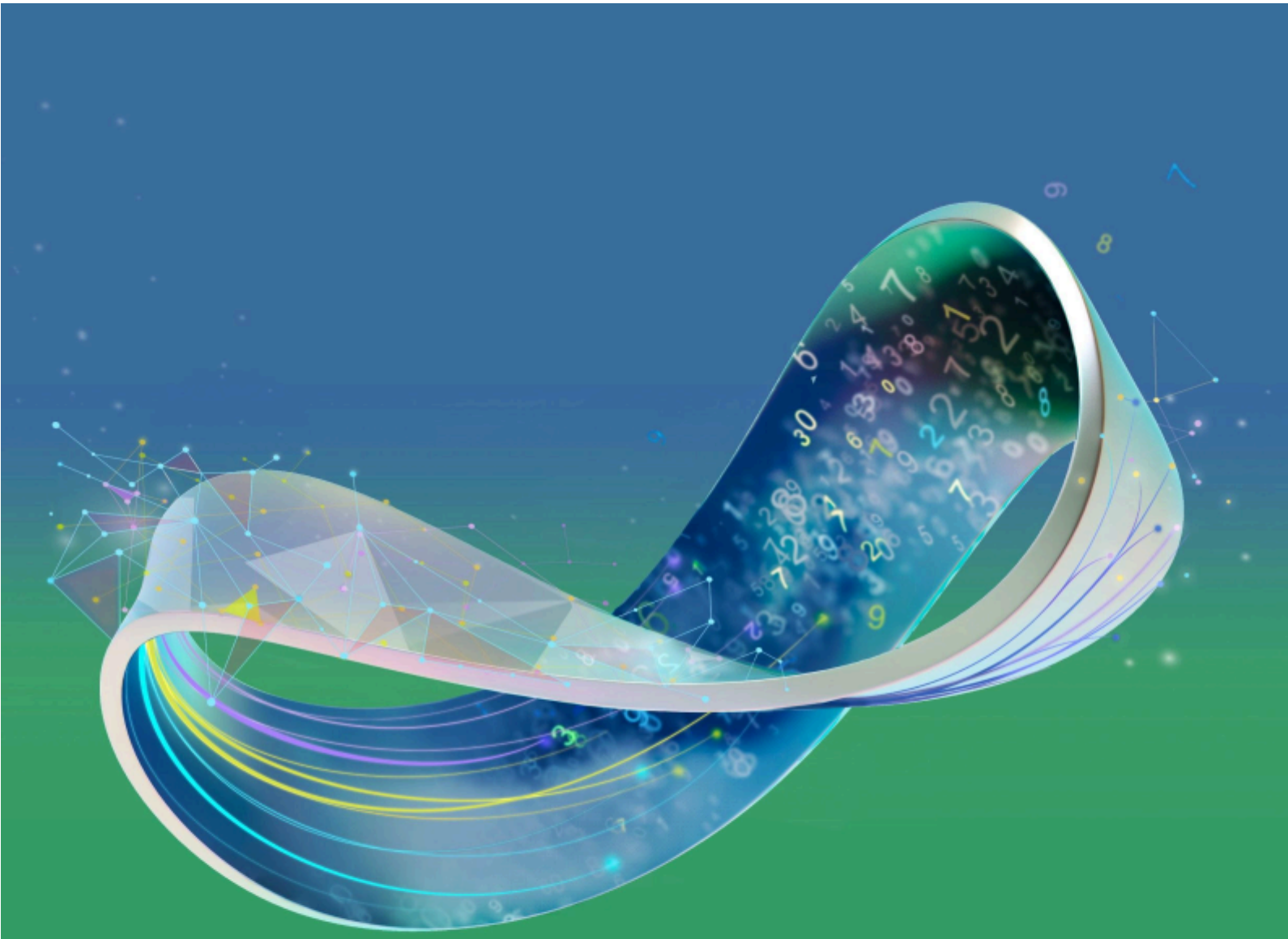
EMBL

EMBL Science AI Strategy

AI for scientific discovery



embl.org European Molecular Biology Laboratory

EMBL Science AI Strategy

2025

Table of Contents

Prepared by the EMBL AI Science Working Group:
Wolfgang Huber, Jan Korbel, Anna Kreshuk, Julio Saez-Rodriguez, Oliver Stegle (co-chair), Sameer Velankar, Jessica Vamathevan (co-chair)

1. Introduction

State of the field

Much like the technological inventions that powered the first industrial revolution, or like the internet, machine learning (ML) and advances in artificial intelligence (AI) constitute a revolution with transformative potential. This potential is particularly large in biology, and it has been argued that AI/ML will be pervasively used, even exceeding the crucial role currently played by bioinformatics and data analytics.

In the life sciences, we have observed a shift from using AI primarily as a tool for technical tasks, such as cell segmentation or identification of molecular signatures in gene expression profiles, to leveraging it for directly producing tangible and biologically interpretable outcomes. AlphaFold, the Nobel prize-winning breakthrough in AI, serves as a prime example, providing an end-to-end solution to a longstanding challenge in molecular biology. Other applications of AI include building foundation models for biomedical imaging, predicting cellular responses to external perturbations, identifying and interpreting pathogenic changes in genome sequences, and applying language models to analyse human phenotype data.

Such breakthroughs are heralding a new phase in biology, with AI/ML turning into the leading modelling discipline for complex biological systems across biological scales, spaces, and time.

