lasers for understanding features and processes in structural biology. He shows that the advance of those electron accelerator-based machines has allowed tremendous progress in the determination of the structures of biomolecules and the understanding of their function.

Aleksandra Walczak, Chiara Poletto, Thierry Mora, and Marta Sales describe physics research against pandemics, a multi-disciplinary problem at the crossing of immunology, evolutionary biology, and networks science. Pandemics are also multi-scale problems at the spatial and temporal levels: from the small pathogen to the large organism; and from the infective process at cellular scale (hours) to its propagation community-wide (months). Simple mathematical models such as SIR (susceptible-infected-recovered) have been a source of inspiration for physicists who model key quantities at an epidemic outbreak, such as the effective reproductive

number R, in situations where a disease has already spread. A prominent example is the recent COVID-19 pandemic that has been more than a health and economic crisis. It illustrates our vulnerability where interdisciplinary and multilateral science play a crucial role addressing a global challenge such as this.

Promise and progress in further diagnostics and therapies are also considered. Lucio Rossi explains the progress in magnetic field strength as it can be achieved with superconducting magnets, while Marco Durante discusses the progress in charged particle therapy for medical physics.

5.2 Accelerators for health

Angeles Faus-Golfe[1] and Andreas Peters[2]
[1]Laboratoire de Physique des 2 Infinis Irene Joliot-Curie,IN2P3-CNRS—Université Paris-Saclay, Orsay, France
[2]HIT GmbH at University Hospital Heidelberg, Germany

Energetic particles, high-energy photons (x-rays and gamma rays), electrons, protons, neutrons, and various atomic nuclei and more exotic species are indispensable tools in improving human health.

The potential of accelerator-reliant therapy and diagnostic techniques has increased considerably over the past decades, playing an increasingly important role in identifying and curing affections, such as cancer, that otherwise are difficult to treat; they also help to understand how major organs such as the brain function and thus to determine the underlying causes of diseases of growing societal significance such as dementia.

5.2.1 Motivation to use and expand x-rays and particle therapy

The use of **x-rays** in radiotherapy (RT) is now the most common method of RT for cancer treatment. While x-ray therapy is a mature technology there is room for improvement. The current challenges are related to the **accurate delivery of x-rays to tumours** involving sophisticated techniques to **combine imaging and therapy**. In particular, the ability to achieve better definition and efficiency in 4D reconstruction (3D over time) distinguishing volumes of functional biological significance. Further technical improvements to reduce the risk of a treatment differs from the prescription and moving towards 'personalised treatment planing' are being made. Some of these techniques such as: image-guided radiation therapy (IGRT), control of the dose administered to the patient (*in vivo* dosimetry) or adaptive RT to take into account the morphology changes in the patient, are the state of the art and are being implemented in the routine operation of these types of facilities. An example is the so-called MR linac, which provides magnetic resonance (MR) and RT treatment at the same time. Finally, the reduction of the accelerator costs and the increase of reliability/availability in challenging environments are also important research challenges to expand this kind of RT in low- and middle-income countries (LMICs).

Low-energy electrons have historically been used to treat cancer for more than five decades, but mostly for the treatment of superficial tumours given their very limited penetration depth. However, this limitation can be overcome if the electron energy is increased between 50 and 200 MeV (i.e., **very high-energy electrons, VHEE;** figure 5.1). With the recent developments of high-gradient normal conducting (NC) radio frequency **(RF) linac technology** (figure 5.2) (CLIC Project n.d.) or even the novel acceleration techniques such as the **laser-plasma accelerator** (LPA) (figure 5.3), VHEE offer a very promising option for anticancer RT. Theoretically, VHEE beams offers several benefits. The ballistic and dosimetry properties of VHEE provide small-diameter beams that could be scanned and focused easily, enabling

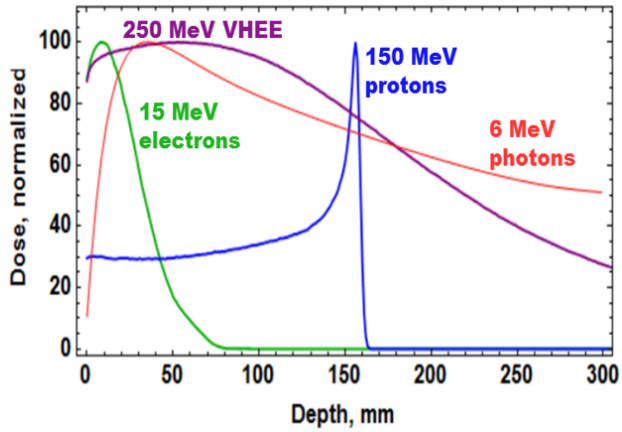


Figure 5.1. Dose profile for various particle beams in water (beam widths $r = 0.5$ cm).

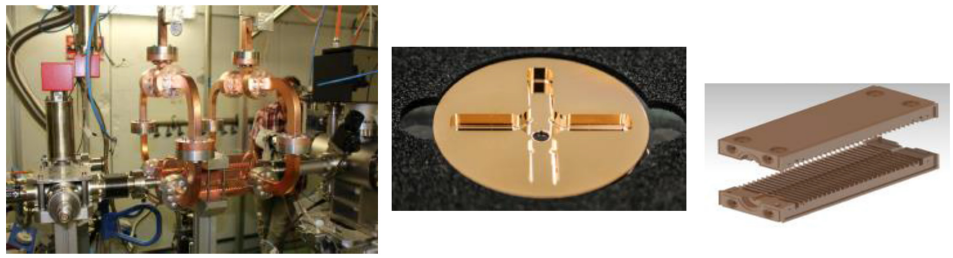


Figure 5.2. CLIC RF X-band cavity prototype (12 GHz, 100 MV m^{-1}).

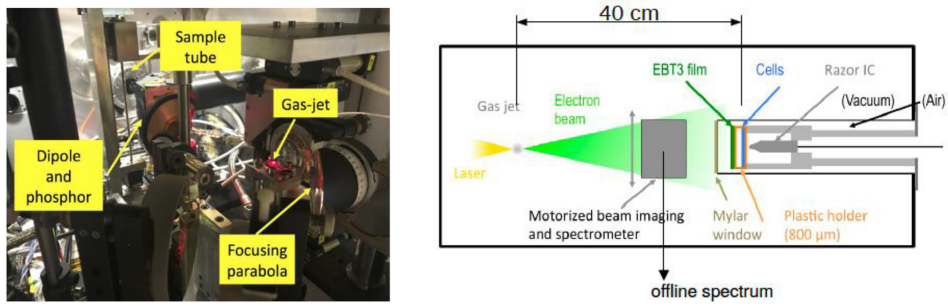


Figure 5.3. Setup of Salle Noire for cell irradiation at LOA-IPP laser-driven wakefield electron accelerator.

finer resolution for intensity-modulated treatments than is possible with photons beams. Electron accelerators are more compact and cheaper than proton therapy accelerators. Finally, VHEE beams can be operated at very high-dose rates and fast electromagnetic scanning providing uniform dose distribution throughout the target and allowing for unforeseen RT modalities in particular the FLASH-RT.

FLASH-RT is a paradigm-shifting method for delivering ultra-high doses within an extremely short irradiation time (tenths of a second). The technique has recently

been shown to preserve normal tissue in various species and organs while still maintaining anti-tumour efficacy equivalent to conventional RT at the same dose level, in part due to decreased production of toxic reactive oxygen species. The 'FLASH effect' has been shown to take place with electron, photon, and more recently for proton beams. However, the potential advantage of using electron beams lies in the intrinsically higher dose that can potentially be reached compared to protons and photons, especially over large areas as would be needed for large tumours. Most of the preclinical data demonstrating the increased therapeutic index of FLASH has used a single fraction and hypo-fractionated regimen of RT and using 4–6 MeV electron beams, which do not allow treatments of deep-seated tumours and trigger large lateral penumbra (figure 5.4). This problem can be solved by increasing the electron energy to values higher than 50 MeV (VHEE), where the penetration depth is larger.

Many challenges, both technological and biological, have to be addressed and overcome for the ultimate goal of using VHEE and VHEE-FLASH as an innovative modality for effective cancer treatment with minimal damage to healthy tissues.

From the accelerator technology point of view the major challenge for VHEE-RT is the demonstration of a suitable high-gradient acceleration system, whether conventional, such a X-band or not, with the stability, reliability, and repeatability required to be operated in a medical environment. In particular, for the VHEE-FLASH is the delivery of very high dose rate, possibly over a large area, providing uniform dose distribution throughout the target (Faus-Golfe 2020).

All this asks for a large beam test activity in order to experimentally characterize VHEE beams and their ability to produce the FLASH effect and provide a test bed for the associated technologies. It is also important to compare the properties of the electron beams depending on the way they are produced (RF linac or LPA technologies). Preliminary VHEE experimental studies have been realized using NC RF accelerator facilities in NLTCA at SLAC and CLEAR at CERN, giving very promising results for the use of VHEE electron beams. Furthermore, some

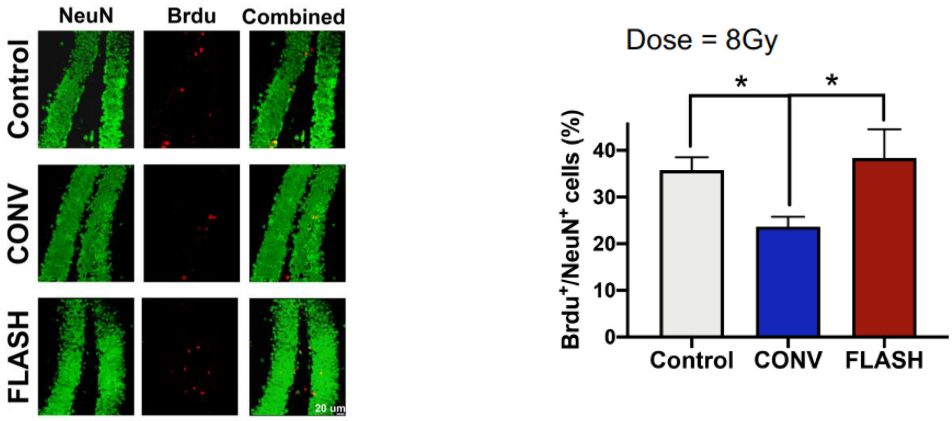


Figure 5.4. FLASH preservation of the neurogenic niche in juvenile mice (courtesy of C Limoli).

experimental tests have been carried out with laser-plasma sources at ELBE-DRACO in HZDR, at LOA in IPP, and at the SCAPA facility at the University of Strathclyde. In particular for the FLASH, experimental studies have been realized at low energies with the Kinetron linac at the Institute Curie at Orsay and the eRt6-Oriatron linac at CHUV and at very high energies at CLEAR at CERN.

Proton and ion beam therapy has growing potential in dealing with difficult-to-treat tumours, for example, because of the risk of damaging neighbouring sensitive tissues such as the brainstem or visual nerves in the case of head tumour treatments. Also, some treatments may benefit from the use of particles that deliver doses with greater radiobiological effectiveness (RBE) and higher local precision, notably carbon, and in the near future also helium ions.

Recent investigations using ultra-short and ultra-high dose rates (called FLASH) of electron beams showed growth retardation of tumours with the same effect as in conventional therapy, but with minimized impact to the surrounding tissue. FLASH with proton and ion beams is expected to offer additional healthy tissue sparing from beam stopping in the tumour—but the research on this topic is still not completed, and the experiments and evaluations are ongoing. Healthy tissue sparing with FLASH would enable a dose increase as well as a significant reduction of treatment time without additional aggravations. These new key findings may influence the accelerator development for particle therapy considerably in the next future.

In x-ray radiotherapy the integration of imaging devices for image-guided radiation therapy is standard. In-room CTs and now the introduction of magnetic resonance imaging (MRI) in the form of the recent clinical adoption of the MR linac show these necessary advances for better positioning of the patients and online observation of the tumours. Whereas CTs are also introduced in particle therapy facilities, the combination of MRI and hadron beams is delicate due to the high magnetic fields of these diagnostic devices diverting the proton/ion beam. In

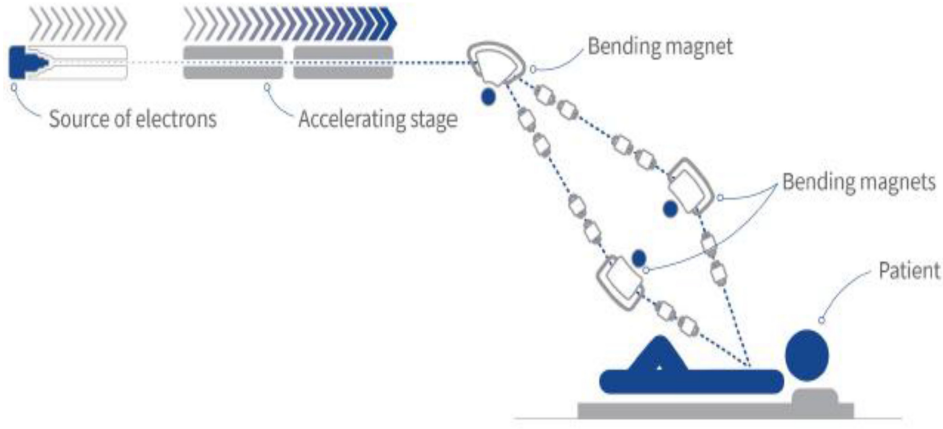


Figure 5.5. FLASH facility cartoon for CHUV—Lausanne.

addition, also the stray fields may lead to effects on the treatment systems in the near future. The investigations on this topic have started (see below).

In the last three decades about 25 proton and 4 hadron therapy facilities were built in Europe (figure 5.6). But only a few more are actually under planning or construction, among them only one hadron facility project (SEEIIST[1]). Thus, new



Figure 5.6. Particle therapy facilities in Europe; see https://www.ptcog.ch/index.php/facilities-world-map. SEEIIST—The South East European International Institute for Sustainable Technologies http://seeiist.eu/.

[1] https://seeiist.eu/; SEEIIST—The South East European International Institute for Sustainable Technologies.

efforts are required to make these techniques smaller, cheaper (in investment and operating costs), and easier to maintain, which will be discussed in a separate chapter below showing potentials for the next three decades.

5.2.2 Further developments in the next decades

5.2.2.1 Introduction of helium for regular treatment

For protons and 3He/4He similar radio-biological properties have been determined, but the lateral scattering is reduced by nearly 50% in the case of helium ions versus protons (figure 5.7). In recent years, helium ions again became of interest for clinical cases where neither protons nor carbon ions are ideally suited, especially for treating paediatric tumours. Currently, patient irradiations with scanned ^{4}He ions at the Heidelberg Ion beam Therapy Center (HIT) in Germany are used only in 'treatment attempts' ('individuelle Heilversuche') and will go into regular operation by mid 2024. Other ion therapy facilities in Europe (e.g., CNAO in Italy and MedAustron in Austria) have also started technical upgrades to produce helium ion beams in the near future.

Recent studies have shown that 3He ions can be a viable alternative to ^{4}He, as they can produce comparable dose profiles, demanding slightly higher kinetic energy per nucleon, but less total kinetic energy. This results in 20% less magnetic rigidity needed for the same penetration depth which may be of importance for the design of future compact therapy accelerators like superconducting synchrotrons or energy-variable cyclotrons.

5.2.2.2 Image-guided hadron therapy using MRI

The observation of the patient's position during treatment has become more and more a standard procedure. 3D camera systems make it possible to control the location of the patient within tenths of millimetres observing the exterior of the body. But organs struck by tumours can move dramatically within the abdomen,

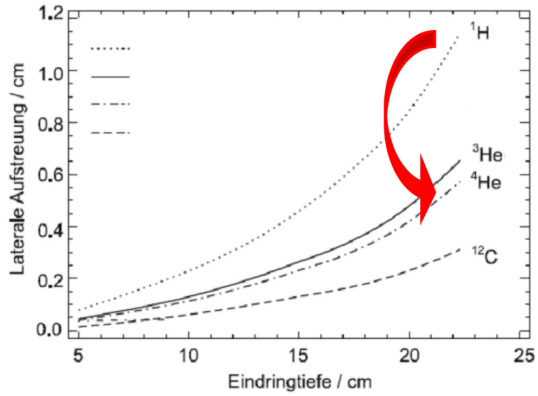


Figure 5.7. Lateral scattering of different light ions, adapted from Fiedler (2008).

which is not visible from outside. To observe the soft tissues an MRI scanner is the preferred diagnostic tool. But due to the moderate to high fields used the charged particle beam is affected. In addition, conventional MRI scanners are not constructed to have an inlet for radiation to be applied in parallel to the diagnostic procedure. Nevertheless, studies like ARTEMIS at HIT in Heidelberg have been started to investigate possible arrangements of open, low-field MRIs (figure 5.8). Still the deflection and/or distortion of the beam can be observed, but algorithms will be developed to cancel out this influence of the MRI's magnetic field. A second goal of such studies is to optimize the design of the measurement coils to spare out space for the beam entrance fields. A further aspect should not be underestimated—the impact of the magnetic (stray) fields on the QA devices and the online monitoring detectors used during treatment. All these have to be examined and possibly adapted to be magnetic field compatible; at worst new or alternative detector principles have to be applied. A lot of technical solutions have to be found in the next years before MRI scanners can be introduced in regular particle beam therapy.

5.2.2.3 The future of compact accelerator concepts

To enhance the coverage of particle therapy in Europe and worldwide and to enlarge the number of patients that can profit from this special treatment, the investment costs of such facilities should be reduced as much as possible, which requires smaller and simpler machines to reduce manufacturing and operating effort. Especially the size of the accelerator has an important influence on the building costs. And the amount of beam losses demands more or less concrete for radiation shielding, and



Figure 5.8. Setup of an MRI scanner from Esaote at HIT's experimental place.

should be minimized by design. In addition, the operating and maintenance team needed should be small, but adequate and well-trained, and sustained by a modern control system, which predicts pre-emptive maintenance measures through AI algorithms and thus guarantees highest availability.

Energy-variable cyclotrons

Most proton therapy facilities use cyclotrons to produce the beam, as these accelerators are compact, have only a few tuning parameters, and are thus simple to operate. Over the last three decades the size and weight have dramatically shrunk by factors 3–10 while the magnetic fields using superconductivity were increased up to a factor 4 (figure 5.9). These improvements still show the potential of cyclotrons while parallel developments like proton linacs are still larger and have a much more complex technique. But the actual cyclotron generation consists still of fixed energy machines, which demands a degrader for energy variation causing two main drawbacks: (a) high local beam losses and (b) relatively low currents for mid-depth and skin-deep tumours, which may be a problem to use the FLASH-based treatment procedure. To overcome these disadvantages first studies of energy-variable cyclotrons were undertaken and published. These show the possibilities to consequently advance the cyclotron systems to lighter, high-field superconducting arrangements, but now iron-free and thus capable to vary the magnetic field and energy in reasonable times. In addition, the produced currents of such a setting do not depend on the energy because no loss mechanism is involved anymore. The resulting machines would be 'FLASH'-ready and have the additional possibility to be enhanced to ^{3}He/^{4}He beam combined with proton therapy in one cyclotron setup. More simulation studies and prototype constructions are needed to reach these goals in the next 5–10 years.

Superconducting fast-ramped synchrotrons

Wherever the flexibility of different ions from protons, helium, carbon to oxygen ions is asked to have the ability to study different dose distributions and linear energy transfer (LETs), a flexible accelerator concept with a high bandwidth of magnetic rigidities is needed. Today small synchrotrons with circumferences of 60–80 m and iron-dominated normal-conducting dipoles and quadrupoles are built for such purposes. Together with the injector linac and a variety of ion sources on one

Table 1 Comparison of Main Parameters of Several Commercial Cyclotrons

	Mevion S250	IBA S2C2	Varian Proscan	IBA C230
Magnet Type	Superconducting	Superconducting	Superconducting	Copper
R pole (m)	0.34	0.50	0.80	1.05
D Yoke (m)	1.80	2.50	3.10	4.30
Height (m)	1.20	1.50	1.60	2.10
B_o (T)	8.9	5.7	2.4	2.2
B_f (T)	8.2	5.0	3.1	2.9
Mass (mt)	***25***	***50***	***100***	***250***
T_f (MeV)	254	230/250	250	235

Figure 5.9. Comparison of main parameters of several commercial cyclotrons; see 'Compact, low-cost, lightweight, superconducting, ironless cyclotrons for hadron radiotherapy', PSFC/RR-19-5.

side and all the high-energy beam transport lines on the other side such a facility has a large required space in contrast to the treatment rooms. As a result, the building costs are high and such facilities were set up only in combination with large university hospitals. To shrink the size of these accelerator arrangements the main idea is to equip the synchrotron with superconducting magnets. This is a big challenge, because curved 3–4 T magnets ramped at 1–2 T s^{-1} are needed for an efficient operation. Studies at Toshiba in Japan and at CERN (NIMMS, HITRIplus[2]) are underway to prepare prototype magnets.

In addition, modern control systems providing the multiple energy extraction method—using several post-accelerations in the same synchrtotron cycle—have to be developed to enhance the duty cycle of such machines. A new approach using time-sensitive networking (TSN), the next generation of (real-time) ethernet in industry, will be implemented at HIT (partly within HITRIplus) in the next years.

However, an open question still exists: Can synchrotrons be used later on for FLASH therapy, as the particle filling is limited by space charge restrictions? In addition, the necessary high dose rates would demand short extraction times (or fast extraction methods) with rapid refilling of the synchrotron and short cycle times (see above). These are all big challenges.

The first proton and ion linac-based facilities

After several years of R&D and developments in research at international laboratories (CERN, TERA, ENEA, INFN, ANL) the first linac-based proton therapy facilities are under construction and commissioning in UK (STFC Daresbury) and in Italy. With the use of high-frequency copper structures, designed to achieve relatively compact solution and high repetition rate operation, linacs will allow the production of beams with fast energy variation (without the need for mechanically moved beam energy degraders), as well as small emittance beams that are potentially suited for the further development of mini-beam dose-delivery techniques. The shielding around the linac can also be reduced compared to other installations allowing a more flexible solution that could be installed in existing buildings and other restricted space settings. Also, dedicated designs for He and C ions are being studied, which would require less power compared to presently operating synchrotrons and allow for flexible pulsed operation. HG RF technology as well as high-efficiency klystrons are key developments for the future and further spread of this approach. Furthermore, the linac technology is attractive as a booster option to increase the output energy of cyclotron-based existing facilities.

Towards a VHEE RT facility

NC RF linac is the technology being used for most of the VHEE research. The main advantages of the linacs are the flexibility and the compactness. Regarding the linac design in the energy range of interest for VHEE applications there are different possibilities offering the desired performances and compactness with different

[2] **NIMMS** (Next Ion Medical Machine Study). https://www.hitriplus.eu/ Heavy Ion Therapy Research Integration plus.

degrees of technology maturity. The S-band technology is the most mature one; HG compact linacs of this type are already available from various industrial partners. The C-band and X-band RF linacs are still less mature and are mainly constructed in labs with the help of industries for machining. Recently a considerable effort is being made from the industrialization point of view. The current and next future available machines for VHEE are the eRT6-Oriatron at Centre Hospitalier Universitaire Vaudois in Laussane; ElectronFlash at IC in Orsay; CLARA at Daresbury; AWA at ANL; and CLEAR at CERN (all based on NC RF linacs). A VHEE-FLASH facility based on a CLIC X-band 100 MeV linac is being designed in collaboration with CHUV to treat large, deep-seated tumours in FLASH conditions. The facility is compact enough to fit on a typical hospital campus (figure 5.5). Another proposal in this sense is the upgraded PHASER proposal at SLAC. Finally, ELBE at HZDR and the next future PITZ at DESY are based on SCRF linac technology.

Recent advances in the high-gradient RF structures where more than 100 MeV m^{-1} are now achievable in the lab environment are transforming the landscape for VHEE RT. VHEE RT requires beam energies between 50 and 200 MeV, an improved dose conformity and scale to higher doses rates, in the case of the FLASH-RT until 50 Gy s^{-1} are needed. Novel high-gradient technologies could enable ultra-compact structures, with higher repetition rates and higher currents. An international R&D global effort is being made by major accelerator laboratories and industry partners and is focused on two aspects: material origin and purity, surface treatments, and manufacturing technology on the one hand and the consistency and reproducibility of the test results on the other. Some promising R&D in the next decade are the distributed coupling accelerator developed at SLAC and the use of cryogenic copper that is transforming the linac design offering a new frontier from beam brightness, efficiency, and cost capability. Another approach for the next generation of compact, efficient, and high-performance VHEE accelerator is the use of higher-frequency millimetric waves (~100 GHz) and higher repetition rates using THz sources.

An important R&D effort to apply these technologies in the medical industry has to be made in the next decade, if successful, this could be a step further in the quest for compact and efficient VHEE RT in the range of hundreds of MeV. For achieving these aims, a synergistic and multidisciplinary research effort based on accelerator technology as well as physical and radiobiological comparisons to see how well VHEE can meet the current assumptions and become a clinical reality is needed (Very High Energy 2020).

Therapy facilities based on laser plasma acceleration

As high-performance lasers have increased greatly in recent years in terms of power and repetition rate their use for particle therapy may be possible in the future. The actual limit of about 100 MeV achieved for the highest proton energies driven by ultra-intense lasers using Target Normal Sheath Acceleration (TNSA) depicts a major milestone on the way to the needed energies. But still the broad energy spread of the accelerated protons is not feasible for treatment modalities. The reached

energies for laser-accelerated ions is still a magnitude lower and thus far from the necessary values. The very short dose peaks may be attractive for FLASH therapy, but then the repetition rate of the Petawatt lasers should reach 100 Hz and more, which is not the case today. In addition, the target configuration has to resist this high load on a long-term basis—a therapy facility runs several 1000 h a year. Furthermore, the reliability of a laser-based proton or ion accelerator must reach 98% or more to be of practical use in a medical facility. But this technique should be explored with high effort in the next decade to identify the long-term potential.

Concerning the VHEE there is an intense R&D effort in LPA to be applied in the next generation of VHHE-RT facilities. The major challenge for the LPA technique is the beam quality, reproducibility, and reliability needed for RT applications. This R&D is being carried out in facilities such as the DRACO at ELBE at Helmholtz-Zentrum Dresden-Rossendorf (HZDR) and in the Laboratoire d'Optique Appliquée (LOA) at Institute Polytechnique de Paris (IPP) where a new beamline dedicated to VHEE medical applications known as IDRA is being constructed. The new beamline will provide stable experimental conditions for radiobiology and dosimetry R&D (Very High Energy 2020).

A wide international R&D programme, in particular we highlight the role of the EU network EUPRAXIA (http://www.eupraxia-project.eu/), will be needed in the next decade in order to convert these 'dream' facilities into reality.

5.3 Bionics and robotics

5.3.1 Bioinspired micro- and nanorobotics

Friedrich Simmel[1]
[1]Technische Universität München, Garching, Germany

5.3.1.1 General overview

Robotic systems transform the way we work and live, and will continue to do so in the future. At the macroscopic scale, robots greatly speed up and enhance manufacturing processes, provide assistance in diverse areas such as healthcare or environmental remediation, and perform robustly in environments that are inaccessible, too harsh, or too dangerous for humans. In the lab they can, in principle, perform large numbers of experiments in parallel, reproducibly and without getting tired. This supports, for example, the search for new chemical compounds in combinatorial approaches, and can also generate the large datasets required by data-hungry machine learning techniques.

Robotic systems are 'reprogrammable multifunctional manipulators' and typically comprise sensors and actuators connected to and coordinated by an information-processing unit. Sensors provide information about the environment, which is evaluated by a computer and then used to decide on the necessary actions—which often means mechanical motion of some sort. At a macroscopic scale, a wide range of sensors, electromechanical components, and powerful—potentially networked—electronic computers are available to realize robotic systems, which are essentially all powered by electricity. Is it possible to realize robotic functions also at the micro- or even nanoscale, where we have to work with molecular components, and onboard electronics is not available?

In fact, over the past years researchers have begun to work on the development of 'molecular robotic systems', in which sensors, computers, and actuators are integrated within molecular-scale systems. Among the many possible applications for molecular-scale machinery and robots are, most prominently, the generation of nanomedical robots that autonomously detect and cure diseases at the earliest stages, and the generation of molecular assembly lines that will enable the programmable synthesis of chemical compounds.

Biology as a guide for molecular and cell-scale robotics

Biology has inspired the development of robotic systems at the macroscale in manifold ways—many robot body plans are derived from those of animals (humanoid, dog, insect-like robots, etc), and the movements and actions of these robots resemble those of their living counterparts. Roboticists are concerned with 'motion planning', 'robotic cognition', etc, and therefore ask similar questions as neuroscientists. The field of swarm robotics is inspired by the observation of social behavior in biology.

But also at the cellular and molecular scale we can find inspiration for robotics—cells, like robots, have sensors and actuators; they store and process information,

they move, manufacture, interact with other cells, they can even self-replicate (which robots cannot, so far). To name but a few, specific examples for biological functions that are of direct relevance for molecular robotics are protein expression, molecular motors, bacterial swimming and swarming, chemotaxis, cell shape changes and the cytoskeleton, cell-cell communication, the immune system, muscle function, etc.

What's important is that biology is a very different 'technology' than electro-mechanics and electronic computers—biological systems are self-organized chemical systems far from thermal equilibrium. If we want to build bioinspired robots at this small scale, we will have to apply other principles than those developed for macroscale robotics.

From molecular machines to robots

Over the past decades, one of the major topics in biophysics (and in supramolecular chemistry as well) was the study of molecular machines and motors. Research in this field has clarified how machines operate at the nanoscale and how they differ from macroscopic machines. The small size of the machines changes everything—these machines operate in a storm of Brownian motion, in which the forces they are able to generate are small compared to the thermal forces. Motion sometimes is achieved via a ratchet mechanism that utilizes and rectifies thermal motion (which is only possible out of equilibrium); in some cases it also involves a power stroke, in which a chemical reaction (ATP hydrolysis) effects a conformational change in the machine or motor, biasing its movement in one direction.

Research on biological molecular machines informs molecular robotics in several ways. First, these machines have taught us how they work (at least in principle), and have given examples for what they can achieve (e.g., transport molecules from point A to B, synthesize molecules, exert forces, pump ions/molecules across membranes, etc). They also have indicated speeds ($\mu m\ s^{-1}$) and forces (piconewton) that can be achieved with molecular machinery. Their spatial organization often plays a role in their function (e.g., huge numbers of myosin molecules acting together in muscle, ATP synthase or flagellar motors embedded in membranes, etc) and thus needs to be controlled.

Even if the construction of powerful synthetic molecular machines will not succeed in the near future, experimental work in biology and biophysics has provided protocols that allow the extraction, purification, and chemical or genetic modification of biological machines and their operation in a non-biological context. Many researchers have already started to harness biological motors such as kinesin and myosin in an artificial context (e.g., for molecular transport, biosensing, agent-based computing, as active components of synthetic cell-like structures or synthetic muscles). Other molecular machines such as ATP synthase have been used to power biochemical reactions within synthetic cells.

But then, not every machine or motor should be called a robot. What we expect from a molecular robot is the programmable execution of multiple and more complex tasks (as opposed to non-programmable, repeated execution of always the same task), potentially with some sort of decision-making or context-dependence. This likely will require the combination of multiple molecular components into a

consistent system that can be continuously operated or driven out of equilibrium. This requirement also defines one of the major challenges for the field, namely systems integration of molecular machines and other components to perform useful tasks, which also comes with challenges for energy supply and interfacing with the environment or non-biological components.

DNA-based robots

DNA molecules turn out to be ideal to experimentally explore ideas in nanoscale biomolecular robotics. DNA intrinsically is an information-encoding molecule, based on which a wide variety of schemes for DNA-based molecular computing have already been developed. Further, DNA nanotechnology—notably the so-called 'DNA origami' technique—has enabled the sequence-programmable self-assembly of almost arbitrarily shaped molecular objects. Various chemical and physical mechanisms have been employed to switch DNA-based molecular objects between different conformations, and to realize linear and rotary molecular motors. Thus, in principle, all the major functional components of a robotic system—sensors, actuators, and computers—can be realized with DNA alone.

As mentioned, in order to realize robot-like systems, these separate functions have to be integrated into consistent multifunctional systems. However, only few experimental examples have convincingly demonstrated such integration so far. In one example by Nadrian Seeman and coworkers, a 'molecular assembly line' was shown to be capable of programmable assembly of metallic nanoparticles by a molecular walker. The walker could collect nanoparticles from three assembly stations, which were controlled to either present a nanoparticle or not. This resulted in a total of $2^3 = 8$ different assemblies that could be 'programmably' realized with the system. In the context of nanomedicine, origami-based molecular containers have been realized that open up only when certain conditions—e.g., the presence of certain molecules on the surface of cells—are met. The containers can then present previously hidden molecules that trigger signaling cascades in the cells, or release drugs. This also exemplifies the two main fields of application that have been envisioned for DNA robots: programmable molecular synthesis with molecular 'assembly lines', and the delivery of drugs by nanomedical robots.

While these prototypes are extremely promising, many challenges remain—and some of these probably pertain also to nanorobots realized with other molecules than DNA. First, information-processing capabilities of individual molecular structures are quite limited—essentially, they are based on switching between a few distinct states (i.e., of similar energy), but separated by a 'high enough' activation barrier), which means that their computational power should be similar to that of finite state machines. Second, current instantiations of DNA robots are quite slow (movements with speeds of nm s^{-1} rather than μm s^{-1}), and do not allow fast operation or response to changes in the environment. Thirdly, molecular robots are, of course, small, which poses a problem in many instances, where we might want to integrate them into larger systems, let them move across larger length scales, and operate many of them in parallel.

Active systems

One of the major visions in bioinspired nano- and microrobotics is the realization of autonomous behaviors, which, as a subtask, involves autonomous motion. In this context, over the past decade there has been huge interest in the realization and study of active matter systems, which include self-propelling colloidal particles, and active biopolymer gels actuated by ATP-consuming molecular motors. In contrast to other, more conventional approaches based on manipulation by external magnetic or electric fields, such systems promise to move 'by themselves' and also display interesting behaviors such as chemotaxis or swarming.

As before, autonomously moving particles or compartments *alone* will not make a robot, and again the challenge will be to integrate such active behavior with other functions. For instance, it would be desirable to find ways to control and program active behavior—the output of a sensor module could be used to control a physicochemical parameter that is important for movement. Active particles that move in chemical gradients need to be asymmetric, and potentially this asymmetry (or some symmetry-breaking event) could be influenced by a decision-making molecular circuit. Another challenge will be to find the 'right chemistry' that allows active processes and other robotic modules to operate under realistic environmental conditions—(e.g., inside a living organism).

Collective dynamics and swarms

If single nanorobots are unavoidably slow and rather dumb, maybe a large collection of such units can do better? In order to overcome their limitations, a conceivable strategy is to couple large numbers of robots via some physical or chemical interaction and let them move, compute, operate collectively. First steps in this direction have been taken by emulating population-based decision-making processes such as the 'quorum sensing' phenomenon known from bacteria, or swarm-like movement of microswimmers. Ideally, for a robotic system one would like to be able to couple such collective behaviors to well-defined environmental inputs and functional outputs—(e.g., if there is light, self-organize into a swarm, move towards the light source and release (or collect) molecules, etc).

There are various challenges associated with these ideas. How can one program the behavior of a swarm? As swarming depends on the interactions between the constituting particles/robots and their density, one could think of changing these control parameters in response to an external input. 'Programming' the behavior of such dynamical systems would mean to choose between different types of behaviors that are realized in different regions of their phase space. Potentially, the behavior of a whole collection of particles could be influenced by a single or a few particles (leader particles) and programming the swarm would amount to programming or selecting these leaders.

Cells as robots

As mentioned above, biological cells really behave a little like microscale robots. Biology has tackled the 'systems integration' challenge and realized out-of-equilibrium systems, in which various functionalities play together, behave in a

context-dependent manner, and which are controlled by genetic programs. From a robotics perspective we can therefore ask whether we can (i) build synthetic systems that imitate cells but perform novel functions and thus act as cell-scale soft robots, or (ii) engineer extant cells to become more like robots?

Essentially the same approaches to engineer biological systems are pursued in synthetic biology, and they come with the same challenges. Regarding the first ('bottom-up') approach—putting together all the necessary parts to generate a synthetic living system is yet another systems engineering challenge. In order to realize synthetic cells, metabolic processes need to be compartmentalized and coupled to information processing, potentially growth, movement, division, etc, which has not succeeded so far (cf the separate EPS challenge section 4.4). The second approach circumvents the challenge of realizing a consistent multifunctional molecular system, but engineering of extant cells is difficult due to the sheer complexity of these systems. Engineered modules put additional load on a cell (whose exclusive goal is to self-sustain, maybe grow and divide), which compromises their fitness, and they also often suffer from unexpected interactions with other cellular components (also known as the 'circuit-chassis problem' in synthetic biology).

Power supply

A major issue that has to be tackled for all of the approaches described above is power supply. How are we going to drive the systems continuously to generate robotic behaviors? Cells come with their own metabolism, which means that cell-based robots would simply have to be fed in the same way as cell or tissue cultures. In the absence of a metabolism, however, molecular robotic systems driven by more complex chemical fuels such as ATP or nucleic acids probably will need to be supplied with these fuels using fluidics. When operation inside of a biological organism is desired, biologically available high-energy molecules might be used as fuels.

In the context of nano- and microrobotics, external driving with magnetic or electric fields via light irradiation or heating is heavily investigated. Here one challenge is to convert these globally applied inputs into local actions—potentially by some local amplification mechanism (e.g., plasmonic field enhancement) and/or by combining actuation with molecular recognition or computation that lead to action only when certain conditions are met. So far, in most cases externally supplied energy was used for mechanical actuation (e.g., motion or opening/closing of containers), but only rarely to power complex behaviors or information processing.

In some cases it is not yet clear whether the energy balance will work out—depending on the mechanism employed and the efficiency of the robots, it may not be possible to deliver enough power to small systems to enable movement or more complex behaviors in the presence of overwhelming Brownian motion and other small-size effects. For instance, rotational diffusion of nanoparticles can be too fast to allow for directional movement, energy dissipation in aqueous environments may be too fast to heat up nanoscale volumes, etc.

Hybrid systems

In light of the limitations in our ability to generate autonomous biomolecular robots with similar capabilities as macroscopic robots, a realistic and potentially powerful approach will be to focus on hybrid systems, in which non-autonomous molecular systems are combined with already established robotic (electromechanical) systems. Such an approach would combine the advantages of the different technologies involved—bionanotechnology and synthetic biology on the one hand, and electronic computing and electromechanical actuation on the other. For instance, it will be very hard if not impossible to outcompete electronic computers using the limited capabilities of molecular systems alone. On the other hand, biomolecular systems and/or cells are at the right scale and 'speak the right language' to interact with other molecular/biological systems.

