European Commission

Joint Research Centre

Artificial Intelligence for Biology: Capabilities, Readiness, and Policy Implications

Ruiz Serra, V., Fortuna, P., Curion, F., Petrillo, M., Leoni, G., Bili, D., Melandri, D., Vangheluwe, N., Comte, V., Bertolini Agnoletto, L., Coecke, S., Ceresa, M., Querci, M., Wiesenthal, T.

2026

Contact information
Name: Fabiola Curion
Address: Via E. Fermi, 2749. TP126
I-21027 Ispra (VA), Italy
Email: fabiola.curion@ec.europa.eu
Tel.: +39 (0) 332785437

Joint Research Centre
https://joint-research-centre.ec.europa.eu

JRC147549
EUR 40815

PDF ISBN 978-92-68-42482-7 ISSN 1831-9424 doi:10.2760/9397036 KJ-01-26-343-EN-N

Luxembourg: Publications Office of the European Union, 2026

Some content was created using the commission internal GenAI tools (GENESIS and GPT@EC)

How to cite this report: Ruiz Serra, V., Fortuna, P., Curion, F., Petrillo, M., Leoni, G. et al., *Artificial Intelligence for Biology: Capabilities, Readiness, and Policy Implications,* Publications Office of the European Union, Luxembourg, 2026, https://data.europa.eu/doi/10.2760/9397036, JRC147549.

Contents

Abstract

The European Commission has prioritised the use of Artificial Intelligence (AI) in scientific research, including life sciences and biotechnology, through various initiatives such as the AI in Science Strategy and the Apply AI Strategy. As AI-enabled life-science solutions in Research and Development (R&D) become a strategic priority for the European Union, it is essential to understand the current landscape of AI in life sciences, which can bring significant benefits but also raises important questions and challenges. This report provides an overview of the capabilities and trends of certain AI models with applications in molecular and cellular biology. Based on a comprehensive dataset of biology AI systems and other literature, the report aims to set realistic expectations, inform research trends, and support informed policy and governance decisions. Additionally, these findings highlight the need for transparent reporting standards and validation frameworks that enable reproducible, relevant and comparable assessment of AI-designed biomolecules, as well as emerging molecular-cellular and bio(techno)logical innovations.

Acknowledgements

The authors would like to express their gratitude to colleagues at DG CNECT, Antoni Mestre Gascón, Evangelina Markidou, Kyriacos Hatzaras and Saila Rinne, for their valuable feedback.

They authors would also like to acknowledge Epoch AI for developing and maintaining the publicly available Biology AI Models dataset, which was essential to this work.

Authors

The following individuals contributed to this report:

Victoria Ruiz-Serra – European Commission, Joint Research Centre (JRC)

Paula Fortuna – European Commission, Joint Research Centre (JRC)

Fabiola Curion – European Commission, Joint Research Centre (JRC)

Mauro Petrillo – SEIDOR Italy S.r.l.

Gabriele Leoni – European Commission, Joint Research Centre (JRC)

Danai Bili – European Commission, Joint Research Centre (JRC)[1]

Daniela Melandri – European Commission, Research and Innovation (RTD)[2]

Nick Vangheluwe – European Commission, Research and Innovation (RTD)

Valentin Comte – European Commission, Joint Research Centre (JRC)

Lorenzo Bertolini – European Commission, Joint Research Centre (JRC)

Sandra Coecke – European Commission, Joint Research Centre (JRC)

Mario Ceresa – European Commission, Joint Research Centre (JRC)

Maddalena Querci – European Commission, Joint Research Centre (JRC)

Tobias Wiesenthal – European Commission, Joint Research Centre (JRC)

Disclaimer

During the preparation of this work, the editors used GENESIS in order to support the integration process of the contributions by the authors to the final report, as well as to harmonise the narrative style. After using this tool, the editors and authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

GENESIS is the evolution of the GPT@JRC platform (JRC143418) into the JRC's Agentic AI platform developed for staff within the European Commission (EC). It is a platform offering secure access to a wide variety of pre-trained Large Language Models to assist with written office tasks and support

[1] Danai Bili is currently with Johnson & Johnson Medical Technology. She contributed to the report while affiliated with the JRC as a trainee.

[2] Daniela Melandri is currently affiliated with DG CNECT and contributed to the report while affiliated with DG RTD.

scientific work. GENESIS is hosted local at the JRC data centre and is part of a JRC-wide study on the potential applications of this new technology within the European Commission. User prompts and generated responses are not shared with third parties.
Note: The last access date for all web sources referenced in this document, unless specified in the individual citation, is 18/06/2026.

Executive summary

AI models for biology are advancing rapidly, reshaping how research is done in protein design, genome analysis, and drug discovery, among other areas. Their applications are relevant in clinical settings but also extend to industrial biotechnology, crop and livestock improvement, and environmental monitoring, as well as to the design of biological systems with dual-use potential. Frameworks for assessing the readiness, misuse potential, and deployment risk of these type of AI models are still being developed. At the same time, capabilities are concentrated in a few organisations and regions, and the EU is not currently in a leading position. This report provides an empirically grounded account of the field's current state and what it means for EU policy.

Policy context

This report was prepared by the European Commission's Joint Research Centre (DG JRC), with contributions from Directorate-General Research and Innovation (DG RTD), to support evidence-informed policymaking on the rapidly evolving landscape of AI for molecular and cellular biology. It is relevant to the implementation of the AI Act and to the ongoing European initiatives on AI in science, the Apply AI strategy, life sciences, biotechnology, the Medical Devices Regulation, and the proposed Biotech Act. It informs also on broader discussions on industrial policy, data governance, research infrastructure, competitiveness, biosecurity and dual-use frameworks. By combining a structured analysis of 480 biological AI systems with an assessment of scientific maturity, deployment readiness and current barriers, the report provides an empirically grounded perspective on opportunities, limitations and governance needs. Its added value lies in distinguishing technical capability from operational readiness and highlighting the importance of robust evidence generation, transparency, reliability, relevance and confidence in scientific assessment. The report also emphasises that effective governance requires consideration not only of technical performance but also evidence quality, context of use, uncertainty and validation practices when interpreting AI-enabled biological research. The findings are relevant beyond biology, informing wider policy discussions on trustworthy AI, strategic autonomy, innovation and preparedness.

Key conclusions

Four headline conclusions emerge.

Progress in biology AI research models is real but concentrated in data-rich areas like protein structure. Biological AI models are advancing fastest where data is abundant and structured (protein structure, molecular design, sequence modelling) and lagging where it is not (single-cell biology, metabolomics, clinical translation). Policy attention should follow this asymmetry rather than benchmark headlines and counteract the concentration dynamic by which landmark successes draw talent and investment away from underserved domains.

The EU has the foundations to develop biology AI models but lacks coordination. The European High-Performance Computing Joint Undertaking (EuroHPC JU) pre-exascale and exascale systems provide compute capacity sufficient to train even the largest biological AI models surveyed. Yet only one European Institution (TU Munich) sits among the top 20 global developers, intra-EU collaboration lags EU-US collaboration, and training data is fragmented across more than 100 specialised repositories with limited interoperability. The gap is one of mobilisation, not capacity.

Scientific maturity of a biological AI model does not imply readiness for deployment: the maturity paradox. The scientific maturity of a biological AI model and its technology readiness

level routinely diverge. AlphaFold, Evo 2 and AlphaGenome are widely celebrated and freely available, yet none has passed through an integrated readiness assessment for clinical, regulatory or industrial deployment. Treating benchmark performance as a proxy for readiness produces only a partial picture of progress and may misallocate regulatory and investment attention. Transparent evaluation frameworks capable of distinguishing these dimensions may improve evidence-informed governance and resource allocation.

Current EU policy frameworks assume assessment methods for biological AI models that do not yet exist. The AI Act, the Medical Device Regulation and the proposed Biotech Act presuppose evaluation frameworks for biological foundation models that have not been developed. Developing EU-level maturity frameworks that combine domain-specific classification with Technology Readiness Levels (TRL) logic is crucial for effective oversight.

These findings reduce a key uncertainty in current policy discussions by separating two questions that are typically conflated: how scientifically mature an AI capability is in a given biological domain, and how ready a specific model is for deployment. Significant uncertainties remain on safety evaluation methods, the dual-use risk profile of openly released models, and the trajectory of EU developer capacity. Policy options arising from these conclusions include coordinated governance of biological training data, coordinated use of European compute infrastructure and incentives for intra-EU collaboration for biological AI, and harmonised maturity assessment integrated into regulatory pathways.

Reported model capabilities should be interpreted together with evidence on reproducibility, reliability, scientific relevance, transparency, uncertainty and validation context. Robust governance and responsible deployment require assessment that extend beyond benchmark performance alone.

Main findings

The report draws on detailed metadata on 480 biological AI models compiled by Epoch AI and supplemented by information extracted and harmonised from the scientific literature by the JRC, providing an empirically grounded overview of the current AI4Bio landscape.

What can biological AI models do? Capability is concentrated in protein-centric work and in predictive rather than generative tasks. Transformer-based architectures account for 66% of surveyed models. Single-cell biology, metabolomics, and multimodal systems are comparatively underrepresented. These findings highlight the uneven distribution of current AI capabilities across biological domains and should be considered when interpreting benchmark achievements and research priorities.

How are they built and resourced? Industry outpaces academia in infrastructure intensity, with the usage of larger datasets, more hardware and longer training durations. Nearly half of surveyed models disclose neither weights nor training code, limiting reproducibility and third-party assessment. Training data is fragmented across more than 100 specialised repositories, with the largest concentrated in the US and EU.

How ready are they for real-world applications? A two-scale assessment combining domain maturity classification with technology readiness levels shows systematic divergence between the two. High-profile models cluster at high domain maturity and low-to-mid TRL: scientifically advanced, but not yet ready for clinical, regulatory or industrial deployment. No surveyed system has completed an integrated readiness assessment.

How should these findings be interpreted for policy? Reported technical capabilities should be interpreted alongside information on scientific maturity, deployment readiness and the underlying evidence base. Throughout the report, benchmark performance alone is shown to provide only a partial picture of progress, highlighting the importance of transparent evaluation, reproducibility and appropriate interpretation when informing governance and policy decisions.

Where does the EU stand? The scientific base is strong and EuroHPC JU compute capacity is in principle sufficient to train the largest models surveyed. Coordination, data governance and intra-EU collaboration are the binding constraints. Only one European institution sits among the top 20 global developers, and intra-EU collaboration lags EU-US collaboration.

Related and future Joint Research Centre work

This report connects to JRC work on AI governance, health technology assessment and digital health. Follow-up work will support implementation of the AI Act and the Biotech Act, refinement of maturity frameworks, and Member State engagement on biological AI infrastructure.

Quick guide

Biological AI models are machine learning systems trained on biological data: protein sequences, genomes, cellular measurements. AlphaFold, which predicts protein structures, is the best-known example. The field is advancing rapidly, driven by competition between academic and industrial actors, raising questions about readiness for clinical or industrial use and the risks of unregulated deployment. The report combines a quantitative landscape analysis of 480 biological AI models examining capabilities, infrastructure, developers, openness, and deployment readiness, to inform EU policy responses. It distinguishes scientific capability from deployment readiness and highlights that reported benchmark performance should be interpreted in the context of available evidence, transparency and current limitations of evaluation.

1. Introduction

1.1. Context, Motivation and Scope

Biology can be read as a layered system of information. DNA stores it, RNA transmits and regulates it, proteins execute it, cells integrate it into function, and tissues and organisms into physiology. Each level has its own vocabulary (the nucleotide alphabet, the amino acid alphabet, gene expression states, morphological features, etc) and its own grammar (regulation, folding, signalling, organisation, etc). These levels are not independent: information flows across them, and the same biological state can be read simultaneously at several scales (**Figure 1**).

This perspective, sometimes referred to as the ***language of life*** (Rao et al., 2026; Topol, 2025), reframes what AI models for biology are doing. They are systems that learn to read, translate between and generate across these biological levels. Protein language models learn the amino acid alphabet from sequence data; genomic foundation models learn the nucleotide alphabet; single-cell models learn cellular states; pathology models learn tissue morphology. Increasingly, the frontier is models that operate across levels, treating biology as a single multi-scale information system rather than a set of separate domains.

It is this shift, from AI as a narrow predictor layered on top of biology, to AI as a system operating on biological information across scales, that makes the field strategically consequential for the economy and society, as recognised and supported also by the EU policy framework.

The European Commission has been explicitly encouraging the use of AI through its policies and various strategies, namely: the *Apply AI* strategy[3], to drive AI uptake across sectors, including healthcare and pharmaceuticals; and a dedicated *AI in Science* strategy (paving the way for the Resource for AI Science in Europe, RAISE)[4], as follow-through on the *AI Continent Action Plan*[5] and complementing the risk-based *EU AI Act*[6]. AI is now acting as a scientific collaborator, automating parts of hypothesis generation, experimentation, and analysis (Purificato, Erasmo et al., 2025).

In parallel, the European Commission is actively promoting strategic life sciences including biotech. For instance, in health and biotechnology, the *Strategy for European Life Sciences*[7], sets a 2030 ambition to make Europe the most attractive place for life sciences Research and Development (R&D) and translation, with data and AI recognised as an essential element of its push for breakthrough innovation. Also, with the aim of accelerating biotechnology and biomanufacturing, discussions are underway to establish a framework of measures through the *EU Biotech Act* [8]. Among its elements, the initiative foresees provisions on artificial intelligence and data as key

[3] https://digital-strategy.ec.europa.eu/en/policies/apply-ai

[4] https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/european-ai-science-strategy/raise-resource-ai-science-europe_en

[5] https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan

[6] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

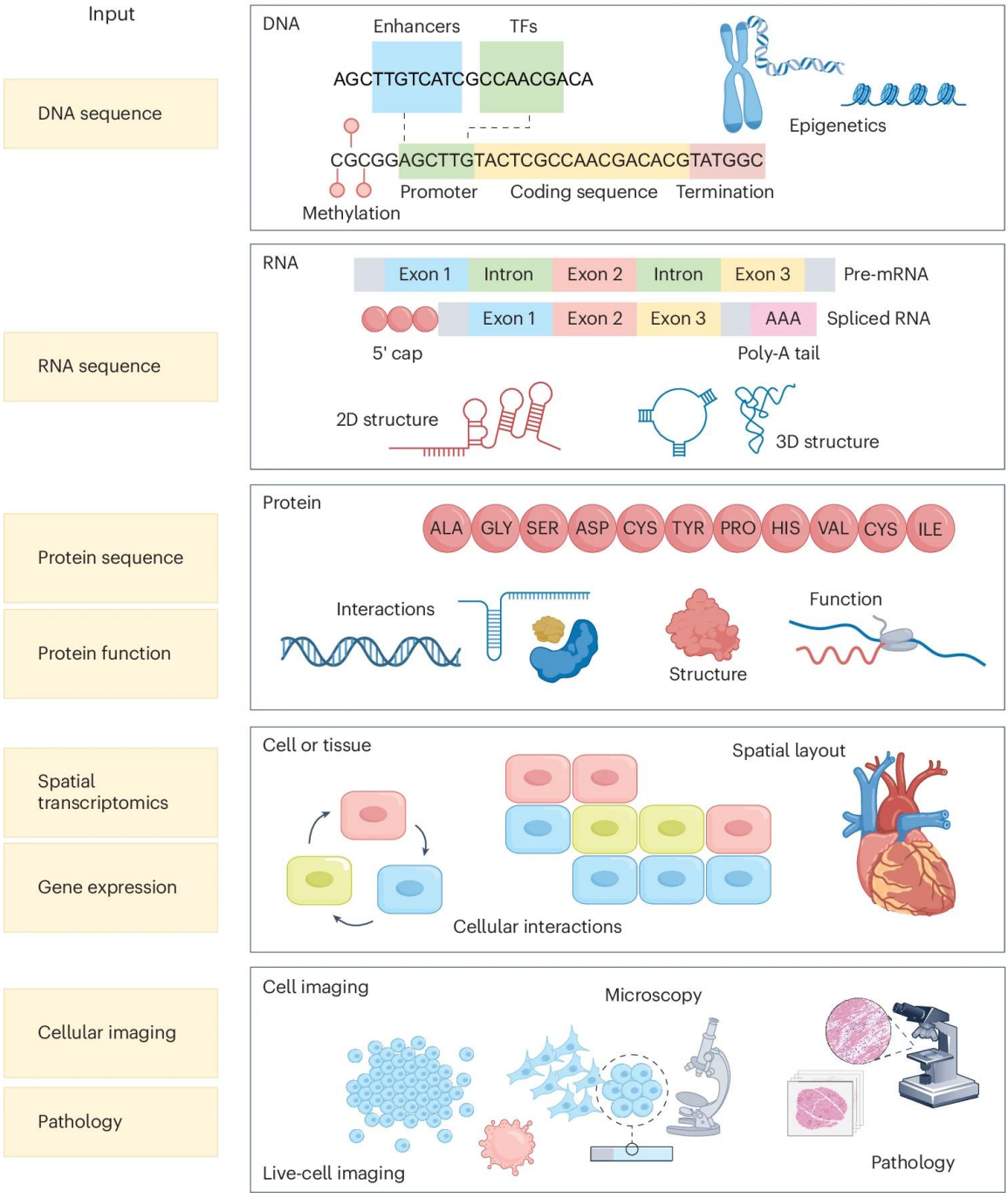
[7] https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/jobs-and-economy/strategy-european-life-sciences_en

[8] https://health.ec.europa.eu/biotechnology/european-biotech-act_en

biotechnology enablers, including guidance on the deployment and use of advanced technologies, such as AI systems, throughout the lifecycle of medicinal products.

As AI and biology-related fields are identified as key priorities, "AI-enabled life-science R&D", the intersection of both, naturally emerges as a strategic priority for competitiveness, public health, and resilience in the EU. Indeed, the global scientific community is advancing rapidly in applying AI to biology, with models and workflows progressively heading in the direction of reading, reasoning over, and even designing biological systems (Bunne et al., 2024; Leslie, 2025) driving transformational changes.

Figure 1. Biological information flow across DNA → RNA → protein → cell → tissue levels.



Source: Adapted from (Rao et al., 2026) .

At the same time, leading international assessments and policy initiatives increasingly recognise that AI-enabled advances in the life sciences bring both transformative benefits and important risks that require structured understanding and preparedness. Recent frameworks and reviews including global AI preparedness initiatives[9], analyses of emerging AI capabilities[10], expert evaluations of biosafety and governance (Committee on Assessing and Navigating Biosecurity Concerns and Benefits of Artificial Intelligence Use in the Life Sciences et al., 2025), and new national-level guidance on securing biological research[11] highlight the need for clearer visibility over capabilities, risks, and responsible practice in AI-driven biology[12].

However, this area of activity is still in its early stages, and the information characterising this cross-disciplinary landscape is scattered and unstructured. In addition, differences in reporting practices, validation strategies and data quality make comparison across studies challenging. This knowledge gap is a significant concern, as accurate and comprehensive information is essential for informed policy and decision-making. Furthermore, the development of AI models across life sciences and biotechnology disciplines in general, and particularly within the biological domain, raises critical questions and challenges. For example, these models may inadvertently perpetuate biases present in experimental data or reflect limitations arising from heterogeneous data generation, annotation or reporting practices or unlock new bioengineering pathways with potential for both beneficial and adverse outcomes (Pannu et al., 2025).

This science-for-policy report seeks to address the existing knowledge gap surrounding what, in this work, is referred to as **AI4Bio** *(AI for Biology)*, defined as AI technologies specifically developed for application in the field of biology or life sciences related sub-areas. It provides a comprehensive overview of the available models, datasets and areas of application, and presents the current barriers, risks and potential policy responses. Given the breadth of the field of biology, **the scope is set on AI models trained on subcellular- and cellular-level biological data** (sequences, structures, cellular measurements, microscopy and pathology images) used to predict, design or simulate biological systems, reflecting the expertise of the authors.

[9] https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

[10] https://www.stateof.ai

[11] https://www.whitehouse.gov/presidential-actions/2025/05/improving-the-safety-and-security-of-biological-research/

[12] https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026#2.1.4.

Box 1. Working definitions used in this report (full set of definitions in Annex 1)**.**

AI4Bio *(AI for Biology)* — AI technologies specifically developed for application in biology and related life sciences (JRC's analysis).

Biology-AI models — models directly and explicitly trained on biological data, including biological sequences, biomolecular structures, biological property labels, or cell-level data (Epoch AI[13,14]).

Foundation model — a model trained on broad data, generally using self-supervision at scale, that can be adapted to a wide range of downstream tasks (Center for Research on Foundation Models (Bommasani et al., 2022)).

1.2. Goals

This report seeks to provide a scientific perspective and comprehensive understanding of the AI4Bio domain with particular focus on the subcellular and cellular level with the intention of informing policy development and decision-making on the most recent developments and potential impacts of this area. Specifically, the report is issued to inform and support policymakers, researchers, and industry in navigating this complex evolving biological digital intelligence field. Particular attention should be paid to the interpretation of evidence, reproducibility, uncertainty and the distinction between technical capability and deployment readiness. To achieve this objective, the report is organised around several key sub-goals, including:

- Map key definitions in the AI4Bio domain.
- Identify reliable resources describing the current developments in AI4Bio.
- Map the core areas of application of AI4Bio.
- Describe models, data and infrastructure used in AI4Bio.
- Identify development trends and current benchmarks in the field.
- Identify risks, future directions, and research priorities.
- Support policy and governance in AI4Bio.
- Provide a set of reproducibility, relevance and confidence metrics and AI4Bio process reporting practices that can be adopted by downstream regulators and funding bodies.