DNA of early breakthroughs

Machine learning is a subfield of AI, but the terms are often used interchangeably. Here, we adopt a **broad definition of AI**, encompassing intelligent systems that learn from data and mimic cognitive functions associated with human intelligence. Machine learning uses algorithms to analyse large amounts of data and provide outputs, and, in some cases, they contain elements of autonomous decision-making. Evaluating past successes of AI in the life sciences highlights the importance of having a **well-defined problem statement** and a clear framework for **quantifying success**. The availability of **large, well-curated training datasets** and **appropriate metrics and benchmarks** using independent **validation data** has been essential (Figure 1). Some of the most notable successes in advancing the virtuous cycle of AI development and evaluation have been driven by global community efforts. For example, in structural biology, the Critical Assessment of Protein Structure Prediction (CASP) initiative has been fostering progress for decades. Increasingly, other communities are setting up similar evaluation mechanisms, such as the DREAM challenges for systems biology and medicine, the Open Problems framework in single-cell genomics, and many others.

Emergence of new stakeholders

The immense potential of AI in the life sciences has attracted a wide array of new contributors. This transformation is now being driven not only by traditional biology institutes but also increasingly by experts from “core AI” fields like computer science. This latter group of stakeholders includes both academic institutions and industry players, such as startups, which have rapidly become integral to the current AI innovation ecosystem. Moreover, the significant

hardware demands for training AI systems have positioned providers of high-performance computing systems, cloud hyperscalers, and GPU manufacturers as key stakeholders.

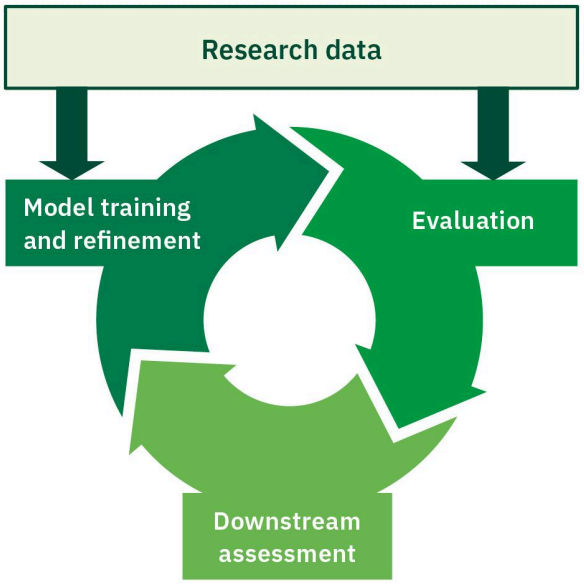


Figure 1: AI model evaluation and development cycle. Accessible research data underpins the continuous evolution and innovation cycle of AI models, combining model training, evaluation, and benchmarking of models, and the downstream application of these models in biology.

Related initiatives

Along with the rapid rise of technology companies that turn to life science questions, many academic institutions are investing in AI initiatives to fully leverage this promising technology to accelerate scientific discovery. These new initiatives are based within life science research institutions, like EMBL, or are centred in engineering and computer science departments (Table 1).

Purpose of this document

EMBL has a special responsibility to be a leader and innovator in European life sciences, not only in terms of scientific results but also through our scientific services, training, and how science is performed.

In this document, we outline EMBL's strategy for responding to the opportunities of AI technologies and the strategic pillars and enablers that will allow EMBL and the broader scientific community to leverage the full potential of AI technologies. This strategy will serve as a framework for EMBL to foster change in this new area, with AI as a major facilitator, to support everyone at EMBL to maximally exploit AI technology for scientific discovery, and to provide a recipe for building an ecosystem with the scientific community to make progress together in this field.

Table 1: Examples of AI life science institutions and initiatives.

Initiative/Institute	AI-related research focus
AI@HMMI Janelia, USA	Neuroscience and imaging technologies
AITHYRA - Research Institute for Biomedical Artificial Intelligence. Austria	AI and biomedical research
CZI, USA	Biomedical research
Helmholtz AI, Germany	Broad applications
IP.AI, Germany	Applied AI
Schmidt Center, Broad Institute, USA	ML + biology, human health
The Alan Turing Institute, UK	AI for health
VIB.AI, Belgium	Plant biology and human disease
Vector Institute, Canada	General AI applications

2. The Opportunity

EMBL's track record and unique strengths

EMBL already incorporates AI across all five missions and has a strong track record of impactful AI applications in the life sciences. EMBL's achievements include using and developing new AI-driven methods for extracting biological insights from imaging, structural biology, and multi-omics data; contributing to the training of AI models through EMBL-EBI's vast biological data resources (e.g. AlphaFold); offering multifaceted training to the scientific community; spinout creation; fostering new collaborations across Europe (e.g. ELLIS); and providing thought leadership.

As such, EMBL can build on its strong foundations to drive the next generation of AI challenges and opportunities, use AI to accelerate its research and services, and define the interface of AI and biology. Additionally, EMBL, as an international organisation and scientific leader, is highly trusted by the global scientific community and member states, and it has vast experience in organising and orchestrating research in life sciences across Europe and globally. These are important attributes for the global collaborations and multilateral partnerships this new arena requires.

Driving the next AI breakthroughs

EMBL's vision is to advance understanding of the molecular basis of life in context. To understand the fundamental principles of life and how it responds and adapts to dynamic environments, the ability to **model biological systems and integrate data across scales** – from molecular machines to cells, self-organising higher-level structures, tissues, organisms, and ecosystems – will be essential. Examples of AI research questions that are well-aligned with EMBL's priorities

include AI-based methods that **bridge machine learning and mechanistic modelling**, uncovering molecular mechanisms by integrating data across biological scales.

Building on the successful collaboration model that underpins AlphaFold, EMBL can act as an **enabler of future AI innovations** by the community at large. EMBL embraces the whole biological data life cycle, from data generation to data dissemination. Leveraging its position as a global hub for biodata, in combination with its expertise in large-scale data curation, processing, and management, EMBL will support academic or industry partners in future AI innovations. Beyond access to data, as well as computational and experimental technologies, impactful innovations will hinge on identifying the right **biological questions that are well-suited for AI approaches** – an area where EMBL has deep expertise. EMBL is thus uniquely positioned to evaluate AI methods within a biological context and to validate AI-based models through well-designed experiments and using data generated by cutting-edge profiling technologies. Finally, EMBL can provision metrics for AI benchmarking, knowledge of scientific processes, open science, and publication standards.