It is obvious that one of the major challenges for this approach is interfacing—finding effective ways to transduce biological into electronic signals, and vice versa. In one direction this coincides with the well-known challenges involved in biosensing, for which a biological signal (the presence of biomolecules) needs to be converted into electrons or photons (which also can be further converted into electrons). In the robotics context, an additional requirement will be the speed of the sensing event, which has to be quick enough to allow responding to changes in the environment in which the robot operates.

In the other signaling direction, efforts have been made to control molecular machines with light, electric, or magnetic fields, which enables potentially fast and computer-controlled external manipulation of these systems. Also various attempts have been made to influence the behavior of cells using external stimuli, notably in areas such as neuroelectronics or optogenetics. Similar approaches might be adopted to control bio-based micro- and nanorobots.

In the case a bidirectional biointerface has been successfully established, one can imagine hybrid robotic systems, in which computer-controlled signals direct the behavior of the biomolecular part of the robot, and sensory information is fed back to the computer, enabling the implementation of feedback or more complex control mechanisms—in this approach, micro- and nanorobots will be all sensors and actuators, with a brain outsourced to an external electronic computer.

Applications envisioned for micro- and nanorobotics

Micro- and nanorobots will be used when a direct physical interaction with the molecular or cellular world is required. As already mentioned above, the main application envisioned for such robotic systems will be in nanomedicine. One instantiation of nanomedical robots are advanced delivery vehicles that can sense their environment, release drugs on demand, or stimulate cell-signaling events. They may potentially be equipped with simple information-processing capabilities that can integrate more complex sensory information (e.g., to evaluate the presence of a certain tissue, cell type, and thus location in the body), and which may also be used to evaluate diagnostic rules based on this information (such as 'if condition X is met, bind to receptor Y, release compound Z', etc). Given the limited capabilities of small-scale systems, it is not clear how programmable such robotic devices will be. It

is well conceivable, however, to come up with modular approaches, in which the same basic chassis is modified with different sensors and actuators, depending on the specific application. Autonomous robots will have to find their location by themselves, which for some applications may be achieved by circulation and targeted localization in the organism. Alternatively, hybrid approaches are conceivable, and allow for active control from the outside (e.g., by magnetic or laser manipulation, depending, of course, on the penetration depth of these stimuli in living tissue). There are many additional challenges for such devices, which are similar to those for conventional drugs (e.g., degradation, allergenicity, dose, circulation time, etc).

Apart from nanomedical robots, for which the first examples are already emerging, a wide range of applications can be envisioned in biomaterials and hybrid robotics. Hybrid robots could use a biomolecular front end that acts as sensor for (bio)chemicals and actuator that allows the release or presentation of molecules (here the overall robot would not be microscale). Potentially, surfaces or soft bulk materials (such as gels) could be modified with robotic devices, resulting in novel materials that can be programmed and change their properties in response to environmental signals in manifold ways (resulting in materials that are smarter than 'smart materials').

One of the most fascinating applications would be programmable, molecular assembly lines. Rather than aiming for a universal assembler, a more modest and achievable goal is the programmable assembly of a finite number of possible assembly outcomes starting from a defined set of components—similar to the DNA-based molecular assembly line mentioned above. This is not unlike a macroscopic assembly line, which is optimized for the production of one defined product with optional variations. Notably, biological processes such as RNA or protein synthesis already look a little like programmable assembly: RNA polymerase and ribosomes read off instructions from a molecular tape (DNA and mRNA, respectively), and use them to assemble other macromolecules. Maybe similar systems can be conceived that allow sequence-programmable synthesis of non-biological products.

In any case, molecular assembly means control over chemical reactions, which also means that we cannot ignore basic chemical rules and simply synthesize anything we want. Another issue is the scale of the process: in order to synthesize appreciable quantities of molecules or molecular assemblies, large numbers of assembly robots will have to be embedded into the active medium of a synthesis machine (similar, maybe, to a DNA or peptide synthesizer), where they can synthesize larger quantities in parallel.

5.3.1.2 Challenges and opportunities

In the sections above, a wide range of challenges and opportunities associated with the development of future bioinspired micro- and nanorobots were mentioned, which are summarized more concisely here. A variety of challenges relate to the way robotic functions will be implemented in the first place:

- Some robotic functions will be realizable with nanoscale systems (composed of supramolecular assemblies, DNA structures, biological motors); others will require larger cell-like systems into which multiple functions are integrated. It will be important to understand the size dependence of these functions and clarify what can be achieved at which length scale, and with which components. Systems integration—combining the components into consistent and functioning systems—will be *the* major challenge for bioinspired small-scale robotics.
- Autonomous behavior versus external control. How complex do robotic systems need to be to act autonomously—and is autonomy required? For many applications, external control will be a more feasible and powerful approach, which also benefits from established macroscale technologies.
- Hybrid robots that combine the advantages of biological systems and traditional robotics seem promising—a major challenge is the realization of efficient interfaces for signal transduction and actuation between biological and conventional robotics.
- Robotic systems should perform useful tasks, which means they need to be operated under realistic conditions. Depending on the field of application, various practical issues will have to be tackled—operation inside a living organism or on the surface of a sensor chip come with very different requirements.

Many physical challenges arise from the question of whether we can create something like a robot at such small scales at all:

- Can small robots sense and respond to small numbers of molecules or photons? There are biological examples of extremely sensitive sensors that require special molecular architectures and amplification processes, which may guide the design of microrobotic sensors. A related question that has been studied extensively in biology is chemotaxis, where cells sense the presence of a concentration gradient.
- What forces can be usefully applied by small robots? Biological molecular motors are known to generate piconewton forces, and the same magnitude is also expected from artificial motors. However, synthetic molecular motors have not yet been used to perform any useful task.
- How fast can a nano- or microrobot move or act? Do we need to achieve directional movement or will diffusion be sufficient? Of course, this again will depend on the application. Interestingly, in biology we do not find molecular motors for transport of molecules in bacteria (diffusion is sufficient), but in the much larger eukaryotic cells. Further, the smallest bacteria tend to be non-motile—apparently motility only pays off above a certain size. There are various further issues such as the dominance of Brownian effects at the nanoscale.
- How much energy is required to perform robotic functions—and how will it be supplied? Should nano- or microrobots be autonomously driven by chemical reactions, or actuated externally via physical stimuli?

- What is the computational power of small robots? How much information can be stored, how fast can it be processed? As the computing power of small systems necessarily is very limited, nanorobotics means 'robotics without a brain'. Nanorobots will tend have 'embodied intelligence', in which input-output relationships between sensor and actuator functions are custom-made and hardwired rather than freely programmable.
- Can one realize collective behaviors of large numbers of nanorobots—and how can one 'program' such systems?

Ultimately, the development of micro- and nanorobotics will lead to advancements in many application areas:

- Nanomedical robots operating in living organisms will be able to deliver potent drugs at the right spot, and in a context-dependent manner. They can also act as sentinels, permanently monitoring, recording, and reporting the presence of disease indicators.
- Small robots will extend the capabilities and enhance the sensory spectrum of larger robots. They will constitute the molecular interface of hybrid robotic systems, which enables bidirectional communication with biological systems. Such systems can be integrated (e.g., with mobile robotic units that perform environmental sensing and monitoring).
- Robotic components extend the functionality of 'smart' materials and surfaces. Embedding robotic functions into materials will enhance their ability to adjust to their environment and change shape and mechanical properties in response to environmental cues. Conceivably, such materials will have something akin to a metabolism that supports these functions and enables continuous operation. Surfaces coated with nanorobotic components will have enhanced sensor properties; they may be able to actively transport matter, take up or release molecules to the surroundings, and change their structure and physicochemical properties.
- Not the least, work on bioinspired small robotic systems will elucidate physical limits of sensors and actuators, and will clarify what it takes to display intelligent (or seemingly intelligent) behavior; it will also result in methodologies that allow programming the behavior of dynamical, out-of-equilibrium systems. Ultimately, the realization of bioinspired robots may also contribute to a better understanding of complex biological phenomena and behaviors.

5.3.2 Robotics in healthcare

Darwin Caldwell[1]
[1]Instituto Italiano di Tecnologia, Genoa, Italy

5.3.2.1 Increasing demands for healthcare

The provision of appropriate healthcare is unquestionably a major worldwide societal challenge that impacts all nations and peoples. The growth in medical robots driven by advances in technologies such as actuation, sensors, control theory, materials, AI, computation, medical imaging, and of course supported by increased doctor/patient acceptance may be a solution to a number of these convergent problems that include:

- Aging society and the increasing burden of dementia. This has been recognized in developed countries for many years, but is now becoming increasingly common globally. According to the OECD around 2% of the global population is currently over 80, and this is expected to reach 4% by 2050. In Europe, people over 80 already represent around 5% of the population and are expected to reach 11% by 2050 (OECD 2020).
- Increased healthcare spending and costs. According to the OECD in almost every country spending on healthcare consistently outstrips GDP growth. It has risen from 6% in 2010 to a predicted 9% of GDP by in 2030 and will continue to increase to 14% by 2060 (https://www.oecd.org/health/health-carecostsunsustainableinadvancedeconomieswithoutreform.htm) (Moses *et al* 2013). The per capita health spending across OECD countries grew in real terms, by an average of 4.1% annually over the 10-year period between 1997 and 2007. By comparison, average economic growth over this period was 2.6%, resulting in an increasing share of the economy being devoted to health in most countries (OECD 2009).
- Shortage of workers in medical and social care professions. A relative decrease in the proportion of healthcare workers is increasing the demands on the active workforce, which is itself ageing in many countries. Data from the World Health Organization (WHO) indicates that the average age of nurses is over 40 in several European countries. Furthermore, the WHO states that the global health worker shortfall is already over 4 million (WHO 2008).
- The rapidly growing population and the need to provide basic medical needs in developing countries.
- Changing family structures. Family sizes are much smaller than in the past and the percentage of elderly people living alone is rapidly increasing.
- Increased acceptance of technology (Abou Allaban *et al* 2020a)—Although there is among many groups (particularly elderly users) an ambivalence about the use of robot technology and a preference for human touch, there is little hostility, and among younger people who are more familiar with computers, AI, smart technology, and advanced communications, there is a growing

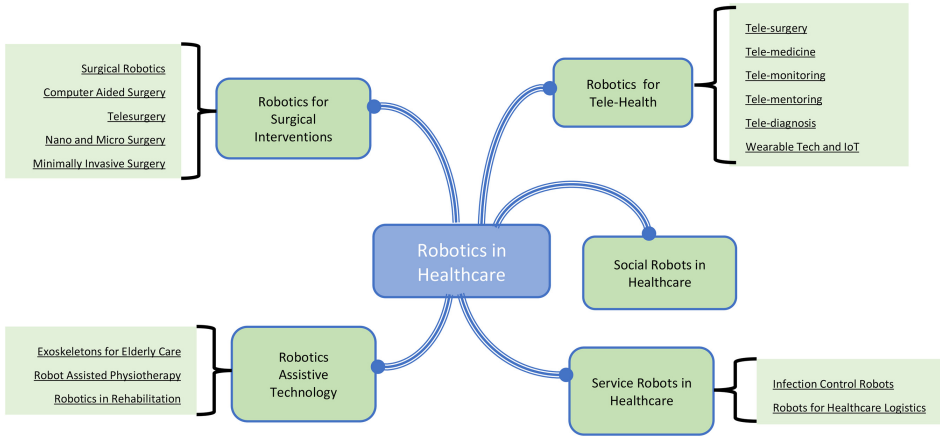


Figure 5.10. Uses of robotics in healthcare.

acceptance of the benefits of robotic and related technologies (Wu *et al* 2014). This increasing acceptance, and indeed in some instance reliance, on robots has been accelerated by COVID-19 (Zemmar *et al* 2020).

Against all of these demands and counter demands there is a widespread belief that medical and healthcare robotics is essential to transform all aspects of medicine—from surgical intervention to targeted therapy, rehabilitation, and hospital automation (figure 5.10).

5.3.2.2 Robotics and sustainable healthcare (Yang et al *2018)*

As the demands on health systems grow, it is perhaps inevitable that we should turn to technology and particularly robotics, both to provide the extra capacity/productivity that will be needed by aging and rapidly increasing populations, and also to continue to enhance and improve the quality of life provided by the healthcare systems. Although robots and robotic systems represent a significant investment cost, experience in others sectors, such as manufacturing, has demonstrated that the use of robotic technology can also offer significant savings and increases in efficiency/productivity, while also contributing to the establishment of high-quality, sustainable, and affordable healthcare systems. Important application domains that could benefit include medical training, rehabilitation, prosthetics, surgery, diagnosis, and physical and social assistance to disabled and elderly people (Stahl *et al* 2016, Wang *et al* 2021).

5.3.2.3 Robotics for medical interventions (Mattos et al *2016)*

Surgical robotics

Robotic surgery involves using a computer-controlled motorised manipulator/arm that has small instruments attached to this assembly. Using artificial sensing this arm can be programmed to move and position tools to carry out surgical tasks. The

surgeon may or may not play a direct role during the procedure. The history of surgical robotics dates back almost 40 years and arose from a convergence of several key advances in technology (Satava 2002).

Minimally invasive surgery (MIS): Driven by the potential to create smaller incisions, with lower risks of infection, reduced pain, less blood loss, shorter hospital stays, faster recovery, and better cosmesis, the first laparoscopic cholecystectomy was performed in the mid-1980s (Antoniou *et al* 2015). The benefits of this approach over conventional open surgery quickly became apparent to surgeons, patients, and healthcare providers. However, although the potential benefits were clear there were also several technical and human factors problems. These included:

- Poor visual access and depth perception, due to the quality of the 2D cameras and displays.
- Difficult hand-eye coordination due to the fulcrum effect of using long instruments inserted through a cannula. This produced motion reversals, and a scaled and limited range of motion that was dependent on the insertion depth of the tooling.
- Little or no haptic feedback and a reduced number of degrees of freedom and dexterity.
- Camera instability and loss of spatial awareness within the body cavity (the which way is up problem).
- Increased transmission of physiologic tremors from the surgeon through the long rigid instruments.

As a consequence, when performing MIS a surgeon must master a new and different set of technical and surgical skills compared to performing a conventional procedure.

VR, telepresence, and telesurgery: Around the time of the first developments in MIS, NASA was studying options for providing medical care to astronauts. Their research teams were particularly interested in using emerging concepts in virtual reality (VR), haptics, and telepresence. Teams within the US Army (Satava 2002) became aware of this work and were interested in the possibility of decreasing battlefield mortality by bringing the surgeon and operating theatre closer to the wounded soldier (figure 5.11).

Robots in surgery: Although the first MIS was only performed in 1987, the history of surgical robotics does in fact slightly pre-date this, with Kwoh *et al* performing neurosurgical biopsies in 1985 (Kwoh *et al* 1988), and in 1988 Davies *et al* performed a transurethral resection of the prostate (Davies 2000). While these and other surgical robots were being developed, clinicians working with the various robotics teams realised that surgical robots and concepts in telepresence/telesurgery had the potential to overcome limitations inherent in MIS by:

- Using software to eliminate the fulcrum effect and restore proper hand-eye coordination. At the same time the software made movement and force scaling possible so that large movements or grasp forces at the surgeon's console could be transformed into micro motions and delicate actions inside the patient.

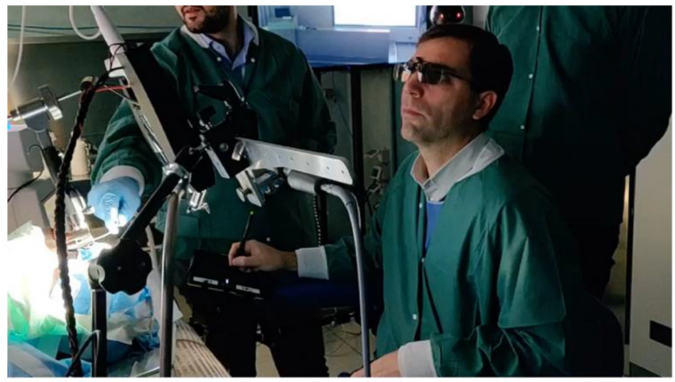


Figure 5.11. VR/AR, advanced user interfaces and telecommunications, and robot technology combine to create an enhanced surgical experience.

- Increasing dexterity using instruments with flexible wrists (designed to at least partially mimic human wrist action) to give increased degrees of freedom which greatly improves the ability to manipulate tissues.
- Using software filtering to compensate for any surgeon induced tremor, making increasingly more delicate operations possible.
- Improving surgeon comfort by designing dedicated ergonomic consoles/workstations that eliminate the need to twist, turn, or maintain awkward positions for extended periods.

Computer-assisted surgery (Buettner et al 2020)

Computer assisted surgery (CAS) is also on occasion known as image-guided surgery (IGS), computer-aided surgery, computer-assisted intervention, 3D computer surgery, or surgical navigation. It is a broad term used to indicate a surgical concept and set of methods whereby the motions of the surgical instruments being manipulated by the clinician are tracked and subsequently integrated with intra-operative and/or preoperative images of the patient. The pre/interoperative images can be produced by a combination of x-rays, computers, and/or other equipment. (e.g., medical ultrasound, ionizing techniques such as fluoroscopy CT, x-ray, and tomography, fixed C-Arms, CT, or MRI scanners). This information is used either directly or indirectly to safely and precisely navigate to and treat a condition: tumour, vascular malformation, lesion, etc, as demanded by the treatment. This increases the efficiency and accuracy of the procedure, and reduces the risk to nearby critical tissues and organs.

Although CAS can be used in traditional open surgery the need for and use of CAS, as with robotic surgery, has been driven by advances in MIS, where the minimal access needed to provide medical benefits results in restricted visualisation of the site and can create difficulties in understanding of the exact spatial location of the tooling within the body. The key difference between CAS and robotic surgery is the lack of a robot!

CAS and IGS systems can use different tracking techniques including mechanical, optical (cameras), ultrasonic, and electromagnetic or some combination of these systems to capture, register, and relay the patient's position/anatomy, the surgeon's precise movements relative to the patient, and the motion of the tooling/instruments, to a computer which displays images of the instruments' exact position inside the body on a monitor (2D or 3D). It is also possible to relay these images to virtual and augmented reality (VR/AR) headsets. This imaging and display are usually (and ideally) performed in real-time, although there can be delays of seconds depending on the modality and application. The tracking technology and the ability to track the location of the instruments within the body is often likened to GPS.

CAS and IGS have become the standard of care in providing navigational assistance during many medical procedures involving the brain, spine, pelvis/hip, knee, lung, breast, liver, prostate, and in otorhinolaryngology, orthopaedic, and cardiovascular systems. The clinical advantages include decreased intraoperative complications, increased surgeon confidence, improved preoperative planning, more complete surgical dissections, and safer junior physician training/mentoring.

Tele and remote surgery (Choi et al 2018)

Telesurgery allows a surgeon to operate on (or be virtually present with) a remote patient, by using a robotic surgical system at the patient's site. This involves the use of teleoperation technologies that use real-time, bidirectional information flow: surgical commands must be sent (in real time) to the remote robot, and everything that is happening in the surgical/patient site must be immediately perceived by the surgeon, through visual, auditory, and on occasion haptic feedback. This separation of the surgeon and the patient is already common with most current surgical robots (e.g., the Da Vinci Surgical system), with the surgeon sitting at an operating console that is a few metres away from the robot and patient, but connected with a dedicated wired connection. Telesurgery, on the other hand, refers to medium- and long-distance teleoperation with distances measured in kilometers or even thousands of kilometers between the surgeon and the robot/patient. The connection may use wired as well as 'conventional' broadband, a dedicated connection, wireless, or some combination of wired and wireless.

Potential benefits of telesurgery include:

- Eliminating the need for long-distance travel, along with associated costs and risks. This will provide for urgent emergency interventions and expert surgical care in underserved regions such as remote/rural areas, underdeveloped countries, in space, at sea, and on the battlefield.
- Surgical training: Telementoring can enhance the training of novice surgeons and bring new skills to more experienced medics. This could revolutionize surgical education.
- Surgical collaboration: Real-time collaboration between surgeons at different medical centers using shared, simultaneous perceived, high-definition visual feedback.
- Surgical data: Telesurgery focuses on moving data instead of the patient or the surgeon. This data is extremely rich, and includes sensory information,

records of the surgical workflow, actions, and decision-making processes. This data can be used to assess surgical quality, or develop AI to enable autonomous surgical supervision, assistance, or even fully robotic surgery.
- Robotic precision: The use of robotic systems increases surgical precision and quality, removing physiologic tremor, reducing adjacent tissue damage, and permitting previously impossibly delicate surgery.

This large separation between the surgeon and surgical robot brings significant medical advantages but also creates many technical challenges, most of which are directly linked to the need to instantaneously transfer, in a safe, secure, and reliable way, massive amounts of data between both ends of the system. Fortunately, advances in data communication, fibre optic broadband, and particularly the increasingly widespread availability of 5G mean that the long-dreamed goal of truly remote telesurgery is now increasingly possible (Acemoglu *et al* 2020a). This will be explored in the later section on general requirements in Robotics in Tele-Healthcare.

***Micro- and nanomedical robots (Nelson et al* 2010, Li *et al* 2017, Soto *et al* 2020)**
Current medical and surgical robots draw most of their design inspiration from conventional industrial robotics, with modifications to suit the particular requirements of operations in, on, and near tissue. Miniaturization of these robotic platforms, and indeed creation of completely novel, versatile micro- and nanoscale robots, will allow access to remote and hard to reach parts of the body, with the potential to advance medical treatment and diagnosis of patients, through cellular level procedures, and localized diagnosis and treatment. This will result in increased precision and efficiency. These advances will have benefits across domains such as therapy, surgery, diagnosis, and medical imaging.

Medical micro/nanorobots are untethered, small-scale structures capable of performing a pre-programmed task using conventional (electric) and unconventional (chemical, biological) energy sources and actuators, to create mechanical actions. Due to their size (sub mm) they face distinct challenges when compared to large (macroscale) robots: with viscous forces dominating over inertial forces, and motion and locomotion being governed by low Reynolds numbers and Brownian motion. Thus, the design requirements, parameters, and operation of these robots is almost unique in the mechanical world, although these features are common in biological systems. Many micro/nanorobots are made of biocompatible materials that can degrade and even disappear upon the completion of their mission.

Despite the challenges caused by the miniature scale, the potential benefits are significant in areas such as targeted delivery, precision surgery, sensing of biological targets, and detoxification.

Targeted drug delivery—Motile micro/nanorobots have the potential to ‘swim’ directly to a very specific target site within any part of the body and deliver a precise dosage of a therapeutic payload deep into diseased tissue. This highly directable approach will retain the therapeutic efficacy of the drug/payload while reducing side effects and damage to other tissue.

Surgery—micro/nanorobots could navigate through complex biological media or narrow capillaries to reach regions of the body not accessible by catheters or invasive surgery. At these sites they can be programmed or teleoperated to perform the required intervention (e.g., take biopsy samples or perform simple surgery). The use of tiny robotic surgeons could help reduce invasive surgical procedures, thus reducing patient discomfort and post-operative recovery time.

Medical diagnosis—isolating pathogens or measuring physical properties of tissue in real-time could be a further function performed by micro/nanorobots. This would provide a much more precise diagnosis of disease and/or vital signals. The integration of micro/nanorobots with medical imaging modalities would provide accurate positioning inside the body.

The use of micro- and nanorobots in precision medicine still faces technical, regulatory, and market challenges before their widespread use in clinical settings. Nevertheless, recent translations from proof of concept to *in vivo* studies demonstrate their potential in the medium to longer term.

Medical capsule robots (MCRs) are smart medical systems that enter the human body through natural orifices (often they are swallowed) or small incisions. They are an already operational form of micro/nanorobot that use task specific sensors, data processing, actuation, and wireless communication to perform imaging and drug delivery operations by interacting with the surrounding biological environment. MCRs face a number of challenges such as size (for non-invasive entry they must be less than 1cm in diameter), power consumption (they must carry on-board batteries), and fail safe operation (they operate deep within the body). MCR design and development focuses on miniaturization of the electronics and mechanical structure, packaging, and software (Beccani *et al* 2016).

5.3.2.4 Robotics in tele-health (De Michieli et al *2020)*

Tele-healthcare involves the remote (tens of km to thousands of km) connection of different subjects involved in the healthcare process. This includes the patients, the clinicians, and others. Tele-health systems allow different kind of interactions: (i) clinician to clinician, (ii) clinician to patient, and (iii) patient to mobile health technology. Each interaction has a different purpose and requirements.

A common scenario where tele-health can provide an important benefit is when the patient or medic is not able to travel to the clinical setup, due to distance limitations, cost, poor transportation links, or other external factors. The COVID-19 pandemic has made this scenario even more real (Chang *et al* 2020, Leochico 2020, Prvu Bettger 2020), yet through the technology and application of tele-medicine, medical treatment, assessment, monitoring, or rehabilitation can continue to be provided without the need to come in person to the clinic. This offers important advantages in terms of infection prevention.

A second application of tele-health is continuous monitoring. Classical tele-health approaches make use of simple technologies such as phones, email, or video-chats, but the latest IoT (Internet of Things) technologies allow continuous monitoring and feedback of many different medical data and processes (Dabiri *et al* 2009), This makes use of developments in ubiquitous computing, cloud storage, and intuitive

human-machine interaction. Tele-Health is also exploring the emerging possibility of using the sensory capabilities of wearable devices that offer a unique opportunity to enhance the monitoring and intervention abilities of medical staff and relatives (De Marchi *et al* 2021). This data can be analysed and sent to nurses, doctors, and public health agencies for monitoring purposes. In addition to providing a greater range of data, wearables can also reduce the demand for staff to make manual observations.

Wearable technologies for health monitoring (Best 2021)

Wearable technologies are patient portable devices that can range from simple systems with a single sensor that measures only one variable to complex devices that can gather a wide range of vital health and lifestyle parameters. They are typically worn on the arm/wrist, or around the chest/waist, but the positioning can be dependent on the variables being measured and each generation of new wearable devices adds greater functionality in increasingly compact packages that have better and better battery life. The information collected by the wearable device is typically stored locally before being relayed directly to the clinicians for monitoring or analysis, but this feed can also on occasion be continuous when the clinical need and device support this functionality.

While many wearable devices were historically custom-made medical devices, the past 10–15 years have seen a vast growth in consumer-grade wearables, such as smart watches and fitness trackers that have increasingly sophisticated and accurate sensing technology (exercise level, sleep patterns, SpO2, heart rate, and even single lead electrocardiographs). These wearables (which generally link to mobile phones) can measure data continuously and can be programmed to alert the wearer and the clinician if values deviate significantly from the norm.

The already existing widespread use and acceptance of wearables means that this technology is likely to have an increasingly important impact on healthcare, providing a much more accurate profile of patient health over a prolonged period.

5G technology and tele-health (Acemoglu et al 2020b)

Improvements in internet speeds and connectivity have meant that even with the ever- increasing data levels demanded by high-resolution systems, wired tele-health connections have been possible for several years. However, there were always constraints when the data had to be transferred wirelessly, and this meant that there were limits on the scope of telemedicine and where this service could be provided. To provide truly flexible operation for remote and tele-health applications away from fixed communication points there needed to be a step forward in wireless communications. This was provided by 5G mobile networks. 5G technology is not just the evolution of 4G. It is a completely new platform capable of enabling new services. 5G provides the following major benefits compared to previous technologies:

- Ultra-high speed broadband up to 10 Gbps.
- Ultra reliability and ultra-low latency down to 1 ms.
- Massive connectivity, able to handle one million devices per square km.

- Quality of service, ensured by multiple input multiple output (MIMO) antennas with beamforming capabilities (i.e., able to direct their power to where it is requested by a user).
- 5G virtual network (VN) element infrastructure, which increases the network's flexibility and resilience, allowing network slicing to create independent virtual networks on the same physical infrastructure. Each VN has its own characteristics in terms of security, traffic routing, and quality of service. Network slicing enables the customization of services for users or applications with specific requirements, making it possible to allocate and reserve resources for mission-critical services such as a telesurgery.

Tele-diagnosis and monitoring (Ding et al 2020)
Tele-diagnosis refers to the ability to make a remote diagnosis using platforms designed to transmit the physical examination records and medical reports to the examining specialist. The tele-diagnosis system should ensure that images and videos preserve the diagnostic quality even after compression for transmission. There can be limitations due to compression, bandwidth issues, and lag, but improvements in internet bandwidths, video imaging, the widespread roll-out of 5G and demands of COVID-19 which has increased user (both medic and patient) acceptance of the technologies, mean that tele-diagnosis and all aspects of tele-health have received a significant boost, and it is expected that this will continue, becoming more common in forthcoming years. This will include fewer network-critical services such as tele-mentoring and tele-assistance.

5.3.2.5 Robotics assistive technology
The elderly population is projected to quadruple between 2000 and 2050, with many experiencing mobility impairments due to physiological muscular decay or associated health conditions, such as stroke, which shows an increasing rate of incidence with the age of the subject. In its '*2030 Agenda for Sustainable Development'* the UN studied future housing, healthcare, employment, and social protection needs (DoE 2017), and identified mobility as being of paramount importance to quality of life, social inclusion, and independence (Richardson *et al* 2015). While walking sticks, wheelchairs, or walking frames can provide reasonable assistance, many elderly/disabled still require additional assistance when walking (Charron *et al* 1995).

Stroke is the leading cause of disability in industrialised countries. Fortunately, over 65% of patients survive, but the majority have residual disabilities, with up to a third having severe disabilities. Hemiplegia, the most common impairment resulting from stroke, leaves the survivor with a stronger unimpaired side and a weaker impaired one (hemiparesis). In addition to stroke, traumatic injuries as well as conditions such as muscular dystrophy, arthritis, and regional pain syndromes also add to the major causes of disability and functional dependence. Deficits in motor control and coordination synergy patterns, spasticity, and pain are some of the most common symptoms of these conditions (Parker *et al* 1986).

Evidence has shown that intensive and repetitive physiotherapy can have substantial benefits (Carr *et al* 1987) including regaining motor control and muscle

strength as well as in restoring/retaining the joints' range of motion. Despite the benefits of intensive physiotherapy, the associated disabilities are seldom considered life-threatening; therefore, they rate relatively low on the priority list for urgent medical assistance. In addition, manipulative physiotherapy is labour-intensive and therefore fatiguing for the therapists as well as the patient; it requires high levels of one-to-one attention in an environment with an international shortage of physiotherapists, and patients must receive individualised treatment.

Against this background, wearable technologies, robots, exoskeletons, and power- assistive techniques seem to offer a promising solution and are increasingly viewed as a potential replacement for the physical labour. This will leave therapists with greater time to develop treatment plans and enhance therapeutic outcomes.

Rehabilitation robots

The goal of rehabilitation is the reintegration of an individual with a disability into work/society. This can be achieved by:

- enhancing existing capabilities (achievable through therapy and training) or by
- providing alternative means to perform various functions or to substitute for specific sensations (achievable using assistive technologies) (Robinson 1995).

Rehabilitation robotics is the application of robotic methods to train or assist an individual with a disability and support their reintegration to work/society. This broad definition covers a variety of different mechatronic machines that support both gait and arm therapy in a clinical setting. It can also include powered orthotics for use in daily life environments, actuated prosthetics, and can even include intelligent wheelchairs (Priplata *et al* 2003).

Many rehabilitation robots are based on traditional robot designs (often collaborative robots used in the industrial sector that have been modified to provide the levels of safety needed for close interaction with patients). They usually connect to a distal segment (wrist, hand, ankle, foot) of the patient and can be used to support rehabilitation of either upper or lower limbs (e.g., Burt [Barrett Technology], ROBERT [Life Science Robotics]). Although traditional arm-inspired designs have been the most common approach in rehabilitation, the specialised/individual nature of the requirements means custom designs are also common (e.g., InMotion ARM [Bionik Laboratories], Hunova [Movendo Tech]; Saglia *et al* 2013, 2019). For both the arm-inspired and custom designs, the approach taken is often similar involving initial assessment and subsequent retraining of the range of motion of a limb (arm, leg) and increasing power functions.

Gait rehabilitation robots

Strokes often result in loss of mobility, which can easily lead to patients becoming house bound and completely dependent on others for many aspects of daily life. Hence, restoration of walking is a vital goal. Robots can play an important part in locomotion therapy and gait rehabilitation. Systems used in gait rehabilitation typically must provide support for the whole body weight of the patient. This can be

achieved using multidirectional overhead body weight support systems (e.g., ZeroG or FLOAT) that require the user to wear a harness that can provide full or partial gravity compensation. Wearing the support, the patients can start training earlier, while muscle control/coordination is poor and muscle strength is low, without the risk of falling. The Andago system provides a similar form of body weight support, but the unit can move freely without the need for ceiling- mounted rails. To take full advantage of the potential of these body weight support systems they can/should be linked to a mobile robot (e.g., a lower limb exoskeleton such as ReWalk or Twin; Laffranchi *et al* 2021) or as part of a stationary gait trainer (e.g., Lokomat, Walk Training Assist; Baronchelli *et al* 2021).

Systems of this type permit patients to relearn walking and safely train after stroke, spinal cord or brain injuries, incomplete paraplegia, or orthopedic patients. For the exoskeleton-based systems this training can in addition be in an unrestricted three-dimensional space. Although free movement is possible and will become more common in the future, the leading robot gait rehabilitation systems such as the Lokomat remain high in cost, and are large, heavy, static structures found in specialised rehabilitation centres.

Balance trainers such as Toyota's Balance Training Assist or the Balance Tutor can be considered an adjunct or subset of gait trainers and can be used for balance training. They often include a platform (fixed, mobile, or in the form of a treadmill) and a weight support system, which can be programmed to disturb the patient's balance allowing them to safely relearn, regain, and enhance their stability.

Exoskeletons for rehabilitation (Pons 2008)

An exoskeleton is a wearable robot that assists with the execution of physical activities by delivering forces/torques at the human joints (Toxiri *et al* 2019). One important and very specific feature of exoskeletons is the physical connection between the human and the robot. This human–robot interaction (HRI) is twofold: firstly, cognitive (the human controls the exoskeleton); secondly, physical (the human and robot are bidirectionally coupled). This physical coupling is one of the most exciting aspects, however, it does bring challenges since it involves the cooperation of two dynamic systems (i.e., human motor control and robot control). Therefore, human-activity recognition (HAR) is required to understand what kind of action the user is performing or wants to perform (Poliero *et al* 2019).

Exoskeletons can be categorized in several ways, but among the most common are:

- *Rigid* or *soft*. Rigid exoskeletons use structural frames and links to support part of the exoskeleton weight and transmit forces and torques. Soft exoskeletons, also known as *exosuits*, are lightweight and use soft/softer materials (fabrics and plastic) that can more easily be integrated with clothing to increase user acceptance. This softness can also be extended to the actuators which may be 'hard' but behave in a soft way such as variable stiffness actuators, series elastic actuators, or pneumatic actuators (Vanderborght *et al* 2013), or can truly follow the soft material paradigm and actually be soft actuators (Di Natali *et al* 2019).

- The assisted joint (e.g., shoulder, neck, wrist, back, hip, knee, ankle, or any combination of these).
- Actuation (passive, active or *quasi-passive*). *Passive* exoskeletons use elastic components such as springs to provide a fixed level of assistance (Huysamen *et al* 2018). Passive actuation is not suited for many impaired users since they have little residual strength to 'charge' the mechanical elements. *Active* exoskeletons represent the optimal fit, exploiting controllable elements like electrical motors or pneumatic actuators (Toxiri *et al* 2018). *Quasi-passive* exoskeletons take advantage of controllable elements not to directly provide assistance to the user but, rather, control the engagement of the connected passive elements (Di Natali *et al* 2020).

Currently most exoskeletons for healthcare are fairly bulky, stationary, systems that target improved rehabilitation. Fewer fully wearable systems are being used due to their complexity, weight, and performance. However, research within industrial and military domains is spilling over into healthcare, and there is now increasing potential for portable, home-based systems both for rehabilitation and as a mobility assist.

5.3.2.6 Social robotics for healthcare (Henschel et al *2021)*

Social robots are often characterised as providing support or training to a person in a caring/social interaction scenario. This may be interpreted by external observers as appearing social. They are physically embodied, often in human form, although animal- and cartoon-like appearances are not uncommon. Although many social robots emulate human form, features, behaviours, and expressions, care must be taken to avoid imitating human appearance or motion too closely to avoid falling into the Uncanny Valley (Pandey *et al* 2018), where behaviours and appearances can become intimidating or disturbing. Social robots have some level of autonomy and can directly interact, communicate, and cooperate with people following social accepted behaviours and rules. They do not typically perform a mechanical task, although they may be capable of movement including on occasion locomotion.

For healthcare-based scenarios many see social robots as playing a part in addressing mental health issues, with the robots acting as social partners in applications such as robot companions, robots as educators for children or children/adults on the autistic spectrum (Cao *et al* 2019, Stower 2019), robots as assistants/companions for the elderly/disabled, and those studying affect, personality, and adaptation (Čaić *et al* 2019, Dawe *et al* 2019, Johanson *et al* 2020). The development of social robots can also create a bidirectional paradigm with the human learning from the robot and the robot learning from the human. This has seen particular relevance in cognitive neuroscience.

A early example of a social robot developed to support developmental psychology, cognitive neuroscience, and embodied cognition was the iCub humanoid. iCub was a child- sized robot designed to manipulate its surroundings, imitate its human partners, and communicate with them (Tsagarakis *et al* 2007). Over the past 20 years iCub and social robots in general have investigated human social concepts such as decision making, intent, trust, attention, perception attachment empathy acceptance, and disclosure.

Although there is a substantial body of research in social robotics with some very positive progress, there is nonetheless a substantial gap between the capabilities of social robots and the expectations of the general public, with potential users expecting s social robot to be 'like a friend' but being disappointed when the robot is not able to go beyond the smart-speaker like single-turn structure of conversation. Nonetheless the ongoing research continues to make important progress and this will only increase, resulting in robots that are more 'social' and personal over time (de Graaf *et al* 2015).

5.3.2.7 Service robots in healthcare

Service robots have expanded rapidly into many areas in the past 20 years and are increasingly common in industrial and even home settings. The healthcare sector has also been seen as an area of potential growth, although pre-COVID-19 many of these opportunities were seen as potential rather than realised goals. This was perhaps due to the challenges associated with the very nature of the services needed in healthcare. However, there is now no doubt that the COVID-19 pandemic has created an environment that has accelerated deployment and use of service robots within hospitals and across the wider healthcare system. Within this context robots are now being used either semi- or fully autonomously to provide services such as effective cleaning in hospitals/care homes and logistics of patients and supplies. This has benefits for patients, healthcare workers, and healthcare systems.