This report builds on, and takes into account, European and global knowledge resources such as the OECD report on Synthetic Biology, AI and Automation report (OECD, 2025), the OECD AI Process Reporting Framework 2.0[15], the WHO Global Health with Responsible Artificial Intelligence report[16] and aims to provide a scientific perspective and **comprehensive understanding of the currently**

[13] https://epoch.ai/data/biology-ai-models-documentation#inclusion

[14] Epoch AI's definitions are adopted here for practical purposes, as their taxonomy directly underpins the model selection criteria used in their dataset, which this report draws upon. It should be noted, however, that these definitions are not universally agreed upon and reflect Epoch AI's own classificatory choices

[15] https://oecd.ai/en/haip-2-launch

[16] https://www.who.int/publications/m/item/leading-the-future-of-global-health-with-responsible-artificial-intelligence

available AI systems in molecular and cellular biology to inform and support policymakers, researchers, and industry in navigating this evolving field

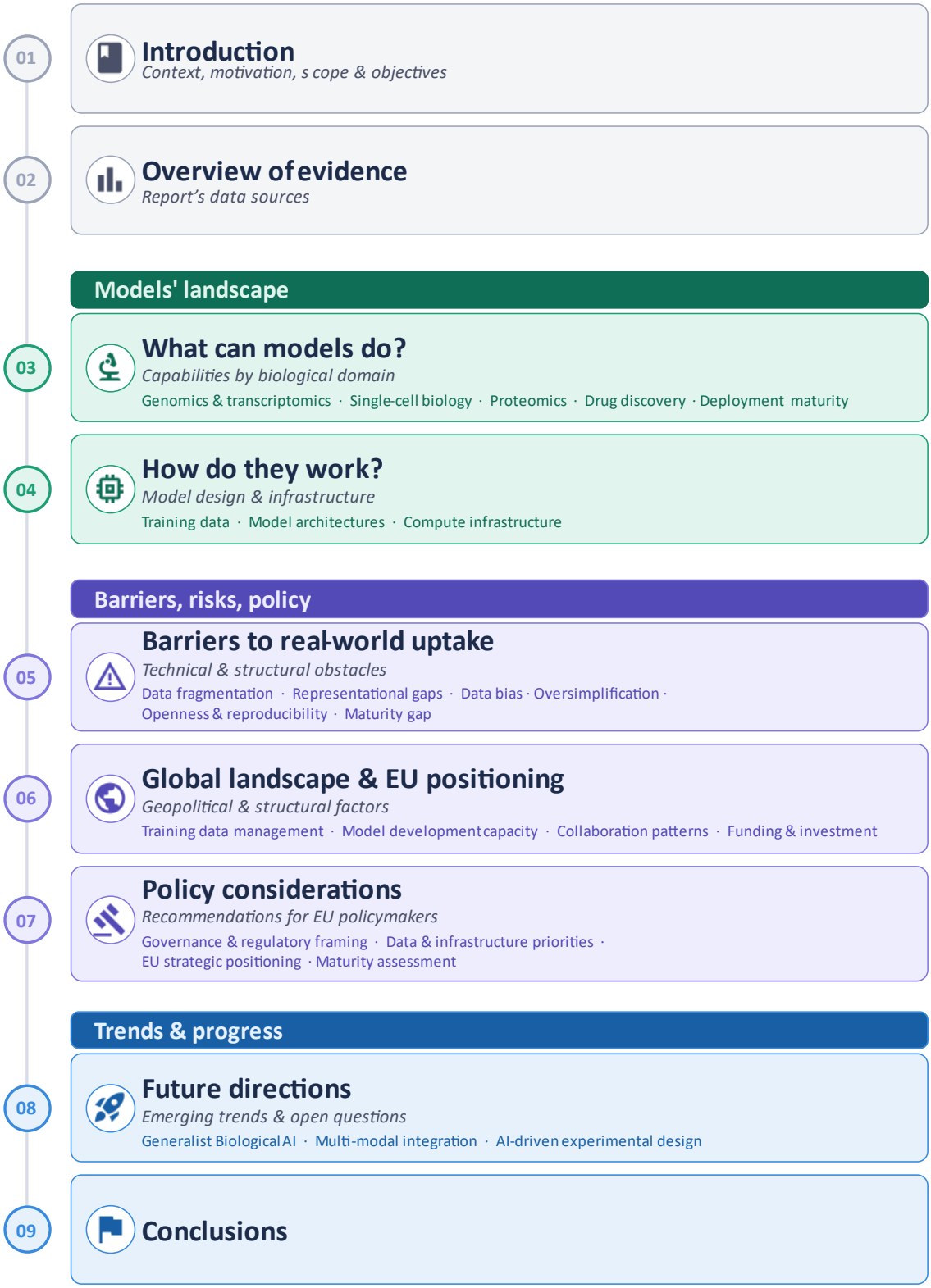
1.3. Report Organisation

The report is organised around a narrative arc that moves from capabilities to barriers and risks to policy implications. It opens by asking what biological AI models are and why they matter, then progressively examines what stands between current capabilities and real-world adoption, before looking ahead to future directions.

- **Section 2** provides a brief overview of the evidence base on which the report draws, including the Epoch AI dataset and metadata extracted from the scientific literature.
- **Section 3** examines what biological AI models can do, surveying capabilities across genomics and transcriptomics, single-cell biology, proteomics, and drug discovery, and closing with an assessment of the technological maturity and readiness of models for real-world application.
- **Section 4** turns to how these models work, covering training data, model design, and computing infrastructure.
- **Section 5** addresses the first layer of barriers to real-world adoption: the scientific and technical limitations that currently prevent biological AI models from being reliably deployed outside research settings, including training data fragmentation, representational gaps, protein-centric bias, interpretability constraints, and the gap between benchmark performance and deployment readiness.
- **Section 6** adds a second, structural layer. Even where technical barriers could be resolved, the capacity to develop, validate, and deploy biological AI models is unevenly distributed across geographies and institutions. This section examines the global and EU landscape of training data governance, model development, infrastructure capacity, collaboration patterns, and funding, showing that these structural conditions have a distinct geographic shape with direct implications for EU strategic positioning.
- **Section 7** draws on the preceding analysis to set out policy considerations for EU action, organised around four themes: the protein-centric and predictive model landscape, training data and infrastructure fragmentation, strategic dependencies and international positioning, and maturity and technological readiness assessment.
- **Section 8** looks ahead to emerging developments, including advances in model transparency, agentic platforms for research, and the trajectory towards Generalist Biological AI.
- **Section 9** concludes the report and distils its main messages for policy audiences.

This structure, illustrated in **Figure 2**, positions the report not only as an assessment of the current state of biological AI but as a forward-looking reference for policymakers seeking to engage with these developments in an informed and strategic manner.

Figure 2. Visual structure of the report.



Source: JRC's analysis.

2. Overview of Evidence

This report draws on a combined corpus of 480 unique publicly available biological AI models, derived from two complementary sources: the Epoch AI Biology AI Models dataset[17] (hereafter the Epoch AI dataset) and a set of literature-curated models extracted from five recent literature reviews.

The Epoch AI dataset is a resource that consolidates information on representative models and their characteristics. Released in early 2025 and regularly updated, it contained 383 biological AI models as of the data retrieval date of 5 January 2026. The dataset emphasises systems that have advanced the field or achieved notable scientific impact, providing a robust sample of historically significant and state-of-the-art models rather than an exhaustive census. Each model is annotated with a structured set of descriptors covering tasks, training data, model architecture, compute use, hardware, accessibility and related characteristics (see Annex 2 on dataset descriptors).

The dataset is complemented by five recent literature review articles on AI in biology, selected for scientific impact and topical alignment with the scope of this report (see Annex 3). Structured tables from these five reviews were extracted and harmonised against the Epoch AI schema, contributing 130 literature-curated models to the corpus; of these, 97 are not present in the Epoch AI dataset and 33 overlap with it, providing additional metadata and allowing cross-checking of annotations across the two sources. The result is an enriched corpus that broadens coverage across biological input types, task types, model architectures and input modalities, and supports transparent reporting of provenance, uncertainty and applicability-domain information in line with OECD GIVIMP principles (OECD, 2018) and the EU Trustworthy AI framework (High Level Expert Group on Artificial Intelligence, 2019).

The analyses presented in subsequent chapters draw on this combined evidence base through descriptive, comparative and qualitative methods, explicitly documenting any epistemic uncertainty sources (including heterogeneous method and model protocols, target-complexity effects and potential publication bias) to enable reproducible, policy-ready conclusions. Each analytical section references the specific methodological choices it relies on.

[17] https://epoch.ai/data/biology-ai-models-documentation, where an AI model is defined as a Machine-learning model that it is reliably documented as a machine-learning model; it contains a learned component (i.e., not a purely hand-crafted or rule-based algorithm); and it has been trained and is supported by experimental results (i.e., not a purely theoretical proposal).

3. What Can Biology AI Models Do?

Biological AI models are commonly classified according to the type of data they are trained on, such as "protein models", "DNA models" or "RNA models". This data-centric terminology does not reflect fundamentally different AI architectures, but rather a community convention grounded in biological interpretation: different input types involve distinct alphabets, for example amino acids versus nucleotides, follow different biological rules, and support different downstream tasks. Models trained on different biological data modalities are therefore presented separately despite often sharing the same underlying machine learning methods. The main data modalities used throughout this report are defined in **Box 2**.

Box 2. Types of models according to their biological training data modality definitions.

DNA — Models trained primarily on DNA data, typically in the format of sequences represented as strings over the nucleotide alphabet {A, C, G, and T}.

RNA — Models trained on ribonucleic acid (RNA) data, including RNA sequences represented as strings over the nucleotide alphabet {A, C, G, and U} and derived representations such as secondary structures or expression measurements

Protein — Models trained on protein data, most commonly amino acid sequences, using the 20-letter alphabet of standard amino acids as well as protein secondary, tertiary and quaternary structures.

Single cell — Models trained on single-cell datasets that measure molecular profiles, such as gene expression, chromatin accessibility, or protein abundance, at the resolution of individual cells, often represented as high-dimensional feature vectors rather than sequences.

Drugs / Small Molecules — Models trained on chemical compounds, including drugs and small molecules, represented using string-based encodings (such as SMILES or SELFIES), molecular graphs, or three-dimensional conformations.

Text/Image — Models trained on biomedical or scientific text (such as literature, clinical notes, or ontologies) and/or biological images (such as microscopy, histopathology slides, or medical imaging data).

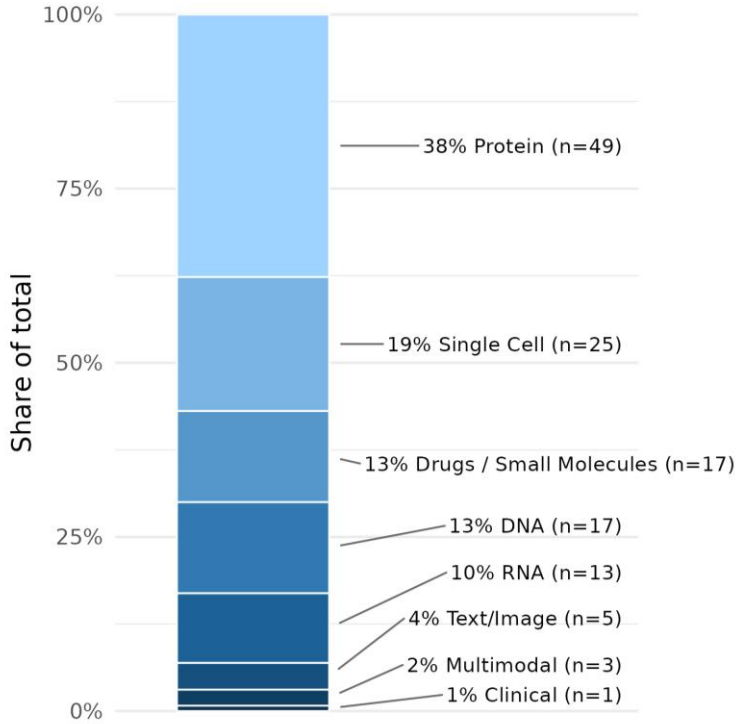
Clinical — Models trained on data derived from patient records or healthcare settings, including electronic health records, diagnostic codes, laboratory results, or clinical narratives.

Multimodal — Models trained jointly on two or more distinct biology data types.

Source: JRC's analysis

Based on literature annotations **Figure 3**, Protein-focused models dominate the literature, underscoring their central role in biological AI research, while RNA-, DNA-, single-cell-, and drug-focused models appear as smaller but clearly established categories. In addition, text/image, multimodal and clinically oriented models are present but remain a modest share of the overall set, reflecting both the higher complexity of integrating heterogeneous data sources and the additional regulatory, data access, and validation requirements associated with clinical applications.

Figure 3. Distribution of models according to their biological training data modality of 130 literature-curated biology AI models.

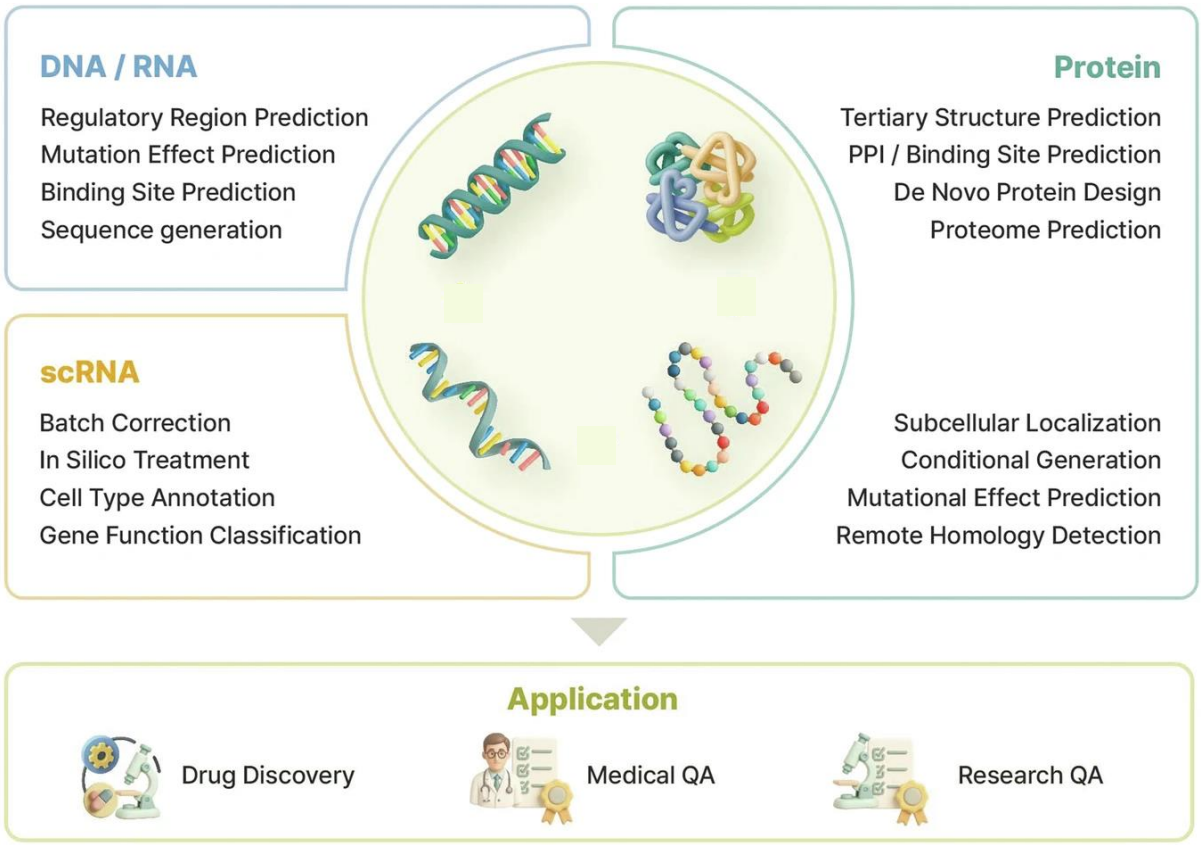


Source: JRC's analysis based on literature

The landscape of biology AI models spanning different data modalities supports a distinct set of downstream tasks and applications. **Figure 4** provides an overview of the principal tasks that biological AI models now perform across the main biomolecular modalities. Organised around DNA/RNA, single-cell transcriptomics (scRNA) and proteins, the figure maps each modality to the downstream capabilities that current models support. For a more granular, quantitative perspective, Annex 4 shows the absolute frequency of downstream tasks across models surveyed in this report. These modality-specific capabilities increasingly converge in integrative applications such as drug discovery, medical question answering and research question answering, where models trained on distinct data types are combined to support end-to-end scientific and clinical workflows.

Building on this overview, this section examines genomic and transcriptomic, single-cell, protein, and drug discovery AI models in more detail. For each application area, representative systems and the capabilities they enable are presented. Rather than providing an exhaustive catalogue, this section focuses on salient trends and influential models that illustrate how biological AI has evolved across domains. Finally, it discusses how mature these models are, both individually and when embedded within industrial platforms.

Figure 4. Downstream tasks of Biology AI models to data modalities (DNA/RNA, protein and single-cell RNA) and their applications in drug discovery, medical QA and scientific research.



Source: adapted from (Ashyrmamatov et al., 2026)

3.1. Genomic and Transcriptomic Models

A new class of AI called genomic language models (also called genomic foundation models or DNA language models) emerged around 2020, where researchers adapted text-processing computational architectures to read and analyse raw genomic sequences: long paragraphs of nucleotides. [18]. These models treat DNA as a "language", learning contextual embeddings via self-supervised objectives (e.g., masking 15–20% of tokens[19]) on large collections of unlabelled reference genomes. Pre-training on hundreds to thousands of genomes, spanning human and non-human species, enables zero- or few-shot performance[20] on downstream tasks, outperforming

[18] The term "genomic foundation model" is increasingly used as a consensus label in the literature (Li et al., 2024a). It is used here as a descriptive term, not a formal classification.

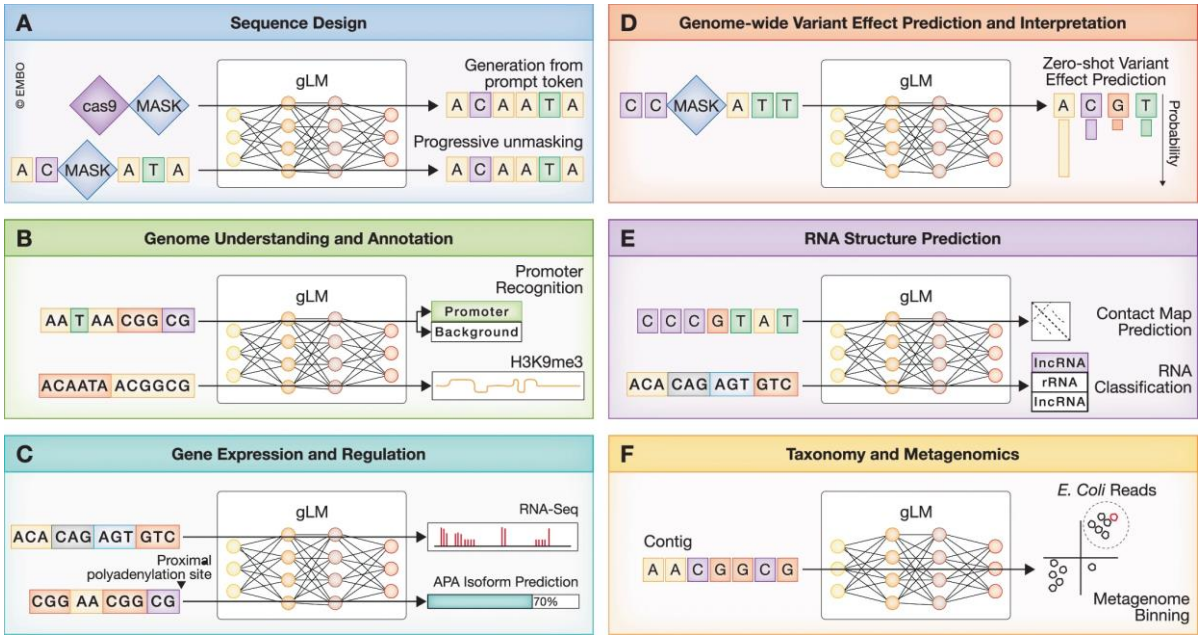
[19] Tokens are short segments of sequence used as the basic unit of input for the model. Masking a proportion of tokens and training the model to predict them from context is a standard self-supervised learning strategy, analogous to predicting a redacted word from the surrounding sentence.

[20] Zero-shot performance refers to a model's ability to perform a task it was not explicitly trained on. Few-shot performance means the model requires only a small number of labelled examples to adapt to a new task, rather than full retraining.

earlier task-specific architectures in variant effect prediction, regulatory element identification, and cross-species transfer (Li et al., 2024a; Guo et al., 2025).