Pioneering lab-in-the-loop within EMBL science

There is considerable potential for incorporating AI into complex experimental processes in the life sciences, to enhance technology components, as well as to enable lab automation and high-throughput experimentation, analysis, and hypothesis testing. The ability to collect and create more informative data more efficiently will help to bring experimental data services to the next level, for example, by increasing the throughput and quality in structural biology, imaging, or genomics facilities.

AI can help open up challenging or currently infeasible experimental frontiers, such as effectively navigating the intractable space of combinatorial genetic perturbations or driving intelligent, self-learning image acquisition. While the concept of incorporating AI into experimental workflows (“lab-in-the-loop”) is widely recognised as a grand challenge of the future, only very few organisations are positioned to embed AI as deeply in the scientific process as EMBL. With EMBL’s rich history of research-driven technology development, EMBL can pioneer settings where AI can act as a co-pilot and decision-support system, pioneering lab-in-the-loop approaches at different time scales, biological scales, and autonomy levels.

Transforming the scientific process

With the rapid advancements in AI and its ability to assist or replicate human reasoning, it is evident that AI will not only influence data analysis and the validation of scientific hypotheses but also play a crucial role in generating new scientific hypotheses. **Future AI systems will change how knowledge is represented,** providing researchers with new tools to access the collective knowledge of all biological papers, high-throughput datasets, and molecular structures, via queryable interfaces that are capable of responding to scientists with scientifically useful results. EMBL needs to embrace these concepts, which can transform the current scientific process and be a driver of these developments, rather than a passenger.

3. Vision for AI at EMBL

The vision

The vision for AI at EMBL is to **foster a transformation** that will enable the organisation to exploit the full potential of AI-based approaches to advance scientific discovery. By embracing AI as a modelling tool to process biodata, enable technological innovations, and comprehend biological systems, EMBL will **advance biological frontiers and discover new scientific concepts.** AI will also accelerate the translation of fundamental science for **tackling societal challenges**, thereby maximising its broad impact.

To realise this vision and tackle this new interdisciplinary challenge, EMBL will forge a **cohesive union of the biological and AI communities** to advance work on the most important life science challenges that are amenable to AI-driven solutions. EMBL will create a place that brings together the AI and biological research communities, fostering **unique interdisciplinary bridges** and new means to engage scientists within different disciplines. Building on EMBL's track record in fostering interdisciplinary modes of working, EMBL will forge a distinct ecosystem where AI and biology synergise, diverse careers flourish, and labs of the future and reimagined industry engagement become a reality.

This vision (Figure 2) will enable EMBL to be a role model for leveraging AI in science and take a leading role in transforming the European life science community to fully exploit AI, contributing to Europe being a world leader in AI.

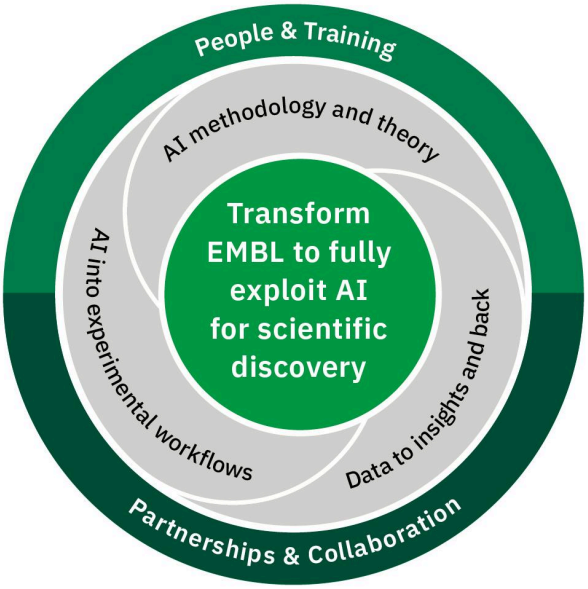


Figure 2: The AI at EMBL vision, strategic pillars and enablers.

Critical success factors

Agility & speed: Owing to the dynamics of AI innovations and current developments, agility and speed in establishing AI at EMBL is indispensable. EMBL will need to grow critical mass and create a protected space for a new community, establish alliances with external stakeholders, and develop a distinct profile in this arena in order to be competitive in the field and maintain its position as Europe's leading life science laboratory.

Purposeful integration with EMBL science: AI can and will impact all areas of EMBL's research and services, perhaps even more so than how we imagine it now. It is crucial that AI is synergistically integrated into EMBL's core scientific areas to harness its potential as a driving force for fundamental research, accelerating the discovery of molecular and mechanistic insights. The EMBL hyper-collaborative approach to science and built-in interdisciplinarity will be critical to ensuring this tight integration of AI within EMBL.

Talent and people: The competition for talent is steep, and institutions such as EMBL need to develop new academic career models to attract talent and compete with high-paying industry roles. Scientists from theoretical computer science backgrounds anticipate a different scientific environment and alternative metrics for measuring success. To be leading, EMBL needs to develop appropriate models to nurture this community, while deeply embedding it in EMBL's core communities and missions. In parallel, internal and external training activities will be a major focus to bring AI expertise to the right level at EMBL and within the broader scientific community.

Advancing interactions with industry: Industry is contributing to AI on multiple levels, from providing compute infrastructure to AI method innovation and their applications in the life sciences. New modes of collaboration with industry will be required, providing further opportunities for pre-competitive alliances, staff exchange programs, and scientific collaborations. EMBL needs to actively develop and continue innovative sustainable models for industry partnerships, for example, by building on existing highly successful setups such as Open Targets or the collaboration with Google DeepMind.