Infection control robots

Infection control and sterilisation has always been a problem within hospitals and healthcare settings (Carling *et al* 2013), and robotic solutions for sterilisation had already been developed prior to the pandemic. However, with the onset of the pandemic several robotics companies have introduced mobile robotic platforms providing disinfection solutions. These robots often have the capacity to navigate autonomously to the desired sterilisation site using lifts, passing through doorways and along corridors. At the target site they can use a variety of sterilisation/disinfection techniques including:

- UVD (ultraviolet disinfection)—disinfection using high-intensity continuous or pulsed UVC light (Beal *et al* 2016, Omron 2022).
- Hydrogen peroxide—surface decontamination using a dry aerosol of hydrogen peroxide disinfectant (Anderson *et al* 2006).
- Physical disinfection of contact surfaces—The human support robot uses a fully autonomous arm manipulator to spray disinfecting liquids on to commonly touched surfaces such as door handles, lift call buttons, handrails, tables, walls, etc The arm and a cleaning brush are subsequently used to clean the region much as would be done by a human cleaner (Ramalingam *et al* 2020).

Robots for healthcare logistics

As with any large and multifaceted organisation, hospitals and care homes have complex logistical requirements ranging from those that are medical centred such as

the transport of medical samples, medicine, and medical supplies, laboratory samples, to the delivery and coordination of daily necessities such as food and linen. Robots with autonomous navigation capabilities that are used in industry and housing/offices offer a pre-existing entry point for this operation, particularly when considering that it is estimated that nurses spend only 70% of their time dealing with their patients with the remaining time being used to track down records, supplies, test results, or in other 'logistical' tasks not directly related to patient care. If robots could perform some or all of these search tasks this would increase overall efficiency and accuracy, and allow the staff to concentrate on their key roles (Bloss 2011).

A number of robots and robotics solutions are already available for logistical operations such as delivery of patient records, pharmacy supplies, diagnostic bloods, and fluid samples and test results.

The use of service robot across the healthcare sector, can and is, having important benefits in terms of removing human cleaning staff from potential disease exposure, preventing the spread of infection, improving levels of cleaning by reducing human error, ensuring 24 h availability, and allowing front-line staff to reduce direct contact, focusing their attention on higher priority tasks and creating separation from direct exposure to infection.

5.3.2.8 Future challenges and trends

Healthcare is a major societal challenge with daunting forecasts for the future, especially given the global population aging trend, the increasing number of patients, the decreasing proportion of healthcare workers, and the increasing costs of care. Fortunately, technological progress and the rise of healthcare robotics offer hope for the establishment of sustainable, affordable, and high-quality healthcare systems.

Robotics can offer significant contributions to progress in disease detection, diagnosis, and treatment at early stages; patient care; earlier more intense personalised rehabilitation; remote treatment; and the training of medical personnel, improving the skill levels of trainees and lengthening the effective career of experienced surgeons/clinicians. All of this can be achieved while enhancing the safety, precision, quality, and efficiency with which care is provided. But these opportunities do not come without challenges and barriers. There needs to be a recognition that while individual technologies can and will bring important but still incremental levels of improvement, major impacts are expected to arise from their full integration into a complete robotic healthcare system.

As noted throughout this chapter the possible areas of application of robotic technology are vast and diverse and therefore identifying generic issues and barriers to implementation is not easy. However, it is clear that there are challenges that encompass economic, social/societal, clinical, and research-related domains.

Technical

Technical challenges are the first and most obvious barriers to the growth, development, and successful use of any and all robotic technologies in healthcare. If the technical issues cannot be resolved subsequent challenges would in all

likelihood not arise. Foremost among the technical challenges are the mechatronics: the hardware, software, and control systems that form the foundation of any robot. With respect to healthcare robotics (as is also true for other robotics sectors) many of these challenges focus on mechanical design (e.g., miniaturisation, use of novel materials, new construction/fabrication techniques, rapid, and customised/personalised prototyping, dexterity, and flexibility to reach difficult parts of the anatomy); actuation (smaller, faster, more precise, and more efficient motors or even other actuation technologies etc); more and better sensors (e.g., smaller, detection of a wider range of parameters); multifunctional sensors; real-time detection of conditions such as cancer; wireless operation, MEMS, haptics, etc; materials (e.g., biocompatible, biodegradable, lighter, MRI friendly, plastics versus metal versus ceramics, 3D printing, biological materials, etc); imaging and display (e.g., VR/AR, 3D visualisation, improved resolution, new imaging techniques and technologies, etc); new robot anatomies and structures (e.g., bioinspired, continuum robots, soft robots, robots capable of controlled access and operation in confined body spaces) and miniaturisation (e.g., chip-on-the-tip stereoscopic imaging); and integration of the most effective imaging technologies (e.g., NBI, fluorescence, autofluorescence, OCT, and MRI) into surgical robotic systems to fully exploit the benefits of real-time disease detection.

There will also, of course, be many challenges for the software. This will impact algorithmic development driven by advances in machine learning, deep learning, and AI; the use and development of assistive and augmentation algorithms to help and guide the user in surgery/rehabilitation/training, etc; virtual bodies (e.g., robust real-time algorithms to perceive the 3D structure of the body/tissue/limb, and realistically model tissue motion/deformation with lifelike visual and haptic sensations); and analysis of vast amounts of wearable data covering ever-evolving aspects of daily life. Alongside these technical/operational challenges of software development will be growing cyber-security concerns that will be further compounded by the growth of tele-health, wearable technology, and issues arising from remote access. These areas, as indeed, many aspects of tele-health, will bring to the forefront previously lower impact problems such as the risks of intrusions, privacy, and security with personal data and even malicious attacks, which must all be addressed by considering legal and ethical perspectives, in addition to technical obstacles.

A further aspect of the debate on software in healthcare robotics that is not so common in other robotics sectors is: should robots be autonomous? In most other areas of robotics full robotic autonomy is the goal, but in healthcare there remain many unresolved issues around the level of autonomy (tele-operated versus semi-autonomous versus full autonomy) that again go beyond the technical to impact ethical, legal, societal, and of course employment issues.

Finally, more than in any other sector there is a very high likelihood that healthcare robots (e.g., surgical systems, wearable devices, exoskeletons, or logistics robots) will on many occasions come close to or into physical contact with a person, and often this person will be unwell. In any such interactions safety will always be paramount. While guidelines are in place for collaborative robots this may not be enough for the health sector and it may also be valuable to take guidance from the

procedures and processes within aerospace or other sectors where there are critical life-threatening interactions between machines/people. Of course, on the positive side robots will also have an increasingly important part to play providing levels of safety and error prevention beyond what is possible with direct human supervision.

Acceptance of healthcare robotics

A major challenge in all aspects of robotics for healthcare is acceptance: by patients, by medical personnel, by healthcare systems, and by society. This acceptance will be influenced by many factors (e.g., age profile, prior experience with and of technology, national/cultural, application area, effectiveness of the technology, attitudes of the physicians, cost, etc). But what is universally clear is that robots will only be accepted if the science and user experience prove the benefits. To achieve this it is important that all users are fully informed with proactive, positive public engagement campaigns that introduce the technology and benefits, effective training of healthcare staff on how to operate these robots, and education that the robots should not be seen as a professional or job-related threat. Furthermore, by giving robots more exposure—making them visible in everyday environments—they are more likely to be accepted.

At the same time the robotics community must recognise that anxieties exist, and understand that the user's (patient or medic) willingness to adopt a system is influenced by the perceived usefulness, the ease of use, avoidance of frustration and embarrassment, managing expectations on the device performance and the delivery against those expectations, trust in reliability, and social/professional pressure. A potential user's intent to use can be further affected by their self-efficacy (i.e., the user's confidence in their ability to face a certain situation). This is critical in many cases of robot acceptance.

It is also critical to note that often this technology will be targeted at older users with reduced motor skills, coordination, muscle strength and/or control, and sensory losses. Designers should pro-actively engage, listening to the user's individual needs, to create age friendly geriatric (Gero)-technologies with recognition that older generations are less comfortable with technology and that older patients might need more time to adapt, while wanting to avoid being stigmatised by using devices that highlight their deteriorating health. Once again there are likely to be ethical considerations for all users, but particularly elderly users who feel that the monitoring (potentially by family members or relatives) is intrusive or dehumanising.

Healthcare robotics and regulatory approval

To ensure patient and clinician safety, robotic devices must obtain regulatory approval. A manufacturer must demonstrate that their device is safe; however, safety validation is complex, time-consuming, and costly. In part this is due to the newness of the area, the lack of examples of best practices, and a lack of clear guidance on applicable safety standards and protocols. Obtaining the required expertise against this background is challenging, especially for start-ups and small-to-medium enterprises. This can and does add to the long lead time for development

from laboratory prototypes to clinical systems and often precipitates the failure of potentially life-enhancing developments. The approach to the development, testing, and approval of vaccines and drugs seen during the COVID-19 pandemic may offer some clues as to how these processes could be improved and streamlined for the benefit of all, without risking safety.

Jobs and healthcare robotics

As in other sectors, those working in healthcare fear that if robots develop the same, similar, or better capabilities to humans, they will replace their human counterparts, resulting in job losses. Concerns around the replacement of people by machines is even more pertinent in this sector because human interpersonal interaction is at the centre of healthcare professions and again ethical and societal issues could be as important and maybe even more important than technical concerns, and each of these topics will need careful consideration.

At the same time there is little doubt that the use of robots will indeed cause changes in the nature of work within the healthcare profession, and there will certainly be a loss of typically low-skilled jobs, as in other sectors. But equally this reduction in more menially and potentially less fulfilling jobs could be seen as a benefit. Further, against a background of increasing demand for healthcare services and reducing numbers of people entering the professions (and willing to undertake these menial jobs) it is certain that there must be a displacement of work profiles. This will create both personal fears and ethical risks that must be addressed by society.

Conclusions

Although the road for healthcare robotic seems long, it is clear the technologies have the potential to bring many benefits to patients, medics, hospitals, and the public healthcare systems in general, bringing better treatment and outcomes, fewer infections, complications, and errors, increasing customer satisfaction and, at the same time, contributing to a more productive healthcare sytem. These are reassuring results offering hope for a brighter healthcare future.

5.4 Physics for health science

5.4.1 Light for health

Thomas G Mayerhöfer[1], Susanne Hellwage[1], Jens Hellwage[1,2], Dana Cialla-May[1], Christoph Krafft[1], Iwan Schie[1,3], Petra Rösch[4], Timo Mappes[4,5], Michael Schmitt[4] and Juergen Popp[1,2,4]
[1]Leibniz Institute of Photonic Technology, Member of Leibniz Health Technologies, Albert-Einstein-Str. 9, 07745 Jena, Germany
[2]InfectoGnostics Research Campus Jena, Centre for Applied Research, Philosophenweg 7, 07743 Jena, Germany
[3]Department for Medical Engineering and Biotechnology, University of Applied Sciences Jena, 07745 Jena, Germany
[4]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany
[5]Stiftung Deutsches Optisches Museum, Carl-Zeiss-Platz 12, 07743 Jena, Germany

5.4.1.1 Introduction (Popp and Strehle 2006, Popp et al *2011)*

Light for Health—what is the meaning of this title? Before we discuss this in detail, we want to introduce another term which is closely connected, namely Biophotonics. Biophotonics derives from two Greek words, namely, $\beta\iota o\sigma$ life, and $\phi\omega\sigma$, light. Correspondingly, under the term Biophotonics we understand the application of optical technologies or methods like microscopy to solve problems in medicine and life sciences. In this sense, the use of light for health applications is nothing new, and has already established for a very long time. One cannot imagine health-related science without light, all the more, because one may extend the definition of light beyond the invisible parts of the electromagnetic spectrum, namely, into the infrared spectral range on the one side and down to the ultraviolet, and even reaching x-rays, on the other side.

Light may be used to diagnose diseases early, if possible before the outbreak, and to prevent or treat diseases (e.g., by identifying infectious bacteria that are resistant to antibiotics). But even more, light helps to understand fundamental processes in cells (e.g., by using conventional or high-resolution microscopes), and thus enables us to elucidate the causes of illnesses. Thereby it can help to fight diseases, for example, by finding the right dose of a pharmaceutical substance or finding the right drug at all. In addition, it can also be employed to track and destroy cancerous cells and tissue, and all that with minimal side effects when compared to systemic drug therapy (figure 5.12).

What is it that renders light unique? Light is able to measure touch-free. In a strict sense, this is not completely accurate—light always changes matter, even if it is not absorbed, but usually such interactions are transient and reversible, so that light does not disturb processes in living matter it illuminates. Light interactions are instant. Instead of waiting a couple of days (e.g., for a hemogram) a result can be gained much faster within hours or, in some cases, in real time. This means it would be possible for a medical doctor to find the cause of a disease immediately and react

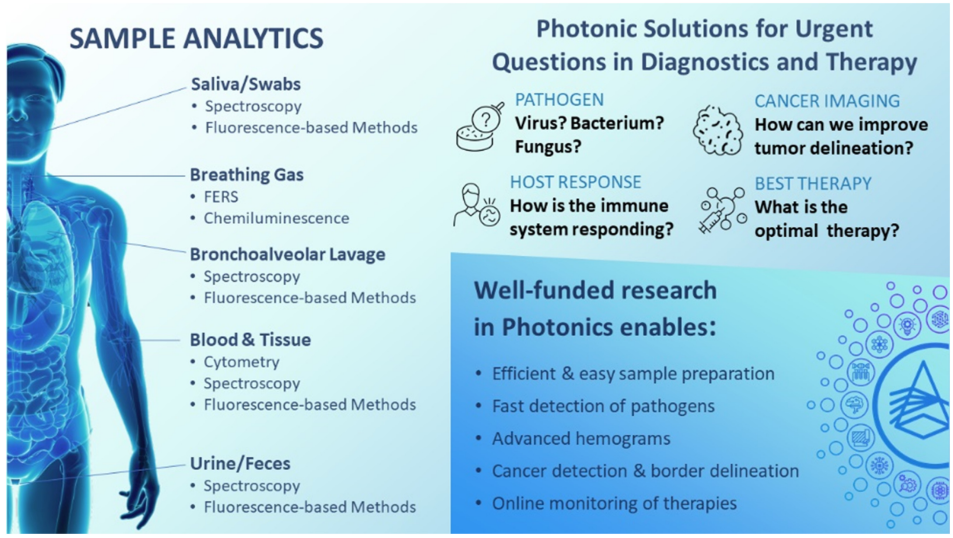


Figure 5.12. Overview of the application of Light for Health (with permission from 'Leibniz Gesundheitstechnologien', Christian Döring). Reproduced from Osibona *et al* (2021) CC BY 4.0.

instantly. In addition, it is the ideal tool to illuminate the world of microbes and cells, down to even sub-cellular structures.

The goals of using light-based technologies and methods for health applications are manifold. First, light-based technologies can solve the urgent problems our civilization is facing be it either strongly aging societies like (e.g., in Europe, North America, or China) or poorer societies such as in parts of Africa. This might be surprising at first glance. However, the focus is not on high-end instrumentation only, but also on portable and wearable instruments that sample data more often or even continuously. These data can be analyzed in rural areas or at the point-of-care or be remotely collected and evaluated. Such technologies can help to reduce costs and enable medical diagnosis and treatments without a medical doctor present. In any case, the direct benefit for the patient is paramount. In addition to this benefit, health-based technologies help to secure existing jobs and to create new ones. And, last but not least, light-based techniques and technologies allow us to satisfy our curiosity to understand life processes at the cellular level, to understand the cause of diseases for the ultimate goal to enable the early diagnosis and treatment of diseases.

Since there are several possibilities to subdivide the vast field of biophotonics and because it is not possible to even cover the most important methods and techniques (this would also go far beyond the purpose of this contribution), we selected examples of what we consider important technological achievements of light for health and offer an outlook of what may be possible in the future. In doing so we focus on examples of photonics approaches which are already used in clinical routines for therapy and diagnosis. Furthermore, we use the example of optical coherence tomography (OCT) to show the challenges one faces when introducing a new method into the clinical workflow. We also present an example of a very

promising point-of-care approach that is close to being ready for clinical use. The latter highlights the potential of photonic approaches to address currently unmet medical needs. Finally, we close this subchapter by giving an outlook on future perspectives. However, we start with a brief introduction into the history of microscopy because modern microscopy, or the achievements in light microscopy, can be seen as one of the drivers of Biophotonics, and many biophotonic approaches involve microscopes or innovative microscopic/imaging concepts.

5.4.1.2 Microscopy (Popp et al *2011, Popp 2014, Meyer* et al *2019)*

Understanding the world in ever smaller dimensions has spurred scientists to develop ever better optical instruments. Our current view of the microscopic world is based to a large extent on discoveries made with the help of the microscope (figure 5.13). This applies to both the animate and inanimate part of nature. Over time, the light microscope has become an outstanding instrument for medicine and the natural sciences. The time of the invention of the microscope cannot be clearly pinpointed. For more than two centuries, starting from the beginning of the 17th century, scientists and craftsmen endeavored to develop the device into the instrument we all know today.



Figure 5.13. A research microscope of 1912 (Stiftung Deutsches Optisches Museum (D.O.M.), License CC0).

While in antique times the use of lenses for the cauterization of wounds is documented, it looks like as if its employment for magnification purposes cannot be attested. After the fall of the Roman empire, Arab scholars took charge of increasing the knowledge in the Middle Ages. As a first milestone, at around 1000 A.D., the invention of a reading stone took place, most likely by Ibn Al Haitham.

It is not clear who first invented the microscope. But for clarity, let's first define the term. If we accept any magnifying instrument as a microscope, then also devices with only one lens fall into this category. A stricter definition of the term requires that a microscope consists of at least two lenses, an objective lens to magnify the object and an eyepiece lens to further enlarge this magnified image. The idea of employing two lenses most likely originated in the beginning of the 17th century, but it is not clear when the first device was actually built. Instruments with two lenses were originally of inferior quality compared to instruments using only one lens, when it comes to high-power magnification. First of all, one problem was the fabrication of glasses without streaks and other defects which increased if a device with two lenses was to be constructed. A further challenge was chromatic aberration (figure 5.14). Since the refraction of light is increasing with decreasing wavelengths (dispersion), a collecting lens will have the focal spots of blue light closer to the glass than for red light. This was a common problem in early microscopes and telescopes until compound optics using differently refracting glasses were engineered. This was a big hurdle for Robert Hooke (1635–1703), who was the first to publish an extensive scientific work about microscopy called 'Micrographia'. His instruments achieved approximately 50× magnification with an incident light setting, using a cobbler's ball for illumination.

While Isaac Newton (1643–1727) had erroneously claimed in 1666 that the problem of achromatism could not be solved, it was Antoni van Leeuwenhoek (1632–1723) who focused his developments on instruments with only one lens, with which he achieved magnifications of up to 260×. One of Leeuwenhoek's secrets was to fabricate lenses virtually free from glass defects, allowing him to reach such high magnifications that he was the first to publish sketches of living bacteria and

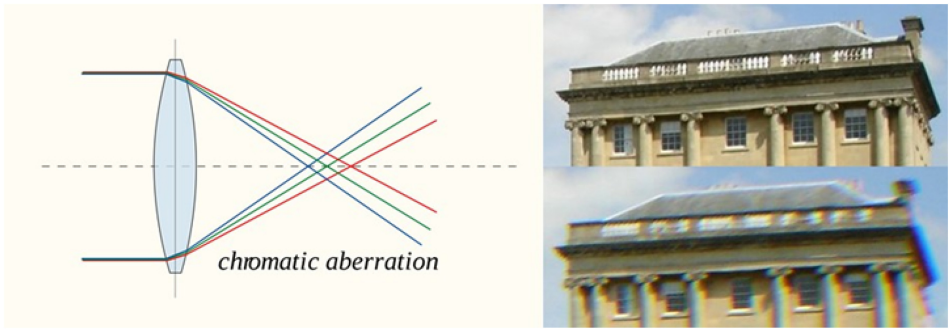


Figure 5.14. Left panel: Chromatic aberration of a single lens causes different wavelengths of light to have differing focal lengths. Right panel: Photographic example showing high-quality lens (top) compared to lower-quality model exhibiting transverse chromatic aberration (seen as a blur and a rainbow edge in areas of contrast). Taken and adapted from https://en.wikipedia.org/wiki/Chromatic_aberration.

spermatozoa. While van Leeuwenhoek did not use today's definition of what he documented, this still can be seen as the beginning of modern life sciences research.

Despite this success, the research for better microscopes lacked systematic efforts, since too many disciplines from glass fabrication to physics and engineering were involved. This changed dramatically thanks to Carl Zeiss (1816–1888), an engineer who started a collaboration with the physicist Ernst Abbe (1840–1905) in Jena, Germany. By intense theoretical studies and systematic practical investigation of the optical properties of lenses Abbe formulated the theory of (microscopic) image formation and revolutionized the entire field. Consequently, Abbe addressed optical aberrations in a holistic manner, asking for optical glass with new refractive indices and dispersion. Therefore, Zeiss and Abbe motivated the chemist Otto Schott (1851–1935) to develop optical glasses in Jena. While achromatic lenses were introduced in the 18th century by merging the images of two wavelengths (red and blue) into one plane, Abbe created the apochromatic lens by using the glasses of Schott. Here, three colors of the spectrum (red, green, and blue) are brought into focus in the same plane, while the spherical aberration is corrected for two wavelengths as well. In 1873 Abbe defined the resolution limit in wide-field microscopy as the wavelength of the light used divided by the sum of the numerical aperture of illumination and observation.

With the collaboration of Zeiss, Abbe and Schott a new age dawned, allowing the reliable fabrication of a comparably large number of high-quality optics. This started the boom of the use of the microscope in medicine and the life sciences. The differences in morphology (e.g., of a bacterium) are optically defined by differences in the refractive index only. These differences are, however, usually comparably small, which renders a lot of features hard to identify and observe. In the 19th century this issue was already solved by chemical staining of the samples. Until today, for bacteria it is common practice to color the samples by crystal violet. This staining technique is called Gram and divides bacteria into two groups: Gram-positive bacteria that become purple by Gram staining, and Gram-negative bacteria that appear pink afterward. Another example is Haematoxylin and eosin (H&E) staining, which is a standard staining technique in pathology.

In order to avoid the chemical treatment of the samples and to observe living cells phase contrast microscopy was invented in 1932 by Frits Zernike (1888–1966) and commercialized in 1941. The Nobel Prize in Physics was awarded to him for this in 1953. A few years later Georges Nomarski (1919–97) invented differential interference microscopy, a method visualizing the local differences of the refractive index of a sample.

The contrast of these methods is based on differences of the refractive index, ranging from roughly 1.35–1.5, depending on the water content, and causing optical inhomogeneities. However, these inhomogeneities often arise from differences in the chemical composition. A different kind of contrast is used by methods that are chemically specific. The most well-known property in relation to this is fluorescence. Here, the sample is illuminated by light with a short wavelength. Chemicals called fluorophores are exited by this light and re-emit the absorbed energy in a lossy process at defined higher wavelengths. This emitted light is then easily detected. The

most famous fluorophore is the green fluorescing protein (GFP). It is naturally expressed in the jellyfish Aequorea Victoria as described by Osamu Shimomura in 1961, but can be prepared also outside the Jellyfish, which was accomplished first by Douglas Prasher. Martin Chalfie and Roger Tsien managed to express the cloned GFP in bacteria in the 1990s for which they received the Nobel Prize in Chemistry together with Osamu Shimomura in 2008.

Further chemical contrast methods are based on the ability of molecules to vibrate. These vibrations can be excited either by visible light (strictly speaking, molecules vibrate also without being excited, and the excitation just increases the amplitude of the vibration), where a small part of the light interacts with the molecules in this way and has a lower frequency afterwards (Raman spectroscopy), or by infrared light the frequency of which is the same as the vibrations (infrared spectroscopy). The spectral signatures allow on the one hand an identification of the molecules as a spectrum can be seen as a fingerprint. On the other hand, changes of the chemical structure can be followed. More importantly, these techniques do not require labeling as in the case of fluorescence, which avoids an alteration of the molecule and a disturbance of cell processes.

To obtain as much information about morphology, structure, and chemical content of a sample as possible, it is of advantage to combine a number of different techniques or modalities within one instrument. Such devices allow multimodal microscopy and imaging and are of particular use for medical applications (e.g., in pathology) where they can complement the classical light microscope or serve as tools for surgical procedures, where they can reveal the borders of cancer and healthy tissue.

In recent years a fundamental limit has been circumvented, which we introduced already, namely the Abbe limit. A resolution limit of about 200 nm in wide-field microscopy often does not allow to identify relevant structures (e.g., to resolve distinct distributions of actin inside dendrites and spines in the brain). One prominent method to go beyond the Abbe limit is called STED, which stands for stimulated emission depletion in confocal microscopy. Put in simple terms, by this technique the fluorescence is suppressed except in the center of the focal spot, which usually leads to a resolution of 30–80 nm (figure 5.15), but values below 4 nm have been reported in literature. For using a stochastic method with significantly less phototoxicity than STED, stochastic optical reconstruction microscopy (STORM) was introduced. For ‘the development of super-resolved fluorescence microscopy’ Stefan Hell, Eric Betzig, and William E Moerner were awarded the Nobel Prize for Chemistry in 2014. In the meantime, a cornucopia of different (scanning) techniques became available like photo-activated localization microscopy (PALM), structured illumination microscopy (SIM), or total internal reflection fluorescence microscopy (TIRFM), the latter making use of the evanescent field of light.

The development of the microscope is still not at its end and further improvements can be expected to allow 3D imaging with high resolution from the components of cells up to whole organ and body imaging in video rate and beyond. It will still improve and with it its value for medicine and the life sciences.

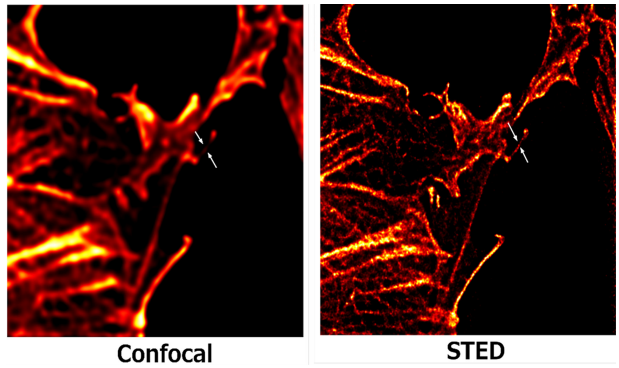


Figure 5.15. Resolution improvement between traditional confocal microscopy and stimulated emission depletion (STED) microscopy (adapted from wikipedia commons, from the page https://en.wikipedia.org/wiki/STED_microscopy under the license CC BY-SA 4.0).

5.4.1.3 Photodynamic therapy (Agostinis et al *2011, Popp* et al *2011)*

Photodynamic therapy (PDT) involves light and a photosensitizing chemical substance used in conjunction with molecular oxygen to induce phototoxicity and elicit cell death in diseased tissue. It is recognized as a treatment strategy that is both minimally invasive and minimally toxic to healthy tissue. PDT's advantages lessen the need for delicate surgery, lengthy recuperation, and formation of scar tissue and disfigurement. PDT is applied to treat a wide range of medical conditions in skin (e.g., acne, psoriasis), macular degeneration in eye retina, atherosclerosis, and malignant cancers (e.g., skin, bladder, lung, breast, head and neck, gastrointestinal), and has also shown some efficacy in anti-viral and anti-microbial treatments.

As depicted in figure 5.16, the workflow of PDT applications involves three components and is a multi-stage process. First, a photosensitizer (PS) with negligible dark toxicity is administered, either systemically or topically, in the absence of light. Second, light activates the PS when a sufficient amount appears in the diseased tissue. The wavelength of the light source needs to be appropriate for exciting the photosensitizer. Third, the PS produces radicals and/or reactive oxygen species (ROS) that are highly cytotoxic and kills the target cells. The light dose is controlled to supply enough energy for stimulation, but not enough to damage neighboring healthy tissue.

The key characteristic of a PS is the ability to preferentially accumulate in diseased tissue and induce a desired biological effect via generation of ROS. Among the specific criteria are also strong absorption with a high-extinction coefficient in the red/near infrared region of the electromagnetic spectrum (600–850 nm) allowing deeper tissue penetration. Whereas some PDT protocols required rapid clearance of PS within 1–2 days to minimize patient photosensitivity, prolonged production of photoreaction was reported for up to 20 days post-administration. Tailored PSs are key to the development of PDT. They are categorized as porphyrins, chlorins, and dyes. The major difference between PSs is the parts of the cells they target. 5-amino-levulinic acid (5-ALA) as member of the porphyrin family localizes in the mitochondria, m-tetrahydroxyphenylchlorin (mTHPC) as member of the chlorin

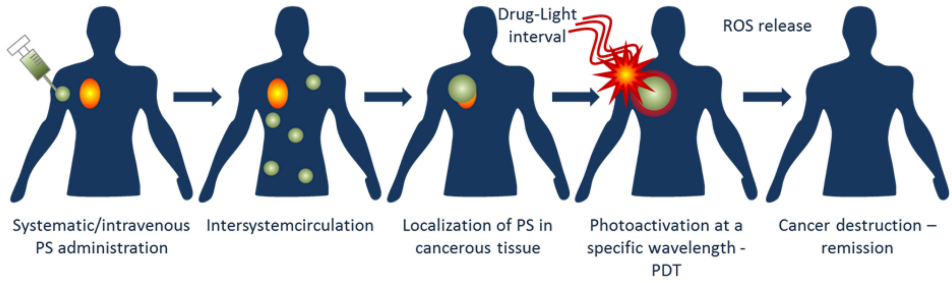


Figure 5.16. Workflow of photodynamic therapy (PDT) after photosensitizer (PS) administration.

family in the nuclear envelope, and the dye methylene blue in the lysosomes. PSs have also been attached to polyethylene glycol (PEG) copolymers that have a hydrodynamic size of 30–40 nm and provide combinatorial phototherapy with photothermal and photodynamic therapeutic mechanisms. To overcome the poor solubility of many PSs in aqueous media, particularly at physiological pH, alternate delivery strategies are used in new generations such as in liposomes and nanoparticles. Antibody-directed PSs have been developed to further enhance targetability.

Illumination systems are another key to the application of PDT. To achieve selected target cell destruction, while protecting normal tissues, the PS can be applied locally to the target area and globally illuminated (e.g., by exposure to daylight) or an external broadband (halogen) or narrow band (halogen with filters, laser, LED) light source. Local illumination offers another way for selected target cell destruction. For internal tissues and cancers, intravenously administered PSs can be illuminated using endoscopes and fiberoptic catheters.

5.4.1.4 Augmented surgery (Gorpas et al *2011, Popp* et al *2011, Krafft* et al *2018)*

Augmented reality (AR) is an interactive experience of the real-world environment where the objects that reside in the real world are enhanced by computer-generated perceptual information, including photonic modalities. AR can be defined as a system that incorporates three basic features: a combination of real and virtual worlds, real-time interaction, and accurate 3D registration of virtual and real objects. Whereas first AR experiences were introduced in entertainment and gaming business, subsequent AR applications have spanned commercial industries such as education, communications, and medicine. A potential medical AR application is to project photonic-generated data about pathological tissue margins into the surgical field of view which assists the surgeon during tumor resection. Neuronavigation is a related technique. Here, the advent of modern neuro-imaging technologies such as computed tomography and magnetic resonance imaging along with the capabilities of digitalization enabled the real-time quantitative spatial fusion of images of the patient's brain. The purpose of this technique was to guide the surgeon's instrument or probe to a selected target during neurosurgery. However, the accuracy of preoperative recorded images was limited due to the shift of soft brain tissue, in particular after removal of larger tumor masses.

Hardware components needed for AR are input devices, sensors, a processor, and a display. Modern mobile computing devices like smartphones and tablet computers contain these elements which often include a camera and microelectromechanical motion tracking sensors such as an accelerometer, GPS, and solid-state compass making them suitable AR platforms. Input device techniques include speech recognition systems that translate user's spoken words into computer instructions, and gesture recognition systems that interpret user's body movements by visual detection or from sensors embedded in a peripheral device such as a pointer, glove, or other body wear. Various display technologies are used in AR including optical projection systems, monitors, handheld devices, and systems that are worn on the human body such as head-mounted displays and eyeglasses; even contact lenses and virtual retinal displays are in development. AR applications often rely on computationally intensive computer vision algorithms. To compensate for the lack of computing power, offloading data processing via networks to a distant machine is often desired.

First research towards AR with biophotonic tools has been reported for fiber probe-coupled autofluorescence intensity and lifetime which enables label-free and real-time identification of cancerous tissue. Figure 5.17 illustrates how the

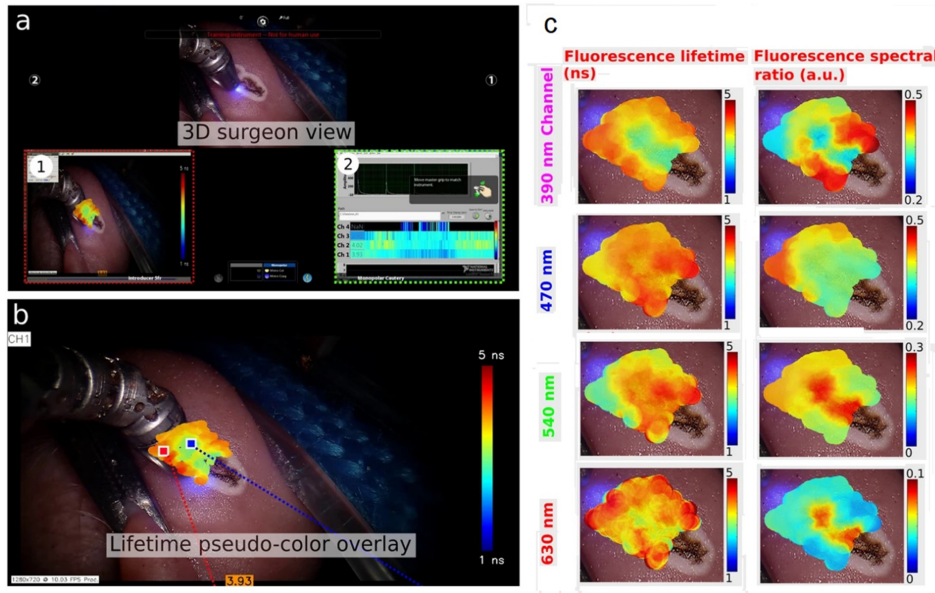


Figure 5.17. Augmentation of multispectral time-resolved fluorescence spectroscopy (ms-TRFS) derived data on the surgeon console during oral cavity surgery. The 445 nm aiming beam used to identify the location probed by ms-TRFS is visible at the distal end of the tool. The panels labeled with (1) and (2) correspond to the lifetime maps that can be visualized and the linear representation of lifetime values, respectively. (b) The white-light image augmented with lifetime values from channel 1 of the instrument. (c) Matrix of the distribution maps for the autofluorescence parameters. Note that each of these maps can be displayed/augmented in real-time during the scanning procedure if needed. Adapted from Gorpas *et al* (2011).

information acquired by a probe measurement technique is transformed into a two-dimensional image resulting in detailed visualization of the fluorescence-based contrast. For each location it takes 40 ms to acquire, process/analyze, and display the AR parameters on the surgeon's console.

Such an AR approach is particularly interesting during robotic surgery which can provide high surgical precision and improve recovery time for the patient and is becoming a preferred means for numerous cancer and disease treatment. Here, due to the loss of tactile feedback, the assessment of tumor margins is based only on visual inspection which is neither significantly sensitive nor specific. The presented augmented surgery can serve as a framework for the clinical translation of other point-scanning label-free optical techniques such as elastic scattering spectroscopy, diffuse reflectance spectroscopy, or Raman spectroscopy.

5.4.1.5 OCT—the challenge to translate an idea into an instrument ready for clinical use (Popp et al *2011, Fujimoto and Swanson 2016, de Boer* et al *2017)*

Optical coherence tomography (OCT) provides structural information of specimens in depth based on the change in the refractive index between microstructures. From a layman's perspective it is akin to ultrasound imaging but using light waves instead of mechanical waves. And while in ultrasound it is possible to electronically measure the time of flight between the excitation and the reflection, to the speed of light a low-coherence interferometry approach is used. The signal is generated through light scattering in structured samples with transition zones of chemically different composed tissue locations, resulting in changes in the refractive index and imaging contrast.

To record a signal, the specimen is illuminated with a broadband illumination source with the central wavelength typically located in the region between 400 and 1500 nm. The bandwidth of the illumination spectrum is proportional to the coherence length, and as such, related to the axial resolution (i.e., shorter central wavelength with broader bandwidth), results in higher axial resolution, while longer central wavelength with narrower bandwidth results in lower axial resolution. In either case the light is considered to have a low coherence, meaning if an interferometric measurement is performed (e.g., using a Michelson interferometer) only photons that travel the same distance will interfere, resulting in a measurable signal. The photons will be ballistically and elastically scattered by microstructures in the sample and detected in the same wavelength range. Typically used sources are broadband lasers, superluminescent diodes (SLED), or halogen lamps. The excitation light is guided into an interferometric detection scheme with a fixed reference arm and, in the case of time-domain OCT, a moving sample arm. Based on the interaction with the sample and the relative length-difference between sample and reference arm interference pattern can be measured or not. The detector is recording the scattered photons from different depths of the tissue and based on a measurable interference a depth profile can be recorded. Depending on the central wavelength and the bandwidth a spatial resolution of typically 2–10 μm can be reached.

Today, time-domain OCT systems are rarely implemented and most work is performed using Fourier-domain OCT, where two possible implementations are

available: spectral-domain and swept-source OCT. In the spectral domain a broadband excitation source is used in combination with a Michelson interferometer scheme. However, instead of a scanning arm spectrometer-based detection is used and a spectral interferogram, which contains the information about the depth profile of the sample, is used. This implementation significantly increases the acquisition speed, as no mechanical sample-arm scanning is required. The other method to acquire an interferogram is based on swept-source excitation, which uses a wavelength-tunable illumination source to generate the interferogram. Here, cheaper detectors can be used and problems, such as signal-roll-off in spectrometer-based implementations, does not occur, resulting in higher penetration depths.

OCT is a great example for the translation of a modality to clinical applications. As a matter of fact, OCT is one of the most widely used modalities for the diagnosis of eye-related disease, including glaucoma, age-related macular degeneration, diabetic retinopathy, diabetic retinopathy, and others. The human retina, which is a light-sensitive, multi-layered tissue located at the back of the eye, provides ideal conditions for the application of OCT. Due to the layered structure with very well-defined bands the retina is perfectly suited for analysis with OCT, and the application of the established method in ophthalmology and routine clinical diagnostics in eye diseases are performed on a daily basis. Thereby the retina is visualized as a three-dimensional, micrometer-resolved cross-sectional image with mm depth. Trained personnel can identify anatomical changes in the retinal layers and diagnose diseases, such as macular degeneration or glaucoma.