Genomic foundation models address a broad and expanding range of tasks across both predictive and generative domains (see **Figure 5**). In essence, these models learn patterns from vast amounts of DNA and RNA sequence data and then apply that learning to interpret or design biological sequences in ways that would be impractical through conventional experimental or computational means alone. The capabilities described below represent the principal application areas that have emerged from this field; **Table 1** presents selected models illustrating how these capabilities are implemented in practice.

Figure 5. Visual summary of application areas of genomic language models.



Source: (Veiner and Supek, 2026).

Genome annotation and regulatory element identification (**Figure 5B**) is the task of pinpointing and labelling the functional regions within a genome. The genome does not consist solely of protein-coding genes; much of its activity is governed by regulatory sequences such as promoters (which switch genes on), enhancers (which amplify gene activity), and splice sites (which determine how genes are read). Identifying these elements accurately and at scale is essential for understanding how organisms function and how disease can arise. Models approach this by learning sequence patterns associated with regulatory activity across large genomic datasets, enabling genome-wide scans without relying solely on evolutionary conservation as a guide. Recent benchmarking indicates that models such as the Nucleotide Transformer (Dalla-Torre et al., 2025) and GENA-LM (Fishman et al., 2025) perform particularly well on promoter and regulatory element recognition tasks, while DNABERT-2 (Zhou et al., 2023) offers strong performance across short-to-medium sequence contexts (Veiner and Supek, 2026; Marin et al., 2024).

Gene expression and chromatin accessibility forecasting (**Figure 5C**) addresses the question of how active a given gene will be in a particular cell type, and how accessible the surrounding DNA is to the molecular machinery that reads it. These are key determinants of cell identity and function. Models trained on large collections of paired sequence and molecular profile data learn to predict

these quantitative outputs from sequence alone, capturing how regulatory elements spread across long stretches of DNA collectively shape gene activity. Supervised sequence-to-function models such as Enformer (Avsec et al., 2021) and Borzoi (Linder et al., 2025) were early landmarks in this area, integrating hundreds of kilobases of context to simultaneously forecast thousands of epigenetic[21] and transcriptional profiles. More recently, AlphaGenome has extended this approach to joint prediction across more than 4,000 molecular tracks at base-pair resolution with a 1 million base context (Avsec et al., 2026).

Genome-wide variant effect prediction and interpretation (Figure 5D) tackle one of the central challenges in genomics and clinical genetics: determining whether a change in the DNA sequence is likely to be harmful, neutral, or beneficial. Many disease-associated variants lie outside protein-coding regions, making them difficult to interpret with traditional tools. Foundation models approach this by scoring the difference in predicted molecular output between the original and altered sequence, enabling systematic prioritisation of potentially pathogenic variants across the entire genome. Performance varies considerably across approaches: models trained with evolutionary constraints, such as the GPN family, tend to show the strongest zero-shot performance for Mendelian traits, while large-scale models such as Evo 2 represent the current frontier among alignment-free approaches (Veiner and Supek, 2026; Benegas et al., 2024).

RNA structure prediction (Figure 5E) concerns the three-dimensional shape that an RNA molecule adopts once it is produced from a DNA template. This shape governs whether and how the RNA molecule carries out its biological role, influencing processes such as gene splicing, protein production, and cellular localisation. Models pre-trained on large RNA sequence databases learn structural and evolutionary constraints that complement traditional computational approaches. RNA-specialised models such as RiNALMo and ERNIE-RNA have demonstrated competitive performance on secondary structure prediction benchmarks, particularly for moderately divergent RNA families (Zablocki et al., 2025; Penić et al., 2024).

Taxonomic classification for metagenomic samples (Figure 5F) is the task of identifying which organism a given DNA sequence fragment comes from, in samples that may contain genetic material from thousands of species simultaneously. This capability is central to environmental monitoring, infectious disease surveillance, and microbiome research. Models such as DNABERT-S (Zhou et al., 2025b) and GenomeOcean (Zhou et al., 2025a) have shown particular strength in metagenomics binning tasks, while specialised models including METAGENE-1 (Liu et al., 2025) have been developed explicitly for public health applications such as pandemic monitoring and pathogen detection from wastewater samples (Veiner and Supek, 2026; Liu et al., 2025)**.**

Beyond prediction, a growing subset of models supports **generative applications (Figure 5A)**, including the design of novel DNA sequences with specified functional properties[22]. Rather than interpreting existing sequences, these models produce new ones that satisfy target criteria, such as encoding a protein with desired biochemical activity, exhibiting specific regulatory behaviour, or adhering to genome-scale structural constraints. This represents a significant shift in how biological design can be approached, moving from laborious trial-and-error experimentation towards

[21] Epigenetics includes any process that alters gene activity without changing the DNA sequence. Epigenomics refers to the global study of epigenetic changes across the entire genome (Fazzari and Greally, 2010).

[22] The models in question do not generate real or concrete sequences, i.e., they do not physically synthesise them; rather, the term 'generation' in this context refers to their *in silico* production.

sequence generation guided by learned biological principles. Models such as Evo and its successor Evo 2 exemplify this dual capacity, generating sequences from bacteria to yeast chromosomes with biologically plausible properties, including coherent operon structures, accurate codon usage, and realistic transcription factor binding motifs, and even components of CRISPR–Cas systems such as functional Cas9 variants, without these features being explicitly taught (Nguyen et al., 2024; Brixi et al., 2026).

Table 1. Examples of genomics and transcriptomics language models. kb: kilobases; Mb: megabases; bp: base pairs; nt: nucleotides. Parameters refer to the number of trainable model weights. Max Context refers to the maximum input sequence length the model can process.

Model	Parameters	Primary Scope	Max Context	Capabilities
Enformer (Avsec et al., 2021)	~250 million	Human	~200 kb	Gene expression; Genome annotation
Borzoi (Linder et al., 2025)	~186 million	Human and mouse	~500 kb	Gene expression; Genome annotation
HyenaDNA (Nguyen et al., 2023)	Up to 6.6 million	Human	1 million bp	Genome annotation; Variant effect prediction
Nucleotide Transformer v2 (Dalla-Torre et al., 2025)	Up to 2.5 billion	Hundreds of species	12-100 kb	Gene expression; Genome annotation; Variant effect prediction
DNABERT-2 (Zhou et al., 2023)	~117 million	Human and multi-species	Variable	Genome annotation; Variant effect prediction
GPN-MSA (Benegas et al., 2024)	~86 million	Human (multi-species alignment)	128 nt	Variant effect prediction
Caduceus (Schiff et al., 2024)	0.5-8 million	Human	131 kb	Genome annotation; Variant effect prediction
Evo (Nguyen et al., 2024)	7 billion	Prokaryotes and phages	~500 kb	Genome annotation; Variant effect prediction; Sequence design
Evo 2 (Brixi et al., 2026)	1-40 billion	All domains of life	1 million bp	Genome annotation; Variant effect prediction; Sequence design; Gene expression and chromatin accessibility forecasting
RiNALMo (Penić et al., 2024)	650 million	Non-coding RNA (multi-species)	1,022 nt	Genome annotation; RNA structure prediction
AlphaGenome (Avsec et al., 2026)	Not disclosed	Human and mouse	1 million bp	Gene expression; Genome annotation; Variant effect prediction
METAGENE-1 (Liu et al., 2025)	7 billion	Metagenomes (wastewater)	512 tokens	Taxonomic classification for metagenomic samples

Source: JRC's analysis.

3.2. Single-Cell Models

Genomic foundation models treat DNA or RNA as a biological sequence and learn from the linear order of nucleotides. Single-cell foundation models (scFMs), the most prominent class of transcriptomic foundation models, take a fundamentally different approach: rather than reading

along a genome, they read across the gene expression profile of an individual cell. Each cell is represented as an ordered list of genes and their associated activity levels, and the model learns, through self-attention, how patterns of gene co-expression define cell types, states, and responses to perturbation (Baek, Song, and Lee, 2025). In this framing, a cell is analogous to a sentence and its genes to words, with the model learning the grammar of cellular identity from millions of such examples.

In a typical scFM, genes within each cell are ranked by their expression level and fed to the model as an ordered sequence. Some models additionally discretise expression into bins to define token positions. Once tokenised, the gene sequence is processed by transformer blocks in which the self-attention mechanism learns which genes are most informative for determining a cell's state. Pre-training is performed in a self-supervised manner using one of two principal strategies: masked language modelling, in which a proportion of gene tokens are randomly hidden and the model learns to predict them from the remaining context; or autoregressive modelling, in which the model learns to predict the next gene token in sequence. In both cases the objective is the same: to learn the co-expression grammar of cell biology from unlabelled data at scale. Once pre-trained, models can be fine-tuned on task-specific labelled datasets, updating only a subset of parameters to adapt general representations to new tasks at reduced computational cost (He et al., 2026).

The field emerged rapidly from 2022 onwards and has since produced a growing number of models that vary considerably in scale, architecture, and intended application, as discussed in the following section.

The analytical journey in single-cell genomics follows a structured two-stage workflow, progressing from raw data to biological insight (**Figure 6**). The first stage, preprocessing and integration, harmonises data by correcting for technical noise, sparsity, and experimental artefacts, producing clean and comparable representations across datasets and experimental conditions.

The second stage addresses downstream biological analysis, covering the identification of cell types through clustering and annotation, the reconstruction of developmental trajectories, and the inference of gene regulatory networks governing cellular identity. Foundation models sit at the intersection of the two stages, providing transferable representations that can support multiple analytical tasks simultaneously without requiring task-specific model development from scratch (Li et al., 2026; Baek, Song, and Lee, 2025).

The principal capability areas are described below, followed by **Table 2**, which maps selected models to these areas.

Denoising and imputation of missing values is necessary because single-cell RNA sequencing data are inherently sparse: many genes appear undetected in a given cell not due to true absence but because of technical limitations in capture efficiency, a phenomenon known as “dropout”. Foundation models leverage learned co-expression patterns to infer likely expression levels for undetected genes. The field has progressed from early autoencoder-based methods to graph-based approaches that exploit cell-cell similarity, and most recently to diffusion-based generative models that represent the current state of the art in imputation accuracy (Li et al., 2026).

Dimensionality reduction is a foundational preprocessing step that transforms the high-dimensional gene expression space, which may span tens of thousands of genes per cell, into a compact, lower-dimensional representation that is tractable for visualisation, clustering, and downstream analysis. Single-cell data are inherently high-dimensional and noisy, and naive compression approaches such as principal component analyses (PCA) struggle to capture the non-

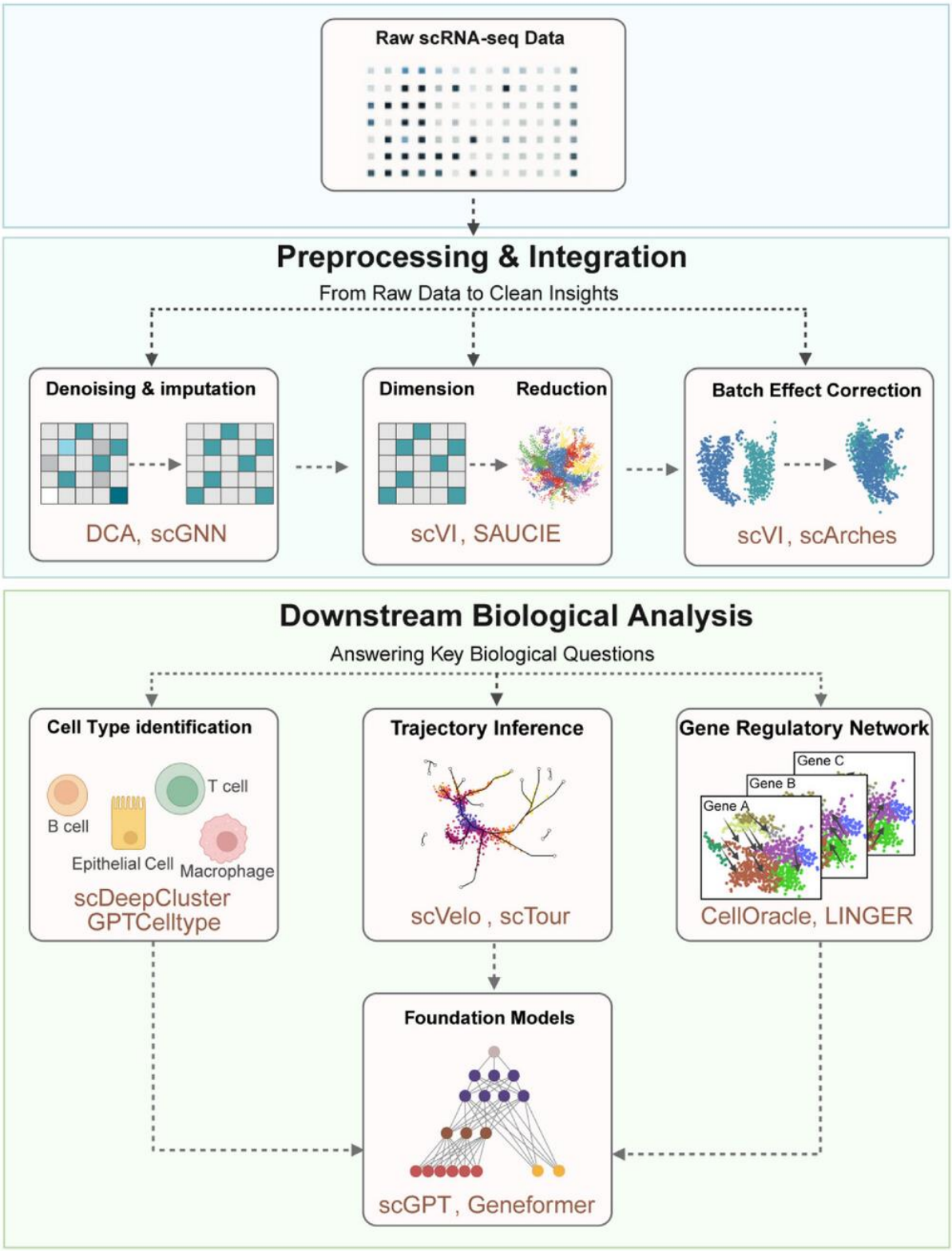
linear structure of biological variation. Deep learning methods, in particular variational autoencoders, learn more expressive latent representations that better preserve the underlying biological variability whilst separating it from technical noise. A recent trend is the integration of external biological knowledge, such as gene regulatory networks or embeddings from language models, into the dimension reduction process to improve the biological interpretability of the resulting latent space (Li et al., 2026). Foundation models contribute here by providing pre-trained cell embeddings that can serve directly as biologically informed low-dimensional representations without task-specific training.

Batch effect correction and data harmonisation addresses the presence of technical variation between datasets generated on different platforms, at different times, or in different laboratories, known as batch effects. These artefacts can cluster cells by their technical origin rather than their biology, obscuring true biological signals. scFMs learn cell representations that are invariant to these sources of variation whilst preserving biological differences, enabling reliable integration of data from multiple sources. Early and highly influential tools such as scVI formalised this by modelling batch identity as a covariate within a probabilistic framework, whilst more recent approaches including scArches enable new datasets to be mapped onto pre-trained references through transfer learning without full retraining (Lopez et al., 2018; Lotfollahi et al., 2022). Foundation models extend this capacity to much larger scales, supporting atlas-level integration across millions of cells from diverse tissues and studies (Li et al., 2026).

Cell-type annotation is the task of assigning each cell to a known or novel category based on its gene expression profile. Before foundation models, this required manual curation and expert review of marker genes, a process that is time-consuming, subjective, and difficult to scale. scFMs embed cells into a shared representational space in which cells of similar biological nature cluster together, enabling automated and reproducible annotation even for unseen datasets. General-purpose large language model approaches such as GPT-4, implemented via the GPTCelltype software package, have demonstrated that cell types can be inferred directly from marker gene lists without task-specific training (Hou and Ji, 2024), whilst specialised models such as RegFormer incorporate gene regulatory network hierarchies into their architecture to improve both accuracy and interpretability (Hu et al., 2026).

Perturbation prediction aims to forecast how a cell's transcriptome will change in response to a genetic or chemical intervention, such as a gene knockout or drug treatment, without running the experiment. This is one of the most ambitious applications of scFMs, as it requires not only knowledge of baseline cellular states but also of how gene regulatory dependencies propagate changes through the expression landscape. Having learned co-expression grammar across vast numbers of cellular states, scFMs can simulate the downstream consequences of perturbations, enabling prioritisation of experimental hypotheses and supporting drug target identification. Geneformer was among the first to demonstrate this capacity at scale, with *in silico* predictions used to identify candidate therapeutic targets in heart disease (Theodoris et al., 2023).

Figure 6. AI-powered single-cell analysis workflow.



Source: (Li et al., 2026)

Trajectory inference and pseudotime analysis address the challenge of reconstructing how cells transition through continuous biological processes such as development, differentiation, or disease progression. Because single-cell sequencing captures only static snapshots of cellular

states, trajectory inference methods must reconstruct the temporal order from cross-sectional data. Foundation model representations have been applied here through RNA velocity approaches such as DeepVelo and veloVI, which extend classical velocity analysis to complex multi-lineage systems and provide uncertainty quantification for inferred dynamics (Gayoso et al., 2024; Chen et al., 2022). More recent generative approaches model cell transitions as continuous vector fields in latent space, offering global trajectory reconstruction with improved handling of branching and cyclical topologies (Li et al., 2026).

Gene regulatory network (GRN) inference involves reconstructing the network of regulatory relationships between genes within a specific cell type or biological context. This is central to understanding how transcription factors and signalling pathways govern cellular identity and response. scFMs learn context-dependent gene embeddings that place genes sharing biological functions or regulatory roles closer together in representational space. Multi-omics approaches such as CellOracle integrate chromatin accessibility data with transcriptomics to ground inferred networks in physical regulatory evidence, whilst foundation model-integrated tools such as scRegNet combine learned embeddings with graph neural network architectures to predict regulatory connections (Kommu et al., 2025; Kamimoto et al., 2023).

Table 2. Examples of single-cell transcriptomic foundation models.

Model	Parameters	Training Data	Key Capabilities
scBERT (Yang et al., 2022)	~10 M	~1.1 M human cells	• Cell-type annotation
Geneformer (Theodoris et al., 2023)	~30 M	~30 M human cells	• Perturbation prediction • Gene regulatory network (GRN) inference • Cell-type annotation
scGPT (Cui et al., 2024)	50 M	~33 M human cells	• Cell-type annotation • Batch integration and data harmonisation • Perturbation prediction • Imputation of missing values
scFoundation (Hao et al., 2024)	~100 M	>50 M human cells	• Batch integration and data harmonisation • Perturbation prediction • Imputation of missing values
GeneCompass (Yang et al., 2024a)	~142 M	>100 M human and mouse cells	• Gene regulatory network (GRN) inference • Cell-type annotation • Perturbation prediction
UCE (Rosen et al., 2026)	650 M	~36 M cells (multi-species)	• Cell-type annotation • Batch integration and data harmonisation
CellFM (Zeng et al., 2025)	800 M	>100 M human cells	• Gene regulatory network (GRN) inference • Perturbation prediction • Cell-type annotation • Batch integration and data harmonisation
TranscriptFormer (Pearce et al., 2026)	~1,077 M	~112 M cells across 12 species	• Cell-type annotation • Batch integration and data harmonisation • Trajectory inference and pseudotime analysis
Nicheformer (Tejada-Lapuerta et al., 2025)	~49 M	~110 M dissociated and spatially resolved cells	• Cell-type annotation • Batch integration and data harmonisation

Source: JRC's analysis.

3.3. Protein Models

Few domains illustrate the transformative power of AI in biology as vividly as proteomics and protein science. Proteins are the molecular machines of life, and understanding their structure, function, and interactions has long been a central ambition of biomedical research. For decades, however, decoding a single protein structure could take years of experimental work.

The arrival of deep learning has fundamentally changed this picture. Models trained on vast repositories of protein sequence and structure data can now predict, design, and interpret proteins at a scale and speed that was simply unimaginable a decade ago.

AI-based protein models now address a broad set of tasks spanning both prediction and design. The capabilities described below represent the principal application areas that have emerged from this field; **Table 3** presents selected models illustrating how these capabilities are implemented in practice.

Protein structure prediction is the task of determining, from the amino acid sequence alone, the three-dimensional shape that a protein adopts. This was long considered one of the hardest problems in biology, since even small proteins adopt extraordinarily specific and complex folds governed by physical forces acting across all their atoms simultaneously. AlphaFold2 released by DeepMind in 2021, changed this picture decisively (Jumper et al., 2021a). By combining evolutionary sequence information with a deep neural network, it achieved accuracy approaching that of experimental methods for many protein families, earning its lead developers (Demis Hassabis and John Jumper, together with David Baker for his work on protein design) a share of the 2024 Nobel Prize in Chemistry[23]. AlphaFold3 extended this further by enabling the prediction of how proteins interact not only with other proteins but also with small molecules, DNA, RNA, and other biomolecules, using a diffusion-based architecture that represents a departure from its predecessor (Abramson et al., 2024). Open and independent alternatives such as RoseTTAFold (Krishna et al., 2024), ESMFold (Lin et al., 2023) and the more recent Chai-1 (Discovery et al., 2024) and Boltz-1 (Wohlwend et al., 2025) have broadened access, with ESMFold in particular offering rapid single-sequence prediction without the need for large sequence alignment databases.