Access to large-scale computing resources: AI applications require new resources and capacity building that is distinct from established computational biology. The successful development and application of AI methods are heavily reliant on access to potent computational resources. A pragmatic and scalable compute strategy and stable partnerships combined with environmentally sustainable approaches will be required. AI innovations will integrate in-house resource development and national and international compute resources.

Data and domain knowledge are key: Access to large, curated, and computable data is essential for harnessing AI. Creating AI-ready data resources is a global effort and includes a large number of academic and commercial entities. Building on its position as the home to the largest biological data resources in the world and as a global leader in open-source bioinformatics software and data standards, along with its long-standing expertise in modern life sciences data generation technologies, EMBL will contribute strong domain knowledge and expertise. As an international organisation embedded in the European life science community, EMBL should also

take on the coordination of or be a pivotal player in the orchestration of future data resources required for AI.

4. Strategic Pillars

To optimally drive EMBL's science through AI, innovation across **EMBL's research**, **technology development, and experimental and data services** needs to be **tightly connected and coordinated**. EMBL's distinctive capacity to host research groups in proximity to experimental facilities for the creation and provision of instrumentation and software and its unique mission to provide open data resources promote a **virtuous circle unique to life science organisations**. AI will be tightly integrated into this circle, enhancing both its research and service via three pillars:

(i) development of AI methodology and theory;

(ii) application of AI to create insights from data; and,

(iii) integration of AI into experimental workflows ("lab-in-the-loop") and creation of new scientific hypotheses.

Pillar I: AI methodology and theory

AI research directions will be defined by scientific questions that are aligned with EMBL's core strengths in molecular biology, and will advance and accelerate existing directions to open up new frontiers. To enable EMBL to **build, adapt, and apply state-of-the-art AI models**, we will expand capacity in AI methodology. EMBL will invest in methodological and conceptual foundations, as well as establish approaches to rigorously benchmark the performance of AI models and characterise their capacity and limitations. AI systems need less of traditional "everything-thought-through" engineering, because they are designed to learn from data. However, there is an essential need for sound theoretical underpinnings of the scientific learning objectives and incentives/rewards encoded in such models, and the choice of metrics to benchmark alternative models. A **strong methodological foundation** will allow us not only to carefully design models for specific scientific objectives but also to inspect trained AI models and their outputs from a biological and mechanistic perspective.

The power of traditional theoretical approaches in biology using pencil-and-paper mathematics and computational simulation (e.g. differential equations, statistical mechanics) is well-known. AI models, in essence, are also mathematical functions, and as such they are already, and will be more in the future, modelling complex biological processes such as protein folding, regulation of gene expression, cellular signalling, tissue development, and ecosystem behaviour. AI in biology will benefit from such theoretical research strands, incorporating biophysical constraints and domain knowledge into the definition of AI models, for example. Conversely, new theoretical modelling approaches will facilitate the distilling of mechanistic insights and biological understanding from trained AI models. **EMBL aims to establish conceptual and mechanistic insights from AI-based models using theoretical approaches.**

Methodological foundations to build and adapt AI models

- **Foundation models across biological domains and modalities:** Conduct research in foundation models of biological knowledge and data, including tailored approaches for data modalities in biology, such as biological networks, cellular images, or tabular data. This will enable EMBL to develop or contribute to expressive and versatile foundation models trained on massive biological datasets across scales and technologies (see Pillar II).

- **Multi-scale modelling of biological systems:** Develop the most impactful biological questions that are tractable for AI and AI frameworks. Examples include new approaches to model biological systems across temporal and physical scales. These conceptual advances will underpin EMBL's goals to study biological systems across scales from molecules to ecosystems.

- **Causal representation learning:** Establish new modelling concepts that allow for harnessing instrumental variables and external (biological) perturbations to infer causal relationships and causally-informed representations from data. These approaches will help bridge the gap between correlations seen in observational data and hard causal evidence from an experimental or interventional study – from genetic perturbations to ecosystem interventions.

- **Frameworks for automated machine learning:** Evaluate and establish automated machine learning (AutoML) frameworks. This will enable EMBL to profit from advances in automation of machine learning, a key technology to address the bottleneck of AI experts, which also allows the blurring of boundaries between users and developers of AI.

- **Metrics & AI benchmarking:** Develop metrics and the concepts to benchmark and assess AI models. This will enable robust and scientifically sound benchmarking of AI models for specific scientific questions or tasks.

- **Support Open Science in AI:** Contribute to open-source software initiatives and projects that underpin or use AI, in particular, where this agrees with and advances the scientific method, enhances reproducibility, and promotes open research principles.

Theory and new concepts to connect AI to mechanism

- **Advance interpretable AI, use of prior knowledge, and formal mathematics/biophysics-based reasoning:** Develop methods and tools to inspect trained AI models for improved interpretation, robustness, and transparency. Solutions to this challenge will also be essential to encode biological domain knowledge and concepts into AI models.

- **Theory and principles of active learning and experimental design:** Advance theory in AI-guided experimental design and agent-based modelling. These concepts will fuel the transformation of data services and allow for the establishment of novel "lab-in-the-loop" concepts, whereby AI tools become co-pilots for experimental design decisions or act as agents in highly automated lab setups.

- **Theory of AI:** Develop solutions to deduce general principles from trained AI models, for example, by creating white-box interpretable approximations of black-box models. White-box representations of AI models will establish connections between AI and theory and formalise trade-offs between predictive capacity versus mechanistic interpretability of models.

- **Emergence:** The behaviour of a biological system is often not directly evident from the properties of its components alone. We aim to understand and predict emergent phenomena. This is a core aim of fundamental science and is equally essential to guiding interventions (e.g. in agriculture or medicine) or designing new (sub)systems (e.g. in immunology).

