EMBL 50 YEARS | 1974–2024

AI in biology and health: opportunities and challenges

High level roundtable discussion

Heidelberg, Germany

European Molecular Biology Laboratory
Laboratoire Européen de Biologie Moléculaire
Europäisches Laboratorium für Molekularbiologie

Life sciences continuously transform our understanding of the living world, with applications spanning human through to planetary health. The use of Artificial Intelligence (AI) is well-established within the life sciences in Europe, thanks, in large part, to the role that the European Molecular Biology Laboratory (EMBL) has played.

Molecules are the building blocks of life. Molecular life sciences help us to understand the living world, and address human and planetary health challenges. As Europe's only intergovernmental organisation dedicated to molecular biology research, and with EMBL's European Bioinformatics Institute (EMBL-EBI), EMBL scientists have, for decades, generated, integrated, stored and shared the high-quality, computable biological data which now underpins Europe's huge AI potential.

In the context of accelerated use of AI across all areas, there is an increasing need for informed discussions about the challenges and opportunities associated with AI in the life sciences. To this end, EMBL convened a roundtable with representatives from governments, industry, and academia identifying ways Europe can respond to and fully realise AI.

Discussion covered both short and long term applications of AI, and included the possible benefits of AI for life sciences, as well as aspects requiring thoughtful consideration - for example, the role of societal biases in health datasets. Development of successful AI requires four key components: (1) talented AI researchers, often working in an interdisciplinary team with computational-aware subject experts; (2) large scale computable data, with appropriate meta-data and formatting; (3) large scale computational infrastructure with appropriate hardware acceleration; and (4) robust governance and policy environments, enabling the use of responsible AI. Europe has, or has the necessary conditions for, all four components - with particular strengths in the first two. Discussion acknowledged that for Europe to respond to the advance of AI, two fundamental needs must be resourced: (1) capacity building, to ensure an adequate workforce, given the demand and presence of, appropriately architected compute; and (2) continued open access to high quality, diverse data.

Threaded throughout the discussion was reference to AlphaFold, an AI system developed by Google DeepMind which makes state-of-the-art, accurate 3D protein structure predictions - which previously took years - in minutes. This work built on decades of EMBL's data management expertise, using molecular data from public resources co-hosted by EMBL-EBI (and including data generated by EMBL scientists). AlphaFold illustrates a transformative application of AI within life sciences, with huge implications for both discovery-driven and early-stage human health research.

The roundtable focused on four key topics, listed below.

1. Computable Data

Life sciences have a long tradition of generating, enriching and using data. From Darwin and Humboldt's notes and drawings, to today's images painstakingly generated through processes such as x-ray crystallography, life sciences require open accessible data and information to progress the field. AI has only deepened this need, placing specific emphasis on computable data (that is, data which has appropriate meta-data and knowledge enrichment; and that is standards-based, structured and complete [with low or limited errors or omissions]), deployed in a scalable manner (e.g. in shared components). AlphaFold typifies the possibilities associated with the development of AI using open, accessible and computable data.

Beyond the kinds of molecular, non-patient data used by AlphaFold, one of the key questions posed by new AI methods is how healthcare information, including patient data, can be better connected to research, in a secure, controlled way. The use of AI will undoubtedly enable tremendous opportunities for improved health delivery. Not every application of AI requires access to sensitive patient data; however, where it does, this calls for safeguards to ensure privacy and other protections - such as approaches which securely and consensually facilitate data collection and access. A considered regulatory approach which supports this is paramount. This way, societal benefits from the exponentially increasing amounts of health and other data being generated can be realised. One possible avenue is the development of 'trusted environments' and platforms - regulatory and technical environments in which AI can be responsibly deployed. There is an opportunity to reform regulation to support the use of data for both healthcare and research as a standard solution.

In order to truly understand human wellbeing, we need to capture and combine complex information, to better understand interactions between the genome and the environment - for example, studying not only genetic predisposition to disease, but also the effect of the environment. One of the exciting potentials for AI is its increasing capacity to handle multi-modal data (different measurement types, such as genomics and imaging, or imaging and environmental measures), as well as high-dimensional datasets (that is, data with many features). To enable this, data must be computable internally (usable by AI within its domain of origin), and also be computable across modes (for example, via standardised meta-data and knowledge linkage across different datasets). Legislative and regulatory environments have a role to play in enabling these 'data-features' as well.