The practical impact of these tools has been amplified by the AlphaFold Protein Structure Database[24], a collaborative resource maintained by EMBL-EBI and Google DeepMind that now provides freely accessible structure predictions for over 200 million protein sequences covering the breadth of catalogued life (Fleming et al., 2025; Varadi et al., 2024). A major recent extension, resulting from a four-way collaboration between EMBL-EBI, Google DeepMind, NVIDIA and Seoul National University, has added millions of AI-predicted protein complex structures (pairs and higher-order assemblies of proteins) to this database, marking the first time such data have been available at this scale[25]. High-confidence homodimer predictions are now openly available, with heterodimer predictions being integrated in subsequent releases. For a comprehensive understanding of the technical and policy implications of this field, a recent JRC report, "The role of artificial intelligence in scientific research: A science for policy, European perspective" (Purificato, Erasmo et al., 2025), provides an in-depth analysis, offering valuable insights into the current state and future directions of protein structure prediction and its applications.

Protein-protein interaction prediction concerns the identification of which proteins bind to each other and, in the structural domain, the prediction of what the resulting complex looks like. Protein language model embeddings have been widely adopted as input features for interaction classifiers, and structure-based tools such as AlphaFold-Multimer (Evans et al., 2022) and AlphaFold3 can now produce atomic-level models of many protein complexes directly from sequence. These tools have

[23] https://www.nobelprize.org/prizes/chemistry/2024/press-release/

[24] https://alphafold.ebi.ac.uk/

[25] https://www.embl.org/news/science-technology/first-complexes-alphafold-database/

enabled systematic prediction of interaction networks at proteome scale, a capability with direct relevance to understanding signalling pathways, designing therapeutic molecules, and interpreting the effects of disease-associated variants.

Variant effect prediction addresses the question of how a change to the amino acid sequence of a protein, as can occur through natural mutation or deliberate engineering, is likely to alter its stability, function, or interactions. Protein language models are particularly well suited to this task: because they are trained to predict masked positions within sequences drawn from across evolutionary history, they implicitly learn which amino acids are tolerated at each position and which are likely to be disruptive. Log-likelihood scores derived from models such as ESM-2 (Lin et al., 2023) and ProtTrans (Elnaggar et al., 2022) can be used in zero-shot fashion (without any task-specific training) to rank mutations by their predicted effect, and benchmarking on the ProteinGym platform, which collects experimental fitness measurements for hundreds of proteins, shows that these approaches are competitive with more specialised methods (Notin et al., 2023). Structure-aware variants such as SaProt, which encode both sequence and three-dimensional context, extend performance further, particularly for stability prediction tasks (Su et al., 2024).

Protein function annotation is the challenge of assigning a biological role to a protein of unknown function. With experimental characterisation of every protein in a proteome being impractical, computational assignment of Gene Ontology terms (a structured vocabulary of molecular functions, biological processes, and cellular components) has long relied on sequence similarity to known proteins. Protein language model representations provide a richer basis for this task: models trained on large corpora encode both evolutionary and structural signal, and the resulting embeddings substantially outperform similarity-based approaches on standard benchmarks including CAFA (Chen et al., 2025). This capability is particularly valuable for the large fraction of proteins from poorly studied organisms for which no close relatives with known function exist.

***De novo* protein design** represents a qualitative shift in what AI-based tools can do: rather than interpreting existing proteins, these models generate new ones. Two approaches have emerged as particularly effective. Backbone generation models such as RFdiffusion, which fine-tunes the RoseTTAFold prediction network on a structure-denoising task analogous to diffusion image generation, can produce novel protein folds, binders to target proteins, enzyme active sites, and symmetric assemblies, often with experimental success rates orders of magnitude higher than earlier physics-based methods (Watson et al., 2023). Sequence design models such as ProteinMPNN (Dauparas et al., 2022) then assign amino acid sequences to these generated backbones, a step known as inverse folding; ProteinMPNN achieves sequence recovery well above earlier computational tools and has been experimentally validated on diverse design challenges using crystallography and cryo-electron microscopy. Multimodal models such as ESM3 integrate sequence, structure, and function into a single generative framework, enabling generation constrained by any combination of these modalities, and have produced functional variants of green fluorescent protein in regions of sequence space far from any natural example (Hayes et al., 2025).

Table 3. Examples of AI-based protein models.

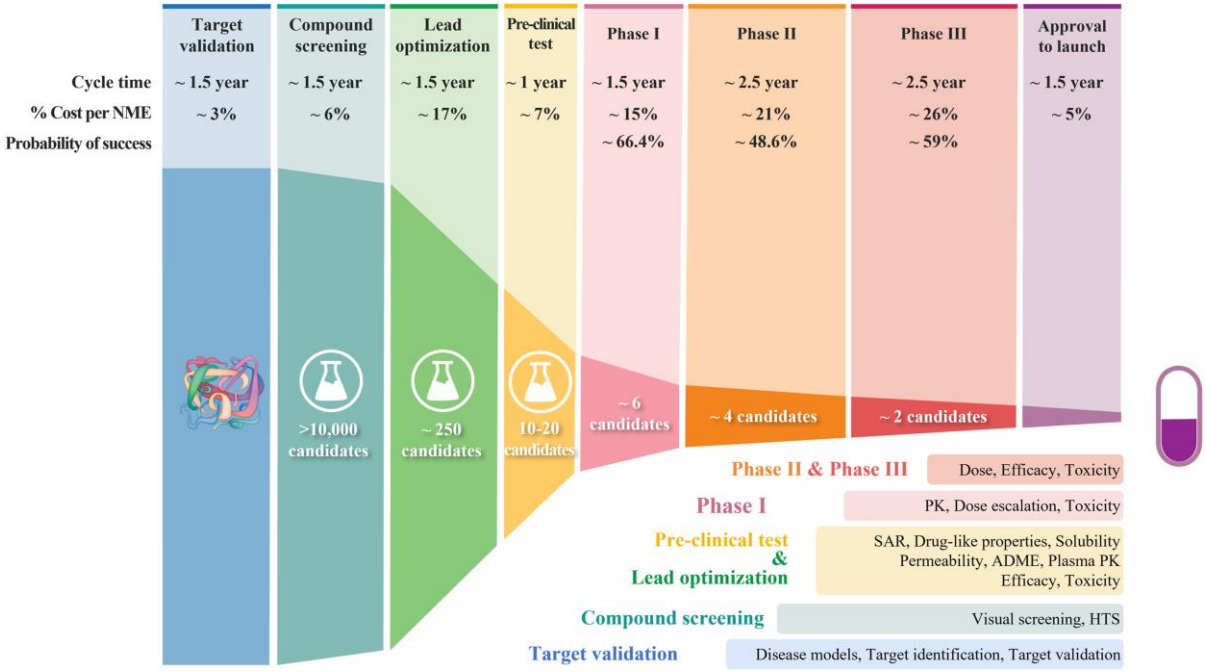
Model	Parameters	Primary scope	Key capabilities
AlphaFold2 (Jumper et al., 2021b)	~93M	Single-chain proteins (all species)	Structure prediction; variant effect prediction
ESMFold (Lin et al., 2023)	690M	Single-chain proteins (all species)	Structure prediction; function annotation; variant effect prediction
AlphaFold3 (Abramson et al., 2024)	Not disclosed	Proteins, nucleic acids, small molecules (all species)	Structure and complex prediction; drug-target modelling
Chai-1 (Discovery et al., 2024)	Not disclosed	Protein complexes and biomolecular assemblies	Structure and complex prediction
Boltz-1 (Wohlwend et al., 2025)	Not disclosed	Biomolecular interactions	Structure and complex prediction; drug-target modelling
ESM-2 (Lin et al., 2023)	Up to 15B	Protein sequences (all species)	Variant effect prediction; function annotation; embeddings for downstream models
ProtTrans / ProtT5 (Elnaggar et al., 2022)	Up to 3B	Protein sequences (all species)	Function annotation; variant effect prediction
ESM3 (Hayes et al., 2025)	98B	Proteins (sequence, structure, function)	Protein design; function annotation; multimodal generation
RFdiffusion (Watson et al., 2023)	~60M (fine-tuned)	Protein backbone structures	*De novo* design; binder design; enzyme scaffolding
ProteinMPNN (Dauparas et al., 2022)	~1.6M	Protein backbones	Sequence design (inverse folding)
SaProt (Su et al., 2024)	~650M	Protein sequences and structures	Variant effect prediction; function annotation

Source: JRC's analysis.

3.4. Drug Discovery and Development Models

Drug discovery and development is the process of identifying and advancing new medicines from initial scientific hypothesis through to clinical use. The process is highly demanding: it typically requires more than a decade (**Figure 7**) and substantial financial investment, on the order of billions of euro, to bring a single compound to market (Wouters, McKee, and Luyten, 2020), and success rates remain limited, with fewer than 14% of candidates that enter clinical trials receiving regulatory approval (Wong, Siah, and Lo, 2019). The scale of the challenge is partly a consequence of the vast chemical space that must be explored, estimated to encompass more than 10^{60} possible drug-like molecules (Bohacek, McMartin, and Guida, 1996). Computational approaches have supported drug discovery for several decades, and the application of artificial intelligence has accelerated their scope and impact across every stage of the pipeline.

Figure 7. Estimate of duration, cost percentage, and probability of success of the steps of drug discovery and development process. NME: new molecular entity; HTS: high-throughput screening; SAR: structure-activity relationship; ADME: absorption, distribution, metabolism and excretion; PK: pharmacokinetics.



Source: (Sun et al., 2022)

AI methods address a wide range of tasks across the drug discovery and development pipeline, from early target identification and virtual screening to lead optimisation, toxicity prediction, and clinical trial design (see **Figure 7**). In essence, these models learn patterns from large collections of chemical, biological, and clinical data, and then apply that learning to evaluate or generate candidate molecules in ways that would be impractical through conventional experimental or computational means alone. The capabilities described below represent the principal application areas that have emerged; **Table 4** lists selected models illustrating how these capabilities are implemented in practice.

Target identification and structure characterisation is the first step of the drug discovery process, in which researchers seek to identify a biological molecule, typically a protein, whose activity is linked to a disease, and to characterise its three-dimensional structure. Deep learning-based structure prediction tools, most notably AlphaFold2 (Jumper et al., 2021b), have substantially accelerated this phase by reducing reliance on time-consuming structural biology experiments, with the AlphaFold Protein Structure Database now providing structural models for over 200 million proteins. Building on this, AF2RAVE (Gu, Aranganathan, and Tiwary, 2024) combines structure prediction with enhanced-sampling molecular dynamics to access alternative protein conformations that are often essential for drug binding but are not captured by structure prediction alone.

Structure-based virtual screening and pocket characterisation involves computationally evaluating candidate binding sites and molecules to identify those most likely to support productive drug binding. PocketVec (Comajuncosa-Creus et al., 2024) generates descriptors for protein binding pockets through inverse virtual screening, and has been used to identify over 32,000 candidate binding pockets across the human proteome. vScreenML 2.0 (Andrianov, Haroldsen, and Karanicolas, 2024) is a machine learning classifier that reduces the high false-positive rates typically seen in

structure-based virtual screening, helping to focus experimental resources on the most promising candidates. More recently, co-folding approaches such as AlphaFold3 and Boltz-1 have shown potential as docking engines, directly predicting protein-ligand bound complexes, though their generalisation to chemically novel compounds remains an active area of evaluation (Masters, Mahmoud, and Lill, 2025). This reduces reliance on high-throughput experimental screening and focuses experimental resources on the most promising candidates.

Structure-based *de novo* molecule generation has become one of the most active areas of AI-driven drug discovery, with models generating novel candidate molecules directly within a target protein pocket, rather than selecting from existing libraries. DecompDiff (Guan et al., 2024) is a diffusion model that decomposes ligands into functional fragments to improve binding affinity and chemical validity, while FlexSBDD (Zhang, Wang, and Liu, 2024) extends this approach by allowing the protein structure itself to flex during generation, better reflecting the dynamic nature of binding.

Lead optimisation and Quantitative Structure-Activity Relationship (QSAR) modelling addresses the refinement of candidate compounds to improve their therapeutic activity, selectivity, and drug-like properties. REINVENT 4 (Loeffler et al., 2024), an open-source generative framework developed at AstraZeneca, uses recurrent neural networks and transformers with reinforcement learning to guide molecule generation towards desired property profiles and is among the most widely deployed tools in industrial lead optimisation. Chemistry42 (Ivanenkov et al., 2023) is an industrial AI platform that integrates generative design with medicinal chemistry filters; it has been used to design and experimentally validate novel inhibitors against targets such as DDR1 and CDK20, illustrating how AI-generated leads can progress to laboratory testing within a substantially compressed timeframe compared with conventional design cycles.

Molecular representation and language-based design. A separate strand of work applies large-scale pretraining to chemical structures, in a manner analogous to language models. MegaMolBART[26], trained on over a billion molecules from the ZINC database, learns general-purpose molecular representations that can be used both for property prediction and for generating novel molecules by sampling its learned chemical space.

Drug repurposing and target discovery identifies new therapeutic applications for existing compounds, or novel targets for diseases with limited treatment options. TxGNN (Huang et al., 2024a) is a graph-based foundation model trained on a medical knowledge graph spanning over 17,000 diseases, supporting zero-shot repurposing predictions whose outputs have been shown to align with off-label prescribing patterns observed in clinical practice. Precious3GPT (Galkin et al., 2024) takes a complementary, multi-omics approach, integrating transcriptomic and other biological data across species and tissues to support target identification and to predict how candidate compounds might affect ageing-related biological processes.

QSAR models, together with dedicated **ADMET** prediction tools assessing **A**bsorption, **D**istribution, **M**etabolism, **E**xcretion, and **T**oxicity, remain foundational to lead optimisation in real-world settings. Nonetheless, these approaches are not well represented in the Epoch AI dataset examined in this report, and so they receive less attention here than their real-world importance would warrant.

[26] https://github.com/NVIDIA/MegaMolBART

Table 4. Examples of drug discovery models.

Model	Parameters	Training Data	Key Capabilities
AF2RAVE (Gu, Aranganathan, and Tiwary, 2024)	Not reported	No new training; uses AlphaFold2 predictions as input for enhanced-sampling MD simulations	Target identification; structure characterisation
PocketVec (Comajuncosa-Creus et al., 2024)	Not reported	Inverse virtual screening of ~32,000 binding pockets across the folded human proteome	Structure-based virtual screening; pocket characterisation
vScreenML 2.0 (Andrianov, Haroldsen, and Karanicolas, 2024)	Not reported	7,224 protein-ligand samples (binary classification)	Structure-based virtual screening; pocket characterisation
DecompDiff (Guan et al., 2024)	Not reported	100,000 protein-ligand complexes (CrossDocked2020)	Structure-based *de novo* molecule generation
FlexSBDD (Zhang, Wang, and Liu, 2024)	Not reported	reported~140,000 protein-ligand pairs, augmented via structure relaxation	Structure-based *de novo* molecule generation
REINVENT 4 (Loeffler et al., 2024)	Not reported	~1M molecules (prior model); ~200B molecular pairs (Mol2Mol)	Lead optimisation and multi-objective design
Chemistry42 (Ivanenkov et al., 2023)	Not reported	Not reported	Lead optimisation and multi-objective design
TxGNN (Huang et al., 2024a)	Not reported	Medical knowledge graph: 123,527 nodes and 8,063,026 edges across 17,080 diseases	Drug repurposing and target discovery

Source: JRC's analysis.

3.5. Technological Maturity and Readiness of Biology AI Models

In the context of this report, the maturity of AI models in biology refers to the extent to which these models are robust, validated, interpretable, scalable, and ready for use beyond exploratory research. A mature AI system is one that not only performs well in controlled settings but can be reliably deployed in real-world environments, whether in drug development pipelines, clinical decision support, or industrial biomanufacturing. Assessment therefore requires consideration not only of reported performance but also of reproducibility, reliability, robustness, transparency and the context of validation. For policymakers, assessing maturity is essential for three reasons: it informs funding priorities, guides regulatory preparedness, and helps anticipate where AI-driven biotechnology will create opportunities and risks in the near term.

Assessing maturity in biology is uniquely challenging because biological systems are noisy, heterogeneous, and often lack comprehensive ground truth. Unlike software systems where correctness can sometimes be formally verified, publicly available biological AI models operate in

domains where "correct" outputs may be context-dependent, probabilistic, or simply unknown until experimentally validated. These characteristics make it difficult to evaluate not only how well a model performs, but also how ready it is for reliable deployment in practice.

The interpretation of maturity may also be influenced by differences in experimental regulatory or non-regulatory laboratory procedures (OECD, 2018), validation endpoints, target complexity and data availability across biological domains.

Currently, no established reference framework exists for assessing the maturity of AI biology models across their diverse applications. This section proposes combining two complementary frameworks: a domain-specific three-tier maturity classification originally developed for AI protein design, and the well-established Technology Readiness Levels (TRLs) used across EU research policy. Together, these provide both an intuitive classification system and standardised developmental milestones that can be applied across the examples presented in Section 3.5.2.

3.5.1. Maturity Assessment Frameworks

3.5.1.1. Domain-Specific Maturity Classification

A useful conceptual framework for assessing the readiness of AI biology models was developed by Koh et al. (2025) in their analysis of AI protein design toolkits. Whilst originally formulated for protein design, the three-tier classification they propose could be applied more broadly across AI biology applications to characterise developmental stage and deployment readiness.

The framework distinguishes **three levels of maturity**: **nascent** (computational proof-of-concept only), **advanced** (experimental validation achieved), and **mature** (production-ready deployment). Specific to the AI-drive protein design applications, this three-tier schema maps onto seven “toolkits” (T1–T7), spanning different sub-toolkits such as protein structure prediction, protein function prediction or protein sequence among others (see **Table 5**).

Table 5. Maturity of AI-driven protein design tools, reflecting real-world validation and deployment readiness.

Maturity	Description	Toolkit (T1-T7)	Sub-toolkits
Nascent	Computational proof-of-concept	T3. Protein function prediction	T3c. Post-translational modification
		T4. Protein sequence generation	T4b. Function-to-sequence generation
		T6. Virtual screening	T6b. Developability assessment
Advanced	Experimental validation	T2. Protein structure prediction	T2d. Conformational dynamics modelling
		T3. Protein function prediction	T3a. Gene Ontology (GO) annotation T3b. Binding site identification
		T4. Protein sequence generation	T4a. Evolution-guided generation
		T5. Protein structure generation	T5a. Template-based structure design T5b. Generative backbone design T5c. Sequence–structure co-design
		T7. DNA synthesis	DNA synthesis
Mature	Production-ready deployment	T1. Protein database search	T1b. Sequence alignment T1b. Structure alignment
		T2. Protein structure prediction	T2a. Protein folding T2b. Biomolecular co-folding T2c. Structure stability prediction
		T4. Protein sequence generation	T4c. Structure-to-sequence generation
		T6. Virtual screening	T6a. Binding and functional activity prediction

Source: adapted from (Koh et al., 2025)

Importantly, maturity is not binary or uniform across a system's components. A single AI platform may combine mature components (such as protein structure prediction) with nascent capabilities (such as post-translational modification prediction). Overall readiness is determined by the weakest link in the chain, particularly for systems that must complete multiple steps to deliver value.

This three-tier framework provides an intuitive way to characterise AI biology models that maps well onto the development lifecycle.

3.5.1.2. Technology Readiness Levels (TRLs)

Technology Readiness Levels (TRLs) provide a complementary staged framework for evaluating how far a technology has progressed from basic research to operational deployment. TRLs situate projects along a staged spectrum, from basic principles (TRL1) to operational deployment in real world (TRL9) (**Table 6**).

Originally developed by NASA for aeronautical and space projects, TRLs were subsequently adopted globally, and extended as a general framework for evaluating the maturity of a wide range of technologies and innovation projects (Martínez-Plumed et al., 2022). However, measures of technological maturity do not necessarily translate in clinical or societal readiness. In this sense,

TRLs can describe how "built" or "functional" a system is, but they do not fully capture whether the system is appropriate for deployment within regulated healthcare ecosystems.

Table 6. Summary of Technology Readiness Levels (TRLs) according to several characteristics.

Environment	Goal	Product / Evaluation	Key Deliverables	TRL	Description
Laboratory	Research	Proof of concept	Scientific articles published on the principles of the new technology	TRL 1	Basic principles observed
			Publications or references highlighting the applications of the new technology	TRL 2	Technology concept formulated
			Measurement of parameters in the laboratory	TRL 3	Experimental proof of concept
Simulation	Development	Prototype	Results of tests carried out in the laboratory	TRL 4	Technology validated in lab
			Components validated in a relevant environment	TRL 5	Technology validated in relevant environment
			Results of tests carried out at the prototype in a relevant environment	TRL 6	Technology demonstrated in relevant environment
Operational	Implementation		Results of prototype-level tests carried out in the operating environment	TRL 7	System prototype demonstrated in operational environment
		Commercial product/service (certified)	Results of system tests in final configuration	TRL 8	System complete and qualified
		Deployment	Final reports in working condition or actual mission	TRL 9	Actual system proven in operational environment

Source: (Martinez, Gomez, and Hernández-Orallo, 2020)

TRLs are routinely used in EU research funding to situate projects along this staged spectrum and provide a common language for comparing technological maturity across different domains. The framework helps identify where additional development effort is needed and provides clear milestones for progression from basic research through to market deployment.