There are significant efforts to translate OCT to other clinical applications and different pathologies, such as intravascular imaging of atherosclerotic plaque, fibrotic diseases, and cancer diagnostics. While the premise of high spatial resolution with depth information is very intriguing there are methodical and technical challenges that need to be addressed first. For example, one of the problems is the access to the regions of interest where a measurement has to be performed. These locations are, except for skin, not directly and easily assessable with the typical systems and additional developments for *in vivo* endoscopy probes are required. There have readily been multiple implementations for *in vivo* endoscopy probes, which are typically based on four types of scanning geometries (i.e., micro-electro-mechanical-systems (MEMS), piezo-tube scanner, micro-motors, and optical rotary joints). The systems are technologically quite challenging, time-consuming, and usually manufactured individually. From the methodical point of view a significant challenge is the low penetration depth and the often heterogeneous, low-contrast data. The typically reached penetration depth of 1 mm is very sufficient to extract the relevant information about the layer structure. In many clinical settings the relevant features occur way below the depth of 1 mm, resulting in hard to interpret results. Moreover, tissue is not as well structured as the well-defined layers and the resulting image information is very homogeneous with very few features, making an analysis challenging. The homogeneity, at least at a depth of 1 mm, also appears when underlying changes in the physiological condition occur. An additional challenge is the interpretation of the data from *in vivo* studies and comparison to pathological reference information. Currently, in research it is standard to compare

OCT data acquired from biopsies with pathological hematoxylin and eosin (H&E) staining of thin section slides and a one-to-one correlation is often hard to achieve. This correlation of the visual information from OCT data becomes even more difficult in *in vivo* applications, as correlation between measurement sites and a removed sample is not easily achieved and has to be addressed, for example, through combination with other optical reference modalities.

In summary, OCT is a good example of the translation of an optical modality to real applications, particularly in ophthalmology. Moreover, there are multiple very promising clinical applications where the method offers great potential. Nevertheless, the translation also has to address methodical and technological challenges. Specifically, the development of novel and high-resolution ultrasound transducers in combination with image processing approaches readily reaching resolutions of sub 50 μm, while providing depths of up to 10 mm. The next few years will be very exciting in terms of technological development and additional translation to clinical applications.

5.4.1.6 *Point-of-care, particularly with respect to the use of light-based techniques for infectious diseases (Popp and Bauer 2015, Tannert* et al *2019, Matanfack Azemtsop* et al *2020)*

As mentioned above the first classification of bacteria was made possible by the advent of the earliest microscopes using different morphologies of the cells. Nevertheless, this information is limited, since only a few basic morphologies are known for bacteria. In a further step, bacteria with different cell-wall properties could be distinguished on the basis of Gram staining. Morphology and Gram staining is still used today in the routine identification of bacteria. However, with Robert Koch, the cultivation of bacteria became possible leading to the discovery of Anthrax and tuberculosis-causing bacteria. Koch also showed for the first time that the differentiation of bacteria (i.e., pathogens) based on biochemical tests was possible. This method is based on the fact that bacteria can metabolize different substances. Therefore, selective agars or enrichment agars are used to make different properties of the bacteria visible by means of indicator substances used during cultivation. The characteristic metabolic profile can then be used to distinguish the bacterial species. Today, the actual identification takes place with the help of automated systems such as the API (analytical profile index) system. The cultivation of bacteria in the presence of antibiotics with Etests or disk diffusion then provides information about potential (multi) resistances of pathogens.

In addition to cultivation-based methods, also nucleic acid-based techniques, such as polymerase chain reaction (PCR), are used to identify pathogens. Here, after biomass enrichment, first the DNA of the pathogens is isolated and then selected discriminative targets of this DNA are amplified in order to identify the species.

Cultivation-based tests are still the gold standard of bacteria identification; nevertheless, they require a lengthy cultivation time for biomass production. To shorten the complete identification process light-based methods are available. Using a short initial cultivation of only several hours, microcolonies (figure 5.18) can be used. Here the biomass of the microcolonies are dried as a homogeneous film on

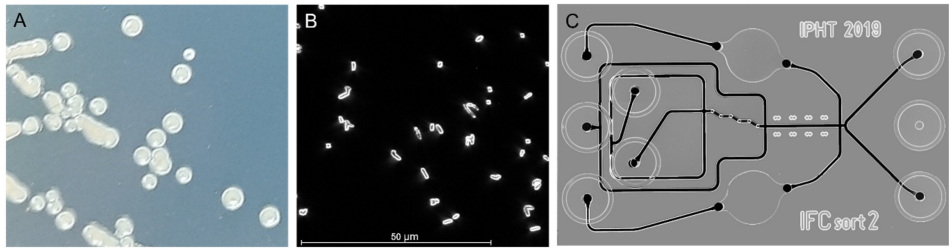


Figure 5.18. (A) Microcolonies on an agar plate; (B) microscopic image of single bacteria cells; and (C) microfluidic chip to sort bacteria.

suitable sample holders and the pathogens can then be identified by means of the molecular fingerprint methods Infrared or Raman spectroscopy (see section 5.4.1.2). Applying a Raman fiber probe even getting measurements directly on an undisturbed microcolony on an agar plate is possible.

When Raman spectroscopy is combined with a microscope, spatial resolutions on the order of bacteria size can be achieved. In figure 5.18(B) a dark field microscopic image of bacteria is presented. Using the excitation laser of the Raman microscope, single bacterial cells can consequently be measured. It is therefore possible to identify bacteria directly by analyzing the fingerprint Raman spectrum, without prior cultivation. Nevertheless, the bacteria cannot always be found as a pure sample. Therefore, to apply Raman microscopy it is necessary to isolate bacteria from their natural habitat. For pathogens this means destruction-free isolation methods need to be acquired for medical samples (e.g., blood, urine, or ascites) but also from environmental samples (e.g., water or soil). Despite the isolation time, identification based on Raman microscopy can be carried out in under two hours. After the identification of the pathogen species discrimination of sensitive and resistant isolates is also possible with this method. However, it is not possible to determine to which antibiotic the isolates are resistant or whether multi-resistant pathogens are present.

Microfluidic chips can be used to obtain further information on the resistance profile of individual pathogens. By combining Raman microscopy with different microfluidic chips, the changes of pathogens in the presence of different antibiotics can be analyzed. In this way, a resistance profile of the pathogens is obtained within approximately 90 min.

Besides the identification of species or the determination of resistances, microfluidic chips in combination with Raman microscopy can also be used for other applications. Since the Raman spectrum of a cell reflects normally the whole biochemical profile of this cell, the different contents in nucleic acids, protein, carbohydrates, and lipids may lead to an overall differentiation of phenotypes, growth states, or physiological states. Using this Raman spectroscopic information inside a microfluidic sorting chip (figure 5.18(C)) the cells can be then sorted into the different classes. The method is especially suitable for bacterial consortia which are heterogenous and/or cannot be cultivated. One application is the differentiation of

cells with or without specific biomarkers such as special pigments, storage material like PHB, or secondary metabolites.

Another application uses stable isotopes which are metabolized during cell growth and can be recognized by characteristic changes inside the Raman spectra. Applying the isotope incorporation technique, it is possible to differentiate microorganisms according to, for example, metabolic activity or carbon metabolic pathways. In addition, the Raman-based cell sorting allows to add subsequent other analytical techniques such as PCR or whole genome analysis to the different sorted sample portions and to gain even more information about the sample.

Overall, the combination of Raman spectroscopy, chip-based sampling strategies as well as chemometric spectroscopic data analysis methods represents a powerful point-of-care approach comprising the entire process chain (i.e., from sampling to the final diagnostic result). By doing so, culture-free isolation and identification of pathogens, their host response, and their antibiotic resistance can be achieved. Thus, this promising approach offers the potential to overcome an unmet medical need by reducing the critical parameter 'time' to initiate a personalized lifesaving therapy as compared to the gold standard microbiology.

5.4.1.7 Outlook and future challenges

Big data—artificial intelligence

The most dramatic leap in the use of light-based concepts for medical diagnostics and therapy will come from advances in digitalization. For a broad/interdisciplinary use of optical/photonic methods for medical diagnostics and therapy, automated evaluation approaches are indispensable. They only allow the researched photonic approaches to be used in the sense of a qualitative or quantitative evaluation. The mere observation of the generated photonic data such as spectrum or image is very often not sufficient). In this context, artificial intelligence (AI) concepts utilizing deep learning approaches in particular are playing an increasingly important role in the successful implementation of light-based methods. Intensive cross-fertilization between optics/photonics and AI will enable experiments and photonic technological concepts to be advanced to such an extent that routine use far beyond specialized optics laboratories becomes conceivable. Because of the fact that the measurements of spectra, images, and time traces becoming easier and faster the speed of the analysis needs to be accelerated as well. This requires fully automatic data analysis pipelines to minimize the user intervention and to speed up the analysis by optimized algorithms. Further leaps in innovation can be expected in particular if research into new AI-based light analysis tools is carried out in line with research into new light-based processes. In this context, multimodal methods are particularly worth mentioning. As mentioned above, it has proven to be very advantageous to combine several optical methods in order to significantly increase the information content or to expand the application range of optical methods. Especially in such multiscale spectroscopy and correlative multimodal imaging, in which two or more optical modalities are used in combination, AI-based concepts can on the one hand guide the design of the light-based measurement methods in (e.g., the sense of

compressed sensing), and on the other hand AI concepts are the key to the application of such new multimodal photonic technologies, in that extreme amounts of data (big data) can be managed and visualized and evaluated in a user-centered manner.

Overall, there are great synergies between light/optics/photonics and AI, which are just being tapped in the context of light-based health technologies. Light or optics/photonics form an ideal platform for the application of AI concepts. In the first place, the automated evaluation of large amounts of data (big data) should be mentioned here. In addition, AI opens up great potential in the derivation of secondary data and conclusions from the primary light-based information, which in turn enables new design possibilities for optics and sensor technology. In doing so the optical process can be revolutionized: if, for example, the core purpose of an imaging system is to quantify a certain application-specific pattern, then it must be optimized above all to optimally illuminate, transmit, image, and recognize this type of correlation. The requirements for pattern recognition must therefore already be represented at all levels in the design process. This is only possible through digital design methods. AI-based algorithms will also play a major role in the recording of image data in the future. In this context, AI-based compressed sensing and AI-based compressed acquisition as well as AI-generated inverse modelling of the measurement processes should be mentioned.

New light applications—smart photonic materials

As explained above, with the help of STED and TIRFM microscopy, it is already possible today to obtain a high-resolution 'live image' of individual body cells. STED and STORM microscopy build on light microscopy and allow an enormous leap in the resolution of images. Until then, the limits of light microscopes were defined by wide-field imaging. STED microscopy is based on findings and advances in quantum physics/optics. Overall, quantum physics has given us a new understanding of the connection between light and information as well as access to new light sources and measuring methods or contrast phenomena for medical diagnostics and therapy in recent decades.

The so-called second quantum revolution will usher in another new era of optical imaging methods for biomedical issues through the use of non-classical light sources in the coming years. By using quantum effects, it may be possible in the near future to create images beyond our current state. The measurability of new characteristics of a light field, beyond the intensity distribution, opens up new conceptual and technological possibilities for optical health technologies. Here, light sources that can generate noise-free quantum states of light, imaging methods that rely on the correlations between different photons, and detector systems going far beyond the concept and capability of classical cameras to measure correlations in light should be mentioned. Furthermore, progress in the generation of extreme light (i.e., ultrashort pulses in the XUV or x-ray) allowing intensities higher than currently achievable will open up completely new possibilities for medical imaging and diagnostics.

The ability to structure materials on the nanoscale range creates the prerequisite for a completely new control of the propagation of light, which does not occur in

natural materials. Particularly outstanding is the ability of nanosystems to concentrate light to regions well below the wavelength. The combination of these nanophotonic materials with elements of classical and quantum optics opens up imaging beyond classical limits (e.g., the Abbe Limit in wide-field imaging).

In summary optical medical research in the future must and will be researched along the entire chain, starting from the light source, via the light–matter interaction and detection to information evaluation, as a holistic and interdisciplinary problem in order to rethink it and design it for the future. By converging photonics, electronics, material design, and big data new optical health technologies will emerge to detect and fight diseases far before their onset. This is particularly important not to lose the fight against rising healthcare costs, requiring a massive restructuring of healthcare. Instead of the mere treatment of acute diseases, the prevention of diseases, the shortening of the time to convalescence, and the improvement of the quality of life must be given more and more priority in the future. This requires, in particular, personalized and precise treatment and an alignment of principles along the 'continuum of integrated care'. Here, new photonics concepts utilizing the above-mentioned approaches will play an important role.

Translational hurdles

In addition to the continuous further development of photonics technology research in the area of developing new, powerful hardware and software components, a particularly great challenge in the coming years will be simplifying and shortening the path of technological innovations into clinical routine. Translating biophotonic research results faces major challenges, especially with regard to, for example, the EU Medical Device Regulation (MDR) for translational research. Currently, the regulation makes it significantly more difficult or impossible to test biophotonic approaches on patients in the form of preclinical or clinical studies. Numerous light-based technologies have proven their potential for certain diagnostic and therapeutic questions in proof-of-principle studies, but the actual performance has not yet been demonstrated under routine clinical conditions in the form of comparative studies on a large cohort of patients. Here, funding for such validation studies is urgently needed to generate a marketable product. It only makes economic sense for industry to support a biophotonic proof-of-concept approach if there is a regulation-compliant study on a large cohort of patients that clearly demonstrates the added value.

What such translational research could look like in concrete terms is demonstrated by the 'Leibniz Centre for Photonics in Infection Research' (LPI) (see https://lpi-jena.de/en/), which was recently included in the national roadmap by the German government. Within the framework of the LPI, diagnostic and targeted therapeutic light-based approaches will emerge through the combination of photonic methods with infection research, which, after appropriate approval, will be transferred directly to industrial production and clinical application. A central element within the framework of the LPI are technology scouts who, at the beginning of the value

chain, work together with clinicians to identify precisely fitting photonic solution approaches that have already successfully proven their potential in proof-of-concept studies for specific medical issues. LPI can thus serve as a blueprint for other medical problems, such as cancer or neurodegenerative diseases, to overcome the Valley of Death of clinical translation in these areas as well.

5.4.2 Synchrotrons and free-electron lasers for structural biology

Henry N. Chapman[1,2,3]
[1]Center for Free-Electron Laser Science (CFEL), Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany
[2]Department of Physics, Universität Hamburg, Hamburg, Germany
[3]Centre for Ultrafast Imaging, Hamburg, Germany.

The field of structural biology studies the forms of macromolecules such as proteins, RNA, and DNA, and the larger complexes that they can produce. Proteins are the nanomachines that drive the processes of life, and DNA and RNA carry the information to produce those machines.

Mammalian cells, for example, usually contain over 40 million macromolecules to catalyse reactions, regulate the flow of molecules for energy or signalling, sense and respond to other stimuli (such as forces), exert forces, repel pathogens, repair damage, or replicate and divide. By revealing the structures of proteins at the atomic scale, the mechanisms and functions of proteins can be uncovered, giving insights into disease, infection, the action of medicines, energy conversion, and food production. More recently, it has given the ability to design and construct new functioning materials. Accelerator-based x-ray light sources, and in particular synchrotron radiation facilities, have been largely responsible for creating and propelling the field of structural biology over the last thirty years by revealing protein structures through the method of macromolecular crystallography, as well as the study of the interactions between proteins via solution scattering techniques. There are presently several large developments in the field that promise profound changes and a more complete understanding of biological processes. The first is the ability to predict the 3D structure of a protein from its genetic code, using artificial intelligence methods based on examining the huge number of structures determined so far. Another is the effort to map out every protein molecule in the cell to understand the entire system, gained primarily through improvements in electron microscopy. Recent breakthroughs in electron microscopy also provide structures of proteins from many images of single uncrystallised molecules, which can be sorted and classified to observe conformational variabilities of proteins. Finally, large-scale, x-ray free-electron laser facilities have come online and are now providing movies of proteins undergoing reactions and responding to stimuli, and providing structures over relevant ranges of physiological environmental conditions that are not accessible in electron microscopy (Schlichting 2015, Spence 2017, Chapman 2019). Together, these developments are ushering in a new technological age of protein design using the raw materials of amino acids to create molecular machines to achieve specific functions (Huang *et al* 2016).

Proteins are linear polymers made out of 20 different amino acid building blocks. These are arranged in the sequence prescribed by the genetic code. Each amino acid type has different physical–chemical properties, such electronegativity and hydrophobicity, which together influence how the polymer chain folds and coils into its unique complex three-dimensional structure that confers the protein with particular

structural properties and functions. Such functions range from catalysing specific reactions, regulating the flow of signalling molecules, sensing and responding to those signals as well as other stimuli (such as forces). By actually imaging the molecular structure, an understanding of the mechanisms of proteins can often be revealed. This necessarily requires radiation of short enough wavelength to resolve interatomic spacings. For x-rays there are no atomic-resolution lenses to directly form this image, so the molecular structure must be inferred from the scattered radiation that interferes to produce a diffraction pattern. One of the earliest and most iconic examples is Rosalind Franklin's Photograph 51 of DNA fibres, taken in 1952 using a laboratory x-ray tube and which led to the proposal of the double- helix structure by Francis Crick and James Watson.

The first protein structures soon followed, applying an earlier discovery that purified proteins can be crystallised. Crystals give strong, measurable diffraction peaks, called Bragg reflections, which can be mapped out in three dimensions as the crystal is rotated. The strength of each reflection is proportional to a particular spatial modulation of the electron density of the molecule. A complete set of these spatial frequencies can be combined to create an image of the molecule, except for the fact that the phase of the diffracting wavefields cannot be recorded (which give the relative alignments of the spatial frequencies). This was first solved by an approach akin to holography, by recording diffraction from crystals with and without a heavy reference atom. Those first structures, hemoglobin and myoglobin, earned Nobel Prizes to Max Perutz and John Bernal. The molecular structure of hemoglobin immediately revealed how this molecule has a high affinity to oxygen when blood cells are in the lung and yet release it in other parts of the body to power many other cellular functions. The protein binds four oxygen molecules but undergoes a conformational change upon binding the first, making affinity dependent on oxygen partial pressure. Now, the protein data bank, an on-line repository of protein structures, has accumulated over 200 000 structures, indicating the enormity of the global efforts of structural biology. The vast majority of those structure were generated in the last two decades using synchrotron radiation facilities with specialised beamlines and instruments dedicated to macro-molecular crystallography.

One reason for the success of crystallography is that by crystallising a protein, the molecules are brought into an alignment and regularity which provides a cooperative increase of the diffraction signal, proportional to the number of molecules or the volume of the crystal. This amplification of the signal enabled large enough crystals to be measured with weak laboratory x-ray sources (sometimes over days) and which now can be accomplished even with microscopic crystals in seconds at a synchrotron radiation facility. Accelerating charged particles radiate, and when those particles are moving near the speed of light (as they are in accelerators such as storage rings and synchrotrons that were built for high-energy physics experiments) they emit very energetic x-ray beams that are directed into a narrow cone in the forward direction due to the relativistic motion of the charges. Although a nuisance for the experiments the machines were built for, this directed radiation was a boon for crystallography which requires a collimated or laser-like beam to impinge on the crystal. The synchrotron radiation from electrons in storage rings was initially

exploited parasitically from the early 1970s, and some of the first uses were to measure the weak scattering from biological fibres and macromolecular crystals. Particle physics laboratories such as DESY in Germany, the SLAC National Accelerator Laboratory and Brookhaven National Lab in the USA, Daresbury in the UK, and KEK in Japan were centers of innovation which designed and constructed dedicated machines, optimised to create bright beams of x-rays, tunable over wide wavelength ranges. A particular innovation was the development of the wiggler and undulator devices—periodic magnetic structures that oscillate electrons side to side as they travel—to provide further amplification factors proportional to the square of the number of periods in the case of undulators. Designs of the storage ring electron optical layout confined the electron beam to ever smaller dimensions to increase the brightness of the x-ray beams they produced. There has been an unabated exponential increase in the brightness output of new or upgraded facilities as a function of time that continues even today, so that we now have an increase of over 10^{22} in average brightness at a so-called fourth-generation facility as compared with the first storage rings. Today, there are more than 50 facilities all over the world, with each facility typically housing 20 or more beamlines, each with specific instrumentation for various analysis techniques for the study of all phases of matter. The instruments serve broad disciplines such as materials science, medical imaging, atomic and plasma physics, and structural biology (Jaeschke *et al* 2020).

Synchrotron radiation facilities began to impact and define the field of structural biology in the 1990s with the construction of dedicated macromolecular crystallography beamlines consisting of a precision goniometer to orient and rotate the crystal and a large-area detector to digitally capture the diffraction patterns that could be directly processed to obtain an electron density map. Unlike laboratory sources which primarily emit at discrete fluorescence photon energies, the wavelength of synchrotron radiation sources can be continuously tuned. This provides a powerful approach to solve the 'phase problem' mentioned above, by exploiting changes in the scattering properties of specific atoms at energies near resonances in their absorption. This method of multiple-wavelength anomalous diffraction (MAD) phasing allowed the structures of native metalloproteins to be obtained, as well as proteins containing the amino acid methionine (which contains a sulphur atom and which can be replaced with selenomethionine containing a more suitable selenium atom). Other proteins not amenable to this method could be solved by a method called molecular replacement, where a structure of a similar protein is used as a starting point to refine the atomic coordinates, subject to many constraints such as bond lengths and angles. As the x-ray brightness of facilities increased, robotic handling of crystals and the inexorable improvements in computing capabilities removed bottlenecks and opened up new possibilities for large-scale screening experiments. Today, as long as the quality of the crystal is sufficient, the protein structure can be obtained in minutes—a task that was worthy of a PhD dissertation at the turn of the twenty-first century.

The short wavelength of x-rays needed to resolve molecules at the atomic scale unavoidably means that the radiation is energetic enough to ionise those atoms. Biological materials cannot withstand too much x-ray exposure, as determined by a maximum tolerable dose (energy deposited per atom or per unit mass) at which the

very structure under investigation is degraded. Obtaining large crystals can be difficult, and while bright x-ray sources can make up for the loss of diffracting strength (proportional to crystal volume) this comes at the cost of increased dose and damage through the larger exposure. The radiation sensitivity of protein crystals can be extended by cryogenically cooling them, to immobilise the products of radiolysis, such as free radicals, and prevent them from reacting with protein molecules.

Without the practice of cryo-crystallography it would not have been possible to obtain the vast majority of the 200 000 structures in the protein data bank. The observations and analyses of the molecular structures revealed using synchrotron radiation have brought fundamental new understanding and knowledge about biology. The structure of the green fluorescent protein (GFP) from a jellyfish showed how a barrel-like arrangement of the protein chain that houses a chromophore confers bright green fluorescence and showed how to mutate the protein to change the emission wavelength, thereby providing a visible marker that can be genetically targeted to other proteins for studying cancer and other diseases, and revolutionising microscopy of the cell. Synchrotron radiation also revealed the structure and mechanism of adenosine triphosphate (ATP) synthase, the remarkable macromolecular machine that produces the molecule ATP that is used as a fuel to drive most protein reactions. The ATP synthase is powered using a potential gradient across the cell membrane. Many proteins are located in the lipid membranes of cells, which are notoriously hard to crystallise, but cryo-cooling gave the possibility to measure some small crystals of ion-channel proteins to reveal how the channel can selectively pass potassium ions and not sodium ions, at the extremely high rates as needed for electrical signalling in the nervous system. And the technique was crucial to obtain the structure of the ribosome—the large complex machine consisting of protein and RNA which assembles new proteins amino acid by amino acid, by reading the genetic code provided by a strand of messenger RNA (mRNA). All these examples garnered separate Nobel Prizes (Jaskolski *et al* 2014). There were two more for the structure of a G-protein coupled receptor and the structure of the protein that transcribes a DNA sequence into the mRNA strand.

Despite all these successes, cryogenic cooling may lead to subtle changes in the molecular structure which may mislead the interpretation of how a protein catalyses a reaction, for example, or how a drug molecule can bind to and inactivate a viral protein. Cooling a protein to liquid nitrogen temperatures also prevents us from studying the evolution of a reaction or studying protein dynamics. In 2009, a revolutionary new machine was brought online that can address this problem: the x-ray free-electron laser, or XFEL. Unlike circular storage rings, an XFEL employs a linear accelerator coupled with a 100 m long undulator device to create x-ray pulses with a billion times higher peak brightness (outpacing the previous trend of synchrotron radiation improvements). Each pulse has a duration of tens of femtoseconds, packed with as many photons as used to generate a protein crystal dataset at a synchrotron. The x-rays are created from accelerated electrons via an amplification process in the undulator. With energies of several giga-electron-volts, the electrons travel at close to the speed of light, and therefore they keep up with the light field that they generate in the undulator. That light field in turn influences the

trajectories of the electrons and forces them to separate into bunches that are a wavelength apart. The bunched electrons radiate in phase, giving an amplification proportional to the square of the number of electrons. The process grows exponentially along the length of the undulator due to this positive feedback. As with synchrotron sources, the machine addresses a plethora of scientific needs, but emphasises ultrafast processes on the timescale of atomic motions and chemical processes. Today there are five operating x-ray FEL facilities in the world, with more in construction. The largest is the European XFEL in Germany, a facility that is three kilometers long with the potential to operate five undulators simultaneously (figures 5.19 and 5.20).

The short intense pulses of XFELs break the dependence between exposure, dose, and crystal size. When focused to a small spot, a single XFEL pulse vaporises any material in its path by stripping electrons from all atoms, creating an expanding plasma. However, it takes time for atoms to begin moving since they have inertia, and by the time atoms have moved a fraction of a bond length the pulse has already

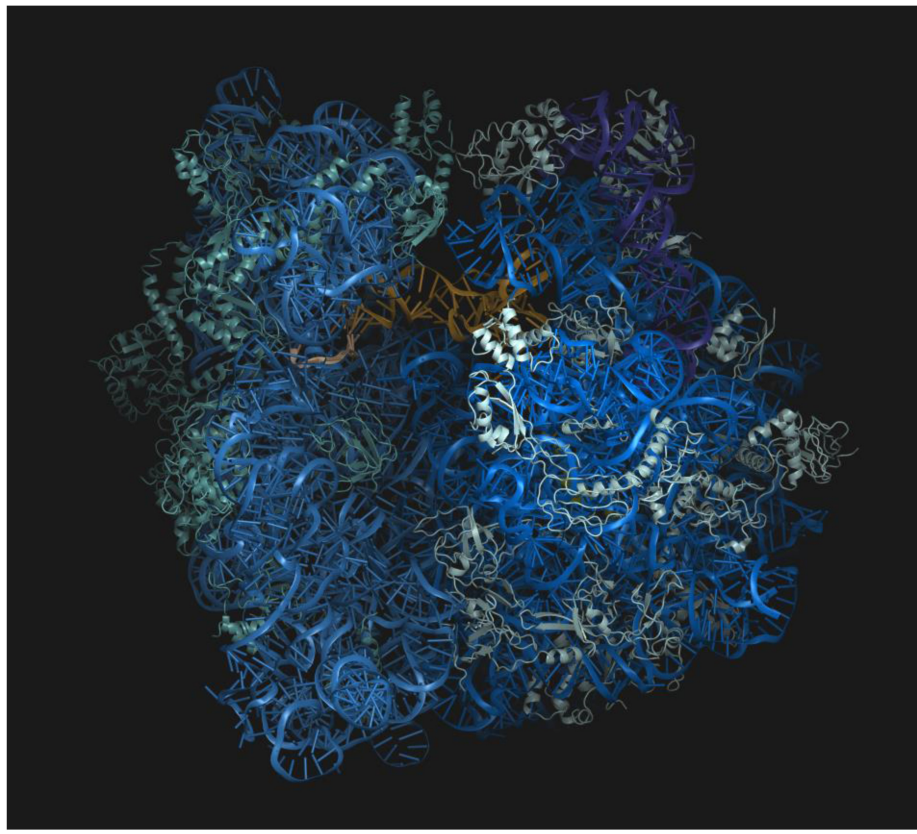


Figure 5.19. The molecular structure of the ribosome, as determined by x-ray crystallography at DESY in Hamburg, showed researchers how this complicated nano-machine synthesises new proteins by reading genetic information encoded in messenger RNA molecules (credit: Joerg Harms, MPSD).

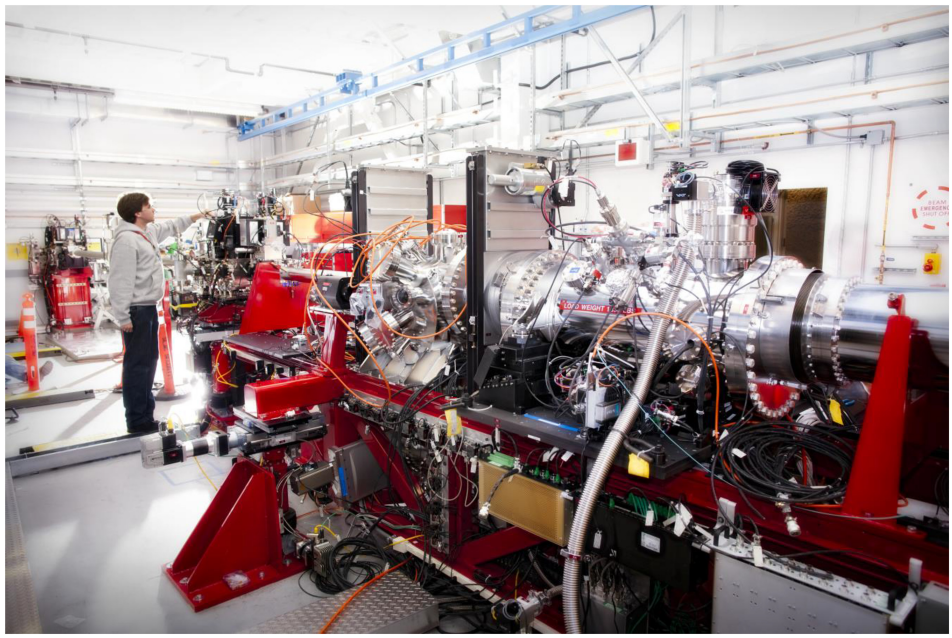


Figure 5.20. A beamline at the Linac Coherent Light Source of the SLAC National Accelerator Center in USA, used for serial femtosecond crystallography measurements (credit: SLAC National Accelerator Laboratory).

passed (Neutze *et al* 2000). Thus, the exposure to biological materials and protein crystals can far exceed the previous dose limits, giving much stronger diffraction. This concept of 'diffraction before destruction' opened up the method of protein nano-crystallography (Chapman 2019). Since a crystal lasts only one pulse, the full three-dimensional diffraction dataset requires a stream of nanocrystals to be rapidly fed into the beam one by one: at the European XFEL located in Hamburg, Germany, pulses arrive with as little as 220 ns between them. One method is to deliver crystals across the path of the x-ray beam in a high-speed liquid microjet. Crystals are exposed in random orientations, which must be inferred from the diffraction pattern itself. This approach, called serial femtosecond crystallography, measures many tens of thousands of individual crystallites in minutes, at the appropriate physiological temperature, and gives structures that are completely free of radiation damage. Most importantly, the short exposure time of each pulse freezes out all motion and can capture various stages in chemical reactions at femtosecond precision. That is most readily achieved in photoactive proteins (such as those involved in vision or photosynthesis) by triggering a reaction with a short optical pulse from a laser that is precisely timed to arrive at a particular time before the x-ray pulse. Datasets can be collected at different delay times and the resulting structures assembled into a 3D movie. The first conformational changes of the chromophore responding to a visible light photon have now been captured in exquisite detail and several groups are racing to understand how the large photosystem II complex breaks down water molecules and creates a separation of charges

(to power ATP synthase), releasing molecular oxygen as a by-product (Brändén and Neutze 2021). Such knowledge will help us to understand how to more efficiently (and cleanly) capture the energy of sunlight.

5.4.2.1 Futures

Nature utilises 20 different amino acids to make proteins and so for an average sized-protein of about 300 amino acid residues there are $20^{300} = 2 \times 10^{390}$ different possibilities of protein structures. There is not enough matter in the Universe to make one copy of each of these, which of course dwarf the number of unique proteins thought to exist in nature (less than 10^{10}), let alone the number of proteins whose structures are solved (200 000 in the protein data bank). While not every random sequence of amino acids would create a protein to do something useful (just as not every random sequence of letters produces interesting literature) it is clear that nature has only sampled a tiny fraction of the possible useful protein structures that can be made. There is therefore a goal to design new proteins to carry out functions of nanomachines or form new materials, and specifically to design new proteins so far from nature that they cannot be thought to have evolved from existing sequences. Using our literature analogy, we can aspire to form completely new genres. This requires the long-sought ability to predict the three-dimensional fold of the protein —the way the polymer chain coils and bends on itself—from the sequence of amino acids that make up that chain.

In the year 2021, that ability was demonstrated to remarkable precision by the computer program AlphaFold 2, developed by Google labs (Jumper *et al* 2021). The program uses a deep neural network that was trained on the structures in the protein data bank—most of which were determined using x-ray crystallography at synchrotron radiation sources (Burley *et al* 2022) Are those sources now obsolete? The answer seems to be: no, and further experimental capabilities are needed. Proteins are not static objects, and as seen from the very first structure of hemoglobin, their conformations change as they carry out their functions, respond to their environment, and interact with other proteins. The macromolecular machines that produce ATP or transcribe DNA are exceedingly complex and intricate in their mechanisms, which are only currently understood at a superficial level. Many such machines are driven by chemical or voltage gradients, Brownian motion, or entropy, and it is not yet possible to predict how a such a system operates or how it responds in changing environments. Even the act of breaking such a mechanism—to inhibit the action of a virus or other pathogen—cannot be predicted from a static structure. This knowledge requires mapping out the conformational energy landscapes of proteins in full detail and under a range of environmental conditions or subjected to stimuli. That is, structures must be measured as they fluctuate or are directed to carry out reactions. As seen above, XFELs are answering this call, using intense femtosecond pulses to collect diffraction snapshots of small crystals, but more is required to expand the scope of these measurements and to extract structures and dynamics from large datasets.

Serial crystallography at XFELs must be expanded to achieve high-throughput multidimensional measurements. The kinetics of drug binding, for example, could be mapped by measuring diffraction at various times after mixing crystals with those

molecules, over a range of temperatures. Currently the time needed per measurement point in this matrix is limited by the repetition rate of the x-ray source (which can potentially reach millions per second) and the matching rate of the detector. It may be feasible to probe tens of conditions per second. Since serial crystallography can be carried out by ejecting crystals across the x-ray beam in liquid jets, the experiment can be controlled and automated using microfluidic technologies. In this way, entire libraries of potential drug molecules or compounds could be studied, greatly expanding current activities of crystallographic screening. During the current COVID-19 pandemic, crystallography-based searches for compounds that inhibit certain viral proteins have taken place, measuring thousands of conditions over month-long campaigns, or several minutes per compound. This potentially could be accomplished in minutes. With automation comes the ability to control and use the analysis of resulting structures to guide measurements and to optimise searches. Development must be made to integrate production, purification, and chemical synthesis with the analytical capabilities of serial crystallography.

Today, molecular movies have been made for proteins undergoing reactions triggered by short light pulses, precisely timed to x-ray pulses. While such studies have given some general insights of the actions of proteins, these photoactive proteins represent but a small fraction of all proteins. The means is required to precisely trigger reactions in systems that are not sensitive to light. One promising method is to engineer caged compounds, which are released by a flash of light to begin the reaction. More generally, there are efforts to use advanced data processing and pattern matching techniques to precisely order a set of snapshots recorded at fluctuating times from experiments where the temporal evolution of a process is only loosely imposed. This relies upon machine learning to place those snapshots on a manifold, in a high-dimensional phase space, which must be discovered in the data itself. Recent tests have shown a remarkable ability to achieve a time resolution much better than the duration of the x-ray pulses, perhaps aided by temporal fluctuations in those pulses (Hosseinizadeh *et al* 2021). Such machine learning methods could be used to feed back into the generation of the temporal profiles of x-ray pulses, perhaps by modulating the electron beam in the accelerator, to optimise the ability to place events according to their sequence in time.

The crystallisation of proteins has been key to obtaining structures of macromolecules using x-rays from the earliest days. An alternative approach is to use a powerful electron microscope to image the molecules directly, but this suffers from a similar dependence of damage on dose as for x-rays. It is now possible to assemble 3D molecular structures at atomic resolution from very noisy low-exposure images, but this too requires cooling them to cryogenic temperatures. Nevertheless, cryo-electron microscopy is revolutionising structural biology by obtaining structures of molecules that cannot crystallise (Kühlbrandt 2014). And while structures may be distorted by the low temperatures, it is possible to sort the images into different structures to observe distinct conformations (such as a ribosome in different stages of constructing a polypeptide). X-ray free-electron lasers promise similar capabilities, but without the need to cool the molecules. This is being pursued by reducing the size of crystals in diffraction experiments down to the smallest they can go—to

single molecules. The diffraction patterns of single molecules are certainly weak without the amplification provided by the crystalline lattice but, by using similar algorithms as those used to sort and assemble data in cryo-electron microscopy, it may be possible to build up diffraction datasets of single molecules that are streamed across the x-ray FEL beam as an aerosol or in a thin liquid jet. As is the case for crystals, it should be possible to trigger reactions and obtain structures as a function of time, but now molecules that follow different reaction pathways could be sorted and examined independently. From large ensembles of single-molecule data it will be possible to capture rare events that happen spontaneously, such as the precise moment a molecule binds.

High-power x-ray FEL pulses are needed for single-molecule diffractive imaging, which might be achieved by adapting chirp-pulse amplification schemes to the x-ray regime. Stronger diffraction signals will require even shorter pulses—at the scale of attoseconds—to ensure 'diffraction before destruction' of single molecules. The total pulse energy may not necessarily be much larger than what is generated today using a kilometer-long linear accelerator. With high enough control of electron beams with the next generation of particle accelerators it may be possible to shrink the facility to a size that can fit into a laboratory. Together with electron microscopes, large-scale x-ray facilities, and computational prediction, these will help to vastly expand the study of the machinery of life and usher in the new technical age of protein design—the amino age.