Pillar II: Data to insights and back

EMBL hosts some of the world's largest biodata resources, representing datasets across all major biological data modalities, from genomics to proteomics, structural biology, and imaging at all scales. EMBL will foster applications of AI in biology, both inside and outside of EMBL, leveraging these unique resources for future AI innovations. This will be achieved by coordinating and contributing to global partnerships to **develop new AI models and enable open science innovations.**

A prerequisite for such innovations is the ability to access these data in AI-ready standards, which includes accessing highly curated data or integrated high-quality datasets across modalities for AI training and benchmarking. By enhancing and adapting existing data resources and frameworks, EMBL will **serve high-quality, openly accessible training and benchmark datasets** across biological domains. In parallel, next-generation data services will be established to provide effective solutions for **sharing trained AI models and their outputs** with the community (Figure 3).

At the same time, AI will provide significant opportunities for enhancing EMBL data resources, delivering more powerful, user-friendly, and cost-efficient data services by building on EMBL's **track record in data curation and management.** Enhancement of EMBL data services to serve AI-ready data will also act as a means to establish new interactions and collaborations with external AI stakeholders, thereby accelerating AI innovations in the field.

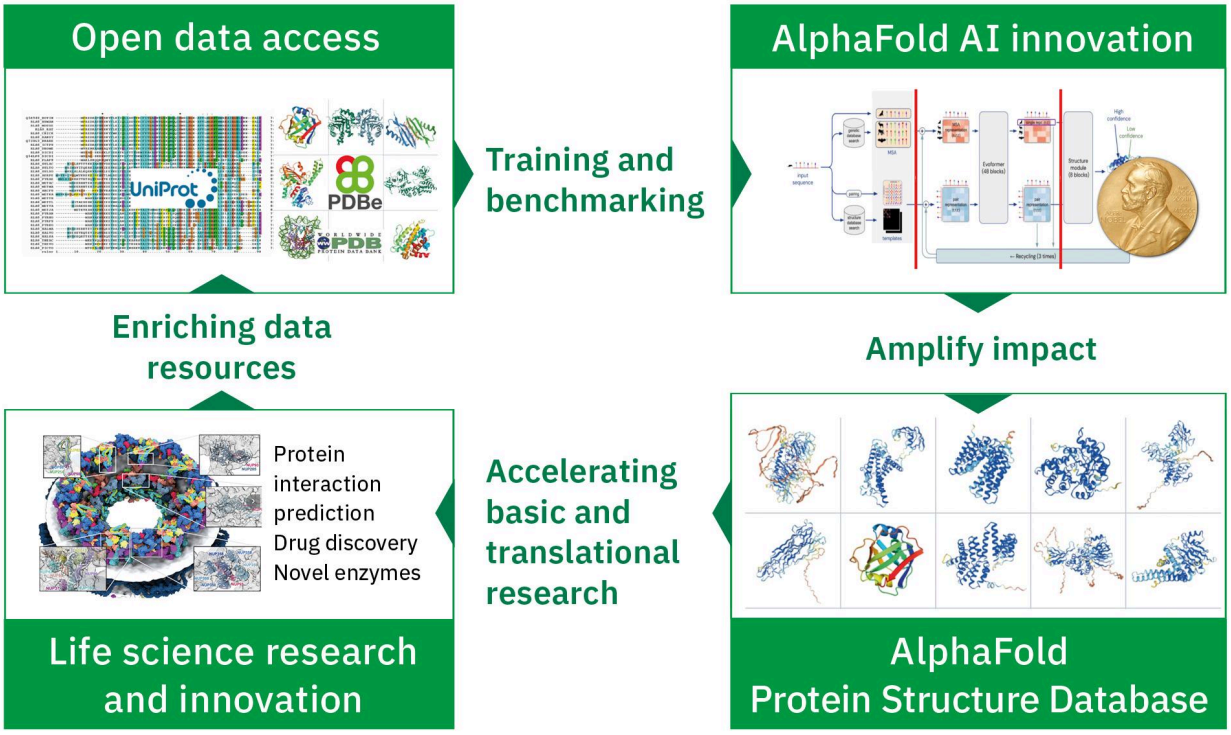


Figure 3: Data AI innovation cycle. Illustration of the data AI innovation cycle with AlphaFold as an example. Open access data repositories at EMBL have enabled AI innovations. New data service innovations amplify the impact of AI models by making data available, enabling and accelerating life science research innovations.

Applied AI innovations building on EMBL biodata

- **Identify biological questions and opportunities for future AI approaches:** Work with biological and AI communities to identify datasets for facilitating impactful opportunities for AI-based solutions. EMBL will leverage its position as the home of biological data and its role in European networks to establish community efforts to generate and integrate training datasets.

- **Develop and train AI models based on EMBL biodata:** Apply the methods established in Pillar I to foster applied AI innovations across biological domains at EMBL. In addition to fostering tailored AI solutions, EMBL will leverage specific opportunities for connecting and integrating different data types and modalities, e.g. establishing foundation models across biological scales and data modalities.

- **AI model challenges and benchmarks:** Develop, host, and organise AI model benchmarks and community challenges. Progress in AI hinges on the ability to assess and compare alternative solutions. EMBL will roll out new concepts developed in Pillar I to underpin public-facing challenges and benchmarks, e.g. building on efforts such as the DREAM challenges.

- **Global partnerships to enable open science AI development:** Contribute to and coordinate international consortia to further open science principles in AI. As part of this

aim, EMBL will support AI development by participating in community benchmarking efforts and making AI models' outputs available via EMBL data services.

Supporting AI innovations through new services and data resources

- **Standards and infrastructure for AI-ready data:** Establish data standards and infrastructure across EMBL data services, in biological domains such as imaging, genomics, and structural biology, as well as amend EMBL's data management frameworks to yield AI-ready datasets. Streamlined data representations and facilitating access to high-quality, integrated and relevant data slices on performance compute infrastructures will enable scientists to more directly use EMBL data services to train, benchmark, and validate AI innovations.