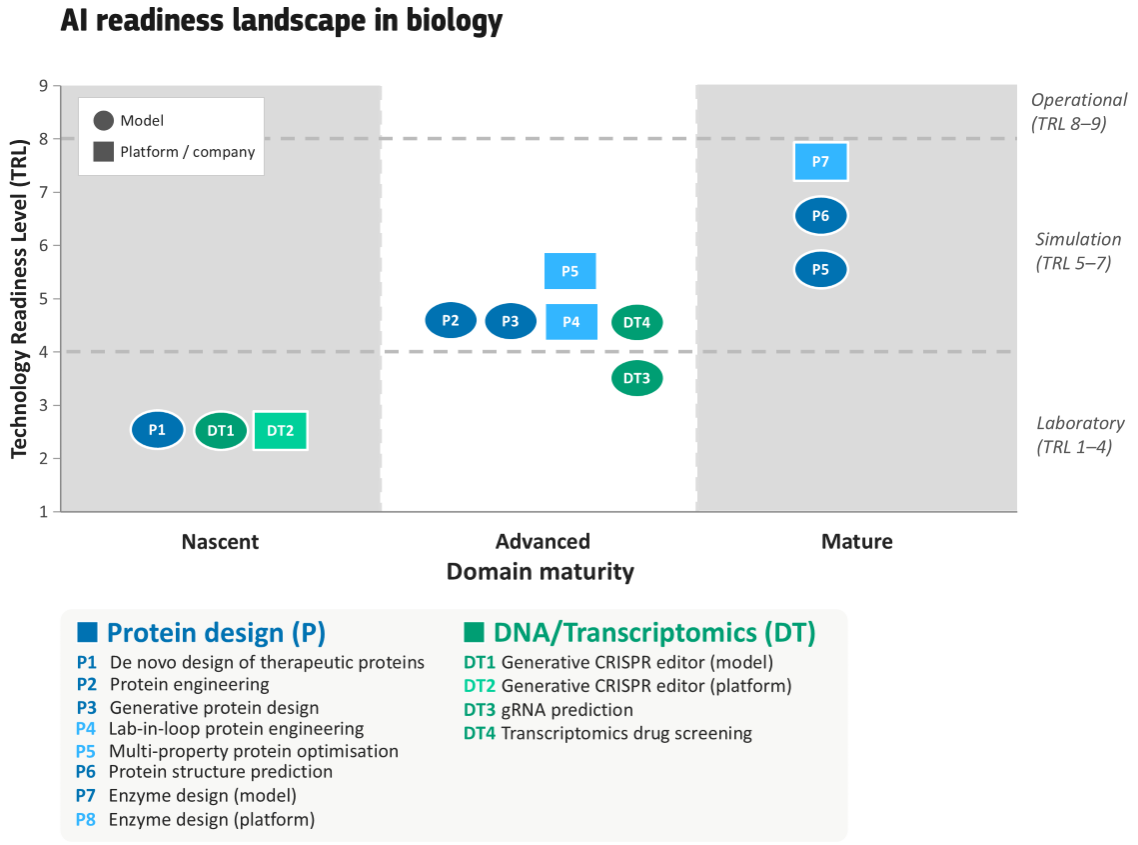
3.5.2. Integrating Domain-Specific Maturity Classification with TRL Frameworks

The domain-specific maturity classification and TRL framework measure different things. The first measures biological validation and deployment readiness within a domain, capturing how far a biological AI model has progressed from computational concept to experimental confirmation to operational integration. TRLs, by contrast, measure technical development stage against standardised operational criteria, independently of the biological domain. This distinction also highlights that benchmark performance (reliability), scientific maturity (relevance), deployment readiness and regulatory readiness should not automatically be interpreted as equivalent concepts.

Because they capture different dimensions of progress, the two scales frequently diverge, giving rise to the **maturity paradox**. A model can be domain-mature, fully embedded in research practice and transformative within its field, whilst sitting at a modest TRL if it has never been formally qualified for regulated deployment. The same divergence appears when distinguishing individual models from the platforms that deploy them: a research model may be domain-mature and widely adopted yet sit at a low TRL, whereas an industrial platform built on that same model may reach a higher TRL by virtue of the engineering, validation, and integration work that surrounds it. **The implications of the maturity paradox are further discussed in Section 5.6.**

Figure 8 illustrates how both frameworks operate in practice across a selection of biological AI models and industrial AI-driven platforms, analysed in Sections 3.5.2.1 and 3.5.2.2 respectively.

Figure 8. Illustrative AI readiness landscape in biology illustrated by tasks from Sections 3.5.2.1 and 3.5.2.2 positioned by domain-specific maturity (x-axis: nascent, advanced, mature) and Technology Readiness Level (y-axis: TRL 1–9). Placements reflect the authors' expert assessment based on the frameworks described in this report and should be interpreted as indicative rather than definitive classifications.



Source: JRC's analysis.

3.5.2.1. Maturity of Biological AI Models

Table 7 reports the integrated maturity of individual AI biology models and span the full maturity spectrum, from nascent proof-of-concept to operational research infrastructure.

The most important pattern is the consistent gap between domain maturity and TRL: models that are transformative within their biological domain frequently sit at a lower TRL than their scientific impact might suggest. AlphaFold2 and AlphaFold3 are the clearest illustration. They are domain-mature, yet at TRL 5–6 because no regulatory authority has formally accepted their outputs as a certified component of a drug development submission.

A second pattern is the clustering of most models in the advanced/TRL 4–5 band. Generative protein design, guide RNA prediction, lab-in-loop engineering, transcriptomics-guided drug screening, and disease progression modelling all occupy this range: experimentally validated and producing measurable biological value but not yet embedded in regulated or commercial operational environments.

The nascent category contains two distinct cases: AI-generated genome editors and *de novo* protein design, both involve generative models producing novel molecular entities with proof-of-concept validation but no clinical pathway.

Table 7. Maturity of selected AI biology models.

Task	Model	Domain maturity	TRL	Key evidence
Protein structure prediction	AlphaFold2 / AlphaFold3 (Jumper et al., 2021b; Abramson et al., 2024)	Mature	5-6	2024 Nobel Prize in Chemistry; >200M structures in open database; routine research use. Not formally qualified for regulated deployment.
Enzyme design	ProteinGAN (Repecka et al., 2021)	Mature	6-7	First peer-reviewed generative AI enzyme design; GAN trained on natural sequence diversity; validated industrial function.
Generative protein sequence design	RFdiffusion / ESM3 (Watson et al., 2023; Hayes et al., 2025)	Advanced	4-5	Double-digit experimental success rates for binder and antibody design; outcomes predominantly in vitro (Kosonocky et al. 2026).
Guide RNA activity and off-target prediction	DeepCRISPR / DeepHF (Chuai et al., 2018; Wang et al., 2019)	Advanced	3-4	Outperform rule-based scoring on benchmarks; peer-reviewed; research use only; not validated in clinical workflows.
Protein engineering	ESM-2 + Multilayer Layer Perceptron fitness predictor (Zhang et al., 2025a)	Advanced	4-5	ESM-2 zero-shot generation + supervised active learning; 2.4-fold enzyme activity improvement in 4 rounds / 10 days.
Transcriptomics-guided drug screening	Transcriptomic active learning model (DeMeo et al., 2025)	Advanced	4-5	Paired phenotypic and transcriptomic measurements with active learning; 2-fold hit-rate increase over standard screening.
AI-generated genome editor	OpenCRISPR-1 model (Ruffolo et al., 2025)	Nascent	2-3	Proof-of-concept in human cells or in vitro; no clinical translation; generative PLMs trained on Cas.
Novel protein design	Chroma (Ingraham et al., 2023)	Nascent	2-3	>300 designs experimentally tested with high soluble-expression rates, two confirmed by X-ray crystallography, though no functional/therapeutic validation.

Source: JRC's analysis.

3.5.2.2. Maturity of Biological AI Models Embedded in Industrial Platforms

Industrial AI-driven platform maturity reflects the degree to which a model has been integrated into a validated, operational workflow, whether commercial or clinical. It is typically higher than the underlying model's TRL because it includes engineering, quality assurance, regulatory engagement, and commercial integration. Where an industrial platform has not yet surpassed the model's own TRL, this reflects early-stage commercialisation.

Table 8 covers four industrial platforms and company deployments. The consistent finding is that platforms sit one TRL tier higher than the underlying models they deploy, because the engineering, quality assurance, and commercial integration surrounding a model adds operational readiness that

the model alone does not possess. Biomatter, dedicated to industrial enzyme design, is the clearest example of full alignment: both domain-mature and TRL 7–8, with revenue-generating commercial contracts constituting genuine operational deployment in an industrial environment.

Cradle-1 and MULTI-evolve illustrate the more typical advanced-stage profile: multi-partner validation and measurable performance gains, but with programmes still in preclinical or research phases. MULTI-evolve's placement at Advanced rather than Mature should also be read with some caution. An independent technical commentary[27] has since questioned how well its predictions actually work when tested on data the model hadn't seen before. MULTI-evolve's predictions reportedly performed no better than a much simpler method that just adds up the effects of individual mutations.

Profluent Bio's OpenCRISPR-1 stands apart as the only nascent platform entry, reflecting the large gap between a proof-of-concept genome editor functional in human cells and any commercially or clinically deployable product.

[27] https://dewitt-lab.github.io/posts/2026-04-28-multievolve-baselines/

Table 8. Maturity of selected industrial AI-driven platforms deploying biological AI models.

Task	Platform / company	Domain maturity	TRL	Key evidence
Enzyme design	Biomatter[28] Intelligent Architecture™ (LT)	Mature	7-8	Commercial contracts with BASF, Thermo Fisher Scientific, Neogen[29]. Only example where domain maturity and TRL fully align; certified revenue-generating deployment.
Multi-property protein optimisation	Cradle Bio[30], Cradle-1 (NL/CH)	Advanced	5-6	(Bixby et al., 2026), bioRxiv; foundation protein language model + active learning; 4-7x faster than rational design; validated across antibodies, enzymes, gene-editing tools; multi-pharma partnerships.
Lab-in-loop protein engineering	Arc Institute MULTI-evolve (US)[31]	Advanced	4-5	(Tran et al., 2026); protein language model + focused experimental testing; compresses engineering from months to weeks; controlled lab setting.
AI-generated genome editor design	Profluent Bio OpenCRISPR-1 (US)[32]	Nascent	2-3	(Ruffolo et al., 2025); first fully AI-generated Cas9-like editor functional in human cells; no clinical translation; platform at pre-commercial stage.

Source: JRC's analysis.

[28] https://biomatter.ai/

[29] https://biomatter.ai/resources/news/biomatter-raises-a-e6-5-million-seed-round-to-unlock-the-power-of-generative-ai-for-enzyme-design/

[30] https://www.cradle.bio/

[31] https://arcinstitute.org/news/multi-evolve

[32] https://www.profluent.bio/

4. How Do Biology AI Models Work?

Having outlined what biology AI models can do, this section turns to how they work. Three elements largely determine the capabilities described in the previous section: the **training data** that shape what a model learns, the model design and **architecture** that determine how it learns, and the **computational infrastructure** that makes training and deployment feasible at scale. These three elements jointly shape questions of access, reproducibility and competitiveness as well as the confidence that can be placed in reported capabilities and their translation into practice that are central to the European policy debate on AI in biology.

4.1. Training Data

Figure 9 provides a high-level overview of 75 biological datasets most frequently used to train contemporary AI4Bio models (see Annex 5 for more details). The figure uses a treemap visualisation, where each rectangle represents a distinct dataset or database and its area reflects the number of models that report using it as a training source. This visualisation is intended as an illustrative snapshot rather than an exhaustive inventory. It captures the relative prominence of some of the most widely used and well-established biological data resources, based on available reporting, and should be interpreted as indicative rather than complete.

The treemap highlights the **central role of protein-focused resources** in current biological AI development. Large clusters correspond to datasets such as the Protein Data Bank (PDB)[33], which archives experimentally determined three-dimensional structures of proteins and other biomolecules and is used by 26 models, and UniProt[34] and its derived sequence clusters (UniRef[35]), which provide curated protein sequence and functional annotation data and are used by 20 models. Notably, the AlphaFold Protein Structure Database[36], which consists entirely of AI-predicted protein structures, is used as a training dataset by 10 models. This illustrates the **growing role of synthetic AI-generated data, which refers to computationally predicted structures rather than experimental laboratory data, in biological AI development**. Together, these datasets form the backbone of training data for a relevant fraction of existing AI4Bio models.

By contrast, datasets representing other biological domains occupy smaller areas in the treemap. These include chemical structure and bioactivity databases (such as chemical bioactivity databases like PubChem[37] and ChEMBL[38], which annotate small molecules with their biological activity against molecular targets), the large repository of reference DNA sequences (RefSeq[39]), therapeutic and clinical annotations, and integrative biological knowledge resources. This disparity highlights the concentration of current AI4Bio development around protein sequence and structure resources. As a result, **progress in the field remains primarily anchored to** well-standardised, high-coverage

[33] https://www.rcsb.org/

[34] https://www.uniprot.org/

[35] https://www.uniprot.org/uniref

[36] https://alphafold.ebi.ac.uk/

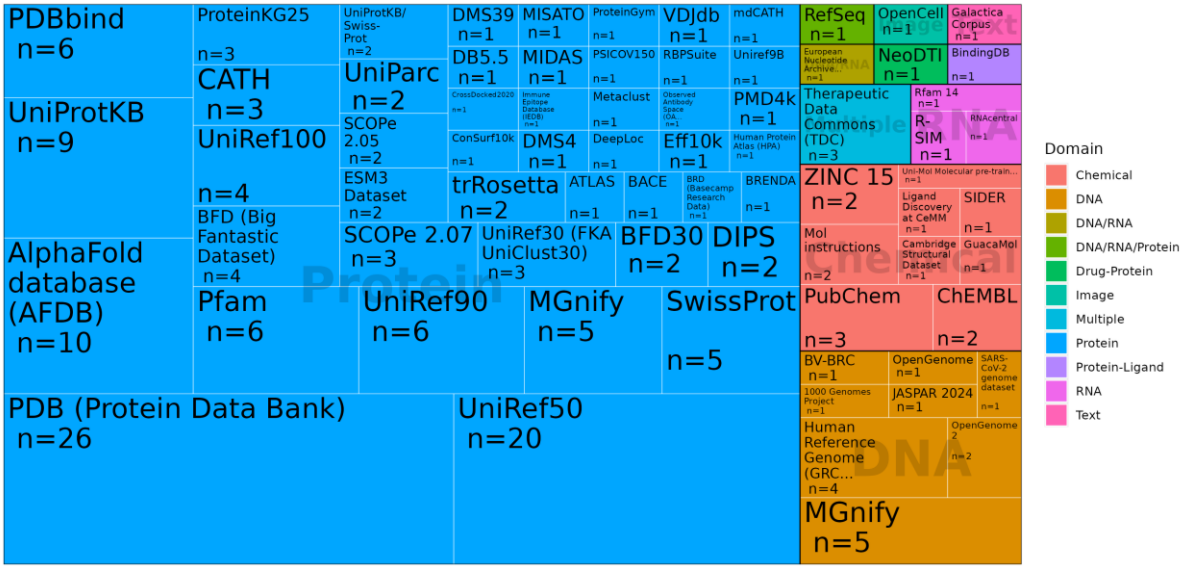
[37] https://pubchem.ncbi.nlm.nih.gov/

[38] https://www.ebi.ac.uk/chembl/

[39] https://www.ncbi.nlm.nih.gov/refseq/

protein data infrastructures, while multi-modal, clinical-grade, and small-molecule datasets remain comparatively under-represented.

Figure 9. Training datasets used by 134 biological AI models retrieved from the Epoch AI dataset.



Source: JRC's analysis.

The most commonly used data sources to train **DNA and RNA AI models** include curated reference genomes such as the human reference genome (GRCh38) (Schneider et al., 2017), multi-species assemblies hosted through databases including NCBI RefSeq[40] and Ensembl[41], and large-scale human variation resources such as the 1000 Genomes Project[42] and gnomAD[43]. These datasets capture the genetic diversity found across human populations and across the tree of life, providing the foundation for models intended to generalise beyond a single species.

Metagenomic data, derived from environmental samples such as ocean water, soil, and gut microbiomes, has become an increasingly important training source, particularly for models targeting microbial diversity and public health applications. Repositories such as MGnify[44] and the European Nucleotide Archive (ENA)[45] have been central to this effort. Evo 2, for instance, was trained on OpenGenome2[46], a curated dataset of approximately 9.3 trillion nucleotides spanning more than 128,000 genomes across prokaryotes, archaea, eukaryotes, and viruses (Brixi et al., 2026). Models intended for regulatory and functional prediction, such as AlphaGenome, additionally incorporate aligned multi-omic experimental tracks from large consortia including ENCODE[47], GTEx[48],

[40] https://www.ncbi.nlm.nih.gov/refseq/

[41] https://www.ensembl.org/

[42] https://www.internationalgenome.org/

[43] https://gnomad.broadinstitute.org/

[44] https://www.ebi.ac.uk/metagenomics/

[45] https://www.ebi.ac.uk/ena/browser/home

[46] https://huggingface.co/datasets/arcinstitute/opengenome2

[47] https://www.encodeproject.org/

[48] https://www.gtexportal.org/home/

and the 4D Nucleome project[49], linking DNA sequence to measurements of gene activity, chromatin state, and three-dimensional genome organisation (Avsec et al., 2026).

Single-cell AI models are pre-trained on large aggregated collections of single-cell RNA sequencing (scRNA-seq) data drawn from public repositories. The principal sources include the Chan Zuckerberg CELLxGENE consortium[50], the Gene Expression Omnibus (GEO)[51], the Human Cell Atlas[52], and specialised collections such as the Cancer Cell Line Encyclopedia (CCLE)[53]. Training corpora range from 30 million cells in the earliest models to over 100 million in the most recent, with token counts after gene- or bin-based tokenisation typically exceeding 10^{12} across the largest datasets.

The choice and curation of training data have significant consequences for model quality. While scale is often presented as a key advantage, downsampling and ablation studies have suggested that performance can plateau beyond a certain corpus size, and that carefully curated datasets with consistent annotation standards may offer greater value than indiscriminately aggregated collections (Nadig et al., 2025; DenAdel et al., 2025). Variable quality, inconsistent cell-type ontologies, and batch effects across repositories are recurring concerns that mirror those identified for genomic models.

Protein AI models draw from a distinct but interlinked set of training resources depending on their objectives. Sequence-based protein language models are typically trained on large-scale repositories of protein sequences, with UniProt and its component databases (UniRef50, UniRef90) serving as the primary sources. UniProt currently archives over 250 million sequences, the vast majority of which are computationally inferred from genome sequencing rather than experimentally verified. ESM-2 , for example, was trained on the full UniRef50 and UniRef90 datasets encompassing hundreds of millions of sequences spanning all known forms of life (Lin et al., 2023). ProtTrans models were trained on the Big Fantastic Database (BFD)[54], a cluster-merged dataset of approximately 2.5 billion sequences (Elnaggar et al., 2022).

Structure prediction models additionally rely on the PDB, the global repository of experimentally determined protein structures, which contains over 220,000 entries. Although this is a large corpus by experimental standards, it represents a tiny fraction of sequence space and is biased towards soluble, well-behaved proteins amenable to crystallography and cryogenic electron microscopy. AlphaFold2 (Jumper et al., 2021b) and RoseTTAFold (Baek et al., 2021) use multiple sequence alignments (MSAs) constructed from public databases alongside PDB structures as training inputs, integrating co-evolutionary information across related proteins to infer structural constraints. AlphaFold3 additionally incorporated data on protein complexes with nucleic acids and small molecules from the PDB, enabling its expanded scope (Abramson et al., 2024).

Generative and design-focused models are often trained on structural data from the PDB supplemented by the AlphaFold Protein Structure Database, which as of 2024 contains over 200 million predicted structures and constitutes by far the largest available collection of protein three-

[49] https://4dnucleome.org/

[50] https://cellxgene.cziscience.com/

[51] https://www.ncbi.nlm.nih.gov/geo/

[52] https://www.humancellatlas.org/

[53] https://sites.broadinstitute.org/ccle

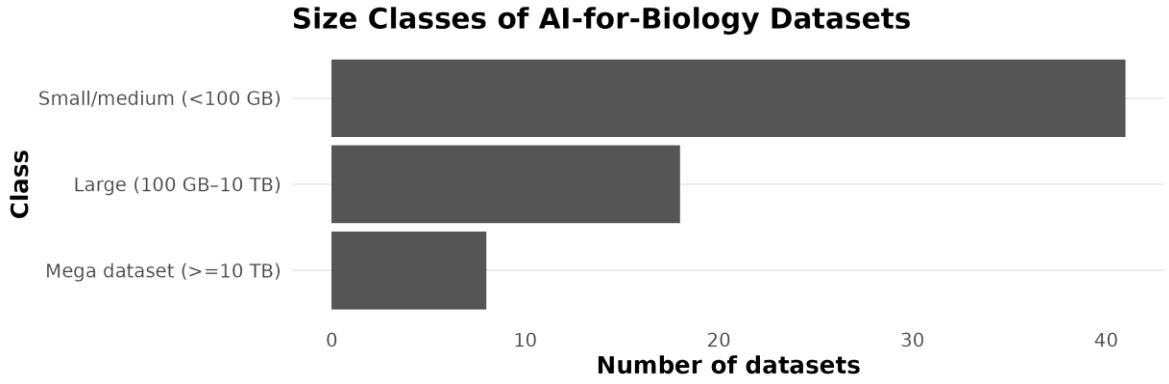
[54] https://bfd.mmseqs.com/

dimensional models (Varadi et al., 2024). RFdiffusion was pre-trained on approximately 140,000 PDB structures and fine-tuned on a structural denoising task (Watson et al., 2023), while ProteinMPNN was trained on around 19,700 high-resolution PDB structures (Dauparas et al., 2022).