5.5 Physics research against pandemics

Thierry Mora[1], Chiara Poletto[2], Marta Sales-Pardo[3] and Aleksandra M Walczak[1]
[1]Laboratoire de physique de l'école normale supérieure, CNRS, PSL University, Sorbonne Université, and Université de Paris, 75005 Paris, France
[2]INSERM, Sorbonne Université, Pierre Louis Institute of Epidemiology and Public Health, Paris, France
[3]Department of Chemical Engineering, Universitat Iovira i Virgili, 43007 Tarragona, Catalonia, Spain

5.5.1 Introduction

Diseases and epidemics have always accompanied humanity. Viruses need hosts to survive and both humans and livestock fit the role. Science has contributed significantly to our current ability to fight infections through vaccination, medicine, and therapy.

Epidemics are emerging collective phenomena. In 2019 the SARS-CoV2 pandemic hit the world. The curve of new cases we were checking every day is the result of interacting phenomena occurring at completely different spatial and temporal scales. More specifically it is the result of interacting systems that co-evolve. Humans react to the epidemic and change their behavior in the attempt to contain it. Viruses, on the other hand, continuously mutate and diversify. Can quantitative science in general, and physics in particular, help us grasp this complexity and better predict and control? We can schematically break down this complex interdependence in four main steps, from the micro to the global scale (figure 5.21): (1) predicting the emergence of a new mutant variant strain; (2) predicting its ability to escape the hosts' immune system; (3) predicting its spread through the population; and (4) predicting the behavior of the population in response to the epidemic. Each one of these scales contributes to whether an outbreak results in an epidemic or not, and how this epidemic will progress and evolve. Physicists have been working on all of these scales, bringing about both important methodological and practical insights that have had a deep impact on biology, medicine, and public health. Here, we walk through these scales, showing how they interact.

5.5.2 The physics of host–pathogen evolution

5.5.2.1 The diversity of immune responses

At any time, each person is surrounded by many pathogens. Our immune systems manage to recognize most of them in time and stop us from getting sick. Our immune systems protect us by producing a wide range of defense mechanisms ranging from mechanical and anatomical barriers, such as our skin, physical reactions such as scratching or crying, to the action of cells of the innate and adaptive immune systems. The innate immune system provides a first line of defense by recognizing stereotyped properties of pathogens. Cells recognize general properties of foreign, potentially harmful organisms such as molecules characteristic of

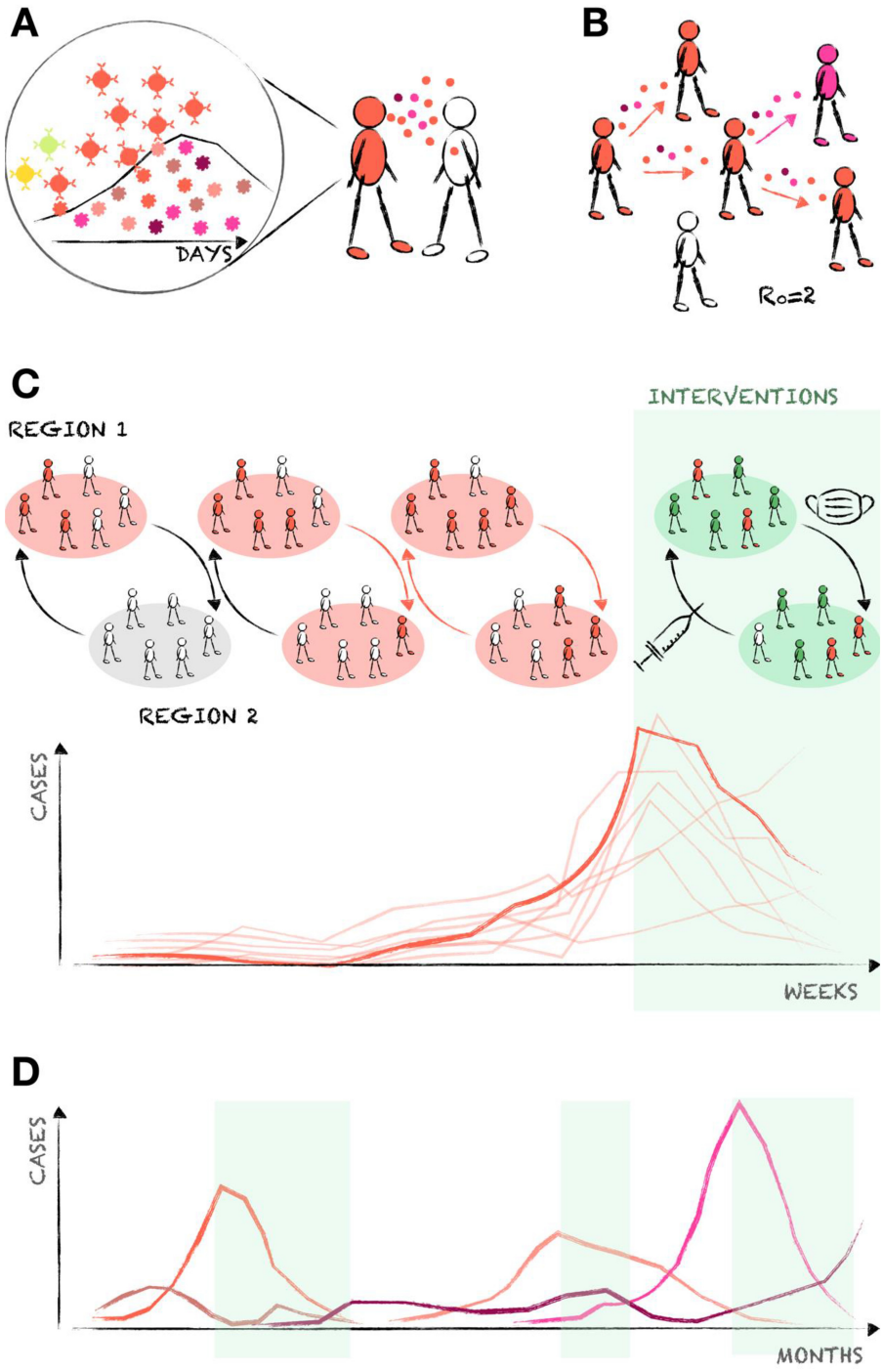


Figure 5.21. The many scales of host–pathogen interactions. (A) Within an infected host the cells of the immune system target existing viruses (and pathogens in general) by binding and neutralizing them. Each host has a vast repertoire of immune cells (denoted by different colors) from which those that best target the

infection are chosen (same colour as the viral strains). In some cases, these cells can further somatically evolve to increase their recognition power. As the virus replicates within the host, it can also mutate within the host such that the strain that infects the next susceptible individual can be different than the one that infects the infecting host. The color of each infected host denotes the most common viral strain within the host. (B) Infected hosts infect on average $\mathbf{R_0}$ other individuals. The immune systems of the infected hosts exert selection pressure on the virus to mutate and diversify so that it can continue infecting. After an infection, hosts develop immunity to the infecting strain (and often similar strains). Through mutations, some new mutants can escape host immunity and re-infect previously recovered individuals. (C) The rate at which the disease spreads within the population depends on the network of interactions between individuals and their behaviour. The disease will first spread within a region of close-knitted interactions and slowly spread to other regions via mobility. Interventions such as vaccinations, face masks, or physical isolation decrease the rate of spread. These dynamics result in the increase and slow down of the number of new cases. An epidemic is one stochastic realisation among many possible ones. Factors such as social-contact structure of the population, immunity profile, strength, kind, and adherence to physical distancing measures, among others, drive and (constrain) the space-phase of possible trajectories. (D) On longer timescales, we see repeated patterns of increase and decrease, governed by events on the smaller scales: interventions, population structure, and emergence of new strains.

bacterial membranes, amoeba, and other parasites. The adaptive immune system is more specific, and uses a large ensemble of cells carrying diverse receptors that are specific for pathogenic molecules. This group of B- and T-cells is called a repertoire. Importantly, the adaptive immune repertoire updates its composition based on the pathogens it encounters, meaning it is a constantly evolving and changing system. The innate system informs the adaptive system of newly detected threats, acts as a messenger between its cells, and the two systems work closely to control all infections.

There are about a billion different B-cells and T-cells, each coming in clones of different sizes, making up together about a trillion different cells that patrol our body. This diversity of cells protects us against pathogens, including those that did not exist when we were born, while at the same time being able to discriminate between molecules natural to our body (good) and those that are foreign (bad). This dynamic ensemble of ever-changing cells is self-organised and distributed, meaning there is no leader cell. Cells communicate between each other to control their numbers and proliferate in case of an infection through signaling molecules, but there is no top-down chain of command. For a few decades now, biological physicists have been studying the equations that describe the interactions between these cells and lead to the composition of the repertoire that protects us against pathogenic threats. These equations are both stochastic, meaning that the forces that drive the repertoire are random and can take different values at different times, but also the interactions are nonlinear, meaning that the existing cells influence the frequency of each other in complex ways. Together these elements make predicting the future frequencies of cells in the repertoire difficult. Physicists have a lot of experience in solving this class of equations and finding the parameters they are sensitive to and those that matter less. These skills are important for predicting future states of repertoires and correctly modelling these systems (Perelson and Weisbuch 1997).

Lastly, pathogen-recognizing receptors of the adaptive immune system are not hard coded in our DNA, like the innate system receptors. They are constantly generated by a highly stochastic process, meaning that each one of us has a set of cells that is unique to us. In fact, it is so unique it can be used to identify us. But random does not mean it can be anything. There are rules that this randomness obeys and physicists have managed to learn these stochastic rules thanks to which the functioning immune repertoire is generated. This uniqueness and randomness imply that we can all have different cells protecting us against the same viral challenge, and that they will protect us equally well. On the one hand, this means there are many good solutions to the same problem; on the other hand it means that we cannot just hope to engineer the exact same solution for everyone. The best protection helps our immune systems find its own solution, which is why both vaccines and immunotherapy work. In summary, this large but finite and constantly changing army of cells controls most of the pathogens we encounter, most of the time.

5.5.2.2 The diversity of viruses

Where do these new mutant viral strains come from? From the point of view of the human population, they either are mutations from existing strains of human viruses, as in the case of seasonal influenza, or come from spillover of a virus from animals (called *zoonosis*) through adaptive mutations, as in the recent H1N1 influenza pandemic or SARS and MERS betacoronavirus outbreaks. Yet, even the strains that come from animal reservoirs come from mutations: before it mutates, the strain that infects animals is not well adapted to humans, and specific mutations allow it to infect and spread through the human population.

The rules underlying evolution are simple. Mutations (small local changes to the organism's DNA) and recombination (rearrangements of different fragments of DNA) generate diversity. For many microbes such as viruses, mutations play a dominant role. These changes are selected according to the survival and reproductive success of organisms that carry them. Mutation first appears in one organism in the population and then spreads with the rate this individual reproduces and has offspring. The first couple of reproductive steps are very perilous since even a very beneficial mutant may not manage to reproduce, simply due to bad luck. Once the individual manages to produce enough of its offspring—and we know how to calculate the chances of that—selection takes over and its final frequency in the population depends on whether it's a beneficial, deleterious, or neutral mutant. The life of an evolving mutant virus, as of any organism, is driven by these initial random events. Due to this initial randomness, even a mutation that is quite deleterious for the organism can rise to prominence and finally be present in every individual in the population.

Mutations and recombination events are very rare. Most of the time DNA gets replicated faithfully, with only occasional errors. For example, the mutation rate when making copies of viral DNA is one every 100 000 base pairs. That number seems small, but if there are 10 000 potentially mutating positions in each virus, and if there are 25 million copies of the virus (a fair estimate for the flu), millions of sites on the viral genome will mutate in each viral replication in each infected person.

Importantly, the number of currently infected hosts increases the probability of a mutation happening. The more mutations that happen, the larger the probability that one of these mutations is beneficial and helps the virus infect new hosts and spread. As a result, although viral evolution is driven by rare and random events, we can see and feel its outcome on large scales.

Ultimately, the fate of every single mutation is to either take over the whole population, or disappear. However, before that happens the organism carrying this mutation can acquire a new mutation but in a different place in its genome. Then the fate of our initial mutation depends on all the other mutations in the organism. For this reason, a viral population carrying many strains with different sets of mutations may co-evolve with the hosts for a very long time, during which the fate of each mutation individually is settled.

One outcome of the evolutionary process is that a single individual and its descendants take over the population—every virus in the population is the offspring of the same ancestor in which the mutation appeared. While this seems very counterintuitive it is an example of the rich-get-richer law and a consequence that individuals with a selective advantage will have more offspring than other individuals. In many practical cases, especially at short times, mutants with similar reproductive success can appear and none of them will completely take over the population, especially if this population is large and not well mixed. This co-existence of several strains often happens in viral populations. While some strains may have a selective advantage, new, fitter mutants are likely to appear on the background of less fit strains, reshuffling the hierarchy of strains and completely changing the race. Of course, this does not mean the new mutant will take over, since it can also share the same fate. Physicists in the last two decades have been instrumental in highlighting and understanding this regime, which is formally analogous to thermodynamic systems driven out of equilibrium (Tsimring *et al* 1996, Desai and Fisher 2007). This work has shifted the focus of evolution from a winner-takes-all to a more complex picture of viral evolution.

5.5.2.3 Viral-immune interactions

Unlike other organisms, viruses need host organisms to reproduce: they do not have their own machinery to replicate their DNA, they need to 'borrow' it from the host. After they infect a host, its immune system attempts to get rid of them, so they need to move on and infect new hosts to survive. This means that one of the ways in which viral mutations are beneficial is by making it easier to infect new hosts, or even new types of hosts as it happens when a virus jumps from one species to another. Other intrinsic properties of the virus can also be increased, such as structural stability of its proteins and capsid or how fast it reproduces. Another type of beneficial mutations, called antigenic mutations, allow the virus to escape the immune response. Beneficial antigenic mutations will change viral proteins in such a way that immune systems of hosts that were previously immunized against the virus will no longer recognize them, because their T-cells, B-cells, or antibodies can no longer bind to them. This gives the virus time to increase the in-host population size, infect new hosts, and in the long run maybe even find better mutations (figure 5.21(A)).

Ultimately, it guarantees the survival of the viral population. In response, individuals infected with this new escape strain will update their immune system, exerting further pressure on the virus to find new antigenic mutations, fueling a continual arms-race (figure 5.21(B)). Biophysicists in recent years have been working on both identifying regions in viral proteins that are prone to antigenic and intrinsic mutations and using stochastic evolutionary models to understand how mutations in different regions influence viral evolution (Nourmohammad *et al* 2016, Chakraborty and Barton 2017). This problem bridges the molecular scale, where the biophysical properties of the amino acid residues that make up viral proteins and underlie their binding properties matter, and the population scale, where a fine understanding of nonlinear stochastic equations is needed. Of course, not every detail matters, and physicists have been working on figuring out which details are important, both through theoretical analysis and quantitative experiments.

We have been focusing on viruses, but the same evolutionary dynamics and co-evolution is valid for antibiotic resistance. Antibiotics are specialised molecules that target and lead to the death of bacteria that infect their hosts. However, bacteria, like all other organisms evolve and beneficial mutations are such that make the bacteria reproduce even in the presence of antibiotics. To overcome this newly evolved resistance, host organisms are either given new types of antibiotics, or higher doses. Physicists are also actively studying this problem using similar approaches as for viral-host co-evolution and have set up fascinating experiments where they identified the mutations that lead to antibiotic resistance even at high doses (Bush *et al* 2011). This kind of quantification is useful for guiding the design of efficient antibiotic regimes to suppress the evolution of antibiotic resistance. Interestingly, often mutations in different places on the DNA work together to produce stronger beneficial effects. This can help us make predictions, as we will see below.

5.5.2.4 Predicting viral-host co-evolution

Since the main events that drive the evolution of pathogens are random, is there any hope in predicting its outcome? In the last couple of years physicists have taken the idea of predicting influenza strains seriously (Łuksza and Lässig 2014, Neher *et al* 2014). They have shown that by taking a probabilistic approach combined with biophysical knowledge at the molecular scale and stochastic equations of evolution, one can predict this year's dominant flu strains, and one can even predict the frequency of these strains compared to other strains. Although viral evolution is driven by random rules, this does not mean everything is equally likely to happen. There are stochastic rules that make certain outcomes more likely than others. We can explore our knowledge of molecular biophysics and stochastic processes to assign probabilities to the outcomes of viral evolution. Specifically, starting with flu strains that have been measured in February of a given year, they use stochastic evolutionary models to calculate the frequency of each strain in the autumn of that year, when the flu starts spreading in the Northern Hemisphere (and analogously in August for the southern hemisphere). These equations account for the competitive interaction of the different strains but also for the interaction with the immune system. Since these are not exactly known, different scenarios must be considered.

The basic idea is similar to the one used in good portfolio design or weather prediction: we do not put all our eggs in one basket but sum over all possible outcomes, weighted by how likely they are. Taken together, the stochastic rules work well and these methods are now used by the WHO to decide which flu strains should be targeted in the flu vaccine for a given year.

Another example of quantitative random rules learned by physicists (Murugan *et al* 2012) describe the stochastic process that drives the huge diversity of B- and T-cell receptors and antibodies that the body can produce. These receptor proteins are encoded in the DNA. However, if we wanted to encode a billion different receptors in the DNA, the DNA would physically not fit into the cell. Instead, the DNA is edited in each B- and T-cell in a way that pre-existing gene building blocks are assembled like legos to produce hundreds or thousands of different receptor combinations. Additional random insertions and deletions of base pairs further increase diversity, resulting in the billion different receptors in a given person. Again, although the rules are random, not every outcome is equally likely. Physicists have quantified these random rules and can calculate the probability of any of us producing a concrete B- and T-cell receptor. Since the process is random, different people have different sets of billions of receptors. But knowing these probability rules allows us to predict that some receptors are more likely to be shared between different people, simply because they are more likely to be produced. These widely shared receptors and antibodies are called 'public'. In general, the machinery can produce much greater diversity than is realised in one person. Even the population of the whole world is not enough to exhaust all the possible receptors that can be generated. However, this is yet another example that random does not mean unpredictable. We can predict exactly how many receptors will be shared between different sets of people, and even which ones (Elhanati *et al* 2018).

In predicting future flu strains we solve one part of the prediction problem: knowing existing viral strains, can we predict their future frequencies. However, can we also predict new mutations that will appear and estimate how likely they are to survive the competition with other mutations and the challenges of the immune system? This is an important question for predicting escape mutations: mutations that are not recognised by the immune system. One approach that has been pioneered by physicists is making quantitative measurements of binding between viruses and B-cell receptors simultaneously in large libraries of immune cell mutants (Adams *et al* 2016). These methods have been extended to measure the binding properties of all possible single mutants in a key SARS-CoV-2 protein (the receptor binding domain of the spike protein) (Starr *et al* 2020). These measurements allow us to predict the ability of the virus to enter human cells, which is linked to disease severity and transmissibility, including for mutant strains that haven't emerged yet. These types of quantitative measurements allow us to explore the evolutionary trajectories of viral evolution and show which mutations close off paths for future mutants. Such knowledge allows us to propose vaccines that can guide the immune system. Physicists have shown that having such knowledge is important to design vaccine schedules that elicit protective immune cells (broadly neutralizing antibodies) that protect us against many viral mutants (Wang *et al* 2015).

5.5.2.5 Quantitative measurements and prediction scales

More generally, an important challenge is to measure or predict the binding affinity of immune receptors and antibodies to their target on pathogens, which in turn determines the efficacy of these receptors to fight the disease. Since both the number of receptors and antigens are extremely large, it is impractical to measure all pairs one by one. Massively parallel methods need to be designed, and biophysics-inspired computational methods need to be developed to fill in the 'gaps' of receptor-antigen pairs that we will not be able to measure. Promising methods that combine droplet microfluidics (Gérard *et al* 2020) and binding assays (Zhang *et al* 2018) are paving the way to that goal. The idea is to encapsulate single cells in small droplets of water suspended in oil, and then manipulate these droplets in a microfluidic chip to perform biochemical reactions in each cell in a compartmentalized way. By combining phenotyping and binding assays in each droplet in this way, and by introducing unique molecular barcodes prior to sequencing, one can associate the full sequence of the receptors to its affinity for a set of antigenic targets, which can themselves be sequenced. To complement these experiments and extrapolate their results for sequences that were not directly tested, we expect machine learning methods to be useful. As they are improved and scaled up in the future, these techniques will help to build a complete specificity map between receptors and pathogens, which will serve as a basis for rational drug and vaccine design.

As we see, the immune system and viruses are constantly co-evolving: the immune systems of the world's hosts exert pressure on viruses and viruses force our immune systems to constantly update itself. Vaccines prepare our immune systems for challenges that we have not yet seen, such that if we encounter the virus we have pre-existing protection. Additionally, the immune system is flexible and protection against one virus can also protect us against similar viruses, as was originally shown by Edward Jerne using the cowpox virus to immunize people against smallpox in the 18th century. Similarly, using sequencing measurements and analysis, physicists have shown that T-cells reactive to the common cold, which is also a virus from the coronavirus family, can recognize SARS-CoV2 proteins (Minervina *et al* 2021). Using quantitative experiments to understand the range of this cross-reactivity is an important challenge that can help us predict and trigger immune responses. This may also allow us to extend the prediction of viral strains to longer timescales.

There are many different molecular solutions to the same challenge, yet we show that we can statistically predict the immune pressures exerted on evolving viruses. How can we reconcile this large molecular diversity with predictability? One answer is that the phenotype is more constrained and reproducible between experiments than the molecular implementation in terms of actual mutations in the DNA (Lässig *et al* 2017). In short, since the selection pressure acts on the phenotype, it is constrained, and thanks to that, predictable. As we showed in the example of the immune system, there are many ways to obtain the same results in terms of mutations. The same results have been observed for antibiotic resistance and *E. coli* evolution under the pressure of the innate immune system in mouse guts.

5.5.2.6 Sub-summary

In summary, host-pathogen co-evolution occurs on many different scales: from the molecular, cellular, organismal to the worldwide population scale. On most of these scales we are dealing with random interactions between many elements. However, thanks to experimental and theoretical work we now have a command of these stochastic rules which allows us to make statistical predictions. Since this co-evolutionary race is constantly happening, why do we not see pandemic outbreaks all the time? We explore this question in the next section.

5.5.3 The physics of epidemics

With COVID-19 the whole world has experienced first hand the exponential growth of a serious infectious disease. This marked feature of many infectious disease epidemics challenges our intuition and leaves us unprepared.

The exponential growth in the number of infected individuals can be explained by simple contagion rules as shown in the influential work of Kermack and McKendrick in 1927 (Kermack *et al* 1927). This work laid down the equations of the SIR (Susceptible-Infected-Recovered) model which has become the core of modern epidemiology because it is able to reproduce fundamental characteristics of epidemics. In particular, the model allows for the clear framing of a key concept in epidemiology, that of the reproductive number $\mathbf{R_0}$. This is the average number of individuals an infected individual infects before recovering when the population is fully susceptible (figure 5.21(B)), and quantifies the transmission potential of an epidemic (see box 5.1).

Box 5.1

Imagine that a few individuals bring a new pathogen into a susceptible population (S). If we assume two people interact at random, the probability for each susceptible individual to become infected at a given time is proportional to the number of infectious individuals in the population (I). The proportionality coefficient depends on two factors: the average number of contacts a person has at each time, $\boldsymbol{k}$, and the rate of transmission per contact, $\boldsymbol{\beta}$. Infected individuals can also recover with rate μ, thus ceasing to be a source of infection (R). It turns out that a simple model that allows transitions between the different states (S, I, and R) results in an exponential growth of the number of infected individuals in the early stages of an epidemic outbreak. The relationship between the generation of new infectious contacts $\boldsymbol{k}\ \boldsymbol{\beta}$ and the recovery rate μ determines the growth rate and the size of the epidemic. Intuitively, the more susceptible individuals an individual can infect before recovering, the larger the epidemic wave will be. The ratio $\boldsymbol{k}\ \boldsymbol{\beta}/\mu$, which can be interpreted as the average number of individuals an infected individual infects before recovering when the population is fully susceptible, is what is called the basic reproductive number $\mathbf{R_0}$. It controls the final size of the outbreak and, together with the rate of recovery μ, the growth rate of the epidemic, $\mu(\mathbf{R_0 - 1}) = \boldsymbol{k}\,\beta - \mu$. If $\mathbf{R_0}$ is greater than one, the epidemic grows exponentially; if it is instead smaller than one, it goes extinct because each infected person on average will infect less than one person. Importantly, a change in number of contacts (e.g., due to implementation or relaxation of social distancing

measures), in transmission per contact (e.g., increased hygiene or face masks), or a reduction in infectious period (e.g., identification and isolation of a case) can change $\mathbf{R_0}$ and as a result have an exponential impact on the unfolding of the epidemic. This is the reason $\mathbf{R_0}$ has important public health implications. It tells us how far we are from suppressing an outbreak, what will be the outbreak impact if left uncontrolled, and how strong is the effort needed to curb it.

The Kermack and McKendrick model relies on the homogeneous mixing assumption (i.e., individuals enter in contact at random). This is very convenient and widely adopted. However, we know that assuming a homogeneous number of contacts per individual is far from reality as the number of connections an individual has is rather heterogeneous. The question is then: How does this heterogeneity affect the spread of an epidemic? It turns out that the precise topology of the network of human interactions can have a dramatic effect on the way an epidemic spreads through the population (Pastor-Satorras *et al* 2015, Kiss *et al* 2017). For instance, in populations in which most individuals have a few connections but there are few with an (outrageously) large number of contacts, the epidemic will almost surely be able to spread (if no measures are put into place) no matter the value of epidemiological parameters.

However, we know that networks of interactions between individuals have more complex structures. Networks of relevant interactions for disease contagion can be quite different: a network of sexually transmitted diseases is not the same as a network of contagion of airborne transmitted diseases (such as influenza, MERS, or SARS viruses). For the former case, only a fraction of the population is at risk which makes it possible to design containment measures that target specific parts of the population. For the latter case in which the majority of the population is at risk of becoming infected, we know that the network of contacts between individuals has a large-scale structure (often called communities) that shapes the spatiotemporal pattern of spread (Guimerà *et al* 2005). For instance, we can visualize this structure as city or country boundaries that are overcome through mobility. Nonetheless, even within the city population we can define some coarse-grained structure of an otherwise very complex network of interactions.

The structure of the network Is extremely infIrmative when it comes to understanding contagion and designing containment and vaccination strategies (Danon *et al* 2011). Containment strategies (especially in the case of airborne transmitted diseases such as SARS or pandemic influenza) aim precisely at preventing contact between different regions (communities/components) of the network and therefore effectively isolating different foci of the epidemic (figure 5.21(C)). An obvious first choice for containment is thus to reduce mobility, including inter-country and inter-city mobility, since these are *de facto* transmission highways within the population. However, quantifying the impact of travel restrictions is not as simple as it may seem, as it requires a deep understanding of the mechanisms driving the epidemic invasion. A convenient framework to tackle this issue is provided by multiscale

models rooted on the metapopulation approach that are able to integrate information on the human mobility network, such as the flight network (Colizza *et al* 2006).

However, there are other aspects that we have to take into account such as the fact that interactions between individuals are neither static (Masuda and Holme 2013) nor happen in the same environment. For instance, the proximity interaction among people who live under the same roof is not the same as that among people who work desk-to-desk in the same office every day, share a bus ride during 15 min, or go for a walk outdoors once a week. In reality, interactions between individuals take place on multiple layers which have different levels of infectiousness associated (or equivalently the networks in different layers have different weights) and are not active at the same time (Kivelä *et al* 2014). Spread is typically driven by the layers with a larger density of connections with high strength. The epidemic then cascades onto other layers, effectively resulting in an increase in the number of infected individuals.

Precise and real-time information on real network structures is impossible to obtain. Still, thanks to extensive data becoming available at all temporal and spatial scales, structural and dynamical properties of these networks are described with increasing levels of detail (Barrat *et al* 2013, Barbosa *et al* 2018, Eames *et al* 2015, Masuda and Holme 2013). This information enables the design of idealised networks of contacts that reproduce these properties for numerical and theoretical exploration. Most of these findings are valuable to understand the effect of social distancing measures that reduce interaction strength. They also highlight how network properties, and their change in adaptation to the epidemic, can qualitatively alter the epidemic progression, by both speeding it up or slowing it down to a polynomial, rather than exponential, growth in the number of cases. Theoretical results are also informative about how to design effective vaccination strategies that reduce epidemiological parameters and can aid in estimating the percentage of the population that needs to be vaccinated to achieve herd immunity (i.e., the number of immune people necessary to prevent the epidemic from spreading throughout the population) as well as which part of the population should be targeted first in vaccination efforts.

5.5.4 Facing outbreaks

Models are able to encode the driving forces of epidemics and as a result provide a coherent framework to make sense of epidemiological data. As such, they are critical tools to face an outbreak. When a new pathogen emerges in the human population, the lack of knowledge of its epidemiological characteristics makes identifying the most effective interventions very hard. In physics terms, the propagation of an infection is all about the interplay between dynamical processes unfolding at different timescales—dynamics of symptoms, infectiousness, and human interactions. By comparing models to data, it is possible to estimate key parameters ruling these processes. The COVID-19 emergency has showcased the crucial role models play in providing epidemiological understanding. Models have allowed for answering critical questions (Ferretti *et al* 2020, Pullano *et al* 2020, Wu *et al* 2020): How

fast does the epidemic spread? To what extent can cases be detected? Can asymptomatic cases transmit the infection? Critical unknowns also stem from the heterogeneous properties of individuals and populations, and their impact on the propagation of the infection (Althouse *et al* 2020, Arenas *et al* 2020, Chinazzi *et al* 2020, Davies *et al* 2020, Gatto *et al* 2020, Schlosser *et al* 2020): What is the role of children in the epidemic? What are the main pathways of epidemic spatial propagation? What is the contribution of superspreaders and superspreading events in the propagation of the outbreak? Answering these questions deepens our fundamental understanding of the epidemic dynamics. More importantly, it has immediate practical consequences. For instance, knowing how fast the epidemic spreads tells us how fast we need to react to avoid peaks in hospitalisations, or excess deaths (Wu *et al* 2020). Estimating the number of unseen cases is essential to identifying the critical gaps in the epidemic surveillance and the best strategies to improve it (Pullano *et al* 2020). Knowing the role of children versus adults in disease transmission tells us up to what extent school closure is effective in preventing transmission (Davies *et al* 2020). More in general the quantification of age-variation in the infection and severe form of the disease inform the design of prioritised vaccination strategies (Bubar *et al* 2021). Understanding the role of asymptomatic versus symptomatic transmission and the heterogeneities in the transmission risk across individuals and settings is essential for planning contact tracing strategies (Aleta *et al* 2020, Ferretti *et al* 2020, López *et al* 2021).

Fitting models to data is a critical step that can become quite daunting. As an example, consider the measurement of the basic reproductive number ($\mathbf{R_0}$) which is often (albeit not exclusively) obtained by fitting an exponential growth model to the initial stages of the spread of an epidemic. However, data from the early stages of spreading are noisy. Moreover, we must account for the fact that data of the same disease in different places can be affected by a number of socio-demographic factors which can result in the variability of the estimates of some epidemiological parameters, including $\mathbf{R_0}$. Thus, relying on a single estimate is problematic. Handling the uncertainty of the estimate is a question which has been tackled reasonably well. However, understanding the context-to-context variability and how this uncertainty propagates within the model is another source of error that can seriously bias *in silico* predictions. Most importantly, this uncertainty should be taken into account when calibrating models to make forecasts (Chowell 2017).

Models can also help provide near real-time information of the current epidemic situation. Models have facilitated the creation of tools for anticipating the epidemic trajectory for both short-term epidemic forecasts (figure 5.21(C)) and long-term scenario analyses (figure 5.21(D)). However, this does not come without challenges. Medical and surveillance data are not collected by a controlled experiment. These data are limited and biased, an issue which is exacerbated during a public health emergency, where testing availability may be limited and surveillance protocols are continuously changing. These issues can seriously hinder our ability to even diagnose the current state of a pandemic. A central quantity to diagnose the real-time status of an epidemic outbreak is the effective reproductive number $\mathbf{R_t}$, which tries to capture which is the number of infected people per infected person in a stage

in which the disease has already spread community-wide. In the literature, different ways for measuring this number have been proposed (Gostic *et al* 2020), which is a *de facto* indicator for authorities as to whether certain epidemic containment measures are being effective, and whether existing measures can be relaxed. Methods to estimate $\mathbf{R_t}$ rely on a very precise estimation of epidemiological factors and on the availability of high-quality, instantaneous data on the number of infected people. For diseases like COVID-19, causing a wide spectrum of symptoms, including asymptomatic and mild symptoms, we know that known cases are an underestimation of real cases. The lack of details on how data have been collected and its possible biases can severely undermine the ability of current approaches to accurately compute $\mathbf{R_t}$, and makes it hard in general to calibrate models with which to build reliable scenarios. Therefore, for models to be of use, it is crucial to know the details about how these data are collected, to understand their biases and limitations. To be helpful, physics approaches need not only to embrace data in all its complexity but also to establish a strong interdisciplinary dialogue with medical, surveillance, and case management experts.

Despite the issues with data, scientists have been able to circumvent some data-related issues by using Bayesian approaches that exploit inherent uncertainty in the parameters to calibrate models using past data on reported number of infected individuals and subsequent deaths (with an appropriate time lag). Indeed, there are many models that can fit past data, but that does not necessarily mean that all such models are able to make reliable future predictions in an appropriate time horizon (usually two weeks in an epidemiological context). This situation has put the stress on the need to develop tools for evaluating forecasts; even for the early epidemic stages, forecast evaluation can help uncover reliable models for a given outbreak and help build meaningful scenarios (Chowell 2018).

One of the cruxes in epidemiological research is precisely forecasting. The inherent uncertainty in parameter estimation makes it impossible to make reliable forecasts of the future evolution of the pandemic in a long time horizon (Castro *et al* 2020). These predictions are uncertain in the same way weather forecasts are, and therefore we need to be able to assign uncertainty to any forecast we make. The recent pandemic has brought together a large community of scientists in a joint effort to use available knowledge to make reliable predictions. Many years of research have produced a wealth of models that can be used to make predictions of the evolution of the pandemic, many of which are quite different in their inner details but whose predictions can be pooled together to make better forecasts (see, e.g., https://covid19forecasthub.org/ and https://covid19forecasthub.eu/). These efforts have shown the power of making ensemble forecasts, but at the same time have put in the spotlight how hard it is to make and interpret forecasts in a constantly changing situation. These difficulties go beyond model uncertainty because they arise from the need to incorporate sudden and profound changes in the epidemic conditions, as the ones caused by the emergence of variants and human response to the epidemic.

One year after COVID-19 emerged, the widespread dissemination of the mutant virus Alpha was like a new pandemic sweeping the world. The variant differed from

the wildtype in several aspects, including transmissibility and severity, generating a rapid change in the trajectory of cases, hospitalizations, and deaths (Volz *et al* 2021). This posed a great obstacle to the epidemic forecast and the planning of mitigation measures and required a prompt effort to re-assess the epidemic situation and re-evaluate key epidemiological parameters. Following Alpha other viral variants of concern were detected (Beta, Gamma, Delta). Several variants were simultaneously circulating, showing complex dynamical patterns often difficult to interpret. The advantage conferred by a mutation critically depends not only on its molecular properties, such as its enhanced ability to infect cells, or to escape existing immunity (resulting from previous infections or vaccination), but also on the characteristics of the environment and the host population. For instance, theoretical studies on multi-strain spread on mobility networks show that the phase space of strain (co-) dominance varies according to the mobility level (Poletto *et al* 2013). Still, theoretical research on the complex interplay between environment, human social behavior, and multi-strain/multi-pathogen interaction is still behind work on single-pathogen/single strain epidemics. The paucity of resolved data to monitor the spread of distinct variants at the population scale was initially a key limitation in studying multi-strain dynamics. This is rapidly changing, thanks to the increasing availability and rapid share of genetic data that is enabling the reconstruction and visualization of the spatiotemporal pattern of strains' co-circulation (e.g., https://nextstrain.org/sars-cov-2, https://covariants.org/, https://cov-lineages.org/, https://www.cogconsortium.uk/).

A second challenge in forecasting eIidemic outbreaks is represenIed by the need to incorporate not only changes in containment measures but also the adoption of those measures by individuals. Public health emergencies trigger a change in human behavior and create a feedback loop between infection dynamics and behavior. Governments may take measures to restrict human interactions and reduce the chance of contagion. For instance, cancellation of events and mass gatherings, travel bans, school closures, and curfews were adopted to curb recent epidemics of Ebola, H1N1, or SARS. To curb the COVID-19 pandemic, these kinds of interventions have been of unprecedented strength and extent, including lockdowns and complete travel bans.

The effect these measures have on human behavior is extremely complex. Adherence to measures varies greatly both geographically and in time and depends on socio-demographic factors. Upon the announcement of a measure people may decide to adopt protective measures beyond what is required by the imposed regulations. In fact, even in the absence of interventions, people may change their behavior to reduce as much as possible the risk of contagion. However, if restrictions last too long, people may find it increasingly difficult to respect both self-imposed and external measures. Importantly, all of these behaviors depend not only on the epidemic situation but also on mass media and social network communication around the epidemic.

The feedback between behIvior and contagion is one of the central problems of the physics of epidemics. A large body of work shows that behavioral change may alter the epidemic spreading profoundly (Funk *et al* 2010). Awareness may suppress

spreading. However, changes in behavior (e.g., by relaxing social distancing) may also generate interesting dynamical features, such as multiple peaks or oscillations. Nonetheless, the great majority of studies from before the COVID-19 epidemic were purely theoretical. This was because very limited data were available on the behavior change of a population in response to an outbreak, thus hindering the validation of such theories with data.

The COVID-19 pandemic has completely changed our outlook on epidemics (Perra 2021). The pandemic has brought about abrupt societal changes which have generated a desperate need to monitor human behavior. Massive efforts have been dedicated to track individual mobility, physical interactions, and attitudes. Large-scale surveys have quantified compliance to recommendations (e.g., face masks and hygiene measures), and changes in social encounters. Through data-for-good programs, tech and communication giants such as Google, Apple, Facebook, Telefonica, Vodafone, and Orange, among others, have put massive amounts of data at the service of the scientific community. These data are extremely valuable because they provide proxy information on human-to-human interactions. As such, they facilitate the modelling of the underlying network of interactions on which a disease propagates and assess adhesion to measures (Perra 2021). Still, while these data are useful, they can only partially address such a multifaceted problem. As always, a deep comprehension of the phenomenon must be based on the combination of diverse and complementary information sources (e.g., digital proxies, surveys). This may require the development of new approaches. The amount of data generated in one year of an epidemic will require a long time to be digested.