- **Curate reference datasets for model training and benchmarking:** Aggregate and curate existing and forthcoming datasets to establish high-quality computable data resources for AI model training. In parallel, we will contribute to frameworks for community benchmarks, opening up EMBL data services to the broader AI community. By working together with biological communities and stakeholders, we will facilitate access to independent validation data, allowing for objective assessment of AI models.

- **Serve AI models and predictions:** Develop and deploy infrastructure solutions to openly and more effectively share AI models and their outputs. This includes the establishment of model repositories or "zoos", combined with dynamic database solutions to host pre-computed model outputs. This aim will build on our experiences with the successful AlphaFold database, enhancing the impact and accessibility of future AI models.

Enhancing the provision of EMBL data services through AI

- **Enhance data curation and harmonisation:** Deploy AI-based tools to improve the efficiency, cost, and utility of EMBL data services. Examples include automated or AI-guided annotation, curation of datasets, and the establishment of integrated datasets across data resources.

- **Interactive query and scientific support systems:** Create interactive user query systems that allow interactive exploration of biological data and implied relationships across resources and biological domains. As AI systems mature, we aim to establish a one-stop shop for all datasets represented in the EMBL archives, which can respond to queries with scientifically useful results, within seconds and at scale.

Pillar III: AI for experimental workflows

Impactful future AI innovations are expected at the interface of AI, experimental data generation, and scientific hypothesis formulation. Consequently, the third pillar aims to **embed AI in experimental technology and workflows** for data acquisition, interpretation, and new experiment conceptualisation (Figure 4)**.** Bespoke AI tools will be developed in tandem with novel lab technologies, to improve their efficiency and enable novel experimental workflows. By combining AI-assisted analysis with **next-generation lab automation**, EMBL will develop and

implement **adaptive experimental designs** – “lab in the loop” – across experimental services and technologies. Beyond the advancement of lab technology, EMBL will **pioneer the integration of AI into all steps of the scientific process**, including generating scientific hypotheses and connecting experimental decisions and data to existing knowledge. These aims will be closely connected with the experimental and data services at EMBL, both internally and externally facing.

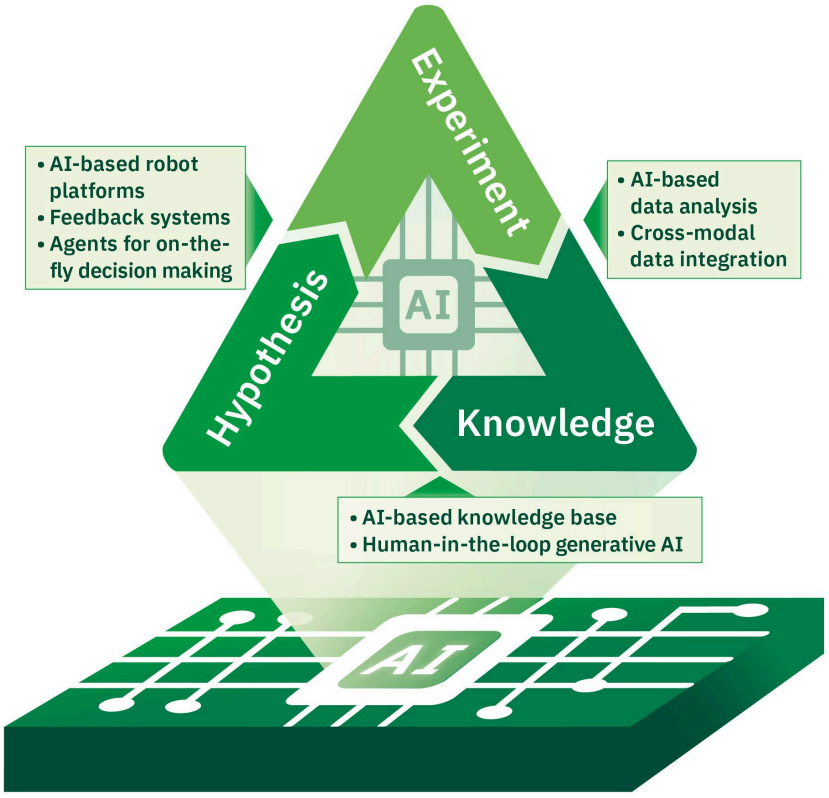


Figure 4: Integration of AI in experimental workflows. The deep integration of AI into experimental workflows will create a virtuous cycle connecting AI-based reasoning from data, experimental design, data generation and data archival.

Incorporation of AI into experimental workflows

- **AI-assisted sample preparation and processing:** Develop AI-based methods for on-the-fly optimisation of experimental protocols, including sample preparation and instrument parameters. Examples include connecting AI to robotised sample preparation for structural biology at EMBL Grenoble or EMBL Hamburg, or the implementation of smart microscopy – integrating automation, (AI) computation, and adaptive technologies.

- **AI-based agents for experimental design:** Establish agent-based approaches where AI assists in larger experiment-planning tasks, such as designing combinatorial perturbations or planning information-efficient expeditions.

- **Interactive exploration tools and on-demand data acquisition:** Pioneer experimental systems where predictive feedback and results drive data generation on demand, e.g. for

experimental validation of data-driven simulations and predictive models, or for efficient exploration of large hypothesis spaces.

- **AI-assisted scientific reasoning**: Leverage AI as a co-pilot to formulate or update scientific hypotheses and experimental parameters based on AI models. Foster human-in-the-loop paradigms with AI-based systems driving evidence integration.

Advancing scale and experimental throughput through AI

- **Bespoke AI tools for new assays and technologies:** Develop AI tools tailored to new data modalities or instrumentation, while leveraging theoretical advances in multi-scale and multi-modal integration. Examples include AI methods for correlative microscopy to streamline workflows, optimise image acquisition, and enable new ways of analysing complex biological phenomena, such as coupling imaging with omics modalities.