Drug discovery AI models draw on a diverse and growing ecosystem of public databases spanning chemical structures, biological activity measurements, protein sequences, toxicological profiles, and clinical outcomes. The most widely used chemical databases include PubChem (~118 million compounds), ChemSpider[55] (~130 million structures), and ZINC (~750 million commercially available compounds for virtual screening) (Sterling and Irwin, 2015). Pharmacological and bioactivity data are primarily sourced from ChEMBL (~2.9 million bioactive compounds) and DrugBank[56] (~500,000 drugs and drug products), while protein structure data are drawn from the PDB. Biological pathway and protein function information is provided by UniProt and KEGG[57]. Together, these resources constitute the data infrastructure underpinning most AI models in drug discovery.

The therapeutics space now extends well beyond small molecules to include peptides, cyclic peptides, oligonucleotides, and antibody-based modalities, each with distinct ADMET determinants and chemical representations (Tanihata and Iwata, 2026). While small molecules represent more than 90% of marketed drugs and constitute the primary focus of most AI-based discovery efforts, AI has also been increasingly applied to biologics and medium-sized molecules, broadening the scope of relevant training data accordingly.

Figure 10. Size classes of 75 biological training datasets.



Source: JRC's analysis.

Figure 10 summarises dataset size classes. Most training datasets fall into the small- to medium-scale category (below 100 GB), while a smaller number qualify as large datasets (100 GB–10 TB), and only a limited subset reach the scale of "mega" datasets (above 10 TB).

[55] https://www.chemspider.com/

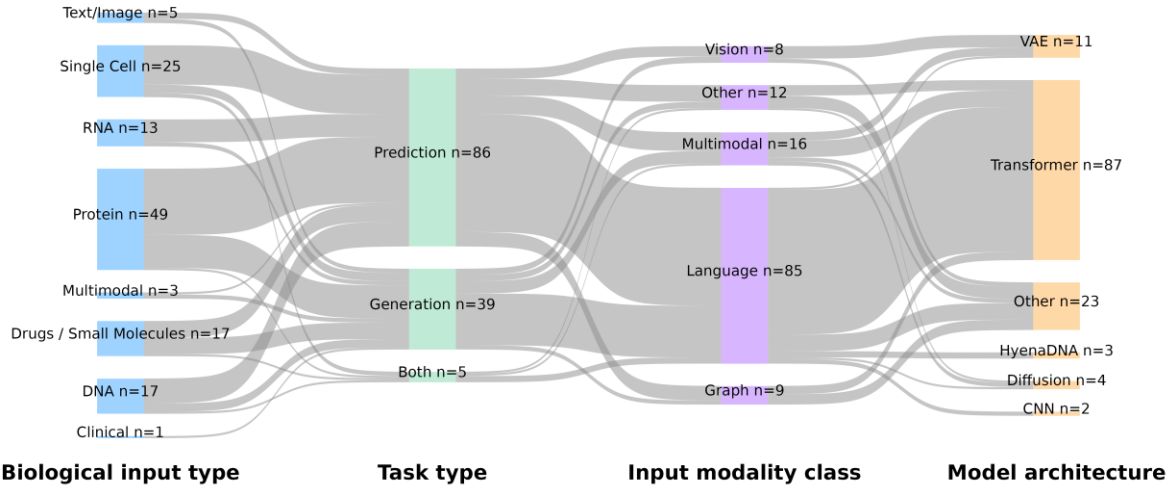
[56] https://go.drugbank.com/

[57] https://www.genome.jp/kegg/

4.2. Model Design and architecture

This section orients the reader to how contemporary AI models in biology are built. Three classification axes are commonly used to describe the design of biological AI models: the **task** they perform (predictive, generative, or hybrid; see **Box 3** for definitions); the **input modality class** (e.g. language-based, vision-based, graph-based, multimodal; see **Box 4** for definitions); and the computational **architecture** that underpins them (transformers, Convolutional Neural Networks (CNNs), Variational Autoencoders (VAEs), diffusion models, long-convolution models such as HyenaDNA, and others; see **Box 5** for definitions). These axes are not independent. In practice, choices co-vary: protein-sequence work tends to use transformer-based language models for predictive tasks, while small-molecule and protein-structure design draw more heavily on diffusion and graph-based approaches for generation. **Figure 11** visualises these co-occurrences across 130 models drawn from recent literature reviews (see Annex 3).

Figure 11. Sankey representation highlights the relative distribution of models across biological input type, task type, input modality class, and model architecture for 130 literature-curated AI Biology models.



Source: JRC's analysis.

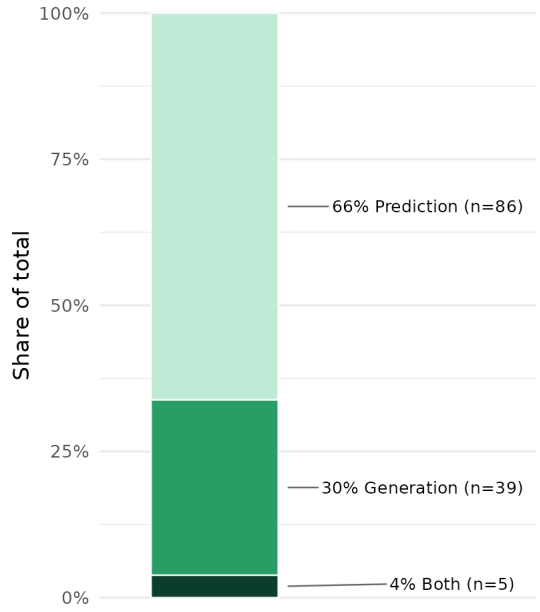
The figure highlights how modelling choices across these dimensions tend to co-occur in practice, revealing prevalent design patterns and structural dependencies in contemporary biological AI systems. The trend protein-prediction-language-transformer, as highlighted in previous sections, is immediately apparent. Models trained on DNA, RNA and single cell data are also associated with prediction tasks, which subsequently, are mostly associated to language models, reflecting the widespread consideration of biological sequences as tokenised strings processed with transformer architectures. In contrast, drug-related models show a higher relative association with generation tasks.

4.2.1. Tasks Type: Predictive, Generative and Hybrid

The analysis of 130 literature-curated biological AI models reveals a clear hierarchy in task-type distribution (**Figure 12**). **Predictive models constitute the dominant category**, accounting for 66% (n=86) of the surveyed literature. This reflects the field's historical orientation toward inference problems that complement or replace resource-intensive experimentation, exemplified by

landmark systems such as AlphaFold2 for protein structure (Jumper et al., 2021b) and Enformer for gene expression from DNA (Avsec et al., 2021).

Figure 12. Task type of 130 literature-curated biology AI models.



Source: JRC's analysis.

Generative models represent 30% (n=39) marking a conceptual shift from *understanding* what exists in nature to *designing* what could exist. RFdiffusion (Watson et al., 2023) for novel protein structures and ProtGPT2 (Ferruz, Schmidt, and Höcker, 2022) for *de novo* protein sequences exemplify this category, with practical applications in enzyme design, therapeutic protein engineering and synthetic biology.

Only 4% (n=5) of models are classified as **hybrid, capable of performing both predictive and generative tasks**. Despite their small share, these systems, i.e. Evo (Nguyen et al., 2024), MoleculeSTM (Liu et al., 2023), ESM-1b (Rives et al., 2021), SCimilarity (Heimberg et al., 2025), and scPRINT (Kalfon et al., 2025), exemplify the foundation-model paradigm, where large-scale pretraining yields representations applicable to diverse downstream tasks[58]. Their limited prevalence reflects technical challenges, balancing competing training objectives, sophisticated optimisation, extensive compute, rather than a lack of scientific interest. As foundation models become more prevalent, the distinction between prediction and generation is likely to blur.

[58] For example, Evo's autoregressive training inherently supports generation, while its learned representations can be leveraged for supervised prediction; ESM-1b's masked-amino-acid pretraining produces representations effective for contact prediction, secondary structure inference and variant effect prediction without task-specific training (Brandes et al., 2022).

Box 3. Task type definitions

Predictive models — These models aim to transform input data, such as biological sequences, into output predictions of properties, structures, or functions that are often experimentally challenging or resource-intensive to determine. This approach has led to advances in various areas of biology, from protein structure prediction to functional annotation and epigenomic profiling. (Hie, 2025)

Generative Models — Generative models are designed to create new data samples from learned patterns. For example, autoregressive models generate data step by step—predicting the next token in a sequence from left to right. Variational Autoencoders (VAEs) learn to map random Gaussian noise into realistic samples through a decoding process. Similarly, diffusion models, both continuous and discrete, gradually transform noisy or corrupted data into clean, structured samples through iterative denoising. Models trained with such sampling-oriented objectives are collectively known as generative models. Based on (Hie, 2025)

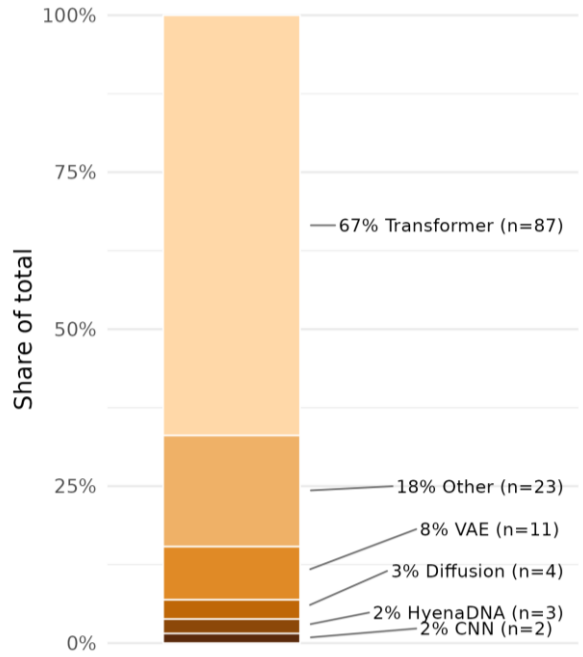
Hybrid — Models that are capable of executing both predictive and generative tasks. (JRC's analysis)

4.2.2. Architecture

Transformers are by far the most common architecture, accounting for **67% (n=87)** of surveyed models (**Figure 13**). Their prevalence rests on two technical advantages: self-attention captures long-range dependencies without the positional limitations of recurrent networks or the locality constraints of CNNs (Vaswani et al., 2023); and the architecture scales efficiently with data and compute (Hoffmann et al., 2022), enabling foundation models trained on massive unlabelled datasets. This scalability comes at a cost, however: transformer inference and training are computationally expensive relative to recurrent or convolutional alternatives.

The transformer category itself encompasses three principal variants. **Encoder-only** models (e.g. the ESM family) use bidirectional attention to learn representations of protein sequences for tasks such as mutation-effect prediction (Lin et al., 2023; Rives et al., 2021). **Decoder-only** models (e.g. ProGen2, scGPT) generate sequences autoregressively (Nijkamp et al., 2023; Cui et al., 2025). **Encoder–decoder** models (e.g. ProtT5) combine bidirectional encoding with autoregressive decoding for sequence-to-sequence tasks (Elnaggar et al., 2022).

Figure 13. Model architecture of the 130 literature-curated biology AI models.



Source: JRC's analysis based on literature.

Non-transformer architectures occupy specialised niches. **VAEs** (8%, n=11) support representation learning and controlled generation, with notable applications in cryo-EM reconstruction (the CryoDRGN family (Zhong et al., 2021; Rangan et al., 2024)) and multi-omics integration (GLUE (Cao and Gao, 2022)). **Diffusion models** (3%, n=4), including EvoDiff (Alamdari et al., 2024), Chroma (Ingraham et al., 2023), and components of AlphaFold3 (Abramson et al., 2024), represent the fastest-growing trend, particularly for structure and sequence design. **CNNs** (2%, n=2) remain relevant for detecting local motifs through hierarchical feature learning. **HyenaDNA** (2%, n=3) replaces self-attention with long-range implicit convolutions, achieving sub-quadratic complexity on genome-scale sequences and motivating derivatives such as Evo (via StripedHyena) and regLM (Nguyen et al., 2023; Nguyen et al., 2024; Lal et al., 2024). The remaining **18% (n=23)** comprises a heterogeneous "Other" category, including graph-based models (DrugBAN, EIHGN (Yang et al., 2024b)), state-space and recurrent sequence models (UniRep (Alley et al., 2019), ProtMamba (Sgarbossa, Malbranke, and Bitbol, 2025)), and task-specific structural systems such as AlphaFold2 (Jumper et al., 2021b) and RoseTTAFold All-Atom (Baek et al., 2021).

The landscape thus reflects a clear consolidation around transformers, balanced by continued experimentation with architectures better suited to long-range efficiency and generative design.

Box 4. Main model architecture classification definitions considered in this report.

Computational architecture determines how information is processed, how models scale with data and computational resources, and how reproducible or interpretable their results may be. While different architectures may be applied to the same biological tasks or data modalities, architectural choices strongly influence model capabilities, training dynamics, and suitability for specific applications.

Transformers — A neural network architecture that uses self-attention mechanisms to process sequential data, allowing for efficient modelling of long-range dependencies (Hie, 2025). The "transformers" category is a broad way to categorise an architecture that encompasses diverse architectural variants like "encoder-only", "decoder-only" and "encoder-decoder" architectures.

Convolutional Neural Networks (CNNs) — Neural networks designed to process data dominated by local interactions among features using convolution operations (Hie, 2025).

Variational AutoEncoders (VAEs) — Generative models that learn to encode inputs into a latent space and decode them back, often used for generating new data (Hie, 2025).

Diffusion — Generative models that learn to transform random noise into structured data through an iterative denoising process, enabling high-quality sample generation in continuous or discrete spaces (Yang et al., 2025).

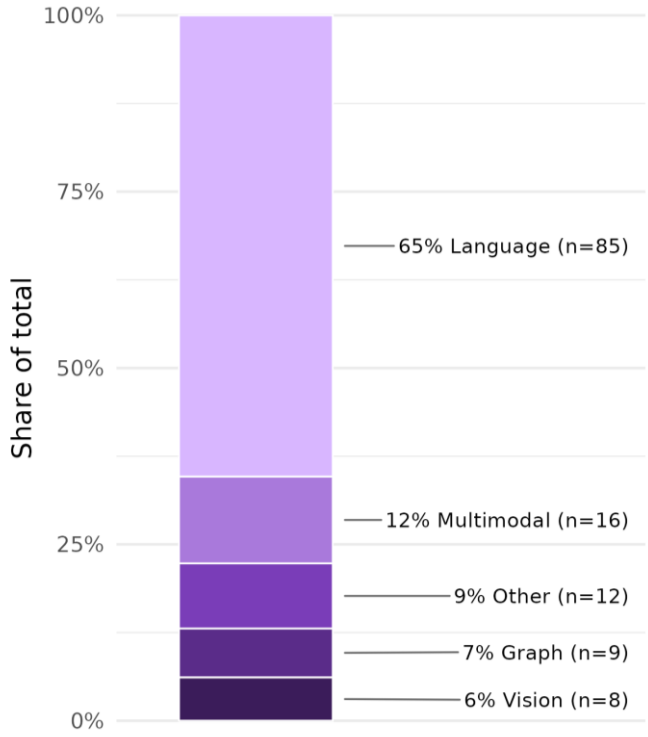
HyenaDNA — A long-sequence neural architecture based on implicit long-range convolutions, designed to efficiently model extremely long DNA sequences (Nguyen et al., 2023).

4.2.3. Input Modality Class

Language-based models dominate the modality axis, accounting for **65% (n=85)** of surveyed models (**Figure 14**), consistent with the Epoch AI task taxonomy (Annex 4, **Figure 52**). Biological sequences are treated as analogues of natural language, with nucleotides or amino acids as tokens and evolutionary or functional constraints creating patterns analogous to grammar and semantics (Heinzinger et al., 2019; Ferruz, Schmidt, and Höcker, 2022). Foundation models such as ESM-2, the Nucleotide Transformer, DNABERT-2 and Evo demonstrate the success of this approach across proteins and genomes, including at scales exceeding 100 kb[59] (Lin et al., 2023; Dalla-Torre et al., 2025; Zhou et al., 2023; Nguyen et al., 2024).

[59] A kilobase (kb) is a unit of measurement for nucleic acid sequences equivalent to 1,000 base pairs of DNA or nucleotides of RNA. Context lengths of 100 kb thus correspond to sequences spanning roughly the size of a small gene or a compact genomic region.

Figure 14. Input modality class of 130 literature-curated biology AI models.



Source: JRC's analysis.

Multimodal models represent **12% (n=16)**, the second most prevalent category, reflecting the recognition that comprehensive biological understanding requires integration across sequences, structures, images and experimental measurements. AlphaFold3 integrates sequence, evolutionary and structural signals (Abramson et al., 2024); Enformer combines DNA sequence with chromatin accessibility and epigenomic context (Avsec et al., 2021). Similar strategies are emerging in molecular design, single-cell biology and protein engineering.

Graph-based models (**7%, n=9**) capture non-sequential dependencies and network topology, with applications in molecular property prediction and generation (EIHGN (Yang et al., 2024b), MolCLR (Wang et al., 2022), Mole-BERT (Xia et al., 2022), Pocket2Mol (Peng et al., 2025)) and in modelling cell–cell or gene regulatory networks (scPROTEIN (Li et al., 2024c), SiGra (Tang et al., 2023)). Their more modest representation reflects greater data-curation demands and limited transfer-learning benefits compared with language models.

Vision-based models (**6%, n=8**) process spatially organised data where geometric and morphological information is essential. Cryo-EM is the principal application domain, with the CryoDRGN family (Zhong et al., 2021; Rangan et al., 2024) reconstructing three-dimensional structures from electron density images. Other examples address DNA visual encodings (VQDNA (Li et al., 2024b)) and molecular geometry (POLYGON (Munson et al., 2024)).

The remaining **9% (n=12)** "Other" category includes integrative structure-prediction systems (AlphaFold2 (Jumper et al., 2021b), ESMFold (Lin et al., 2023), OmegaFold (Wu et al., 2022), RoseTTAFold All-Atom (Baek et al., 2021)) that combine sequence with spatial and evolutionary constraints in ways that transcend simple multimodal fusion.

The distribution reflects both the technical maturity of language models and the practical availability of large-scale sequence databases. Looking forward, increased convergence toward

multimodal foundation models is likely, though sequence-based approaches will remain central where sequences provide sufficient information and computational efficiency matters.

Box 5. Main input modality classes considered in this report.

The input modality class describes the representational approach a model uses to process and learn from biological data. The input modality determines not only what type of biological information a model can process, but also how that information must be structured, what biases may be introduced through representation choices, and what kinds of biological insights the model can ultimately provide. Different modalities privilege different aspects of biological systems: sequences capture evolutionary and functional information encoded in linear order, graphs represent relational and topological structure, vision-based approaches preserve spatial organisation, and multimodal systems attempt to bridge these complementary perspectives.

Language — Refers to inputs represented as symbolic, discrete sequences, where biological entities are encoded as ordered strings drawn from a finite alphabet. Models process these sequences analogously to natural language, learning contextual and positional dependencies. Examples include DNA sequences (A, C, G, T), RNA sequences (A, C, G, U) or protein sequences (amino-acid letters). JRC's analysis based on (Rives et al., 2021; Ji et al., 2021).

Vision — Consists of spatially organised, continuous inputs where biological information is encoded through pixel intensities or spatial grids. The spatial arrangement itself carries semantic meaning. Examples include histopathology and microscopy images, protein contact maps or cell morphology images JRC's analysis based on (Moen et al., 2019) and (Stringer et al., 2021).

Graph — Represents data as structured relational inputs, where nodes correspond to biological entities and edges encode physical, functional, or spatial relationships. Examples include molecular graphs (atoms–bonds), protein–protein interaction networks, gene regulatory networks and cell–cell interaction graphs. Own JRC's analysis based on (Zhang et al., 2021) and (Xia et al., 2022).

Multimodal — Processes two or more different input modalities simultaneously, learning to integrate information from each modality to solve tasks that require cross-modal reasoning, alignment, or generation. JRC's analysis based on (Baltrušaitis, Ahuja, and Morency, 2017; Townshend et al., 2021; Liu et al., 2023).

4.3. Computational Infrastructure

This section examines the infrastructure required to develop and deploy AI4Bio models, focusing on hardware, data, compute, training time, model size and energy use. Together, these dimensions define what is technically feasible in the field.