5.5.5 Conclusions

Adapting to questions of current interest with flexibility and openness is one of the cornerstones of the scientific endeavour. Through interactions with scientists in other fields, physicists studying random processes and network science are in the position to contribute to important societal questions in the face of a pandemic. Statistical physics offers tools to describe rare random processes and complex nonlinear patterns of interaction. Importantly, in the same way that machine learning tools do, physics tools can learn from data with the added bonus of providing interpretable, testable generative models that enable us to build theories that bridge scales and incorporate knowledge from other disciplines. We have argued that effects on the microscopic scale, such as single molecular mutations, can influence observable outcomes, such as the spread of a deadly epidemic at the global population scale. Clearly, a challenge for everyone from scientists to policymakers is first to prevent pandemics from happening and, if they happen, to contain them. Physicists can and do contribute meaningfully to this cause. One of the most remarkable examples is physicists' contribution to developing the three-drug combination, commonly known as a triple cocktail against HIV (Perelson *et al* 1996), which consists of administering three different drugs at carefully timed intervals. Through understanding the interplay between the timescales of mutation of the rapidly evolving HIV viruses and those of the immune response, they

contributed to quenching the spread of the AIDS epidemic, at least in countries where the triple cocktail was economically available.

Working closely with specialists in other fields, physicists are continuing to offer solutions to different pathogenic threats, with different molecular and mutational properties. The major long-term challenge is being able to predict the emergence of a major epidemic, such as HIV or Sars-CoV-2 even before the virus emerges. This involves knowing the physics of processes occurring on multiple inter-related scales: estimating the probability for a variant to mutate and cross species barriers, the probability that it has high enough transmission rates to start spreading, and what physical contagion mechanisms and social behaviours will help it spread, which in turn depend on the molecular properties of the pathogen. It is extremely important to emphasise that model predictions are statistical and have associated uncertainties. Physics approaches describe viral evolution, immune response, and disease spread by associating probabilities to plausible events in the same way weather forecasts do. It is therefore critical to understand the properties of the probability distributions, and especially to understand rare, disruptive events in the tail of the distributions because these can have a dramatic, systems-wide effect that is not well-represented by the typical or average events.

A prominent example of the importance of the tail of the distribution are superspreading events. Many of them were documented during the COVID-19 pandemic (e.g., the epidemic cluster of ~700 people in the Diamond Princess cruise ship that quickly rose from a Hong Kong resident visiting the ship). In South Korea a single infected person was able to generate a cluster of more than 5000 cases (Althouse *et al* 2020). The central role of events of this kind in the epidemic trajectory has urged a paradigm shift in predictive modelling from the concept of $\mathbf{R_0}$, the average number of secondary infections (i.e., infections generated by an infectious case), to the concept of secondary infection distributions. This calls for new methods to better reconstruct this distribution, identify transmission hotspots, and understand how these can be targeted by interventions to rapidly control the epidemic (Althouse *et al* 2020).

At the molecular scale, rare events in the tail of the distribution also drive dramatic changes in the epidemic. For instance, the emergence of particular variants and substrains require the occurrence of a distinct set of beneficial mutations, each occurring randomly as the result of a replication error of a single nucleotide in the genome of a single virus. This mutation is then amplified to affect the entire population through selection or genetic drift. Which mutations occur and when is largely determined by these rare, partly unpredictable events.

Prediction involves not just predicting the most likely outcome at a given time but understanding limits and best/worst case scenarios. Since we are dealing with stochastic dynamical processes, small differences at one time or one scale can lead to big differences in the global outcome. Current research efforts are being dedicated to improving forecasting by better considering the whole spectrum of possibilities and encompassing a wide range of plausible scenarios. This could be done by incorporating not only bottom-up approaches, but also top-down approaches that look blindly at data

and identify important variables that are predictive. At the same time, improvements in forecast evaluation tools will increase our capacities to compare models and learn which assumptions work and which might fail in our models.

Most importantly, ensemble predictions and scenario analysis may be hard to interpret and integrate in the decision-making process. Policymakers must interpret model results in light of their assumptions and limitations. Long-lasting interdisciplinary collaborations between researchers and policymakers are needed to create a culture of outbreak analysis and predictive modelling. Finally, a critical challenge is represented by the communication with the general public and journalists. This is essential to build trust in science and contrast misinformation (Gallotti *et al* 2020). At the same time, communicating epidemic assessment and predictions as well as the uncertainty around them is difficult. Science does not follow a straight line; early hypotheses may be revised as new knowledge accumulates. This may confuse society and lead them to think that changes in containment measures or vaccination strategies happen on a whim as opposed to being the result of careful processing of new knowledge.

There are many other challenges that affect specific scales. At the level of host–pathogen interactions, a big current challenge is combining theory and quantitative experiments to elucidate the so-called genotype–phenotype map: can we find statistical rules that translate the molecular sequence to relevant phenotypes such as immunogenicity? This is a critical step for drug design and for predicting immune escape in disease spread.

At the level of human populations, the challenge is to fully exploit the potential of epidemiological and behavioral data. Physics models would not be useful without the availability of high-quality and rich data. Data should be shared rapidly and widely and made openly available to researchers, accompanied by standardized protocols about data collection and processing and privacy requirements. During the COVID-19 pandemic a wealth of data was made available by government and public health agencies. In addition, initiatives of data gathering, aggregating, and sharing have made these data ready to use by the scientific community—see, e.g., https://coronavirus.jhu.edu/data and https://global.health/. At the same time data for good initiatives have provided anonymised and aggregated information on human behavior. All of this has enabled critical discoveries. Still a strong coordination effort is needed (Oliver *et al* 2020), to get modellers and data scientists involved in the data collection process—thus ensuring appropriate data are collected—and to ensure that researchers working with data know how the data were obtained and what the data actually represent. With this data in hand we can minimize and correct data biases and make more realistic forecasts and scenarios.

As a final point, a ubiquitous problem that we face at all scales is that of heterogeneity: in immunological response and in human reaction, which include gender, age, social status, and geographical diversity. Heterogeneity is indeed an inherent element of random processes. Quantifying the role of heterogeneity at all scales is an open problem important for treatment, vaccination, and response. Again, physics has the tools for addressing and linking heterogeneity at the

molecular, cellular, and organismal scales to population-level processes. Our challenge is to make it happen.

The utilitarian goal of predicting evolution and disease spread is within our reach. As the positive example of dominant flu strain prediction shows, combining the processes at different scales, theory, and data results in successful predictions that can be used by policymakers such as the WHO.

5.6 Further diagnostics and therapies

Marco Durante[1]
[1]GSI Helmholtzentrum für Schwerionenforschung and Technische Universität, Darmstadt, Germany

5.6.1 High-resolution imaging

Marco Durante[1]
[1]GSI Helmholtzentrum für Schwerionenforschung and Technische Universität, Darmstadt, Germany

5.6.1.1 General overview

Imaging was the first application of x-rays, with the famous pictures of Mrs Röntgen's hand published on *The New York Times* on January 16, 1896. In over a century of progress, figure 5.22 shows the progress in imaging from the planar x-ray image produced with the orthovoltage tube of Röntgen to the modern CT angiography.

Imaging has therefore changed medicine maybe more than any other technical advances. Therapies like surgery, radiotherapy, and any other medical and pharmacological intervention strongly depend on the imaging obtained with physics techniques. Imaging today follows the pathway described in figure 5.23 and uses many different techniques, as described in table 5.1.

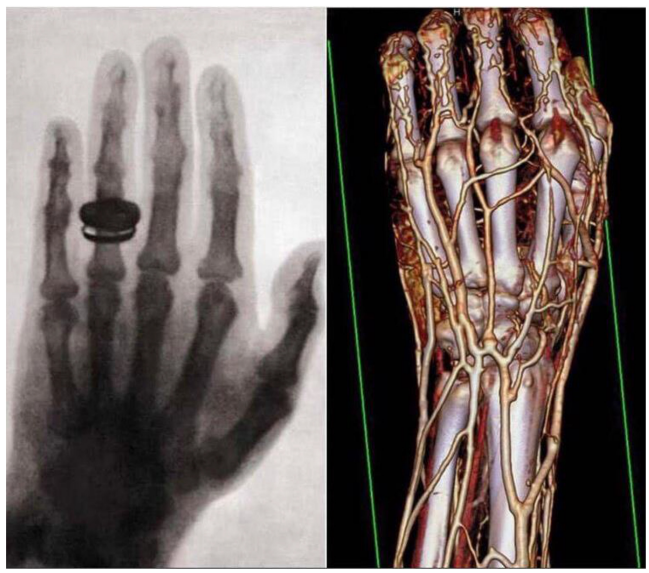


Figure 5.22. X-ray image of a hand obtained by Wilhelm Conrad Röntgen in December 1895 (left), showing the bones and the wedding ring of the wife. A modern CT angiography of a hand (right), showing high-resolution details of the joints and vessels.

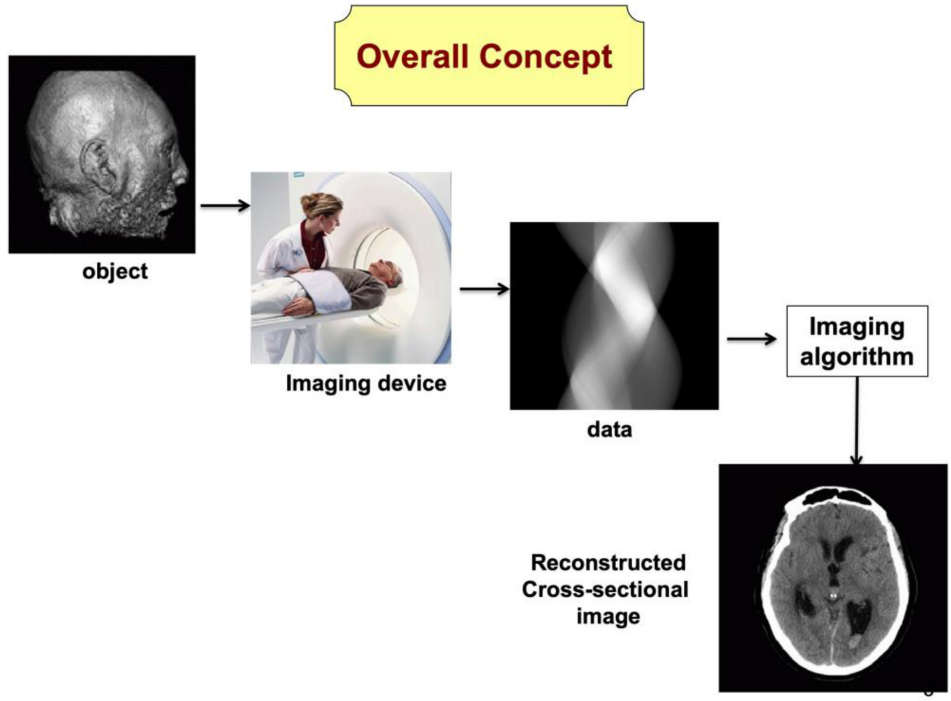


Figure 5.23. Imaging workflow in medicine.

Table 5.1. Main medical imaging techniques in use today.

	Radiation	Techniques
Radiology with external sources	X-rays	Radiography
		Computer tomography (CT)
		Mammography
		Angiography
		Fluoroscopy
	Magnetic field and radiofrequency	Magnetic resonance imaging (MRI)
	Ultrasounds	Echography
Endoradiology	Gamma-rays	Scintigraphy
		Single-photon emission computed tomography (SPECT)
	Positron annihilation	Positron emission tomography (PET)

Different methods are used in different branches of medicine. Mammography is, for instance, specialized for the detection of breast cancer, and angiography for cardiovascular diseases analyses. Other methods are used for many different

methods. Echography is for instance popular during pregnancy, but is also widely used to detect abdominal diseases. SPECT and PET are largely used both in oncology (often associated to CT) and cardiology in myocardial perfusion tests. CT and MRI are largely used in oncology, with complementary targets: while CT has a superior resolution of bones, MRI shows excellent images of the soft tissues (figure 5.24).

When using endoradiology, image fusion is generally necessary to combine the functional and anatomical images. Figure 5.25 shows a fusion of CT and PET imaging in oncology. A whole-body PET-CT scanner has been recently built by the EXPLORER consortium, led by the University of California at Davis (figure 5.26).

5.6.1.2 Challenges and opportunities

The new frontiers of imaging is the improvement of both *hardware* and *imaging* parts. The main hardware challenge is the increase of the image resolution. The new ultra-high resolution CT scanners reached resolution of 150 μm, starting to touch the area of microscopy (figure 5.27).

Similarly, for MRI there is a rush to increase the intensity of the magnetic field, which immediately reflects into increased resolution (figure 5.28).

The cost of MR scanners largely increasing with the field intensity, ranging from less than 500 k€ for a conventional 1.5 T scanner to over 5000 k€ for a 7 T scanner. But for preclinical studies at CEA in Saclay researchers have already installed a 11.7 T MRI magnet. Whether these high magnetic fields are really needed in clinics is a big question for the coming years, where it will be defined what the 'optimal' scanner will be.

The perspectives are even more exciting for software, where we are witnessing the invasion of AI in medicine and in diagnostics. AI systems have demonstrated an

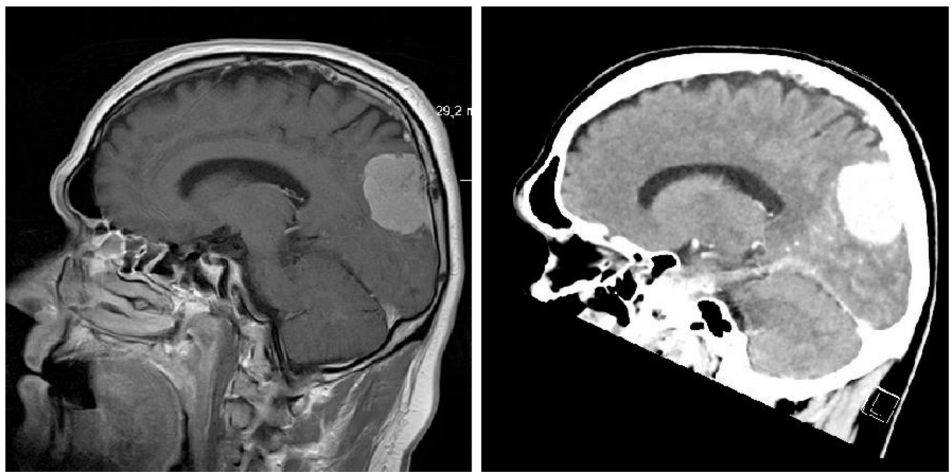


Figure 5.24. MRI (left) and CT (right) images of the same brain tumour.

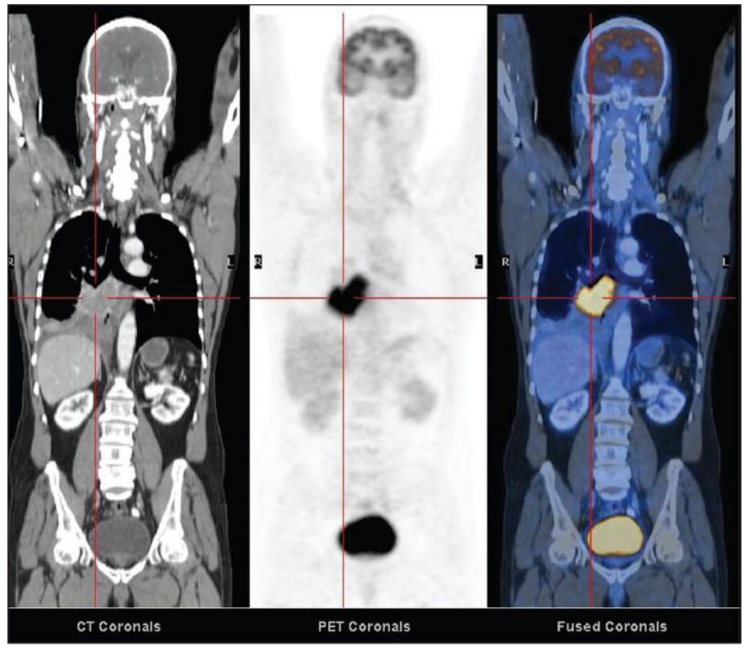


Figure 5.25. Fusion image of a CT and a 18FDG PET clearly identifying a tumor with high aerobic glycolysis.

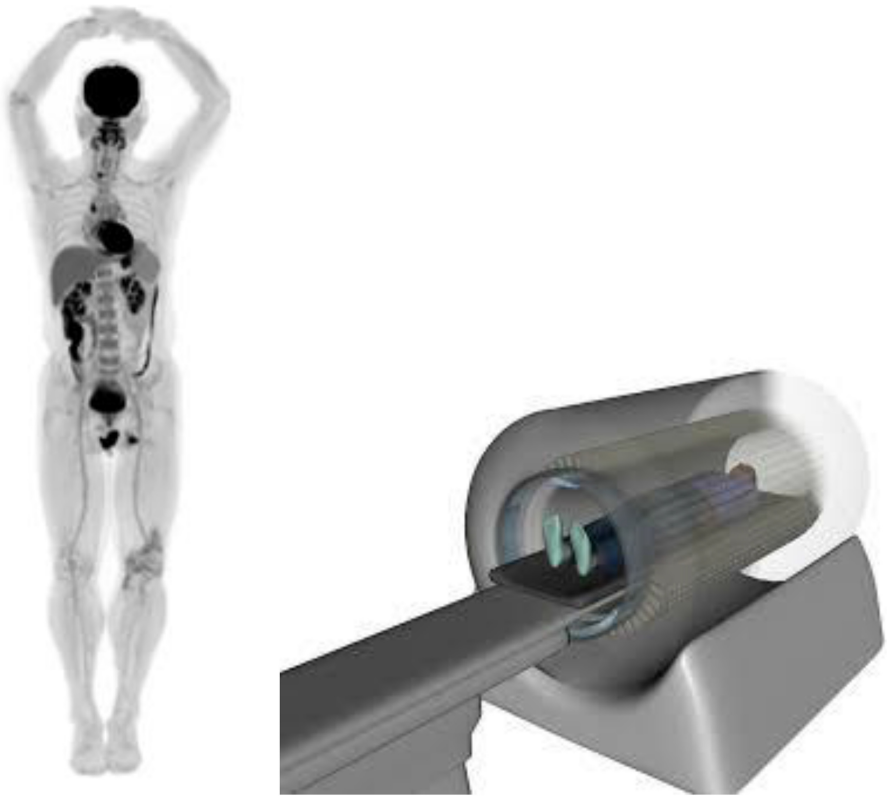


Figure 5.26. The EXPLORERE whole-body PET-CT and a whole-body image produced by the new scanner.

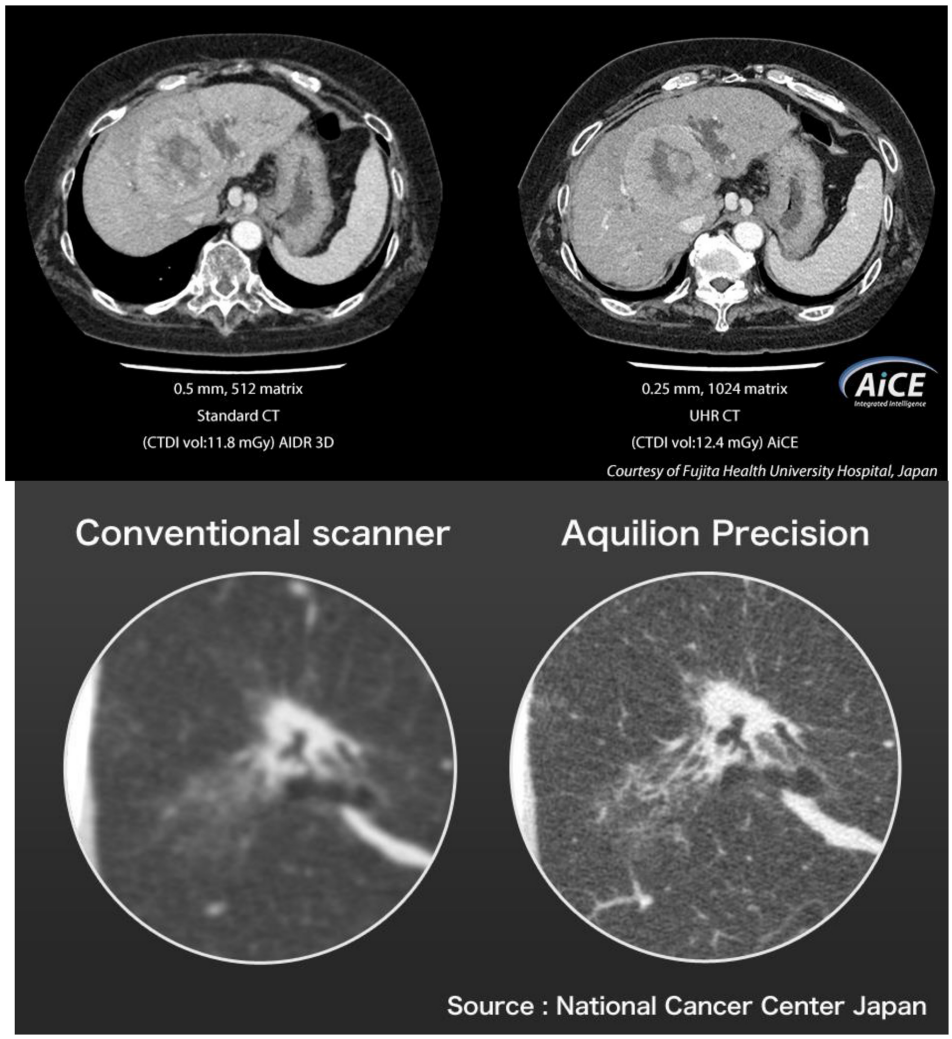


Figure 5.27. Differences between a conventional scanner and ultra-high resolution CT.

extraordinary ability to capture diseases in images notoriously affected by many artifacts such as mammography. Nevertheless, the initial concerns that AI will replace radiologists have gradually vanished, considering that the answer to a single question is not sufficient to draw a solid diagnosis. However, it is clear that future radiologists will all use AI as a sharp-eyed partner to assist physicians in diagnostic procedures.

Another large impact of AI in diagnostics is in the field of radiomics (i.e., the method to extract molecular and histological features from radiographic medical images using data-characterisation algorithms) (figure 5.29). The procedure is very intriguing, because in principle it can overrule the invasive and slow biopsies replacing them with deep learning from conventional radiology. Shape, intensity,

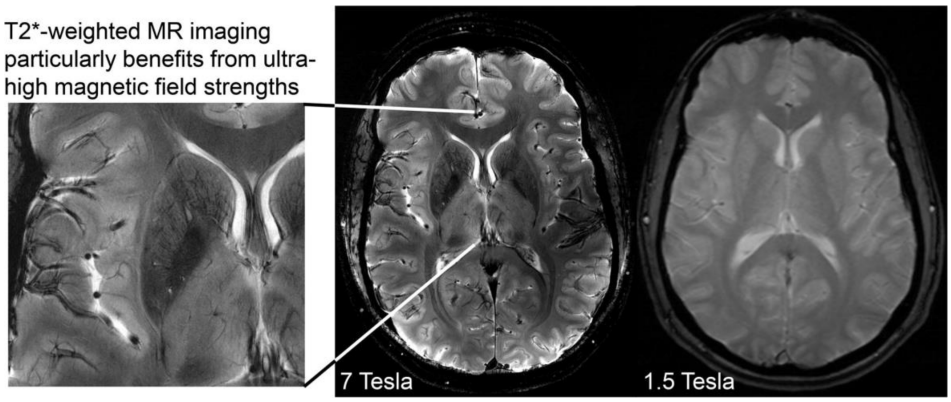


Figure 5.28. Comparison of brain images (T2-weighted) using a conventional 1.5 T and a new 7 T scanner.

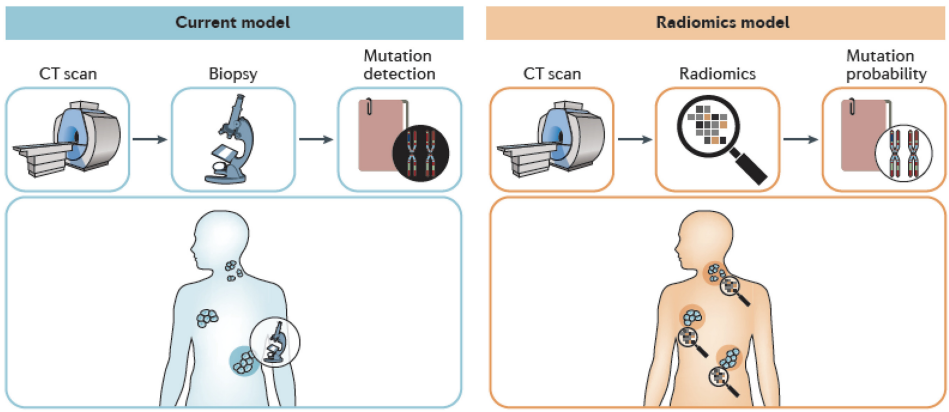


Figure 5.29. An image of how radiomics can change the diagnostic procedure, currently needing biopsies and molecular tests to reach diagnosis.

and texture are typical image features that can be analysed to extract biomarkers using radiomics.

The field of functional imaging is also the one with the highest potential. Functional imaging in oncology generally uses the cancer metabolisms, mostly using 18FDG. However, recent studies are targeting cancer-associated fibroblasts, using fibroblast activation protein inhibitor (FAPI) labelled with ^{68}Ga. The shift from tumour metabolism to the tumour microenvironment may provide new, important information for appropriate treatment. Other compounds (e.g., 18F-FMISO or Cu-ATMS) target hypoxic regions in tumors that are treatment-resistant.

The other popular functional imaging method is fMRI that measures changes in blood oxygenation levels, which are correlated with the activity of a specific part of the brain. Apart from initial studies on brain anatomy, fMRI is now becoming a powerful tool in diagnostics and psychiatry. SPECT can also be used for cerebral

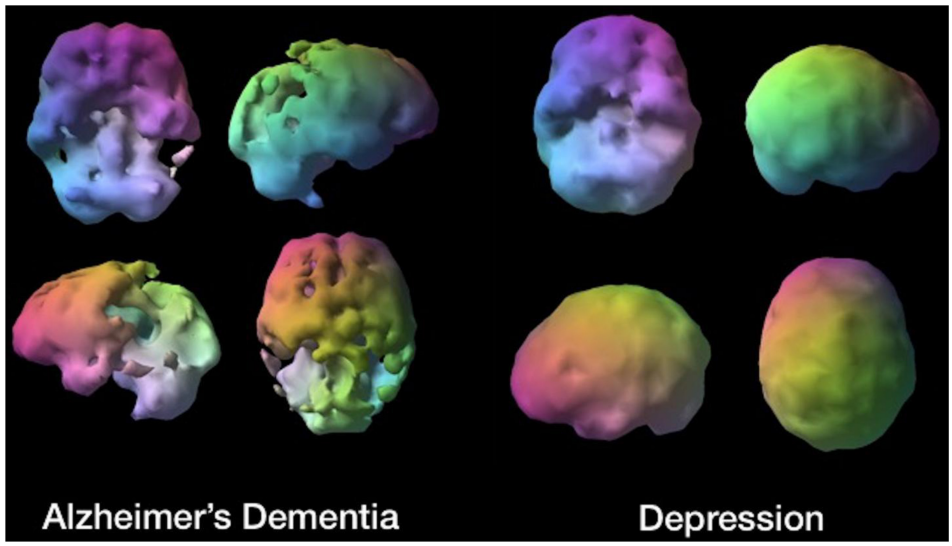


Figure 5.30. 99mTc-HMPAO for cerebral blood flow measurements by SPECT can reveal diseases such as Alzheimer's and depression.

blood flow measurements using compounds such as HMPAO that converts rapidly from the lipophilic form, which passes the blood-brain barrier (BBB), to the hydrophilic form, which is unable to pass the BBB, and is trapped in the brain (figure 5.30).

5.6.2 Innovative therapy

Marco Durante[1]
[1]GSI Helmholtzentrum für Schwerionenforschung and Technische Universität, Darmstadt, Germany

5.6.2.1 General overview

Soon after the discovery of x-rays it became clear that radiation could have been used to destroy tumors, thus starting what we call radiotherapy. The goal of radiotherapy is very simple: to provide a dose as high as possible to the tumour, in order to sterilize it, but simultaneously to minimize the dose to the normal tissue, so that side effects are acceptable. The curves representing the tumour control probability (TCP) and the normal tissue complication probability (NTCP) as a function of the dose must stay separated (figure 5.31). The role of the physics is to enlarge their distance (therapeutic window), thus making the treatment effective and safe.

Figure 5.32 gives a schematic view of the progress in radiotherapy in the past century focussing on the most common male tumour treated in the clinics: prostate cancer. Radiotherapy starts with orthovoltage x-rays that cross-fire the tumour target in a box. In fact, the depth–dose-distribution of x-rays is unfortunate: the entrance dose is always higher than the dose deep in the body, where the tumour is situated (figure 5.33).

Cross-firing from different angles is therefore necessary to enhance the target avoiding unacceptable toxicity in the entrance.

To make the depth–dose distribution more 'flat' it is necessary to increase the photon energy, and therefore the dose distribution improves with the introduction of ^{60}Co γ-rays (about 1 MeV energy) and later with linear electron accelerators (linacs),

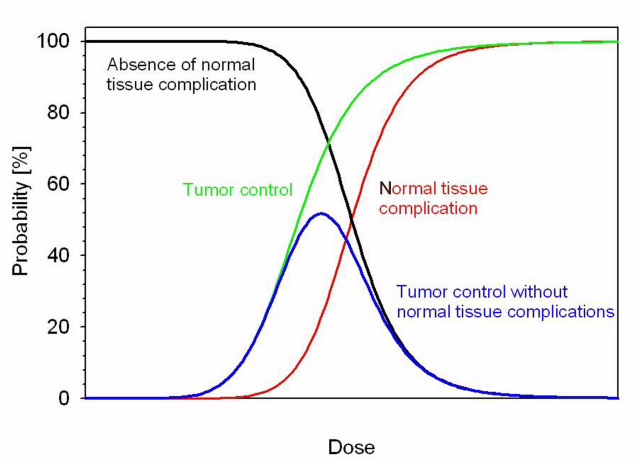


Figure 5.31. The therapeutic window in radiotherapy. The green and red curve should be as separated as possible, so that the blue curve can have a high peak.

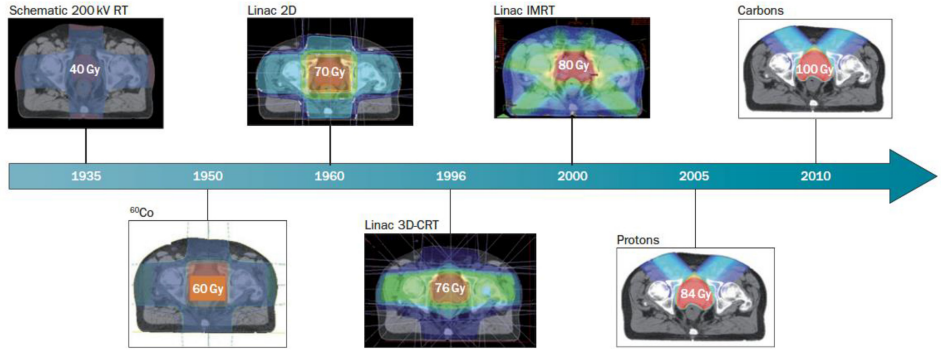


Figure 5.32. The evolution of prostate cancer radiotherapy in the past century. Abbreviations: RT = radiotherapy, kV = kilovolt, 60Co = cobalt-60 γ-rays, linac = linear electron accelerator, CRT = conformal RT, IMRT = intensity-modulated RT. The dose is measured in Gy (1 Gy = 1 J kg^{-1}), and should be as high as possible in the target.

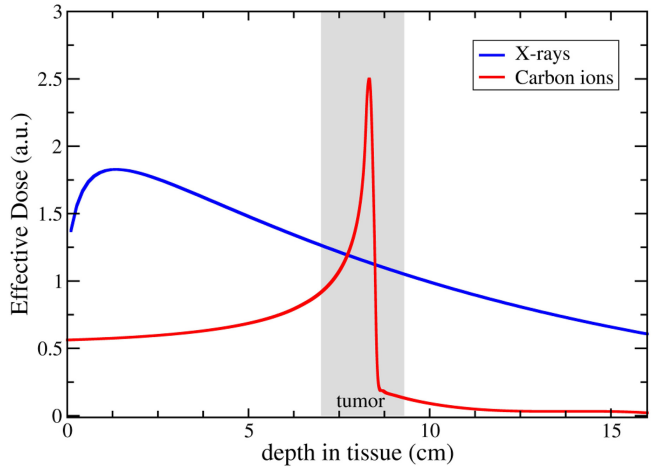


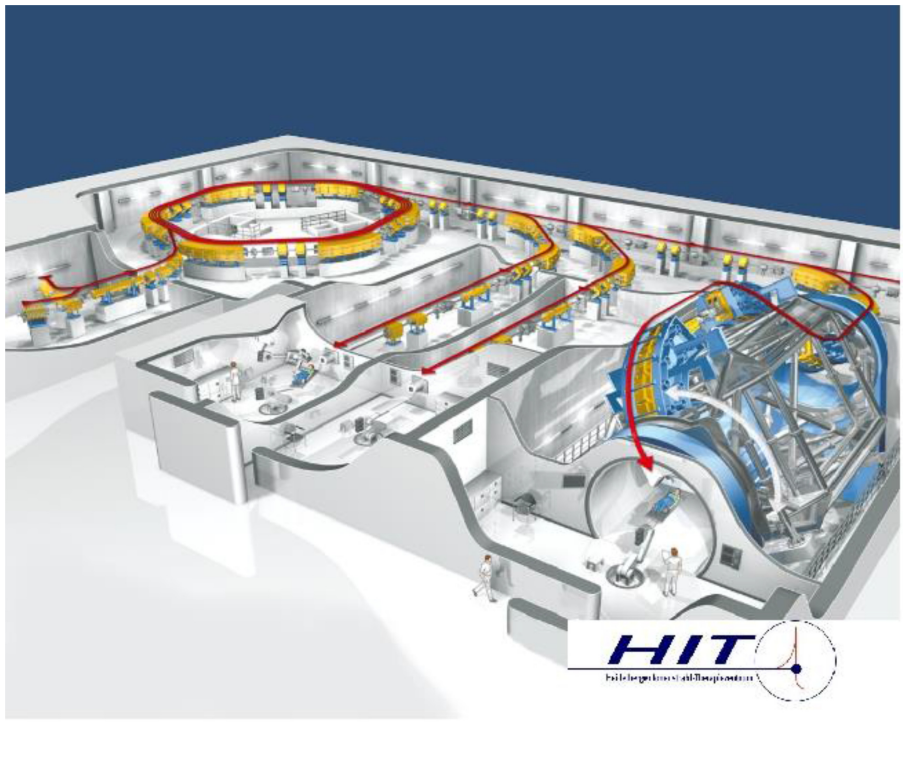
Figure 5.33. Depth–dose distribution of x-rays and charged partciles (here ^{12}C-ions) in human tissue.

which can reach electron energies around 25 MeV, even if the current linacs are normally operated at 6 MeV. However, the cutting-edge technology is the use of protons and heavier ions, because the depth–dose distribution is favourable (figure 5.33). Charged particles deposit indeed most of their energy at the end of the range (the Bragg peak), where the tumour is situated. The exact position of the Bragg peak can be modulated by changing the accelerator energy, and the Bragg peak can be widened (spread-out-Bragg-peak, SOBP) to cover the whole tumour size (generally several cm) by superimposition of beams of different energy.

In addition to the physical advantages, slow ions in the Bragg peak are biologically more effective than fast ions in the entrance channels, whose radiobiology is more similar to x-rays. For this reason, heavy ions are particularly indicated for radioresistant, hypoxic tumors.

As a consequence of their physical and biological advantages, particle therapy is growing worldwide. According to the statistics of the Particle Therapy Co-operative group (PTCOG), in January 2021 there were 111 particle therapy centers actively treating cancer patients, but only 12 using ions heavier than protons (C-ions). Despite the favorable dose distribution shown in figure 5.32, only about 1% of prostate cancer patients are treated with particles. The main problem remains the high cost. Particle therapy centers require indeed large circular accelerators (cyclotrons or synchrotrons) and heavy gantries for beam delivery at different angles (figure 5.34). As a result, the cost of a heavy ion therapy can be as high as 200 M€, versus only 5 M€ for a conventional linac.

In addition to external beam therapy (teletherapy) there has been strong progress in target radionuclide therapy, an internal therapy based on the idea of delivering molecules labeled with radioisotopes to the tumor. The radiopharmaceutical is comprised of a radioisotope with appropriate half-life, a linker, and a small molecule (peptide, antibody, etc) able to target cancer cells. A gold standard for targeted β-therapy is ^{177}Lu, which has been used with success in many tumors including neuroendocrine and advanced prostate cancer.



C BY 3.0)

Figure 5.34. Layout of the HIT facility in Heidelberg, where patients can be treated with protons or C-ions. Two treatment rooms have fixed horizontal lines, while one is equipped with a 670-ton rotating gantry.

5.6.2.2 Challenges and opportunities

It is generally acknowledged that the number of cancer patients receiving radiotherapy (slightly more than 50%) is too low and should be increased in the coming years. While most patients will be treated with IMRT, particle therapy is at a tipping point. There are many efforts to reduce the cost of using superconducting magnets for synchrotrons with reduced footprints. Laser-driven particle accelerators are still immature for medical applications, but they bear the potential of having table-top accelerators for particle therapy, even if this goal goes probably beyond 2050. The same seems to be true for other innovative accelerators such as dielectric wall or fixed-field alternating gradient (FFAG) accelerators.

Beyond the cost, particle therapy has to meet the challenges posed by innovative methods in conventional radiotherapy, which is rapidly improving thanks to image guidance. MRI-linacs are already in use in clinics, and allow treatment of patients with full control of organ movements using MRI imaging. Particle therapy is still lagging behind in image guidance, even if the nuclear interaction of charged particle offer beam monitoring and range verification opportunities that are not possible with photons, such as secondary radiation, proton radiography, and PET.

This includes prompt γ-ray detection to assess the beam range; dedicated monitors are already commercially available. Particle radiography is another attractive possibility, both for imaging and for the possibility of reducing the uncertainty in the conversion between the Hounsfield units coming from the CT and the water-equivalent path length needed for particle treatment planning. PET has been used already for C-ion therapy and would greatly benefit if therapy could be done using radioactive ion beams (e.g., ^{11}C or ^{10}C), positron emitters that can be produced by fragmentation of stable ^{12}C. The results would be an enormous increase in the signal-to-noise ratio and the possibility of monitoring online the beam delivery in the tumour (figure 5.35). The use of radioactive beams in therapy had already been proposed in the 70s at the Lawrence Berkeley Laboratory, but has always been hampered by the low intensity of the radioactive beams. The recent intensity upgrades in the accelerators and the push for improving precision of particle therapy has raised interest in this project. CERN has a project for a cyclotron producing radioisotopes to be injected in conventional medical synchrotrons, whilst GSI, exploiting the intensity upgrade for the construction of the new FAIR accelerator, is actively measuring these beams in the framework of an EU ERC Advanced Grant.