- **Radical throughput improvement:** Elevate experimental platforms by integrating versatile AI-based sorting and targeting mechanisms, including the coupling of modelling-based outputs. As part of this aim, we will develop experimental strategies for built-in validation and on-the-fly evaluation by means of orthogonal readouts. Examples include hybrid instruments that can sparsely acquire their own validation data, e.g. for image restoration and enhancement.

- **Data generation for training and validation of AI models**: Joining forces with the foundation model activities in Pillars I and II, we will develop strategies to acquire data for optimal training and – most importantly – validation of foundation models. Validation activities include not only the generation of “new data” but also the definition of biological tasks and real-world use cases for such models in an experimental and biological context.

5. The Enablers

People and training

The successful implementation of AI at EMBL requires attracting new talent at all levels and backgrounds – from predocs to postdocs to PIs, as well as data and AI engineers, who can bring in new expertise and make this vision a reality. Existing role profiles at EMBL will also need to evolve, and new ones will need to be created. EMBL also needs to establish and promote a value system to specifically attract these new communities.

Adapting existing career models

The current academic career model will remain, with the classical levels of predocs, postdocs and group leaders. However, some adaptations will be needed. For scientists from computer science backgrounds, the transitions between career stages are slightly different, with PhD students

frequently directly transitioning to PI positions. The larger role of PhD students motivates the creation of **new interdisciplinary training programs**, by building on the success of EMBL's postdoctoral programs such as the EIPOD-LinC program, for example. In order to enhance its attractiveness for talent, EMBL should also consider evolving the current model for postdocs or exploring the feasibility of concepts such as **independent fellows**. Such fellowship positions could help to expose highly accomplished young scientists from other fields to the life sciences, providing a stepping-stone towards full independence. Scientists from computer science also require new metrics for success, such as considering conference publications as key scientific outputs, thus motivating **changes to research assessment parameters**.

New career models

In addition to changes to existing career models, new models and modes of recognition will be required in order to scale AI capacity to match EMBL's ambitions. **AI engineering teams** are one such new community, who will facilitate **team science**, enabling new modes of working together with research groups, data science experts, and scientific service teams towards common goals. These teams will also add to **internal consultancy and training** capacity across EMBL and ensure that the latest advances in AI technology are widely accessible across the organisation. Finally, they will establish new externally facing data services, such as model repositories or benchmarking platforms, thus playing a significant role in opening up EMBL's data resources and expertise to the external scientific community. EMBL will explore new approaches to attract and recruit these teams, including from industry and disciplines without a prominent biology track record, as well as establish high-profile fellowship schemes that are geared towards experts from industry.

As EMBL transforms, EMBL will be proactive in consistently promoting intrinsic values such as **openness, inclusion, integrity, transparency, and fairness** to create a workplace that is amongst the best life science institutes in the world.

Training in AI

Training is one of EMBL's core missions and training of EMBL staff and the wider scientific community is an essential driver for spearheading EMBL's AI transformation and the associated scientific, technical, and cultural changes.

EMBL training will underpin AI at EMBL by ensuring that all staff and fellows are equipped to incorporate AI and make sound judgements as to when and how to apply it, as relevant to their roles. To advance the training program, EMBL aims to:

- **Create a resource library to familiarise all staff with AI concepts.**
- **Provide foundational knowledge to EMBL managers and leaders,** enabling them to support their groups/teams in making innovative and appropriate use of AI, while balancing risks and ethical considerations. These measures will empower research groups to use and develop new AI-based approaches for understanding biology.
- **Provide technical training** to enable service teams to enhance their services through AI, and to generate/provide AI-ready datasets.

- **Support the continuing professional development of AI engineers,** to build and maintain AI environments, positioning EMBL as a leader in AI-led biology

EMBL will blur the lines between internal and external training; by considering EMBL staff to be a microcosm of the bioscience research community at large, training materials can be developed and tested to be readily used in, or adapted for, both contexts. AI training will be coordinated in the context of the wider EMBL training curriculum, for example, creating bridges between training in data management, creation of AI-ready data resources, and applied building of AI models.

It is envisaged that the EMBL Course and Conference Programme will be enhanced to include AI, such as the 'AI in Biology' conference, and the training of users of EMBL-EBI data resources, such as those offered when the AlphaFold database was released. The EMBL Scientific Visitor Programme, which includes sabbaticals, will also help foster AI activities, allowing scientists and students of all levels and backgrounds to come to EMBL, promoting active exchange with EMBL scientists as well as allowing them to benefit from new technologies and state-of-the-art equipment.

Partnerships and collaboration

Partnerships and collaborations have always been important to the scientific endeavour but are even more important in the context of AI. New alliances will be required to enable rapid advances and unleash the dynamics required to deploy AI across EMBL. Strategic partnerships will satisfy a range of needs, from accelerating the successful transition to AI-driven life science research to enabling EMBL to engage at the European level to develop the appropriate structures and increase its visibility and credibility in the AI community.

The AI transformation of the life sciences is bringing in new players and demands new requirements. Future collaborations will include interactions with research networks on the regional, national, and European levels (e.g. ELLIS), as well as collaborations and partnerships with AI innovators. An additional set of attractive opportunities will emerge in application domains. Partnerships with industry will require different modes of interactions, from pre-competitive partnerships to joint research projects and the initiation of new spin-offs.

The goals for partnering include:

- **Access to training data**: Develop collaborations with entities who can generate data at scale or companies that have similar data generation and curation needs for biological questions as EMBL.
- **Training AI models:** Foster relationships with academic groups and networks where EMBL can determine the types of models and biological questions amenable for AI and benchmark the models.
- **Training in AI:** Create new interdisciplinary bridges to bring computer science and AI expertise to EMBL and molecular biology to AI institutes. To this end, links with ELLIS or related initiatives will be particularly beneficial.
- **Compute and engineering resources:** Implement stable partnerships with technology companies and national supercomputing centres, creating an ecosystem of collaboration for large AI projects.