Figure 15 places AI4Bio within the wider AI landscape across seven training-resource dimensions:

1. **Number of Epochs:** The number of complete passes the model makes through the entire training dataset.
2. **Hardware Quantity:** The scale of compute infrastructure used, measured by the physical count of accelerators (e.g., GPUs or TPUs) deployed in parallel.
3. **Training Time:** The wall-clock duration required to complete the training process, measured in hours.
4. **Training Power Draw:** The rate of electrical power consumed during training, measured in watts (W), reflecting energy intensity.

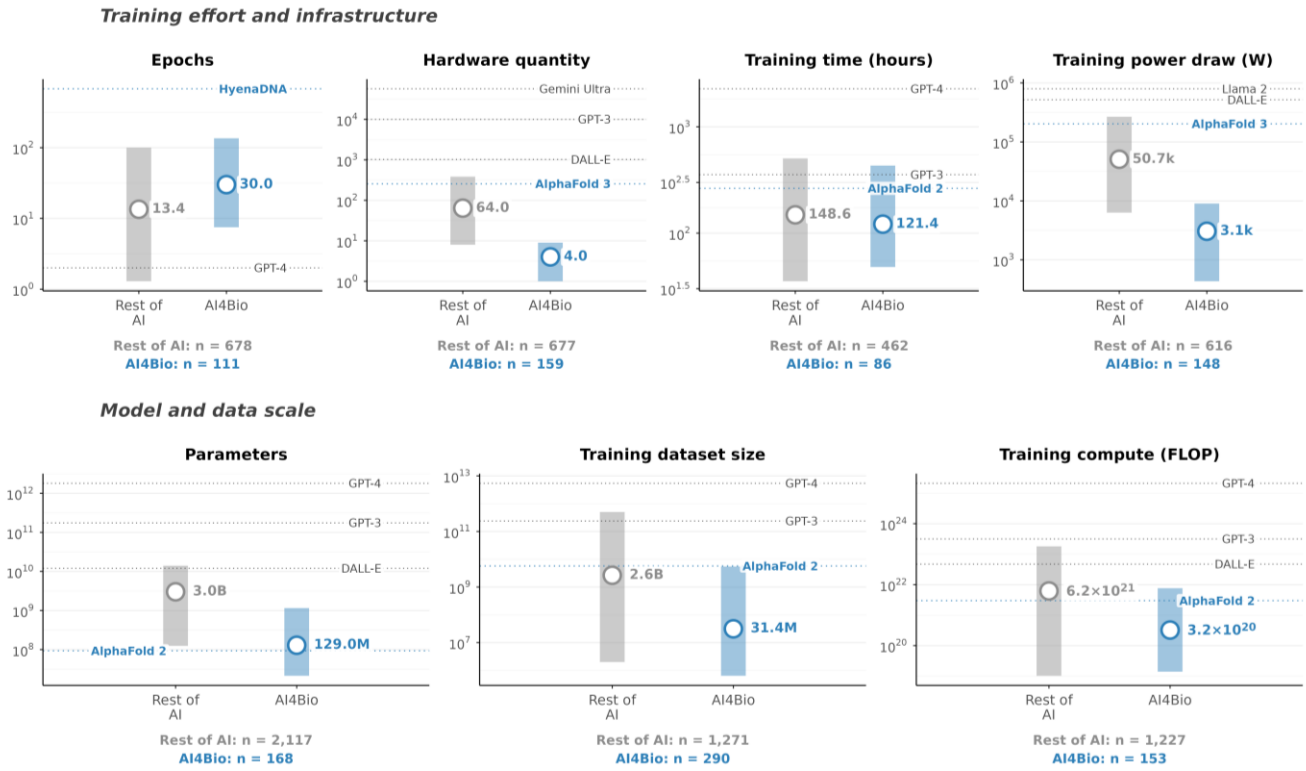
5. **Parameters:** The total number of trainable weights and biases in the model architecture, reflecting model capacity.
6. **Training Dataset Size:** The volume of data used for training, measured in total datapoints, sequences (e.g., number of protein sequences, molecular graphs, or atomic coordinates) excluding duplicates or repeated exposures across training epochs.
7. **Training Compute (FLOPs):** The total floating-point operations (FLOPs) performed across the entire training run, measuring total computational work.

At the median, a biology AI model is trained on 4 accelerators for approximately 121 hours, consumes around 3.1 kW of power, contains 129 million parameters and requires 3.2×10^{20} FLOPs of training compute. **Compared with the broader AI field**[60], AI4Bio operates well below the frontier across most dimensions. **The median AI4Bio model uses**:

- **16 times fewer accelerators** (4 vs 64)
- **16 times less power consumption** (3.1 kW vs 50.7 kW)
- **23 times fewer parameters** (129M vs 3.0B)
- **19 times less training compute** (3.2×10^{20} vs 6.2×10^{21} FLOP)
- **~83 times smaller training datasets** (31.4M vs 2.6B datapoints)
- **Comparable median training time** (121 hours versus 149 hours)
- **~2x higher in epochs (30 vs 13.4)**

[60] The broader AI field refers to the selection of general-purpose models available in the Epoch AI dataset (retrieved June 2026). Training-resource variables for some recent frontier models (e.g. GPT-5.5) are absent from the dataset and could not be included in this comparison. For this reason, older but representative frontier models for which these data are available, such as GPT-4, Gemini Ultra, LLaMA 2, and DALL-E, were therefore used as reference points.

Figure 15. How AI in biology compares with the rest of AI models[61], across seven training-resource variables. Each panel shows one variable on its own logarithmic scale; the dot marks the median model in each group and the vertical bar spans the 25th–75th percentile range. Blue dotted lines and bold labels identify biological AI reference models; grey dotted lines mark major general AI models.



Source: JRC's analysis.

These differences align with the nature of biological modelling. Biological tasks are typically narrow and domain-specific: a model predicting protein structure or classifying cell types from single-cell RNA sequencing data operates over a constrained, structured input space that does not require frontier-scale parameterisation. For example, both AlphaFold2 model for protein structure prediction (Jumper et al., 2021b) and Geneformer model for single-cell modelling illustrate high performance at sub-billion parameter counts). The higher epoch counts, combined with smaller datasets, indicate a potential extensive reuse of curated, domain-specific corpora, a training pattern documented across genomics (e.g. HyenaDNA, trained for 50 to 200 epochs during fine-tunning (Nguyen et al., 2023)) and protein models (e.g. DNCON2, trained for 1600 epochs). Taken together, **Figure 15** positions AI4Bio in a distinct scaling regime from frontier general-purpose AI: one shaped by biological task constraints rather than compute maximisation.

Figure 16 shows how these seven dimensions relate to one another within AI4Bio. Three variables form a tight cluster: hardware quantity, training time and power draw co-vary strongly (pairwise Pearson r = 0.76 to 0.95), indicating that infrastructure scaling tends to expand all three simultaneously. Model parameters and training compute show moderate correlations with this cluster (r = 0.41 to-0.70) and with each other (r = 0.51). Training dataset size relates only weakly to

[61] https://epoch.ai/data/ai-models?view=graph&tab=all

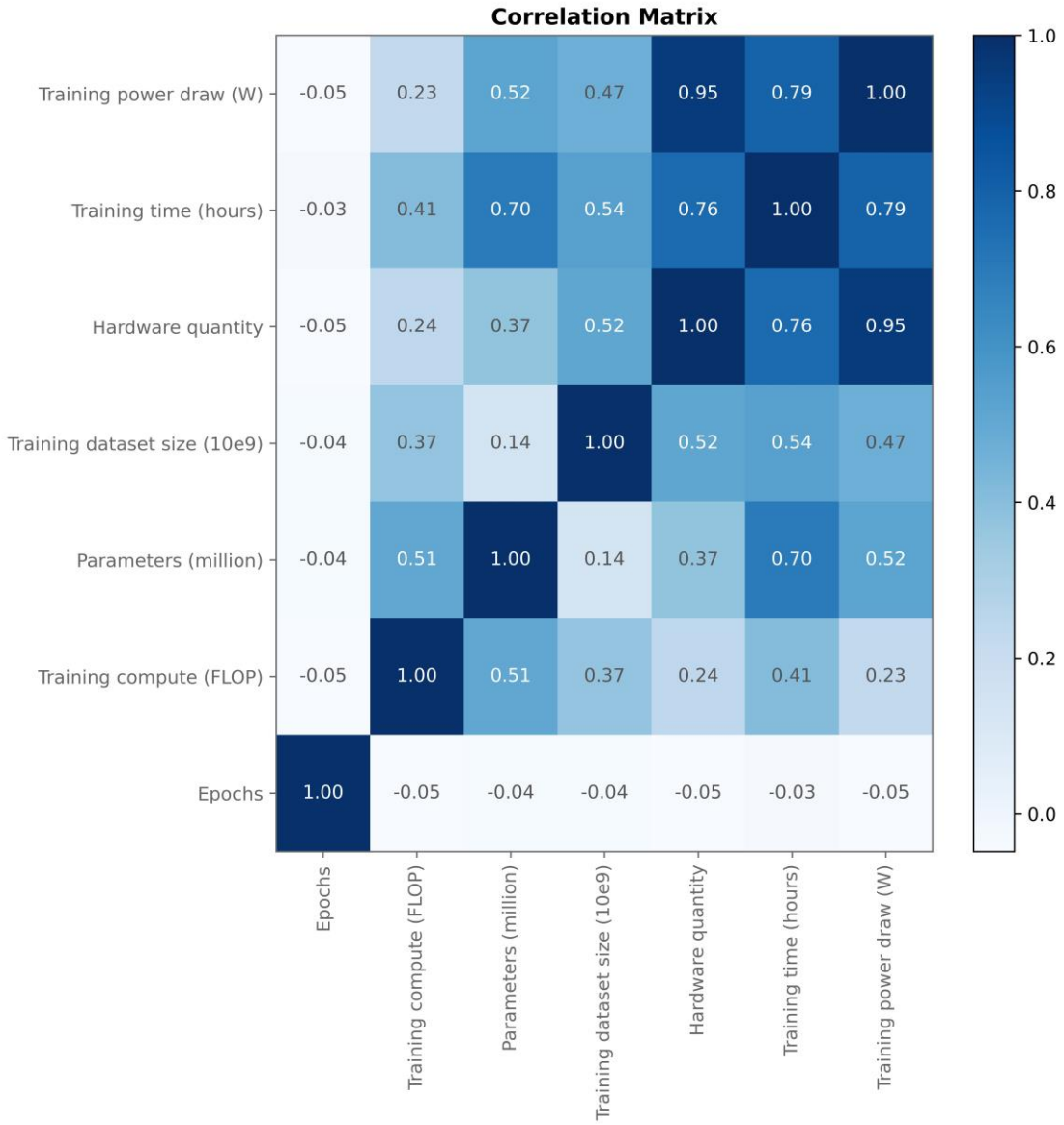
the remaining variables (r = 0.14 to 0.54), and epochs are effectively uncorrelated with all other measures.

Taken together, Figures 15 and 16 yield four observations relevant to both the technical landscape and policy context:

1. **Infrastructure bottleneck is effectively one-dimensional.** Hardware, training time and energy use scale together, meaning GPU access is typically the single constraint governing all three. This pattern is consistent with broader analyses of AI resource concentration (Ahmed and Wahed, 2020; Sevilla et al., 2022).
2. **Biological AI follows a different scaling logic from language modelling.** The weaker parameter--compute coupling (r = 0.51) is lower than expected under standard scaling laws documented for large language models. This likely reflects, at least in part, the architectural diversity of the field, which spans diffusion models, equivariant networks and graph neural networks.
3. **Training data behaves differently from infrastructure variables.** Curated biological datasets such as annotated genomes or high-quality clinical records are orders of magnitude smaller than internet-scale text corpora (e.g. PDB, Uniprot, ENCODE). Training regimes are therefore shaped more by data characteristics than by compute budgets, and epoch reuse seem to be a structural feature of the field rather than an optimisation choice.
4. **Many AI4Bio models are developed with modest resources.** Smaller, resource-efficient models have demonstrated competitive performance on targeted biological tasks and may be better suited to deployment in resource-constrained settings such as public health laboratories or clinical environments outside major research centres.

The subsections below analyse each infrastructure dimension in turn. Methods are described in Annex 6.

Figure 16. Pearson correlation matrix of training power draw and selected technical variables in biological models (236 models).



Source: JRC's analysis.

4.3.1. Data Infrastructure

The development of AI4Bio models relies on training with large datasets (see Section 4.1). This requires investments in scalable data storage and management solutions, including on-premises or cloud-based databases, data warehouses, and distributed file systems, similar to those used in other AI-intensive fields[62].

[62] https://www.ibm.com/think/topics/ai-infrastructure

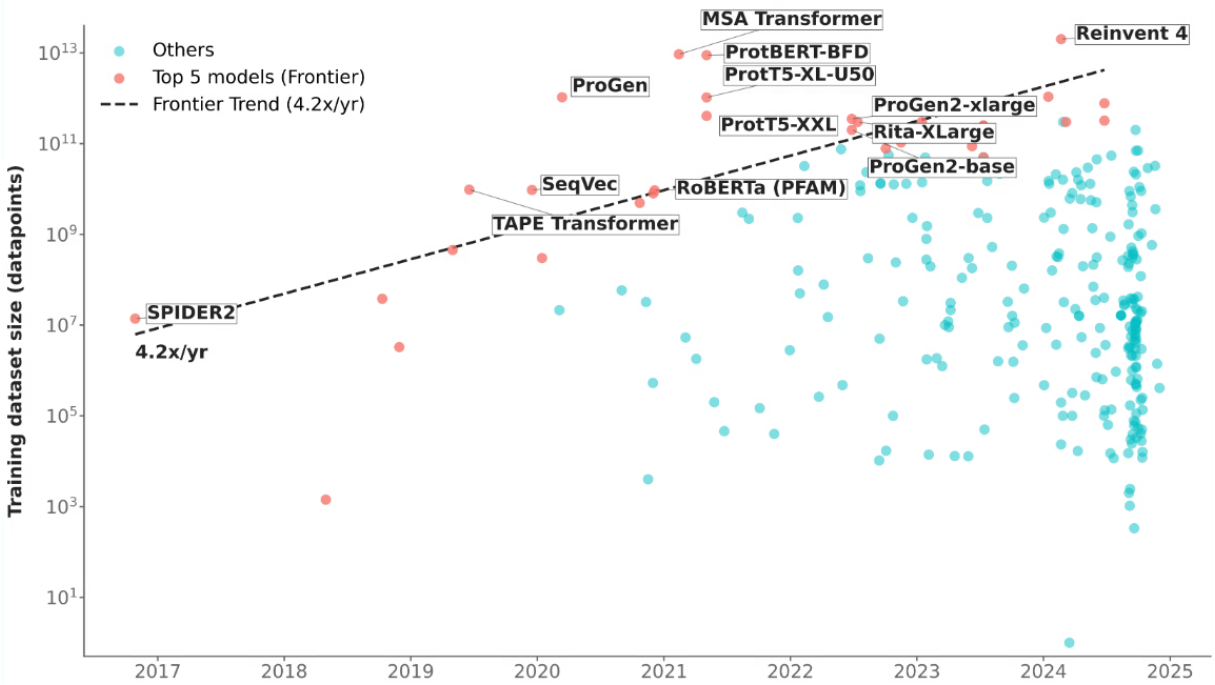
The following results draws on data available for 274 models in the Epoch AI dataset. Across these models, the mean training dataset size is 213 billion data points (see **Table 9** for descriptive statistics). *Reinvent 4* was the model trained on the largest number of data points (20,000 billion). It is an open-source generative AI framework for the design of small molecules to support drug discovery (Loeffler et al., 2024). The frontier analysis of the 5 models utilising larger datasets each year reveals that from 2017 to 2024, the dataset size has grown at a rate of 4.2 times per year (see **Figure 17**).

Table 9. Descriptives for the training dataset size.

Descriptive	Value
Total models	274
Mean	2.13 x 10^11 datapoints
Standard Deviation	1.54 x 10^12 datapoints
Minimum	333 datapoints
Quartile 1 (25%)	1.56 x 10^06 datapoints
Quartile 2 (50%) - median	3.43 x 10^07 datapoints
Quartile 3 (75%)	5.60 x 10^09 datapoints
Maximum	2.00 x 10^13 datapoints

Source: JRC's analysis.

Figure 17. Frontier analysis of the training dataset size (270 models).



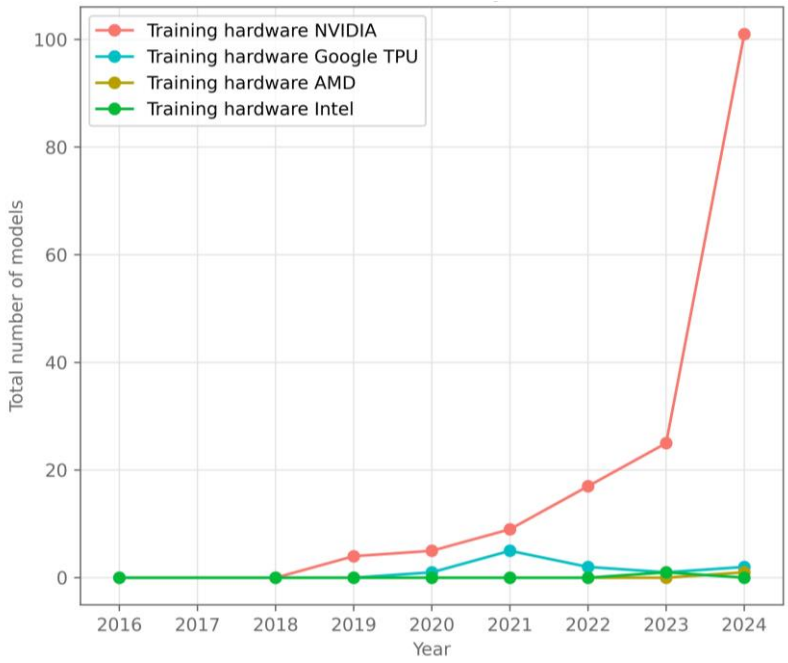
Source: JRC's analysis.

4.3.2. AI Specialised Hardware and Software

Processing large-scale data for AI tasks demands substantial computational power. To meet this requirement, AI infrastructure often relies on specialised hardware that goes beyond traditional central processing units (CPUs) commonly found in conventional IT environments. Specifically, graphics processing units (GPUs) are frequently employed to accelerate matrix and vector computations, while tensor processing units (TPUs) are optimised to handle tensor computations, delivering high throughput and low latency (Nikolić et al., 2022). This pattern is also evident in AI applied to biology. NVIDIA is the most widely used hardware brand among the 179 models for which hardware data are available (see **Figure 18**).

In more detail, information on specific hardware used in 2024 is available for 227 models. The most used NVIDIA GPUs and TPUs in 2024 are presented in **Table 10**. In 2024, the most frequently used GPU was the NVIDIA A100 (n=39 models), with 80GB of memory, released in 2020. Among the top 10 most used hardware devices in 2024, the newest models were released in 2022 and were used to train 3.52% of the AI4Bio models that year. By contrast, hardware released in 2021 accounted for only 0.88% of the AI4Bio models, while 31.71% of models in 2024 were run on hardware released in 2020 or earlier, meaning equipment at least four years old.

Figure 18. GPU and TPU brand use over time (179 models).



Source: JRC's analysis.

Table 10. Top 10 Hardware GPU's and TPU's used in 2024, with 227 analysed models.

Training Hardware	Memory	Release year	Number of AI4Bio models in 2024	Percentage of AI4Bio models in 2024
NVIDIA A100	80GB	2020	39	17.18%
NVIDIA RTX A6000	48GB	2020	10	4.41%
NVIDIA GeForce RTX 3090	24GB	2020	8	3.52%
NVIDIA V100	12GB	2017	6	2.64%
NVIDIA H100 SXM5 80GB	80GB	2022	5	2.20%
NVIDIA A100 SXM4 80 GB	80GB	2020	3	1.32%
NVIDIA A40 PCIe	48GB	2020	3	1.32%
NVIDIA GeForce RTX 4090	24GB	2022	3	1.32%
NVIDIA A100 SXM4 40 GB	40GB	2020	3	1.32%
Google TPU v4	32GB	2021	2	0.88%

Source: JRC's analysis.

Additionally, multiple GPUs and TPUs are often utilised in parallel to enhance processing capabilities and accelerate computational workloads. AI4Bio models use an average of 50 GPUs or TPUs (see **Table 11** for descriptive statistics). ProtBERT-BFD (Elnaggar et al., 2022) was the model run with the maximum number of hardware components (n=1024). This model is a transformer-based protein language model trained on the BFD-100 database (Big Fantastic Database[63]) which contains over two billion protein sequences. The model was trained using self-supervised learning, enabling it to learn contextual representations of amino acid sequences without labelled data. As a result, ProtBert-BFD can generate high-quality protein embeddings that support a wide range of downstream tasks, including protein classification, function prediction, structure-related inference, and mutation effect prediction. It is notable that half of the models were run with 4 pieces of hardware, whereas 49 models were run with only one. This suggests that, despite the attention given to large-scale compute-intensive models, a substantial portion of AI4Bio research remains accessible without specialised hardware infrastructure.

[63] https://bfd.mmseqs.com/

Table 11. Descriptives for the hardware quantity used to train AI4Bio models.