In targeted radionuclide therapy, the new challenge is probably the use of α-emitters. Being short range and with high biological effectiveness, α-particles are indeed ideal for effective tumour cell killing with reduced damage to the surrounding normal tissue compared to the long-range β-particles. Some of the clinical results obtained with radioisotopes such as ^{211}At, ^{213}Bi, ^{223}Ra, ^{212}Bi, ^{225}Ac in metastatic patients have been stunning and have attracted great interest in this target therapy modality. The main challenge for this therapy is actually the radioisotope production. Availability of radioisotopes is a general problem for both diagnostic and therapy because many of them are produced in nuclear reactors that are in the process of being shut down in many countries. Even ^{99m}Tc, the most commonly used radioisotope in the world for radiodiagnostics, is insufficient. The problem is even

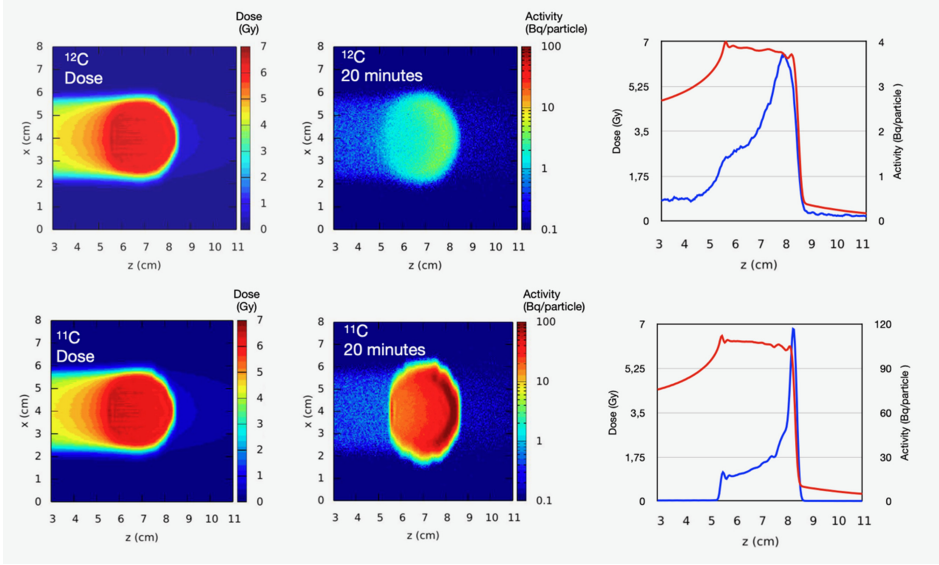


Figure 5.35. Monte Carlo simulations of PET images of stable and radioactive carbon beams. The two columns refer to two different signal acquisition times: 20 s (left) and 20 min (right).

more grave for a-emitters: for instance, ^{211}At is only produced at ARRONAX, the radioisotope factory in Nantes (France). Radioisotope production will certainly be a major challenge in the coming years.

5.6.3 Superconducting magnets from high-energy physics hadron colliders to hadron therapy

Lucio Rossi[1]
[1]Università di Milano—Dipartimento di Fisica and INFN-Department of Milano, LASA Laboratory, via Fratelli Cervi 201, 20056 Segrate, Milano, Italy

Superconducting magnet development has accompanied the energy and luminosity increase of the hadron colliders in the last 50 years. The first-generation superconducting magnets, in particular the study and R&D for the pioneering Tevatron at Fermilab, were instrumental for the development of the MRI technology. The present accelerator magnet technology, following the industrially built HERA, RHIC, and LHC colliders, is mature for direct use in medical environments and in particular for heavy particle (hadrons) therapy. The first application in prototyping phase is a gantry (i.e., the beam delivery system allowing irradiation of the patient from multiple directions). In this contribution, after a review of the achievement of superconducting magnet technology and of the development for next- generation colliders, we discuss the main development lines that should enable a superconducting ion gantry to be commissioned in this decade.

5.6.3.1 Introduction

Magnets have been the most important application of superconductivity, and accelerators can certainly be credited to be among the key drivers of the development of this technology.

Accelerator technologies, and thus accelerator magnets, need long R&D with programs spanning over decades, which allows investigation and R&D to be done in a coherent way and thus to be fruitful. The latest example, the LHC, is the summit of over 30 years of development of superconducting magnets (SCM). Its enormous size and its long-awaited goal, the Higgs particle, whose discovery about ten years ago at the Large Hadron Collider (LHC) at CERN (Last update in the search for the Higgs boson 2012) have been heralded worldwide, and the possible unveiling of the new world beyond the Standard Model make the LHC (LHC DESIGN Report 2004, The Large Hadron Collider 2018) the crossroad between past and future of high-energy physics (HEP). But LHC may be the crossroad also for medical sector superconducting technology. Thanks to the experience gained with superconductivity for LHC and its predecessors, now the accelerator magnets community is exploring the possibility of using superconducting accelerator magnet technology for particle irradiation-based cancer therapy.

5.6.3.2 Accelerator magnets and hadron therapy

Since the pioneering time of Ernest O Lawrence in Berkeley, where he designed, built, and put in operation the first cyclotrons in the 1930s, the demand for higher energy accelerators has driven the quest for larger size rings and higher field magnets. Not surprisingly superconductivity has become a key technology for

modern accelerators and, conversely, particle accelerators have promoted superconducting (SC) magnet technology (Wilson 1999). The timeline of record accelerator energy is shown in the compilation reported in figure 5.36, where we have collected the main accelerators for nuclear and particle physics, classifying them according to the center-of-mass energy and highlighting the projects that are based on superconductivity. Many of them, the most powerful ones actually, feature SC magnets, or superconducting RF cavities, as the main technology and performance driver.

HEP has been pursuing superconductivity for many large projects, the most recent one being the already cited LHC at CERN, and plans are being made for further stronger and larger colliders, as shown on the right side of the timeline of figure 5.36. We wish in particular to refer to the some of the monographs (Wilson 1983, Mess *et al* 1996, Ašner 1999) as well as to the more recent reviews on SC magnets and conductors for accelerators (Tollestrup and Todesco 2008, Rossi and Todesco 2011, Rossi and Bottura 2012, Bottura and Rossi 2015).

As far as the medical sector is concerned, superconductivity has been limited to a few types of low-energy cyclotrons or synchrocyclotrons, mostly for hadron (protons or ions) therapy. The use of particles accelerated at a few hundreds of MeV per mass unit for cancer therapy is a well-established technique (Degiovanni and Amaldi 2015). Particle therapy can provide a powerful tool for the treatment of tumours that are not curable with conventional x-rays or that are better treated by particle beams. There are 24 particle therapy centers in Europe and 105 in all the world. However, only four of them in Europe and only 12 in the world employ heavy ions (Carbon ions), the rest employing only protons, despite the fact that heavy ions, like carbon, are effective for some tumours resistant to x-rays and protons, and are superior in preserving the healthy tissue surrounding the cancerous zone. This is due to the

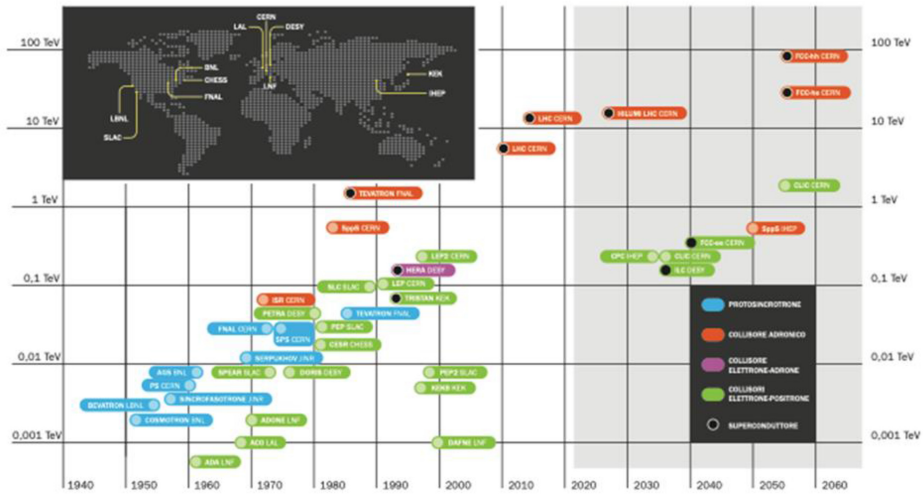


Figure 5.36. The history of accelerators and colliders with center of mass energy versus time. Black dots indicate the superconducting machines. Courtesy of INFN-Communication Office.

much larger size and cost of the required infrastructure for ions than for protons (not to mention x-ray).

As a further consideration, almost all proton centers are equipped with a gantry, allowing beam delivery from multiple directions, boosting the treatment effectiveness. On the contrary, very few ion centers are equipped with a gantry. The first, and so far unique in Europe, has been the center of HIT (Heidelberg Ion Therapy center) (Haberer *et al* 2004) that has been in operation since 2012 and employs classical resistive magnets, which makes it have a length of 26 m and a weight of about 600 tons (rotatable part). A step forward has been carried out by the HIMAC center that from 2018 has been routinely using a more compact ion gantry based on SC magnets, designed and manufactured by Toshiba (Iwata *et al* 2012). This has allowed the Japanese center to reduce the size and weight to about 13 m and 300 tons.

A further reduction of size and weight, which eventually would reflect in reduction of cost, would certainly enable a wider diffusion of ion gantries, thus contributing to better coverage of the territory by ion therapy centers. A large collaboration in Europe is looking to leverage the experience gained in the HEP laboratories for collider SC magnets to try to make a bold step and design a superconducting gantry for ion less than 100 tons in weight. This would allow to reduce the direct cost of the gantry and footprint of the infrastructure. This application would be the second, more direct, beneficial spin-off from HEP accelerators into the medical sector, after the development of the superconductor for the Tevatron that was essential for enabling superconducting MRI magnets. Today MRI is a 5 billion dollars business, and thousands of the large magnets fabricated each year for large size imaging are superconducting ones, thanks to the high-quality, multistrand, high current density superconductor developed first for Tevatron.

5.6.3.3 Accelerator magnets

HEP synchrotrons and especially HEP collider rings are the most challenging applications for SC magnets, as field quality, stability, and cost are very demanding and by far superior to other systems (like Fusion, MRI, high-field solenoids, etc). They are superconducting, with a few exceptions such as fast-cycling synchrotrons like the J-PARC complex; however, the fast-cycling FAIR-SIS100, for example, is superconducting as well. Magnets for superconducting linear colliders such as ILC, XFEL, and ESS are superconducting, too, despite the very low-field/gradient, because of the big gain in compactness and integration within the cryo-module (cryostat) hosting the superconducting radio-frequency cavities. There has been considerable effort in SC magnets for low-energy accelerators for nuclear physics, like cyclotrons or synchrocyclotrons, an effort that continues mainly driven today by medical applications. These SC magnets are similar in design and layout to magnets for MRI or particle detectors for momentum spectrometry, having the shape of solenoid or circular split coils. Here we will limit our consideration only to SC magnets for HEP colliders and their implication for the medical sector.

Superconducting magnets become of interest when the required field is above the iron saturation limit (i.e., approximately 2 T). In some instances, however, SC

magnets are becoming a choice of preference for their compactness and their low-energy consumption also at low field, a topic that is increasingly important in the design of new accelerators.

Dipoles

The first function of a magnet is to guide and steer charges particles (i.e., to provide an adequate centripetal force to keep them in orbit in a circular accelerator or just to bend them in a transfer line; see figure 5.37, left). The adequate force can be given only by a transverse magnetic field, except for very low-energy beams, where solenoids and sometimes electrostatic deflectors are used. Transverse field means a magnetic field whose main component is perpendicular to the particle trajectory. For high-energy accelerators, the beam region is a cylinder that follows the beam path and with the smallest practical dimension, as shown in figure 5.37, right. As discussed later, the cost and technical complexity of the magnetic system are proportional to the energy stored in the magnetic field, which explains why it is important to minimize the size of the magnet bore.

Even though static magnetic fields do not accelerate, in circular accelerators the bending (dipole) field eventually determines the final energy reach. In relativistic conditions the relation between the beam energy, E_{beam} in TeV, the dipole field B in T, and the radius of the beam trajectory inside the bending field R in km takes a very simple form:

$$E_{\text{beam}} \cong 0.3BR$$

Since the dipole field typically covers 2/3 of the accelerator, R is about 2/3 of the average radius of the ring. The above relation clearly shows the interest for the highest possible field for a given tunnel.

Figure 5.38 shows an artistic sketch of the LHC dipole, which is called the Twin-Dipole since it hosts two opposite dipole fields in the same magnet, to bend the two counter circulating proton beams of the colliders. There are 1232 main dipoles installed in the LHC ring.

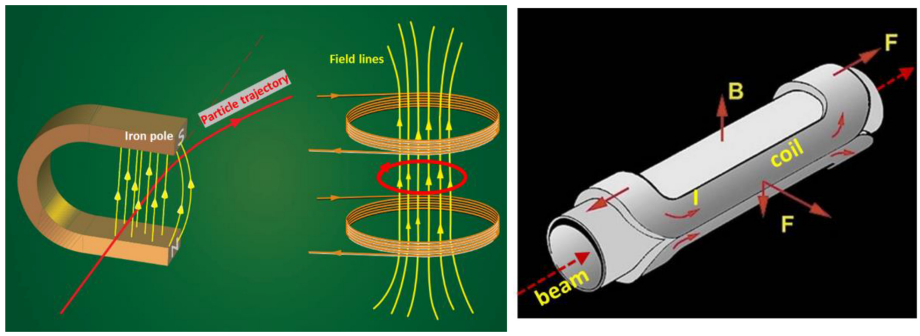


Figure 5.37. Effect of dipole field on charged particle (left, CERN/AC archives) and schematic of an accelerator magnet (right).

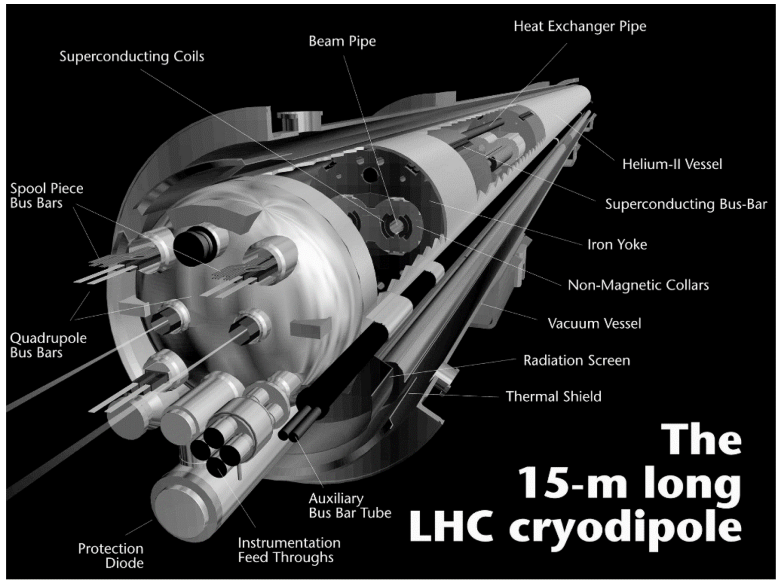


Figure 5.38. The LHC main dipole (CERN/AC archives).

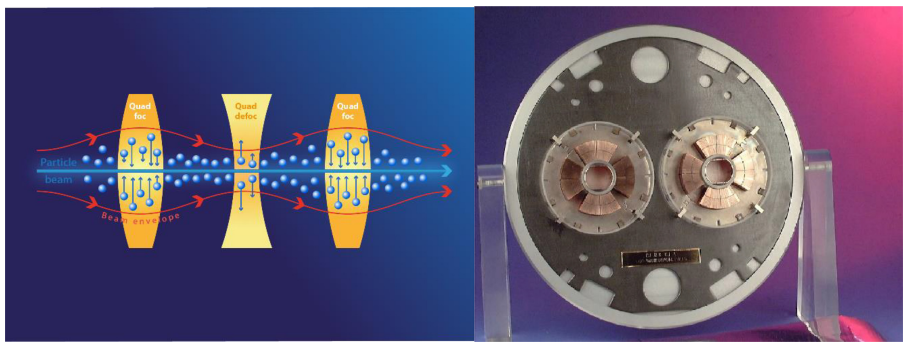


Figure 5.39. Left: Effect of a triplet array of quadrupoles (strong focusing principle). Right: Cross-section of an LHC main quadrupole (CERN archives).

Quadrupoles and high-order magnets

The other key function of magnets is to provide the stability of the beam in the plane transverse to the particle motion. This is realized by the use of quadrupoles, sextupoles, octupoles, and in certain cases of higher-order harmonic magnets; in the LHC, magnets up to the dodecapole order are employed to fulfil the request of the sophisticated beam dynamics. For example, the action of a quadrupole on a particle beam is schematized in figure 5.39, left, showing the effect of the so-called strong focusing on a particle beam. In figure 5.39, right, a picture of the cross-section of one of the 400 main quadrupoles is shown.

While the main quadrupoles are usually of similar—although reduced—size and complexity with respect to the main dipoles, sextupoles and octupoles are of smaller

size, peak field, and stored energy, thus featuring less complexity than the main dipoles. It is worth noticing that the quadrupoles that are in the so-called insertion regions, just before the collision points, usually called the low-β section, are necessary for the optics manipulation to focus the beam to the smallest attainable size (i.e., the minimum β-function) at the collision point. This set of quadrupoles must have an aperture significantly larger than the lattice quadrupoles and requires an integrated field gradient much larger than the regular lattice quadrupoles. Larger aperture and higher gradient result in a peak field that could be close to that of the main dipoles. In exceptional cases, like the High Luminosity LHC project, actually the low-β insertions around the ATLAS and CMS experiment would be equipped with quadrupoles of much larger peak field than the lattice LHC dipoles.

Accelerator dipoles and quadrupoles: design and field evolution

Accelerator magnets are very demanding, as mentioned above. Before examining in detail these unique characteristics, we want to underline that such characteristics must be obtained in hundreds or thousands magnets. The solutions to the required specifications must have a cost within reasonable limits. The HEP community cannot afford the luxury of very expensive perfect magnets. For example, each one of the 1232 lHC dipoles cost about 1 MCHF (value of 2001). This explains why one single component, the main LHC dipole, accounted for 70% of the cost of the whole magnetic system (including all other magnets, protection, powering, cryostat, interconnections, etc) and, alone, for 40% of the *total LHC collider project cost* (significantly the whole magnet system accounts for 1700 MCHF of the 3400 MCHF total LHC cost, 2008 price). However, 1 MCHF is surprisingly small for a 15 m long, 30 tons magnets with 7 MJ stored energy and rated for 9 T, 1.9 K, 13 kA operative conditions. Let's now consider the main design topics of accelerator magnets (Rossi and Todesco 2007): superconductor, critical current density, electromagnetic design, and quench-protection. For brevity we do not discuss structural design and stability design, which are also quite important.

Superconductor

There are tens of thousands of superconducting materials. However, only a few of them have characteristics that make them usable in real devices (i.e., with a critical temperature well above 4.2 K, LHe normal boiling temperature), critical fields above 10 T, critical current density above 1000 A mm^{-2} at the operating temperature and field [e.g., for LHC is 1.9 K and nearly 9 T]). In addition to these characteristics, the superconductor for accelerator magnets needs: (i) 3–20 μm effective filament diameters, to maintain a good field quality during the ramp from injection to flat top; (ii) a geometry and a ductility such as to be assembled in compact cables; and (iii) to be robust enough to be wound in a coil and to withstand mechanical stresses largely exceeding 100 MPa. In figure 5.40 the engineering critical current density J_e of the few available practical superconductor is plotted versus magnetic field. J_e is defined as $J_e = J_c \cdot ff$ where J_c is the critical current density of the superconducting material and $ff = A_{SC}/A_{tot}$ is the volumetric fraction of superconductor in the wire (that contains also stabilizing copper and other passive

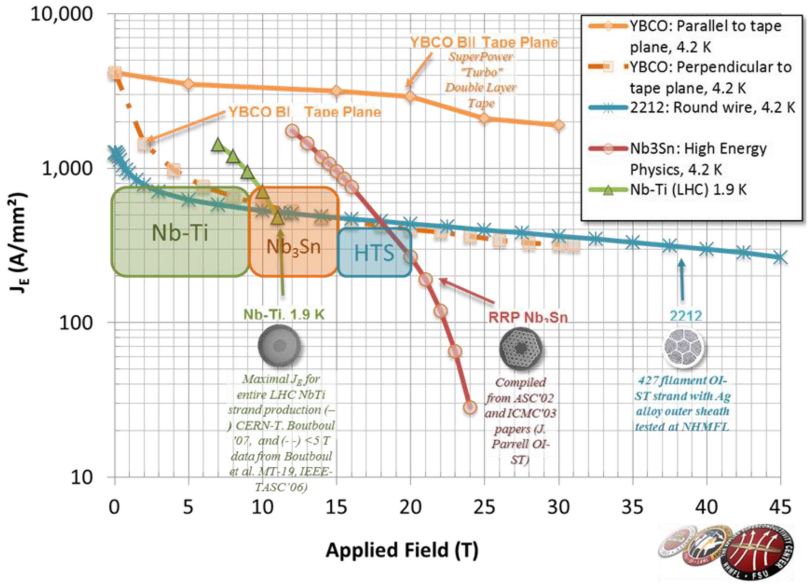


Figure 5.40. Engineering current density of practical superconductor versus field. The useful range for accelerator magnets is 200–800 A mm^{-2}. The data and curves are courtesy of Peter Lee (NHHMFL-FSU) and refer to year 2013 when the 16 T FCC baseline was decided. The working range is indicated by coloured boxes for Nb–Ti, for the Nb_3Sn developed for HL-LHC, and for HTS that is under development.

materials, like barriers and alloys). As is known, all superconducting magnets in existing accelerators are wound with Nb–Ti superconductor, High Luminosity LHC being the first project breaking the 10 T-wall (i.e., attempting to use Nb_3Sn for collider magnets rated for 11.5 T, and MgB_2 for the powering line; Bottura *et al* (2012), Rossi and Bruning (2016)).

In figure 5.41 a cross-section of the Nb–Ti/copper wires used in the LHC magnet is shown, together with the cable grouping many wires to reach at least 13 kA at 11 T, 1.9 K. The flat cable shown in figure 5.41 whose aspect ratio is ~10 and has 20–40 wires has a slightly trapezoidal cross-section with the two large faces making a 0.5°–1.5° angle, which is called the 'Rutherford cable'.

High critical current density

The average current density in the whole coil section, J, is the basic parameter since the field $B \propto J{\cdot}\cos\vartheta{\cdot}t$, with t being the thickness of the coil. Since the coil thickness, differently from other applications like solenoids, detector magnets, etc, cannot be made very large, the only way to produce high field is to use high J. This implies both high critical current density, J_c, the most relevant material property and obliges using at least 75%–80% of the available J_c, with the consequence that J_c has to be high and also very uniform. The LHC dipoles work at 86% of the crossing of the load line with the critical current curve, and they have been designed and tested to 93% of J_c, so a variation larger than 5% directly affects the performance of the entire accelerators, while it is negligible in other systems (Hull *et al* 2015).

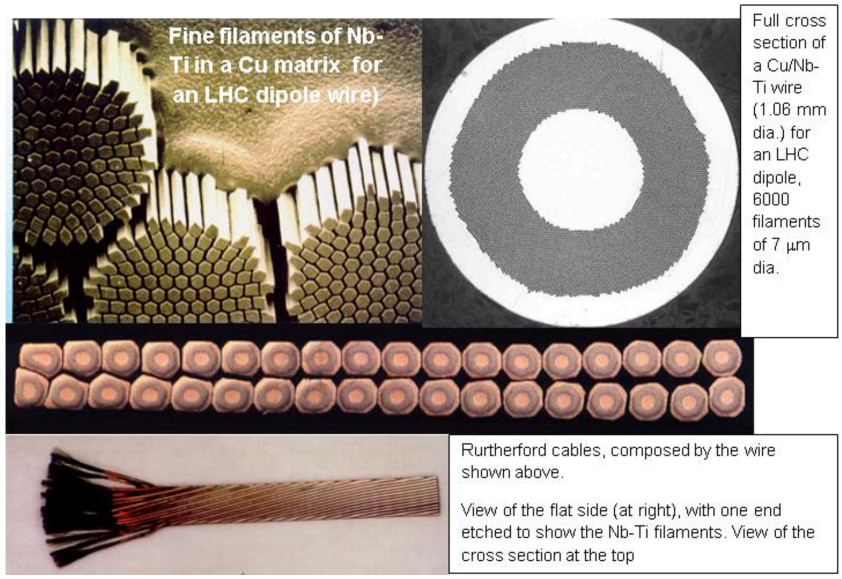


Figure 5.41. LHC superconducting wire and cable (CERN/AC archives).

The high J_c must be preserved by using low stabilizing content (i.e., the fraction of copper in the superconducting wire; see figure 5.41). A minimum amount of copper is necessary, but we need to keep it at minimum compatible with the stabilization and protection requirement, keeping high the filling factor *ff*, to avoid excessive lowering of *J*. Lack of sufficient stabilizing copper can generate big damage as in the case of the well-known LHC incident (Rossi 2010). Usually Cu/SC (or Cu/non-Cu) ranges between 1.5 and 2 for Nb–Ti-based coils and 0.9–1.5 for Nb_3Sn magnets, just the minimum for stabilization and dangerously near to the minimum for protection: we need to protect magnets with such low copper content despite strong force density, high stresses, and a stored energy in the 1–10 MJ ranges. Finally, the filling factor of the cables also needs to be high, keeping voids at 10%–12%: only Rutherford cable can achieve this compaction level. Insulation also must be minimized in volume, despite the fact that voltage could be as high as a hundred volts between coil sections, and a few kV to ground. Usually, the insulation fraction is 10%–15% of the total of that for thin cables, a rather small value.

Where all the factors are folded together, one realizes that the effective (overall) current density is $J_{overall} \sim 0.3\ J_c$. This allows our magnets to work at $J_{overall}$ of 300–600 A mm^{-2} (i.e., *3–10 times higher than any other magnet applications*).

Electromagnetic design and field quality

The coil cross-section is usually a circular sector filled with Rutherford cable arranged with spacers to mimic a distribution of current $J_0\cos\vartheta$ (i.e., a current density that is maximum at the midplane, J_0, decreasing like cosine function when increasing azimuth, and becoming zero at the pole [vertical axis]; Rossi and Todesco 2007, 2006). This is the most efficient distribution to generate transverse field in an

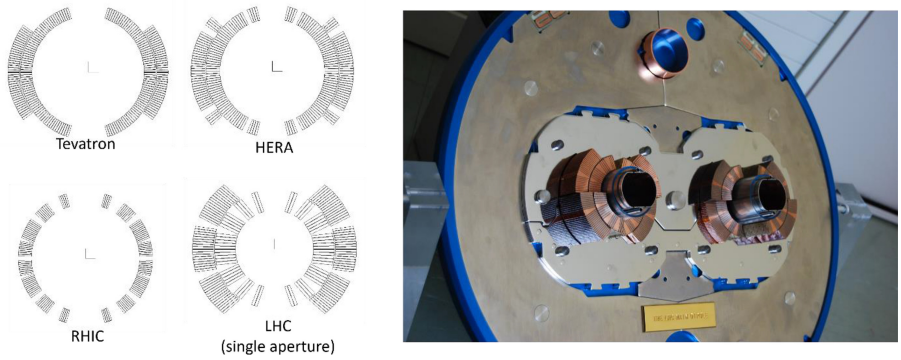


Figure 5.42. Cross-section of the main dipoles of the most important Hadron Colliders (left) and picture of a mock-up of the cross-section of a real LHC main dipole (Twin-Dipole).

infinitely long cylinder. The coil aperture being 50–100 mm and the magnet length being 5–15 m, the approximation is very good, indeed!

In figure 5.42 we show the cross-section of the main dipoles for the most important hadron colliders and a figure of a real LHC main dipole cross-section. A perfect $J_0\cos\vartheta$ would yield a pure dipole field. A perfect $J_0\cos 2\vartheta$ (i.e., with maximum at 0°, 90°, etc) would yield a perfect quadrupolar field, a $J_0\cos 3\vartheta$ current distribution would yield a perfect sextupole field, etc. The winding cannot be perfect, of course: however, the requested relative field accuracy is of the order of 10^{-5}. This calls for very accurate design and tight control of the tolerances. In the LHC all components have very strict tolerances from a few μm (superconductor) to 15–30 μm for the biggest mechanical components.

In addition, the bending strength of each dipole must be equal to nearly 0.01%, because a full sector (154 dipoles in the LHC) is powered in series, This is actually one of the main constraints of the use of SC magnets in accelerators (e.g., the RF cavities in a linac may differ from each other in integrated gradient by 5% without serious drawbacks). In hadron collider the single weakest dipole determines the energy of the beam.

Protection and quench

The main issues for HEP magnets are the large current density in the conductor and the large magnetic energy stored per unit of coil mass. When a magnet undergoes a sudden irreversible transition from superconducting to normal conducting state, which is called 'quench', the magnetic stored energy is turned into heat via Joule effect. SC accelerator magnets can tolerate temperature increases up to 300 K, if well-conceived and manufactured. The hot-spot temperature $T_{\text{hot spot}}$ can be estimated using an adiabatic heat balance, equating the joule heat produced during the discharge to the enthalpy rise of the conductor:

$$\int_{T_{\text{op}}}^{T_{\text{hot spot}}} \frac{C}{\rho} dT = \frac{1}{A_{\text{Cu}} A_{\text{tot}}} \int_{t_{\text{quench}}}^{\infty} I^2 dt$$

where C is the total volumetric heat capacity of the superconductor composite, and ρ is the resistivity. The speed at which the energy is dumped depends on the magnet inductance and the resistance of the circuit formed by the quenching magnet and the external circuit $R = R_{\text{quench}} + R_{\text{ext}}$. The characteristic time of the dump is then $\tau \approx L/R$, which can be made short by decreasing the magnet inductance by use of large current cables. Protection (i.e., fast dump to avoid excessive T_{hotspot}) is then achieved by triggering heaters embedded in the winding pack, and fired at the moment a quench is detected to spread the normal zone over the whole magnet mass. In the LHC the typical time to detect a quench is 10–20 ms and the heaters must activate the quench in 100 ms or less. A further complication in collider is due to the fact that the main magnets are powered in series, and the stored energy in a single circuit is much larger than the few megajoule energy stored in a single magnet. In the case of each of the eight circuits formed by the series of 154 LHC main dipoles, the stored magnetic energy attains the gigajoule level, which was the reason for the gravity of the above-mentioned LHC incident of 2008 (Rossi 2010).

5.6.3.4 Superconducting accelerator magnets: outlook to future

As previously mentioned, LHC has been the summit of 30 years of development of accelerator magnets, all wound with Nb–Ti superconductor. For more than ten years a considerable amount of effort has been occuring in the US laboratories, BNL, FNAL, and LBNL, as well as in Europe at CERN, CEA (France), and—more recently—at CIEMAT (Es), INFN (It), PSI (Ch), to develop accelerator magnets wound with Nb_3Sn. As can be seen in figure 5.40, Nb_3Sn extends the attainable region from 8–9 to 15–16 T. The High Luminosity LHC project at CERN (Brüning and Rossi 2019, Rossi and Brüning 2019) has the main technological scope of design and manufacture, and operated the first set of Nb_3Sn magnets in a collider, ever. To go beyond this field level in the 20 T region use of HTS (high-temperature superconductor) is necessary (Rossi and Senatore 2021), as can be seen in figures 5.40 and 5.43, which summarizes the past and the possible future of HEP accelerator magnets.

However, these technologies, Nb_3Sn and HTS, are not yet mature enough for societal applications, and in the next chapter we will deal with how the technology used for present colliders can spill into the medical sector.

5.6.3.5 From HEP to hadron therapy: the new journey of accelerator magnets

As mentioned in the first section of this contribution, the use of the slim, tubular shape accelerator magnets can open the way to ion gantries of limited size and weight. Two points need to be specifically addressed for such a step into medical applications: (1) the need to ramp the field in coils that must be indirectly cooled (no use of liquid because gantry is rotating), with a ramp rate of 0.1–1 T s^{-1} which is from 10 to 100 times faster than the sweeping rate of HEP colliders; (2) the fact that the relatively low magnetic rigidity of the beam, together with fields of 3 T or more, implies the necessity to build strongly curved magnets, which has never been attempted for accelerator magnets.

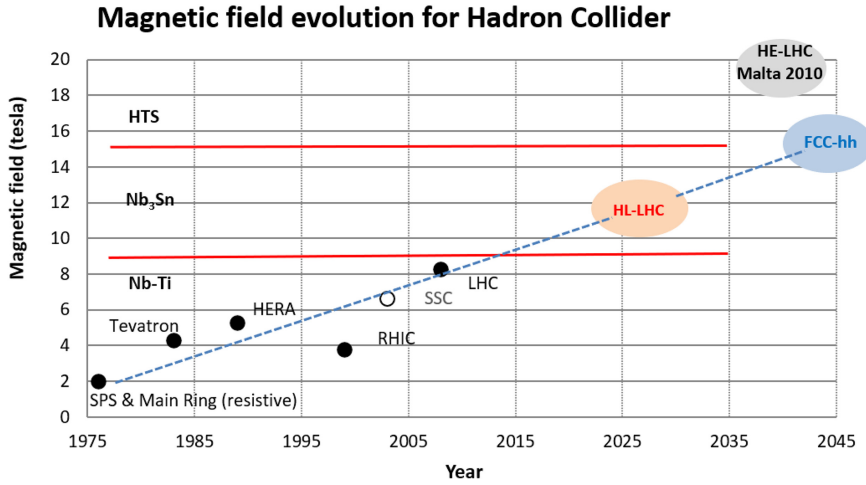


Figure 5.43. Evolution of the field in the SC magnets of the main past and future hadron colliders.

The recent rush for superconducting gantry

The TERA foundation (http://www.tera.it/) with the support of CERN in the last three years worked out a preliminary design for a light (~50 tons) rotatable gantry for ions. The gantry features SC magnets of cosϑ layout, with ramping operation, and has been named SIGRUM (Amaldi *et al* 2021). In 2020, following the phase of preliminary study, a larger collaboration led by CERN and extended to CNAO (CNAO n.d.), MedAustron (MedAustron n.d.), and INFN (departments of Milano-LASA and Genova) took over the SIGRUM design study and called an international review to assess the design. The outcome of the review was very positive, endorsing the basic design of SIGRUM. However, the review panel questioned the choice of the 3 T dipole field on which SIGRUM is based (Karppinen *et al* 2022) and has recommended, besides the construction of a demonstrator magnet, to make an effort toward increasing the dipole field and also to explore the feasibility of switching to a CCT (Canted Cosine Theta) dipole design.

The CCT design

The CCT is a new magnet type that has been pursued at LBNL for HEP and for a superconducting gantry for proton beams (Brouwer *et al* 2020). The schematic of a CCT dipole, basically composed of two slanted solenoids, is shown in figure 5.44, together with a picture of the construction at CERN of 2 m-long prototype (Kirby *et al* 2020), designed and built at CERN for the High Luminosity Project (Brüning and Rossi 2015). The CCT for High Luminosity LHC is the first application of CCT magnets in a collider. Superconducting CCT magnets as possible backbone of SC gantry for ions is the investigation topic of a large European collaboration, formed by CERN, CEA (Saclay, Fr), CIEMAT (Madrid), INFN (Milano and Genova), PSI (Ch), Uppsala University (Se), and Wigner Research Center for Physics (Hu). This collaboration has received two grants, one as part of the H2020-HITRI*plus* program (HITRIplus n.d.) and the other as part of the H2020-I.FAST program (I.

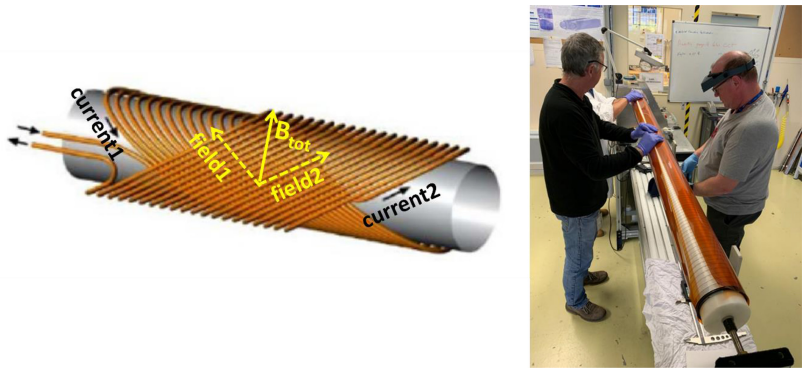


Figure 5.44. Left: CCT principle where oppositely slanted solenoids generate pure dipole field. Right: The High Luminosity CCT prototype during winding at CERN (Courtesy of Glyn Kirby, CERN).

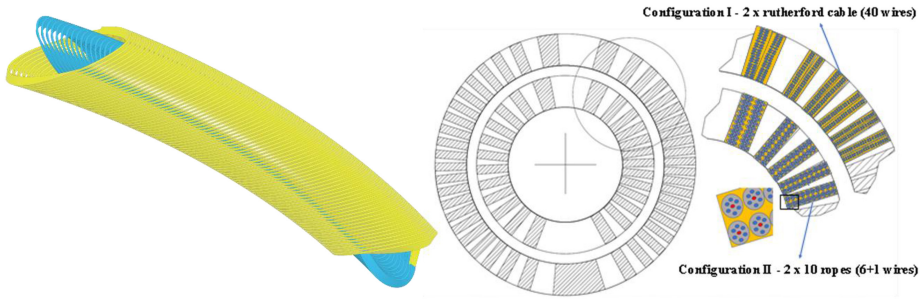


Figure 5.45. Sketch of the 4 T CCT dipole for HITRIplus. Left: 3D view along the length (1 m). Right: Cross-section showing the positioning of the conductor in the former grooves.