- **Dissemination of models and outputs:** With repositories of computable data on the rise, e.g. cellXgene from CZI and Hugging Face, EMBL can provide long-standing expertise and define best practices for organising the review process and making research outputs useful for users. EMBL will consider hosting bespoke model repositories in selected areas where such repositories can be strategically connected and aligned with data services and resources.

Going forward, individual partnerships and relationships will be essential to enabling EMBL to create new AI models or make strong contributions. Key considerations for the decision on specific partnerships include aligned and mutually beneficial goals to ensure an equitable partnership and partner alignment with EMBL's open science and open data practices.

6. Embedding AI within EMBL

Data Science Centre

The pan-EMBL Data Science Centre (DSC) plays a pivotal role in enabling data-centric research and services across EMBL's multiple sites by enabling efficient data management, integration, utilisation, and support throughout the biodata lifecycle. The interaction points between the DSC and AI at EMBL offer great potential for both data science and AI-driven research. One foreseen major area of synergy will be the generation of AI-ready data structures, where the DSC's expertise in data management will allow it to ensure that datasets across EMBL are curated, annotated, and formatted optimally to make data "AI-ready". By ensuring that datasets adhere to high standards of metadata, accessibility, and interoperability, this will lay the foundation for AI models to be trained on clean, structured, and relevant data.

Theory Transversal Theme

The Theory Transversal Theme has facilitated the incorporation of theoretical approaches across all areas of EMBL science. AI at EMBL will build on theory concepts where appropriate, for example, to guide the construction of AI models (see Pillar I). Conversely, theory research will profit from AI as a tool to develop modern theoretical concepts and frameworks. A synergistic interaction between AI and theory will entail mutual benefit and will help to ensure that EMBL will develop scientifically sound and principled AI approaches.

EMBL-European Bioinformatics Institute

As the home of the world's most comprehensive collection of data resources spanning all major biological data modalities, EMBL-EBI will play a central role in facilitating data-driven AI developments and transforming the EMBL-EBI data resources to deliver next-generation services to its global user base. EMBL's AI activities will benefit from extensive expertise at EMBL-EBI in developing data standards, curation tools, and data management and integration, thereby supporting the AI innovation cycle. EMBL-EBI, in coordination with the DSC, will be pivotal in implementing the strategic activities related to Pillar II by leveraging existing and newly generated data.

AI research at EMBL and globally and EMBL-EBI services will also work synergistically. As the AlphaFold case illustrates, development of AI models requires high-quality curated data. The outputs of these models can be provided back to the community as freely accessible online resources, thus augmenting the value of deposited data. EMBL-EBI research will play a significant role in building and deploying these models and thus serve as a bridge between research and service within AI at EMBL. Additionally the transformation of EMBL-EBI's data services and research will require the evolution of new career paths to enable the implementation of EMBL's AI strategy.

Through its data services and research, EMBL-EBI is uniquely positioned to provide curated training and benchmark datasets, develop AI models and the necessary infrastructure to share AI models, support global partnerships, and deliver AI training to accelerate AI advancements and contribute to EMBL's AI vision.

Lighthouse projects

To understand the scale of the ambition and the scientific areas where AI could be applied at EMBL, EMBL scientists, through a structured ideation process, came up with ideas for projects of varying sizes for AI, which were later named Lighthouse projects. The Lighthouse projects serve as examples of biological questions amenable to AI, which are aligned with EMBL priorities and expertise. In developing this strategy, they have been used to identify the requirements for EMBL to conduct world-leading science in AI and biology. They foster concepts to pilot via external and philanthropic funding and open doors for collaboration with member state institutes and industry. The projects to date are listed below. However, it should be noted this is not an exhaustive list and it is likely that other projects will emerge and/or these may evolve through their realisation.

Currently envisioned Lighthouse projects

- Reference Cell
- Learning biophysical dynamics
- AI for tissue engineering
- Planet Earth
- Human GPE (Genome-Phenome+Exposures) Foundational models
- Leveraging artificial intelligence approaches for improving EMBL-EBI Services
- Towards an automated lab of the future
- Experimental services, automated sample to structure workflow in Cryo-EM
- Gene regulation and variant effect prediction
- TrAIning for All

AI Ethics

Data ethics and governance is an area where the DSC and AI initiatives will closely work with EMBL's Bioethics Services. As AI-driven analyses become more prevalent in the life sciences, the ethical considerations surrounding data governance, sharing, and reuse – which includes the reuse of internal data – become increasingly important and relevant for AI model training, AI model transparency, and AI-generated predictions relevant to the life sciences, such as in human health. By embedding key ethical frameworks, EMBL will be well-positioned to lead in the responsible use of data and AI in the life sciences. These will ensure that AI applications at EMBL align with responsible scientific practices, respect individual rights, and aim for transparency and reproducibility. EMBL's research practices can serve as a template for best practices and principles for EMBL member states, and may aid the creation of EU policies for AI in science.

Open science

Adherence to open science principles has been key to the rapid innovation and adoption of AI. Open source data and software have been a key enabler of agility and scientific reproducibility, yet are currently under threat, as the mode of open publication and sharing of innovations is no longer a given for some stakeholders. Major commercial stakeholders are moving towards more closed rather than open research approaches, which is likely motivated by the fact that commercial opportunities have become more visible and important. EMBL can and should contribute to **defending open science and open source principles**, including continuing to lead by example in sharing fully open AI models. Mechanisms could also include promoting incentives for open data producers, including the development of new models of collaboration with commercial stakeholders, to balance the demand for (ideally time-limited) closed source approaches with ensuring a long-term open science trajectory.