Most diseases arise from intersections between genomes and the environment. Connecting health data with environmental data could improve outcomes in both domains - for example, by mapping air pollution data against medical and public health research data, showing links between the environment and our wellbeing. For this, connectivity between data sets needs to be supported through investment in the infrastructure producing and storing this data, as well as the technology linking these sets of data together. Common ways of understanding data across different settings are also important, with a need for shared vocabularies and approaches.

2. Societal Engagement

AI is already being used across society, including in the life sciences. This includes in practical ways, like automating previously time-consuming tasks. However, the huge potential that AI brings is sometimes accompanied by confusion, misinterpretation and distrust. In part, this is because conversations about AI are frequently centred around Large Language Models (LLM), and often fail to distinguish between different AI types and their associated risks. For example, AI-based models have been used in medical applications and devices for over a decade, with hundreds obtaining regulatory clearances as 'SAAMD' (Software as a Medical Device). These models are trained and then perform specific tasks with high predictability; different from the large multi-modal models introduced in recent years.

Distrust in AI may limit an individual's willingness to share their personal data - health or otherwise - limiting the representativeness of the data AI is applied to and limiting applications of AI to specific domains . Scientific research and healthcare are areas where AI can have a large impact, with huge societal benefits[1]. To more accurately represent potential applications of AI, there is a need for case studies that can be easily understood at all levels of society. These might illustrate the impact AI is already having - for example, augmenting stroke research, or improving dementia detection[2]. Use studies can illustrate the established presence of AI, demystifying the technology and linking data-sharing to tangible benefits. One possible result of sharing these studies might be to encourage information-sharing at an individual level, to deliver population-level benefits. It is also important to set aside the 'hype' or media sensationalism associated with AI - separating long term, large scale applications from short term uses. This is because more nuanced language will enable a proportionate consideration of risk. Regulation of AI is being considered at multiple levels, including nationally and internationally[3]. Use studies can also encourage and enable a risk-based regulatory approach, as better differentiation between various types of AI can support the development of proportionate regulatory levers. Part of what makes AI and life sciences such a good fit is the 'puzzle-like' nature of the life sciences. Using the right terminology enables questions to be framed in a way that supports the most effective use of AI.

Relatedly, the roundtable also examined the issue of how to usefully define problems AI might be applied to - not only where AI is being used, but how we arrive at the problem itself. One way of conceiving of a problem is to work backwards from a model, figuring out which processes require interrogation. Another might be contemplating what the next big challenge for human and planetary health could be, and framing questions to address this challenge. Like use cases, unifying 'flagship' projects, such as AlphaFold have the potential to guide responsible AI development and reduce fragmentation across science.

There is a need for consideration of AI-related questions that aren't necessarily scalable or attractive to investment, but which may impact us in twenty years time. The challenge we face currently is that we must use AI in order to achieve gains now, and simultaneously consider ways

[1] https://www.who.int/publications/i/item/9789240029200
[2] https://www.nature.com/articles/s41467-022-31037-5
[3] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

to prepare for the future; for example, with regulatory frameworks or technical set-up to support longer-term uses and benefits of AI.

3. Training

A major challenge discussed in the roundtable was the insufficient AI workforce and lack of trained individuals needed to enable huge AI needs[4]. To respond to the advance of AI, Europe requires significant investment in training.

There is a huge disparity in the number and diversity of those training in AI and related fields, and the anticipated needs of the AI workforce. Already, there is an issue with the ability to fill high-level (and other) digitally oriented roles in Europe. The European Commission found that, in 2021, more than seventy percent of businesses cited a lack of staff with adequate digital skills as an investment obstacle[5]; an issue further compounded by a lack of specialist, advanced training in areas such as AI.

To address these and other shortfalls, consideration should be given to how AI is addressed in schools. This includes primary through to tertiary education levels, and also making AI training available to 'non-AI' professionals. Teachers and trainers should be upskilled in AI. Simultaneously, consideration should be given to entry points for groups outside of those traditionally attracted to the AI field; and how international talent might best be leveraged. However, Europe cannot rely only on attracting external resources, and must also build capacity internally.