FAST n.d.). Both programs started in spring 2021. HITRI*plus* aims at building a 4 T demonstrator of a CCT SC dipole for ion gantry, with the necessary curved shape, with Nb–Ti conductor. I.FAST comprises also three companies in the consortium, in addition to the above-mentioned laboratories: BNG (De), Elytt (Es), and Scanditronics (Se). I.FAST plans to test a straight CCT wound with HTS in the 4–5 T range. These programs are extensively reported (Rossi *et al* 2022a). It is worth noticing that in the frame of this collaboration a working group is considering how to adapt the field description used both by magnet designers and by beam optics designers, since the usual 2D field multipole decomposition is no longer valid for strongly curved magnets. In figure 5.45 a 3D view of the CCT magnet considered for HITRI*plus* is shown together with a cross-section of the coil-former package. In figure 5.45 it is shown that conductors lay in the grooves of the formers (the dashed sectors in the two annular formers) and that two solutions are being considered for the conductor shape: a Rutherford cable, similar to the one used to wind HEP magnets, and a rope cable, of the type 6 strands around a central core.

HITRIPlus and IFAST should be able to assess in three years if CCT is a viable design to compete and maybe replace the classical cosϑ layout magnets.br

The cosϑ design—INFN SIG project

The above-mentioned collaboration among CERN, CNAO, INFN, and MedAustron is moving along two lines below described:

(1) CERN has refined the design of the 3 T dipole SIGRUM gantry and in (Karppinen *et al* 2022) a fairly detailed design is presented for a 70 mm aperture combined function 3 T dipole with superimposed a 5 T m^{-1} gradient, capable of withstanding a ramp rate of 0.1 T s^{-1}. Figure 5.46, left side, shows a sketch of the magnet.

(2) INFN is investigating the possibility to improve the magnet performance, as recommended by the review panel. The study is supported by a 4-year grant of 1 M€ value that has been awarded by INFN management to study various technologies for the ion gantry, and had an effective start on the 1st of January 2022. The grant is complemented by contributions by CERN (250 k€) and CNAO (350 k€). The grant, called *SIG,* includes various items:
   a. Design, manufacture, and testing of a cosi dipole short prototype.
   b. Design of a scanning magnet system.
   c. Design and prototyping of the main components of a new dose-delivery system (DDS), integrated with a new range verification system (RVS) for ions.

The superconducting demonstrator is the most critical item of the SIG program, taking 600 k€ of the INFN budget, matched by the 600 k€ contribution of CNAO and CERN.

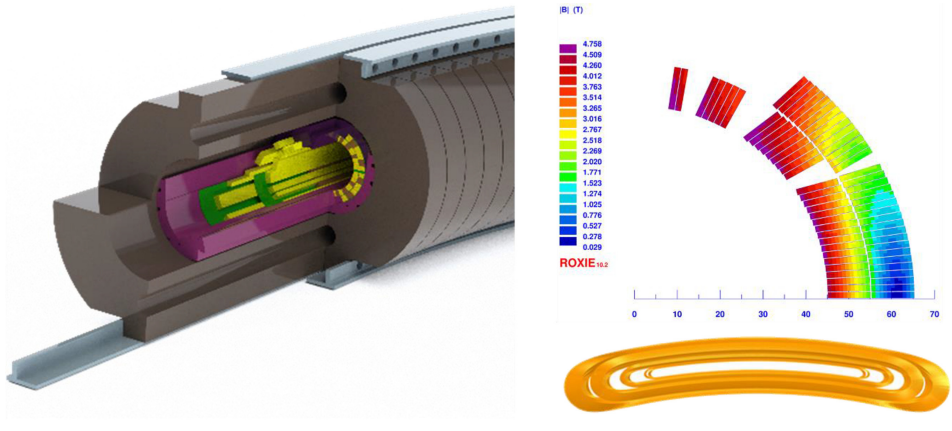


Figure 5.46. Sketch of the SIGRUM magnets (left, courtesy of M Karppinen, CERN) and of the SIG demonstrator coil cross-section with field map and coil top view (right).

For SIG we are investigating the following parameters range (Rossi *et al* 2022b):

- 4–5 T dipole field, which entails a bending radius of 1.35–1.65 m.
- 70–90 mm coil diameter (aperture).
- 0.4 T s^{-1} field continuous ramp rate up-down, with indirect cooling of the coils.

In figure 5.46, on the right side, a coil cross-section with field map for a possible SIG design is shown, together with a top view of the coil, well showing the 'banana shape'.

The Tevatron, HERA, and LHC (at 4.2 K) dipoles feature similar field and aperture (only RHIC has a lower field) (Rossi and Bottura 2012, Bottura and Rossi 2015), but these colliders have a very low-field ramp (0.01–0.03 T s^{-1}). Only the INFN-Discorap prototype dipole (Fabbricatore *et al* 2011, Sorbi *et al* 2013), a built-in collaboration with GSI for FAIR-SIS300, has a faster ramp rate, at about 0.6–1 T s^{-1}, but requires direct cooling, with forced flow supercritical Helium.

The main challenge of the SIG prototype is the strong curvature, of 1.65 m. For comparison, DISCORAP has a 65 m bending radius, with the other projects featuring almost straight magnets.

The HIMAC gantry by Toshiba (Iwata *et al* 2012) is based on dipoles of 2.9 and 2.3 T, with a particular shape (curved race-track coils), with curvature radius of about 3 m. To go beyond this achievement of HIMAC the route of accelerator magnets, of tubular shape and rather low weight, seems the most promising one: with SIG 4 T dipoles, and a new mechanical structure under study at CERN, the weight of the gantry can be of 50–80 tons only.

To conclude this overview, we refer to the plot of figure 5.47, where magnets of different projects are reported in a chart where the squared product of field times the magnet aperture (this parameter gives the class of the magnet since it is proportional

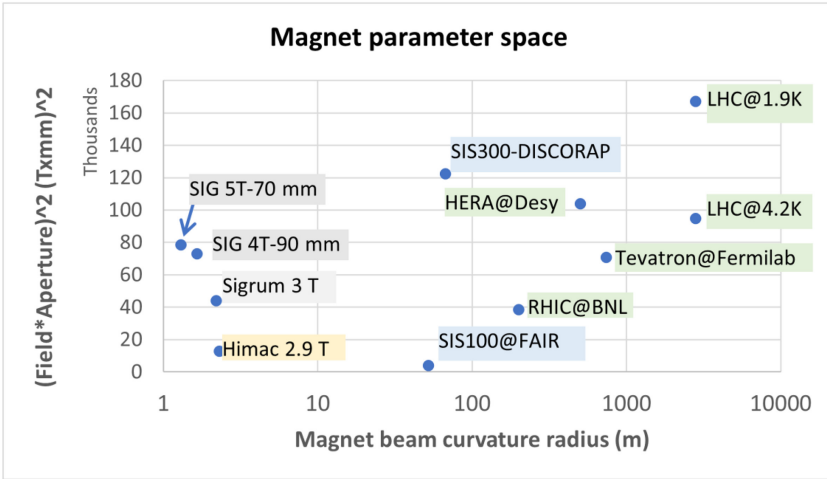


Figure 5.47. Collections of various past and future (SIG and Sigrum) magnets in a special parameter space (see text for details).

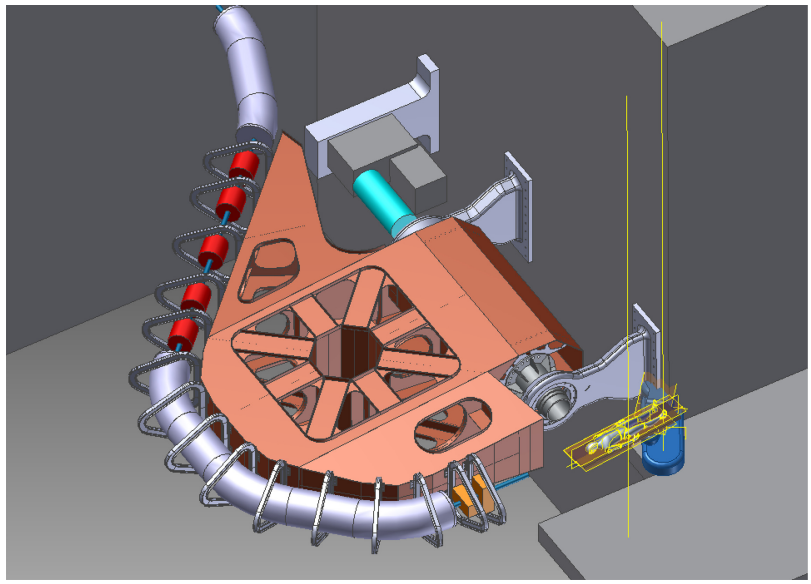


Figure 5.48. Artistic view of SIGRUM in operation (courtesy of M Pullia, CNAO foundation).

to the stored energy) is plotted versus the bending radius. Here the breakthrough made by the HIMAC magnets toward smaller bending radius, and the further step required by the SIG project both in bending radius and in stored energy, is rather evident in this magnet panorama.

The SIG project, as part of the global SIGRUM collaboration, aims at testing the first prototype by 2025 and then to be able to proceed to the gantry construction. A possible sketch of the SIGRUM gantry installed in the CNAO facility is shown in figure 5.48.

5.6.3.6 Conclusion

Superconducting magnets have accompanied the progress of HEP colliders. The community is pursuing the R&D on new more advanced technology for next-generation magnets for 15 T or more, necessary for the post-High Luminosity LHC collider. However, the same community is exploring if the presently available accelerator magnet technology can make a key contribution to cancer therapy, by enabling light and less expensive ion gantries. If successful, this development would give a significant boost to the performance of present hadron therapy centers based on heavy ions.

Acknowledgements

The author is grateful to the many collaborators in the LHC and High luminosity LHC times and, for the discussion on magnet design to, Luca Bottura, Mikko Karppinen, Glyn Kirby, Diego Perini, Davide Tommasini, and Ezio Todesco of the

CERN Technology department. He wants to thank also Gabriele Ceruti, Ernesto De Matteis, Samuele Mariotto, Marco Prioli, Massimo Sorbi, Marco Statera, and Riccardo Valente of the LASA Laboratory (University of Milano and INFN-Milano) for their contributions to the gantry magnet design. This work is partly supported by the European Commission via the H2020-HITRIPlus grant no. 101008548 and H2020-I.FAST grant no. 101004730, and by INFN via the grant CSN5-Call2021-SIG.

References

Abou Allaban A, Wang W and Padr T 2020a A systematic review of robotics research in support of in-home care for older adults *Information* **11** 75

Acemoglu A, Kriegstein J and Caldwell D G *et al* 2020b 5G robotic telesurgery: remote transoral laser microsurgeries on a Cadaver *IEEE Trans. Med. Robot. Bionics* **2** 511–8

Acemoglu A, Peretti G, Hysenbelli J, Trimarchi M and Caldwell D G *et al* 2020a Operating from a distance: the first 5G telesurgery experience *Ann. Intern. Med.* **173** 940–1

Adams R M, Mora T, Walczak A M and Kinney J B 2016 Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves *eLife* **5** e23156

Agostinis P *et al* 2011 Photodynamic therapy of cancer: an update *CA. A Cancer J. Clin.* 250–81

Aleta A *et al* 2020 Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19 *Nat. Human Behav.* **4** 964–71

Althouse B M *et al* 2020 Superspreading events in the transmission dynamics of SARS-CoV-2: opportunities for interventions and control *PLoS Biol.* **18** e3000897

Amaldi U *et al* 2021 Sigrum-A Superconducting Ion Gantry with Riboni's Unconventional Mechanics. No. CERN-ACC-NOTE-2021-0014

Anderson B, Rasch M, Hochlin K, Jensen F H, Wismar P and Fredriksen J E 2006 Decontamination of rooms, medical equipment and ambulances using an aerosol of hydrogen peroxide disinfectant *J. Hosp. Infect.* **62** 149–55

Antoniou S A, Antoniou G A, Antoniou A I and Granderath F A 2015 Past, present, and future of minimally invasive abdominal surgery *JSLS* **19** e2015.00052

Arenas A *et al* 2020 Modeling the spatiotemporal epidemic spreading of COVID-19 and the impact of mobility and social distancing interventions *Phys. Rev.* X **10** 041055

Ašner F M 1999 *High Field Superconducting Magnets* (Oxford: Oxford University Press)

Barbosa H *et al* 2018 Human mobility: models and applications *Phys. Rep.* **734** 1–74

Baronchelli F, Zucchella C, Serrao I D and Bartolo M 2021 The effect of robotic assisted gait training with Lokomat® on balance control after stroke: systematic review and meta-analysis *Front. Neurol.*

Barrat A *et al* 2013 Empirical temporal networks of face-to-face human interactions *Eur. Phys. J. Spec. Top.* **222** 1295–309

Beal A, Mahida N, Staniforth K, Vaughan N, Clarke M and Boswell T 2016 First UK trial of Xenex PX-UV, an automated ultraviolet room decontamination device in a clinical haematology and bone marrow transplantation unit *J. Hosp. Infect.* **93** 164–8

Beccani M, Aiello G, Gkotsis N and Tunc H *et al* 2016 Component based design of a drug delivery capsule robot *Sens. Actuators* A **245** 180–8

Best J 2021 Wearable technology: COVID-19 and the rise of remote clinical monitoring *Brit. Med. J.* **372** n413

Bloss R 2011 Mobile hospital robots cure numerous logistic needs *Ind. Robot* **38** 567–71

Bottura L, de Rijk G, Rossi L and Todesco E 2012 Advanced accelerator magnets for upgrading the LHC *IEEE Trans. Appl. Supercond.* **22** 8

Bottura L and Rossi L 2015 Magnets for particle accelerators *Applied Superconductivity—Handbook on Devices and Applications* ed P Seidel (New York: Wiley-VCH) pp 448–86

Brändén G and Neutze R 2021 Advances and challenges in time-resolved macromolecular crystallography *Science* **373** eaba0954

Brouwer L *et al* 2020 Design and test of a curved superconducting dipole magnet for proton therapy *Nucl. Instrum. Meth. Phys. Res. Sect. A, Accel., Spect., Detect. Assoc. Equip.* **957** 163414

Brüning O and Rossi L 2015 *The High Luminosity Large Hadron Collider Advanced Series on Direction in High Energy Physics* (Singapore: World Scientific) vol 24

Brüning O and Rossi L 2019 The high-luminosity large hadron collider *Nat. Rev.—Phys.* **1** 241–3

Bubar K M *et al* 2021 Model-informed COVID-19 vaccine prioritization strategies by age and serostatus *Science* **371** 916–21

Buettner R, Wannenwetsch K and Loskan D 2020 A systematic literature review of computer support for surgical interventions *2020 IEEE 44th Annual Computers, Software, and Applications Conf. (COMPSAC)* pp 729–34

Burley S K *et al* 2022 A comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students *Biomolecules* **12** 1425

Bush K *et al* 2011 Tackling antibiotic resistance *Nat. Rev. Microbiol.* **9** 894–6

Čaić M, Mahr D and Oderkerken-Schröder G 2019 Value of social robots in services: social cognition perspective *J. Serv. Mark.* **33** 463–78

Cao W, Song W, Li X, Zheng S, Zhang G and Wu Y *et al* 2019 Interaction with social robots: improving gaze toward face but not necessarily joint attention in children with autism spectrum disorder *Front. Psychol.* **10** 1503

Carling P C and Huang S S 2013 Improving healthcare environmental cleaning and disinfection current and evolving issues *Infect. Control Hosp. Epidemiol.* **34** 507–13

Carr J H and Shepherd R B 1987 *A Motor Relearning Programme for Stroke* (Oxford: Butterworth Heinemann)

Castro M, Ares S, Cuesta J A and Manrubia S 2020 The turning point and end of an expanding epidemic cannot be precisely forecast *Proc. Natl Acad. Sci.* **117** 26190–6

Chakraborty A K and Barton J P 2017 Rational design of vaccine targets and strategies for HIV: a crossroad of statistical physics, biology, and medicine *Rep. Prog. Phys.* **80** 032601

Chang M C and Boudier-Revéret M 2020 Usefulness of telerehabilitation for stroke patients during the COVID-19 pandemic *Am. J. Phys. Med. Rehabil.* **99** 582

Chapman H N 2019 X-ray free-electron lasers for the structure and dynamics of macromolecules *Annu. Rev. Biochem.* **88** 35–58

Charron P M, Kirby R L and MacLeod D A 1995 Epidemiology of walker-related injuries and deaths in the United States *Am. J. Phys. Med. Rehabil.* **74** 237–9

Chinazzi M *et al* 2020 The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) *Science* **368** 395–400

Choi P J, Oskouian R J and Tubbs R S 2018 Telesurgery: past, present, and future *Cureus* **10** e2716

Chowell G 2017 Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts *Infect. Dis. Model.* **2** 379–98

Chowell G, Vespignani A, Viboud C and Simonsen L 2018 The RAPIDD Ebola Forecasting Challenge, Special Issue *Epidemics* **22** 1

CLIC Project n.d. (http://clic-study.web.cern.ch/clic-study/)

CNAO n.d. Fondazione CNAO, Pavia, Italy (https://fondazionecnao.it/)

Colizza V, Barrat A, Barthélemy M and Vespignani A 2006 The role of the airline transportation network in the prediction and predictability of global epidemics *Proc. Natl Acad. Sci.* **103** 2015–20

Dabiri F, Massey T, Noshadi H and Hagopian H *et al* 2009 A telehealth architecture for networked embedded systems: a case study in *in vivo* health monitoring *IEEE Trans. Inf. Technol. Biomed.* **13** 351–9

Danon L *et al* 2011 Networks and the epidemiology of infectious disease *Interdisciplinary Perspectives on Infectious Diseases* **2011** e284909

Davies B 2000 A review of robotics in surgery *Proc. Inst. Mech. Eng.* **214** 129–40

Davies N G *et al* 2020 Age-dependent effects in the transmission and control of COVID-19 epidemics *Nat. Med.* **26** 1205–11

Dawe J, Sutherland C, Barco A and Broadbent E 2019 Can social robots help children in healthcare contexts? A scoping review *BMJ Paediatr. Open.* **3** e000371

de Boer J F, Leitgeb R and Wojtkowski M 2017 Twenty-five years of optical coherence tomography: the paradigm shift in sensitivity and speed provided by Fourier domai—OCT *Biomed. Opt. Express* **8** 3248–80

de Graaf M M A, Ben Allouch S and van Dijk J A G M 2015 What makes robots social? A user's perspective on characteristics for social human–robot interaction *Social Robot* ed A Tapus, E André, J-C Martin, F Ferland and M Ammi (Cham: Springer International Publishing)

De Marchi F, Contaldi E, Magistrelli L and Cantello R *et al* 2021 Telehealth in neuro-degenerative diseases: opportunities and challenges for patients and physicians *Brain Sci.* **11** 237

De Michieli L, Mattos L S, Caldwell D G, Metta G and Cingolani R 2020 Technology and telemedicine *Encyclopedia of Gerontology and Population Aging* ed D Gu and M E Dupre (Cham: Springer International Publishing)

Degiovanni O A and Amaldi U 2015 History of Hadron therapy accelerators *Physica Med.* **31** 322–32

Desai M M and Fisher D S 2007 Beneficial mutation–selection balance and the effect of linkage on positive selection *Genetics* **176** 1759–98

Di Natali C and Poliero *et al* 2019 Design and evaluation of a soft assistive lower limb exoskeleton *Robotica* **37** 2014–34

Di Natali C, Sadeghi A and Mondini A *et al* 2020 Pneumatic quasi-passive actuation for soft assistive lower limbs exoskeleton *Front. Neurorobot.* **14** 31

Ding X and Clifton D *et al* 2020 Wearable sensing and telehealth technology with potential applications in the coronavirus pandemic *IEEE Rev. Biomed. Eng.* **14** 48–70

D. O. E. United NationsAffairs S 2017 World population ageing 2017 highlights (Accessed 11 March 2021)

Eames K, Bansal S, Frost S and Riley S 2015 Six challenges in measuring contact networks for use in modelling *Epidemics* **10** 72–7

Elhanati Y, Sethna Z, Callan C G, Mora T and Walczak A M 2018 Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination *Immunol. Rev.* **284** 167–79

Evans L 2018 *The Large Hadron Collider: A Marvel of Technology* (EPFL Press) 2nd edn

Fabbricatore P *et al* 2011 The construction of the model of the curved fast ramped superconducting dipole for FAIR SIS300 synchrotron *IEEE Trans. Appl. Supercond.* **21** 1863–7 Part: 2 Published: Jun

Faus-Golfe A 2020 Design of a compact 140 MeV electron linear accelerator *ARIES D3.4* (https://edms.cern.ch/ui/file/1817163/1.0/ARIES-Del-D3.4-Final.pdf)

Ferretti L *et al* 2020 Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing *Science*

Fiedler F 2008 *PhD Thesis* FZD Rossendorf/TU Dresden

Fujimoto J and Swanson E 2016 The development, commercialization, and impact of optical coherence tomography *Invest—Ophthalmol. Vis. Sci.* **57** OCT1–OCT13

Funk S, Salathé M and Jansen V A A 2010 Modelling the influence of human behaviour on the spread of infectious diseases: a review *J. R. Soc. Interface* **7** 1247–56

Gallotti R, Valle F, Castaldo N, Sacco P and De Domenico M 2020 Assessing the risks of 'infodemics' in response to COVID-19 epidemics *Nat. Hum. Behav.* **4** 1285–93

Gatto M *et al* 2020 Spread and dynamics of the COVID-19 epidemic in Italy: effects of emergency containment measures *Proc. Natl Acad. Sci.* **117** 10484–91

Gérard A *et al* 2020 High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics *Nat. Biotechnol.* **38** 715–21

Gorpas D *et al* 2011 Autofluorescence lifetime augmented reality as a means for real-time robotic surgery guidance in human patients *Sci. Rep.* **9** 1187

Gostic K M *et al* 2020 Practical considerations for measuring the effective reproductive number, Rt *PLoS Comput. Biol.* **16** e1008409

Guimerà R, Mossa S, Turtschi A and Amaral L a N 2005 The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles *Proc. Natl Acad. Sci.* **102** 7794–9

Haberer T *et al* 2004 The Heidelberg ion therapy center *Radiother. Oncol.* **73** S186–90

Henschel A, Laban G and Cross E S 2021 What makes a robot social? A review of social robots from science fiction to a home or hospital near you *Curr Robot Rep.* **2** 9–19

HITRIplus n.d. *Heavy Ion Therapy Research Integration* (https://hitriplus.eu/)

Hosseinizadeh O A, Breckwoldt N, Fung R, Sepehr R, Schmidt M, Schwander P, Santra R and Ourmazd A 2021 Few-fs resolution of a photoactive protein traversing a conical intersection *Nature* **599** 697–701

Huang P-S, Boyken S E and Baker D 2016 The coming of age of de novo protein design *Nature* **537** 320–7

Hull J R, Wilson M N, Bottura L, Rossi L, Green M A, Iwasa Y, Hahn S, Duchateau J and Kalsi S S 2015 *Superconducting Magnets Applied Superconductivity—Handbook on Devices and Applications* (Germany: Wiley- VCH Verlag Gmbh & Co. KGa) ch 4

Huysamen K, Bosch T, de Looze M, Stadler K S, Graf E and O'Sullivan L W 2018 Evaluation of a passive exoskeleton for static upper limb activities *Appl. Ergon.* **70** 148–55

I.FAST n.d. Innovation Fostering in Accelerator Science and Technology (https://ifast-project.eu/home)

Iwata Y *et al* 2012 Design of a superconducting rotating gantry for heavy-ion therapy *Phys. Rev. Spec. Top. PRST-AB. Accel. Beams.* **15** 044701

Jaeschke E, Khan S, Schneider J R and Hastings J B 2020 *Synchrotron Light Sources and Free-Electron Lasers* (Berlin: Springer)

Jaskolski M, Dauter Z and Wlodawer A 2014 A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits *FEBS J.* **281** 3985–4009

Johanson D L, Ho S A, Sutherland C J, Brown B, MacDonald B A and Jong Y L *et al* 2020 Smiling and use of first-name by a healthcare receptionist robot: effects on user perceptions, attitudes, and behaviours *Paladyn J. Behav. Robot.* **11** 40–51

Jumper J *et al* 2021 Highly accurate protein structure prediction with alphafold *Nature* **596** 583–9

Karppinen M, Ferrantino V, Kokkinos C and Ravaioli E 2022 Design of a curved superconducting combined function bending magnet demonstrator for hadron therapy *IEEE Trans. Appl. Supercond.* **32** 1–5

Kermack W O, McKendrick A G and Walker G T 1927 A contribution to the mathematical theory of epidemics *Proc. R. Soc.* A **115** 700–21

Kirby G *et al* 2020 Assembly and test of the HL-LHC twin aperture orbit corrector based on canted cos-theta design *J. Phys.: Conf. Ser.* **1559** 012070

Kiss I Z, Miller J and Simon P L 2017 *Mathematics of Epidemics on Networks: From Exact to Approximate Models* (Cham: Springer International Publishing)

Kivelä M *et al* 2014 Multilayer networks *J. Complex Netw.* **2** 203–71

Krafft C, von Eggeling F, Guntinas-Lichius O, Hartmann A, Waldner M J, Neurath M F and Popp J 2018 Perspectives, potentials and trends of *ex vivo* and *in vivo* optical molecular pathology *J. Biophotonics* **11** e201700236

Kühlbrandt W 2014 Cryo-EM enters a new era *eLife* **3** e03678

Kwoh Y S, Hou J and Jonckheere E A *et al* 1988 A robot with improved absolute positioning accuracy for CT guided stereotactic brain surgery *IEEE Trans. Biomed. Eng.* **35** 53–161

Laffranchi M and D'Angella S *et al* 2021 User-centered design and development of the modular TWIN lower limb exoskeleton *Front. Neurorobot.* **15**

Lässig M, Mustonen V and Walczak A M 2017 Predicting evolution *Nat Ecol Evol* **1** 1–9

Last update in the search for the Higgs boson 2012 Seminar held at CERN, 4 July (https://indico.cern.ch/conferenceDisplay.py?confId=197461)

Leochico C F D 2020 Adoption of telerehabilitation in a developing country before and during the COVID-19 pandemic *Ann. Phys. Rehabil. Med.* **63** 563–4

LHC DESIGN Report 2004 Vol. I The LHC Main Ring Design Report, CERN Report, CERN-2004-003, 4 June

Li J, Esteban-Fernández de Ávila B, Gao W, Zhang L and Wang J 2017 Micro/nanorobots for biomedicine: delivery, surgery, sensing, and detoxification *Sci. Robot.* **2** eaam6431

López J A M *et al* 2021 Anatomy of digital contact tracing: role of age, transmission setting, adoption and case detection *Sci. Adv.* **7** eabd8750

Łuksza M and Lässig M 2014 A predictive fitness model for influenza *Nature* **507** 57–61

Masuda N and Holme P 2013 Predicting and controlling infectious disease epidemics using temporal networks *F1000Prime Reports* **5** 6

Matanfack Azemtsop G, Rüger J, Stiebing C, Schmitt M and Popp J 2020 Imaging the invisible—bioorthogonal Raman probes for imaging of cells and tissues *J. Biophotonics* **13** e202000129

Mattos L S, Caldwell D G, Peretti G, Mora F, Guastini L and Cingolani R 2016 Microsurgery robots: addressing the needs of high-precision surgical interventions *Swiss Med. Wkly.* **146** w14375

MedAustron n.d. *The Center for Ion Therapy and Research, Wiener Neustadt, Austry* (https://medaustron.at/en)

Mess K-H, Schmüser P and Wolff S 1996 *Superconducting Accelerator Magnets* (Singapore: World Scientific)

Meyer T, Schmitt M, Guntinas-Lichius O and Popp J 2019 Toward an all-optical bi—sy *Opt. Photonics News* **30** 26–33

Minervina A A *et al* 2021 Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection *eLife* **10** e63502

Moses H, Matheson D H, Dorsey E R, George E R, Sadoff D and Yoshimura S 2013 The anatomy of health care in the United States *JAMA* **310** 1947–63

Murugan A, Mora T, Walczak A M and Callan C G 2012 Statistical inference of the generation probability of T-cell receptors from sequence repertoires *Proc. Natl Acad. Sci.* **109** 16161–6

Neher R A, Russell C A and Shraiman B I 2014 Predicting evolution from the shape of genealogical trees *eLife* **3** e03568

Nelson B J, Kaliakatsos I K and Abbot J J 2010 Microrobots for minimally invasive medicine *Annu. Rev. Biomed. Eng.* **12** 55–85

Neutze R, Wouts R, van der Spoel D, Weckert E and Hajdu J 2000 Potential for biomolecular imaging with femtosecond x-ray pulses *Nature* **406** 753–7

Nourmohammad A, Otwinowski J and Plotkin J B 2016 Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections *PLoS Genet.* **12** e1006171

Oliver N *et al* 2020 Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle *Sci. Adv.* **6** eabc0764

OMRON 2022 Automating UV Disinfection Process Safely and Wisely (https://web.omron-ap.com/th/ld-uvc) Accessed 10 January 2022

Organisation for Economic Co-operation and Development (OECD) 2009 *Health at a Glance 2009: OECD Indicators* (Paris: OECD Publishing)

Organisation for Economic Co-operation and Development (OECD) 2020 *Historical Population Data and Projections (1950–2050)* (https://stats.oecd.org/) (Accessed 1 Aug 2020)

Osibona O, Solomon B D and Fecht D 2021 Lighting in the home and health: a systematic review *Int. J. Environ. Res. Public Health* **18** 609

Pandey A K and Gelin R 2018 A mass-produced sociable humanoid robot: pepper: the first machine of its kind *IEEE Robot. Autom. Mag.* **25** 40–8

Parker V M, Wade D T and Langton H R 1986 Loss of arm function after stroke: measurement, frequency, and recovery *Int. Rehab. Med.* **8** 69–73

Pastor-Satorras R, Castellano C, Van Mieghem P and Vespignani A 2015 Epidemic processes in complex networks *Rev. Mod. Phys.* **87** 925–79

Perelson A S, Neumann A U, Markowitz M, Leonard J M and Ho D D 1996 HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time *Science* **271** 1582–6

Perelson A S and Weisbuch G 1997 Immunology for physicists *Rev. Mod. Phys.* **69** 1219–68

Perra N 2021 Non-pharmaceutical interventions during the COVID-19 pandemic: a review *Phys. Rep.* **913** 1–52

Poletto C, Meloni S, Colizza V, Moreno Y and Vespignani A 2013 Host mobility drives pathogen competition in spatially structured populations *PLoS Comput. Biol.* **9** e1003169

Poliero T, Mancini L, Caldwell D G and Ortiz J 2019 Enhancing backsupport exoskeleton versatility based on human activity recognition *2019 Wearable Robotics Association Conf. (WearRAcon)* (IEEE) pp 86–91

Pons J L 2008 *Wearable Robots: Biomechatronic Exoskeletons* (New York: Wiley)

Popp J (ed) 2014 *Ex-vivo and In-vivo Optical Molecular Pathology* (Weinheim: Wiley-VCH)

Popp J and Bauer M (ed) 2015 *Modern techniques for Pathogen Detection* (Weinheim: Wiley-VCH)

Popp J and Strehle M (ed) 2006 *Biophotonics* (Weinheim: Wiley-VCH)

Popp J, Tuchin V V, Chiou A and Heinemann S H (ed) 2011 *Handbook of Biophotonics* **vol 1–3** (Weinheim: Wiley-VCH)

Priplata A A, Niemi J B, Harry J D, Lipsitz L A and Collins J J 2003 Vibrating insoles and balance control in elderly people *Lancet* **362** 1123–4

Prvu Bettger J and Resnik L J 2020 Tele-rehabilitation in the age of COVID-19: an opportunity for learning health system research *Phys. Ther.* **100** 1913–6

Pullano G *et al* 2020 Underdetection of COVID-19 cases in France threatens epidemic control *Nature* **590** 1–9

Ramalingam B, Yin J, Rajesh Elara M, Tamilselvam Y K, Mohan Rayguru M, Muthugala M and Félix Gómez B 2020 A human support robot for the cleaning and maintenance of door handles using a deep-learning framework *Sensors* **20** 3543

Richardson C A, Glynn N W, Ferrucci L G and Mackey D C 2015 Walking energetics, fatigability, and fatigue in older adults: the study of energy and aging pilot *J. Gerontol. Ser. A: Biomed. Sci. Med. Sci.* **70** 487–94

Robinson C 1995 Rehabilitation engineering, science and technology ed J Bronzino *Biomedical Engineering Handbook* (Boca Raton, FL: CRC Press) pp 2045–54

Rossi L 2010 Superconductivity: its role, its success and its setbacks in the Large Hadron Collider of CERN *Supercond. Sci. Technol.* **23** 034001

Rossi L *et al* 2022a Preliminary study of 4 T superconducting dipole for a light rotating gantry for ion-therapy *IEEE-Trans. Appl. Supercon.* **32** 4400506

Rossi L *et al* 2022b A European collaboration to investigate superconducting magnets for next generation heavy ion therapy *IEEE-Trans. Appl. Supercon.* **32** 4400207

Rossi L and Bottura L 2012 Superconducting magnets for particle accelerators *Rev. Accel. Sci. Technol.* **5** 51–89

Rossi L and Bruning O 2016 The LHC upgrade plan and technology challenges *Challenges and Goals for ACCELERATORS in the XXI Century* ed O Bruning and S Myers (Singapore: World Scientific Publisher) pp 467–97

Rossi L and Brüning O 2019 *Progress with the High Luminosity LHC Project at Cern Proceedings of IPAC2019 (Melbourne, 19–24 May)* (Jacow Publisher) pp 17–22

Rossi L and Senatore C 2021 HTS accelerator magnet and conductor development in Europe *Instruments* **5** 33

Rossi L and Todesco E 2006 Electromagnetic design of superconducting quadrupoles *Phys. Rev. Spec. Top. Accel Beams* **9** 102401 1–20

Rossi L and Todesco E 2007 Electromagnetic design of superconducting dipoles based on sector coils *Phys. Rev. Spec. Top. Accel. Beams* **10** 112401

Rossi L and Todesco E 2011 Conceptual design of the 20 T dipoles for high-energy LHC *Proc. of the High-Energy Large Hadron Collider Workshop (Malta, Oct. 2010)* ; E Todesco and F Zimmermann CERN-2011-003 (8 April 2011) pp 13–9

Saglia J A *et al* 2019 Design and development of a novel core, balance and lower limb rehabilitation robot: Hunova® *2019 IEEE 16th Int. Conf. on Rehabilitation Robotics (ICORR) (2019)* pp 417–22

Saglia J A, Tsagarakis N, Dai J S and Caldwell D G 2013 Control strategies for patient-assisted training using the ankle rehabilitation robot (ARBOT) *IEEE/ASME Trans. Mechatron.* **18** 1799–808

Satava R 2002 Surgical robotics: the early chronicles *Surg. Laparosc., Endosc. Percutan. Techn.* **12** 6–16

Schlichting I 2015 Serial femtosecond crystallography: the first five years *IUCrJ* **2** 246–55

Schlosser F *et al* 2020 COVID-19 lockdown induces disease-mitigating structural changes in mobility networks *Proc. Natl Acad. Sci.* **117** 32883–90

Sorbi M *et al* 2013 Measurements and analysis of the SIS-300 dipole prototype during the functional test at LASA *IEEE Trans. Appl. Supercond.* **23** 400205

Soto F, Wang J, Ahmed R and Demirci U 2020 Medical micro/nanorobots in precision medicine *Adv. Sci.* **7** 2002203

Spence J C H 2017 XFELs for structure and dynamics in biology *IUCrJ* **4** 322–39

Stahl B C and Coeckelbergh M 2016 Ethics of healthcare robotics: towards responsible research and innovation *Robot. Auton. Syst.* **86** 152–61

Starr T N *et al* 2020 Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding *Cell* **182** 1295–1310.e20

Stower R 2019 The role of trust and social behaviours in children's learning from social robots *8th Int. Conf. Affect. Comput. Intell. Interact. Work Demos (Cambridge, UK, 2019)* pp 1–5

Tannert A, Grohs R, Popp J and Neugebauer U 2019 Phenotypic antibiotic susceptibility testing of pathogenic bacteria using photonic readout methods: recent achievements and impact *Appl. Microbiol. Biotechnol.* **103** 549–66

Tollestrup A and Todesco E 2008 *Review of Accelerator Science and Technology* ed A Chao and W Chou (Singapore: World Scientific) vol 11 pp 185–210

Toxiri S, Koopman A S, Lazzaroni M, Ortiz J, Power V, de Looze M P, O'Sullivan L and Caldwell D G 2018 Rationale, implementation and evaluation of assistive strategies for an active back-support exoskeleton *Front. Robot. AI* **5** 53

Toxiri S, Naf M B, Lazzaroni M, Fernandez J, Sposito M, Poliero T, Monica L, Anastasi S, Caldwell D G and Ortiz J 2019 Back-support exoskeletons for occupational use: an overview of technological advances and trends *IISE Trans. Occup. Ergon. Hum. Factors* **7** 237–49

Tsagarakis N G, Metta G, Sandini G and Vernon C D G 2007 iCub: the design and realization of an open humanoid platform for cognitive and neuroscience research *Adv. Robot* **21** 1151–75

Tsimring L S, Levine H and Kessler D A 1996 RNA virus evolution via a fitness-space model *Phys. Rev. Lett.* **76** 4440–3

Vanderborght B *et al* 2013 Variable impedance actuators: a review *Rob. Autom. Syst.* **61** 1601–14

Very High Energy 2020 Electron Radiotherapy Workshop (VHEE'2020) (https://indico.cern.ch/event/939012/)

Volz E *et al* 2021 Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England *Nature* 1–17

Wang S *et al* 2015 Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies *Cell* **160** 785–97

Wang X V and Wang L 2021 A literature survey of the robotic technologies during the COVID-19 pandemic *J. Manuf. Syst.* **60** 823–36

Wilson M N 1983 *Superconducting Magnets* (Oxford: Oxford University Press)

Wilson M N 1999 Superconductivity and accelerators: the good companions *IEEE Trans. Appl. Supercond.* **9** 111–21

World Health Organization (WHO) 2008 10 facts on health workforce crisis(http://who.int/features/factfiles/health_workforce/en/) Accessed 6 Aug 2020

Wu J T, Leung K and Leung G M 2020 Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study *Lancet* **395** 689–97

Wu Y H, Wrobel J, Cornuet M, Kerhervé H, Damnée S and Rigaud A S 2014 Acceptance of an assistive robot in older adults: a mixed-method study of human–robot interaction over a 1-month period in the Living Lab setting *Clin. Interv. Aging.* **8** 801–11

Yang G Z, Bellingham J and Dupont P *et al* 2018 The grand challenges of science robotics *Sci. Robot.* **3**

Zemmar A, Lozano A M and Nelson B J 2020 The rise of robots in surgical environments during COVID-19 *Nat. Mach. Intell.* **2** 566–72

Zhang S-Q *et al* 2018 High-throughput determination of the antigen specificities of T cell receptors in single cells *Nat. Biotechnol.* **36** 1156–9