While AI presents its own workforce challenges, it can also provide many workforce solutions. This includes in AI fields and other sectors, such as the health sector. AI can also enable innovative solutions to bridge capacity gaps - for example, using AI to train future AI-trainers.

AI could support productivity increases, enabling people to increase both the volume and types of work they can do, and even potentially address some skills shortages in health and other sectors. However, this may also require the reconfiguration of existing roles, to better incorporate the skills needed to effectively deploy AI - for example, in a health setting. One challenge is that training professionals and restructuring sectors to incorporate AI use is likely to encounter resistance - as has been the case with increasing the use of other digital health technologies.

In many ways, the current demand for AI-skills mirrors the same explosive demand bioinformatics experienced thirty years ago - including the need to scale alongside growth. One lesson to take from the field of bioinformatics is that any capacity building activities will need to simultaneously incorporate modes of collaborating. Europe has the advantage of being able to draw resources from many different settings - this necessitates consideration of international sharing and standards setting.

[4] https://www.oecd.org/publications/the-supply-demand-and-characteristics-of-the-ai-workforce-across-oecd-countries-bb17314a-en.htm
[5] https://ec.europa.eu/eurostat/en/web/products-eurostat-news/w/ddn-20230712-1

4. Maximising the impact of AI in biology and health for everyone

Future AI approaches in healthcare, science and beyond - including regulatory approaches - require careful consideration, especially in the contexts of diversity and equity.

Given the improved outcomes available through the application of AI to large-scale datasets, Europe's decades-long human data collection is a significant advantage. However, there is a need for caution in considering this data as wholly representative. Databases which more accurately reflect the diversity in population groups (be it data pertaining to human patients, or to microbiomes) will take time and resources to establish. As AI-enabling systems and frameworks - including those that support computable data curation - are being set up and further developed, diversity and access-equity must be central considerations. This will ensure inputs are as representative as possible, so that any AI-generated outputs are more widely applicable, can be shared equitably, and will ultimately benefit more people.

Decisions being made in Europe and beyond must be done in a way that best equips non-European countries to receive benefits generated. This necessitates working closely with non-European societies, to ensure these groups can receive, and benefit from, the impacts AI will generate. This includes investing in the skills needed to produce quality data, as well as appropriate governance structures. There is increasing consensus from regulatory agencies around the globe that a risk-based approach is one of the best ways to ensure AI-related benefits are shared. Countries commencing establishment of AI-related systems require particular attention and resourcing, given the greenfield opportunities, at national and community levels, to set up the systems and frameworks that are the necessary precursor to high-quality data - and to avoid the need for dysfunctional retrofitting in the future.

Trusted partnerships are required both to establish these systems, and to maximise and equitably distribute AI-derived benefits. Simply solving any life sciences 'puzzle' with AI won't generate impact; platforms and ecosystems which bring together AI, life sciences, industry and government are critical to maximising benefits.

Open Access is a key related amplifier[6]. The open sharing of data greatly accelerates research and discovery. The cumulative power of both trusted partnerships and Open Access is again exemplified in the AlphaFold project. Using AI to generate accurate, 3D protein structure predictions would never have been possible without the initial input of high-quality protein sequence datasets developed, maintained and made freely available by EMBL-EBI. In turn, sharing the outputs of the project - in line with Open Access - on a freely accessible database has magnified the initial impact of AlphaFold, as researchers and users employ these structure predictions in a myriad of ways. While openly sharing AI breakthroughs should be considered case by case - taking into account multiple considerations, including ethics and safety - in the right scenario, it can help to benefit diverse communities and address broader systemic barriers.

[6] https://www.embl.org/news/science/open-data-sharing-accelerates-covid-19-research

Conclusions

AI has the capacity to meaningfully enhance life sciences research, and create significant benefits on an unprecedented scale. Europe's realisation of this potential requires resourcing the generation and storage of high quality data; investing in capacity-building across Europe; and needs ongoing conversations between government, academia and industry. Key legislative and regulatory choices being made concerning AI, in Europe and across the world, are shaping the world's AI ecosystem. The decisions around these systems must be made intentionally, with an understanding their impact will be felt for decades. Research infrastructures such as EMBL will continue to serve as an important connector for these necessary conversations. EMBL stands ready to engage.