Descriptive	Value
Total models	155
Mean	50.53
Standard Deviation	137.57
Minimum	1
Quartile 1 (25%)	1
Quartile 2 (50%) - median	4
Quartile 3 (75%)	9
Maximum	1024

Source: JRC's analysis.

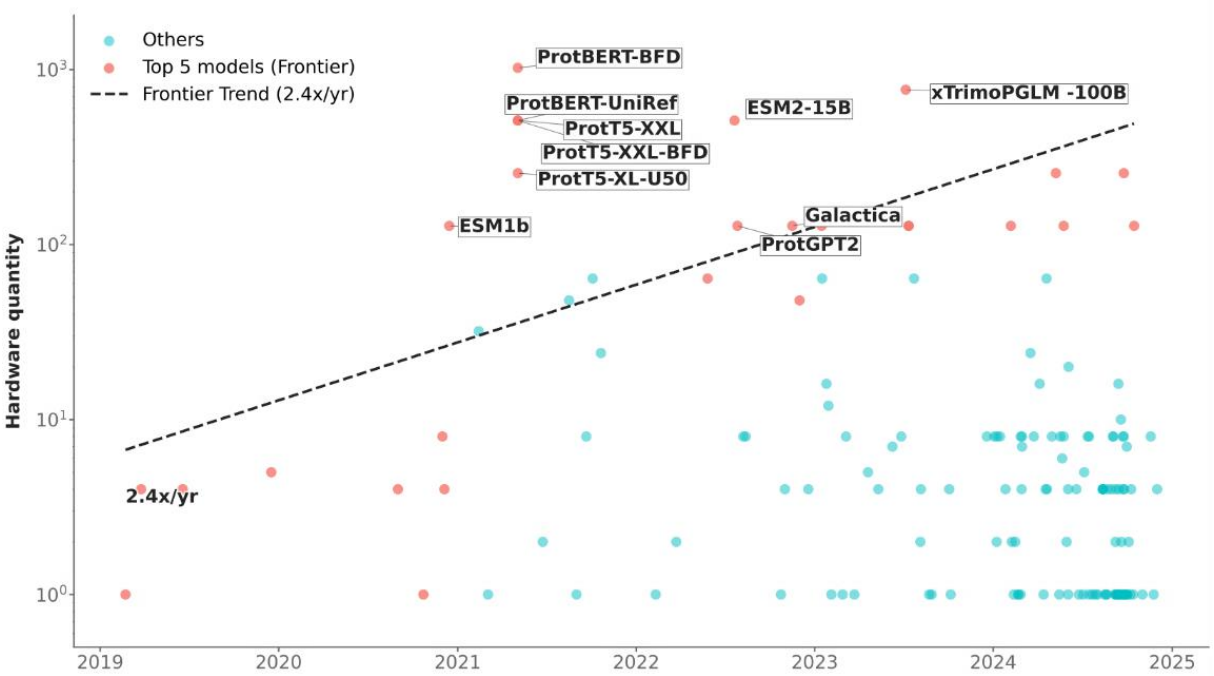
The frontier analysis of the top models utilising highest hardware quantity reveals that from 2017 to 2024, the hardware quantity has grown at a rate of 2.4 times per year (see **Figure 19**).

While purchasing hardware remains an important investment strategy, another growing trend in AI applications is the use of cloud computing. Cloud computing allows users to access computational resources and applications via the internet, without installing and maintaining systems on premises. Cloud solutions are more flexible and easier to scale than traditional IT infrastructure, enabling rapid adaptation to changing computational requirements. Another advantage is that it facilitates real-time collaborations across global research teams which become more frequent after the pandemics period[64]. However, there are risks of relying on cloud providers, including vendor lock-in, concerns regarding ownership, and exposure to security breaches, data loss, and compliance issues. To mitigate these risks, it is a priority to invest in reputable and reliable cloud providers.

In recent years, beyond providing increased storage and processing capacity, dedicated cloud platforms have emerged to support specific biology-related tasks (see **Table 12**), including genomics, multi-omics integration, microscopy, proteomics, metagenomics and clinical genomics.

[64] https://bitesizebio.com/84263/cloud-computing-for-biology/

Figure 19. Frontier analysis of the hardware quantity (151 models).



Source: JRC's analysis.

Table 12. Cloud platforms by research areas.

Research Area	Services	Why
Genomics/NGS	Terra, AWS, HealthOmics, DNAnexus	Optimised for sequence pipelines, population-scale data
Multi-omics	Seven Bridges, Google Cloud Life Sciences	Strong integration of data types and collaboration tools
Microscopy	OMERO (hosted), Google Cloud + TensorFlow	Built for imaging data and machine learning
Proteomics	Galaxy, OpenMS (cloud-deployable)	Community-supported, customisable workflows
Metagenomics	Terra, AWS HealthOmics, MG-RAST (cloud-based)	Designed for microbiome, environmental sequencing workflows
Clinical genomics	DNAnexus, Seven Bridges	HIPAA/GDPR-compliant, supports FDA and CLIA regulations65

Source: online article[61]

[65] Source lists US frameworks; relevant EU equivalents include GDPR, EMA, the EU Clinical Trials Regulation, IVDR and the European Health Data Space.

4.3.3. Computation Performance

Building on the previous discussion of hardware infrastructure and capacity, the present results consider the actual computational intensity required to train and fine-tune these models. Assessing training compute provides a more precise understanding of how infrastructure is being utilised and how model development is scaling over time.

To measure computation performance, this report relies on floating point operations per second (FLOP). Training compute data are available for 150 biological models in the Epoch AI dataset. These models require, on average, 6.0 x 10^22 FLOP (see **Table 13** for descriptives), while the model with highest demand requires 2.25 x 10^24 FLOP. This model is Evo 2 40B (Brixi et al., 2026), a biological foundation model trained on 9.3 trillion DNA base pairs from a genomic atlas spanning all domains of life. In contrast, with the minimum requirements (3.92 x 10^10 FLOP) there is the DeepTX inference framework that leverages deep learning methods to connect mechanistic models and scRNA-seq data and designed to be computationally efficient (Huang et al., 2024c). The low computational footprint of DeepTX reflects its task-specific design: the model approximates a well-defined mathematical mapping rather than learning general biological representations, and amortises training costs across subsequent inference calls.

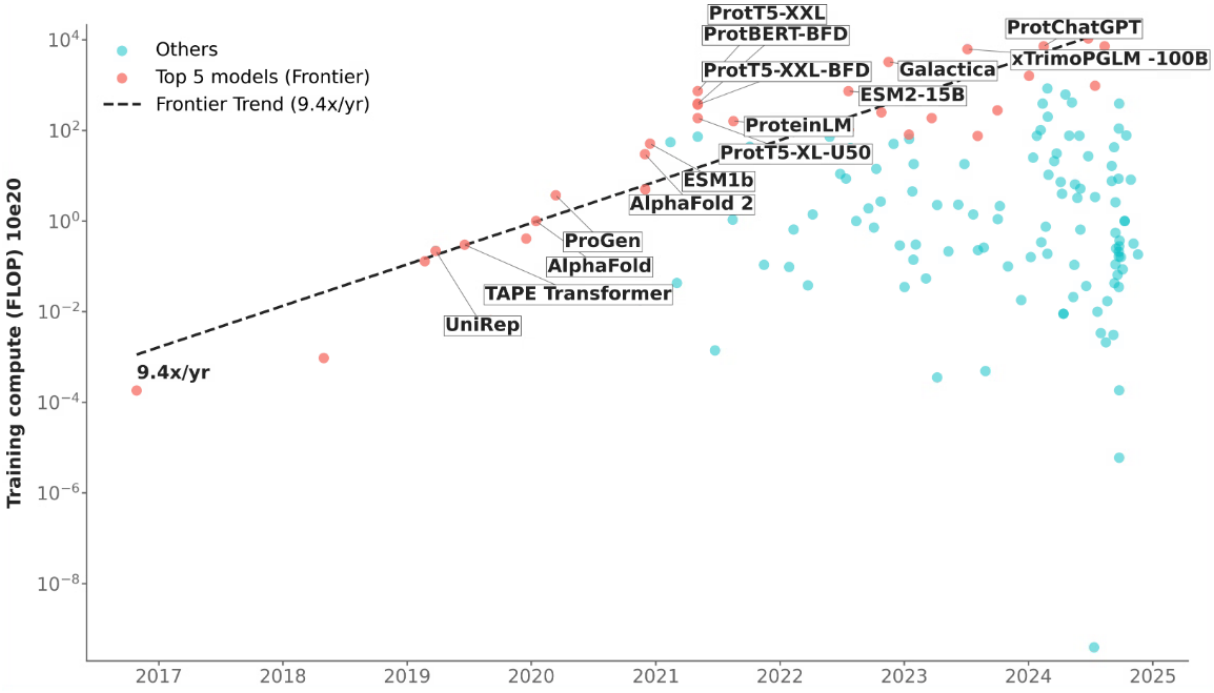
The frontier analysis revealed that the training compute (FLOP) grew at a rate of 9.4 times per year (see Figure 20). Nevertheless, most AI4Bio models currently fall well below the 10^{25} FLOPs threshold used as an indicative criterion for systemic risk under the AI Act. However, at the current growth rate, a subset of frontier biological AI models could approach or exceed this threshold within a relatively short timeframe, particularly as foundation models trained on large multi-modal biological datasets continue to scale. This trajectory warrants proactive monitoring, even if systemic risk designations do not yet apply to most models in this domain.

Table 13. Descriptives for the Training Compute FLOP.

Descriptive	Value
Total models	150
Mean	6.30 x 10^22
Standard Deviation	2.82 x 10^23
Minimum	3.92 x 10^10
Quartile 1 (25%)	1.45 x 10^19
Quartile 2 (50%) - median	2.65 x 10^20
Quartile 3 (75%)	7.49 x 10^21
Maximum	2.25 x 10^24

Source: JRC's analysis.

Figure 20. Frontier analysis of the training compute (FLOP) (n=144 models).



Source: JRC's analysis.

4.3.4. Training Time

This subsection examines training time in order to assess the practical duration required to train AI4Bio models. While training compute reflects the scale of computational resources used, training time provides insight into development cycles, infrastructure demands and the speed at which models can be produced and updated.

This analysis is based on 84 AI4Bio models for which training time data were available in the Epoch AI dataset. On average, AI4Bio models take about 350 hours runoff training time, which corresponds approximately to 14 and a half days (see **Table 14** for descriptive statistics). The maximum training time recorded is 3912 hours (163 days). This value corresponds to the xTrimoPGLM 100B model (Chen et al., 2024), trained on a cluster consisting of 96 DGX-A100 GPU servers (8×40G). During this time, the model has consumed 1 trillion tokens from the dataset consisting of Uniref90 and ColabFoldDB datasets, resulting in state-of-the-art performance on diverse protein understanding tasks. By comparison, general-purpose large language models operate at substantially greater compute scales: Meta's LLaMA 3 family required between 1.3 million GPU hours for the 8B parameter model and 7.7 million GPU hours for the 70B parameter model[66], orders of magnitude above the AI4Bio dataset average. This contextualises the AI4Bio models as computationally modest.

Training times vary substantially across models. The minimum recorded training time is 0.85 hours, while the interquartile range shows that 50 percent of models required between 48 and 432 hours

[66] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

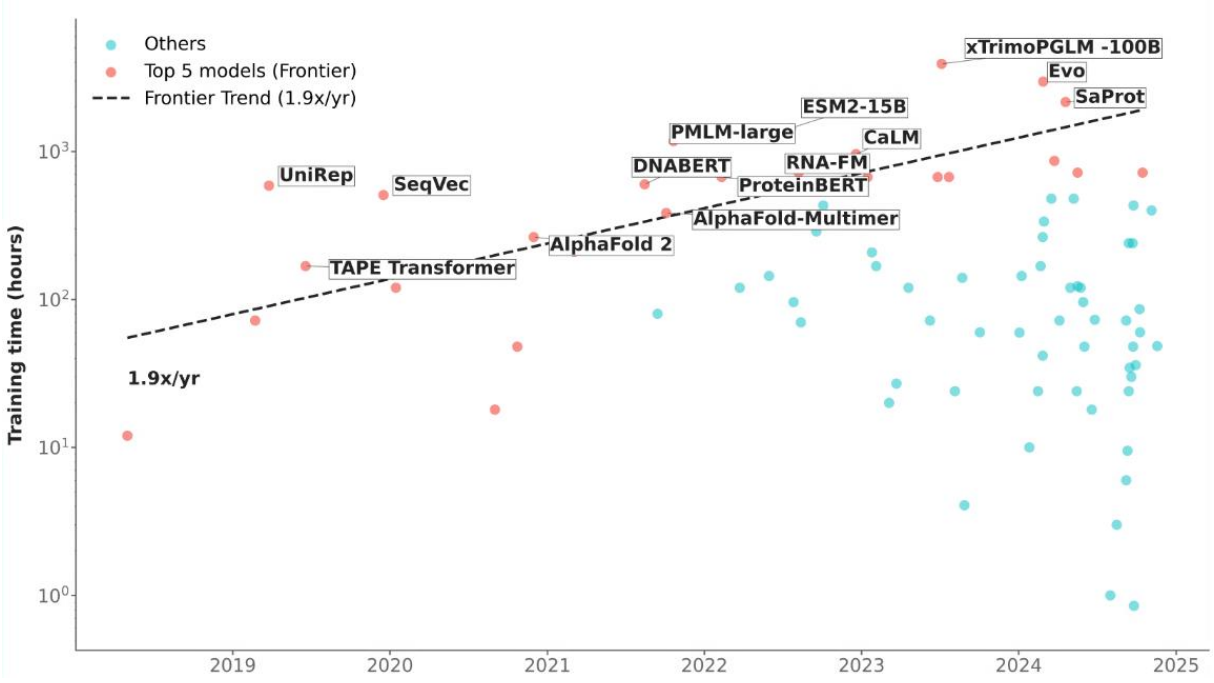
of training. The frontier analysis shows that the training time of the most demanding models is growing at a pace of 1.9 times per year (see **Figure 21**).

Table 14. Descriptives for the training time (Hours) used to train AI4Bio models.

Descriptive	Value
Total models	84
Mean	350 h
Standard Deviation	607 h
Minimum	0.85 h
Quartile 1 (25%)	48 h
Quartile 2 (50%) - median	121 h
Quartile 3 (75%)	432 h
Maximum	3912 h

Source: JRC's analysis.

Figure 21. Frontier analysis of the Training time (Hours) (84 models).



Source: JRC's analysis.

4.3.5. Model Parameters

Following the analysis of training compute and training time, this subsection examines model size in terms of parameters. While compute and time reflect resource use, the number of parameters indicates model capacity and structural scale. This analysis is based on 159 biological AI models for which parameter data were available in the Epoch AI dataset.

On average, AI4Bio models contain around 6 billion parameters (see **Table 15** for descriptive statistics). The model with the maximum number of parameters is EXAONE 2.0 (Pyeon et al., 2025)[67], with 300 billion parameters, and analyses histopathological images to help in early diagnosis and prognosis of cancer. EXAONE 2.0 reportedly achieves "*world-leading accuracy in predicting genetic mutation*" and near real-time performance in clinical settings[64].

Model size varies substantially across the dataset. The smallest model contains 2,176 parameters, while the interquartile range shows that 50 percent of models contain between 20 million and 855 million parameters. Furthermore, models at the technological frontier exhibit a growth rate in number of parameters of 3.6 times annually (**Figure 22**).

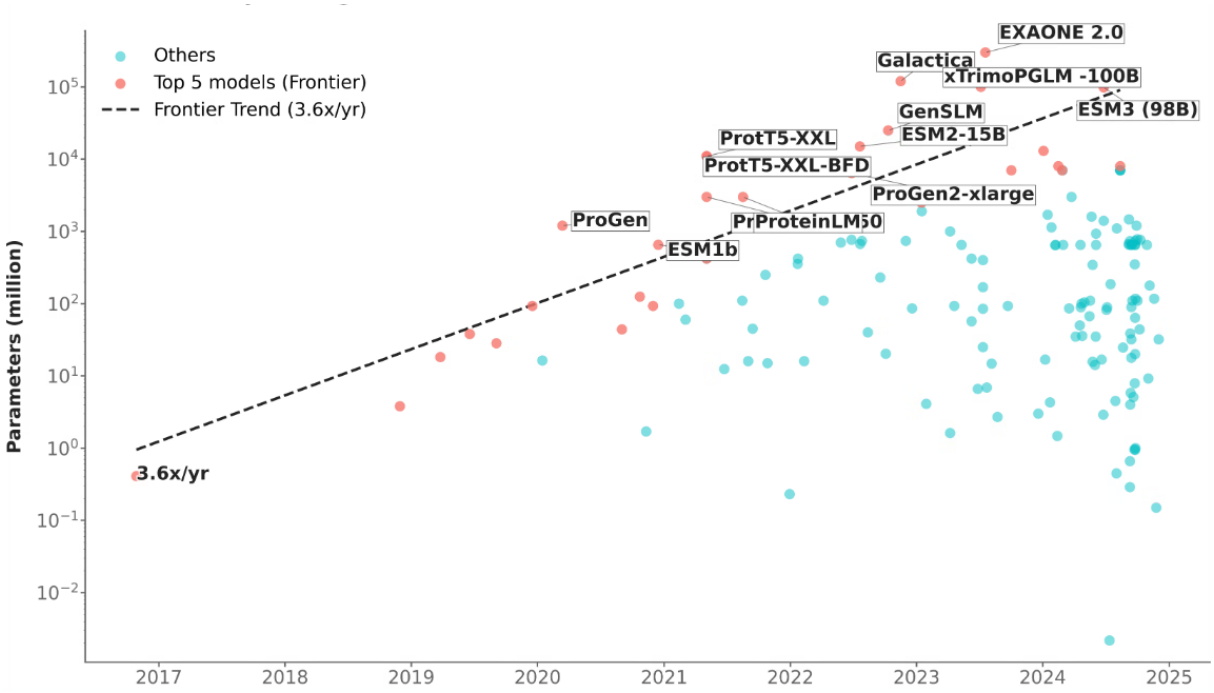
Table 15. Descriptives for the model parameters in the AI4Bio models.

Descriptive	Value
Total models	159
Mean	5646 million
Standard Deviation	28 billion
Minimum	2176
Quartile 1 (25%)	20 million
Quartile 2 (50%) - median	117 million
Quartile 3 (75%)	855 million
Maximum	300 billion

Source: JRC's analysis.

[67] https://www.lgresearch.ai/blog/view?seq=574

Figure 22. Frontier analysis of the Parameters (million) (159 models).



Source: JRC's analysis.

4.3.6. Energy Consumption

Energy consumption in AI systems has become a growing policy concern at both European and international levels due to the direct environmental and economic implications. The European Commission has highlighted the environmental footprint of digital technologies in the European Green Deal[68] and in the Digital Decade policy framework[69]. In parallel, international research has documented the significant carbon and electricity impacts associated with large scale AI training, including studies by (Strubell, Ganesh, and McCallum, 2019); (Patterson et al., 2021); and more recent analyses by the International Energy Agency[70].

Analysing training power draw therefore provides insight into both sustainability and financial aspects of AI4Bio model development. This analysis is based on 144 AI4Bio models for which training power draw data in Watts (W) were available in the Epoch AI dataset.

On average, the AI4Bio models consume 28 kW (see **Table 16** for the descriptive statistics). This is equivalent to the instantaneous electricity use of around 8 to 10 average EU households combined, assuming an average household demand of roughly 3 to 4 kW. The model consuming more energy (611151 W) is xTrimoPGLM – 100B (Chen et al., 2024). This is consistent with the long training time required for the same model (Section 4.3.4).

Energy consumption varies significantly across models. The minimum recorded power draw is approximately 75W, comparable to a traditional incandescent light bulb, corresponding to the

[68] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0102

[69] https://digital-strategy.ec.europa.eu/en/policies/europes-digital-decade

[70] https://www.iea.org/reports/energy-and-ai

CLR_ESP biology AI model (Du et al., 2024), a model for enzyme-substrate pair prediction using contrastive learning from 2024. This low energy footprint reflects the model's architecture: rather than training large representations from scratch, *CLR_ESP* fine-tunes existing protein and chemistry language models using a contrastive learning strategy, achieving state-of-the-art accuracy of 94.70% on independent test data while requiring comparatively few computational resources (Du et al., 2024). The interquartile range shows that 50 percent of models consume between approximately 400 and 9000 W during training.

Table 16. Descriptives for the training power draw (W) in the AI4Bio models.

Descriptive	Value
Total models	144
Mean	28021 W
Standard Deviation	78379 W
Minimum	75.83 W
Quartile 1 (25%)	433 W
Quartile 2 (50%) - median	3149 W
Quartile 3 (75%)	9013 W
Maximum	611151 W

Source: JRC's analysis.

The frontier analysis, of the 5 models utilising more energy each year, reveals that from 2019 to 2024, the energy consumption has grown at a rate of 2.7 times per year (see **Figure 23**). However, this trend is not uniform across the landscape. Several recent models continue to be developed using smaller datasets, fewer hardware components, and more modest compute requirements. This coexistence of high-demand frontier systems and resource-efficient smaller models reflects the heterogeneity of the AI4Bio field, though it should be noted that more modest models typically address narrower tasks and do not replicate the generalisation capacity of frontier-scale systems.

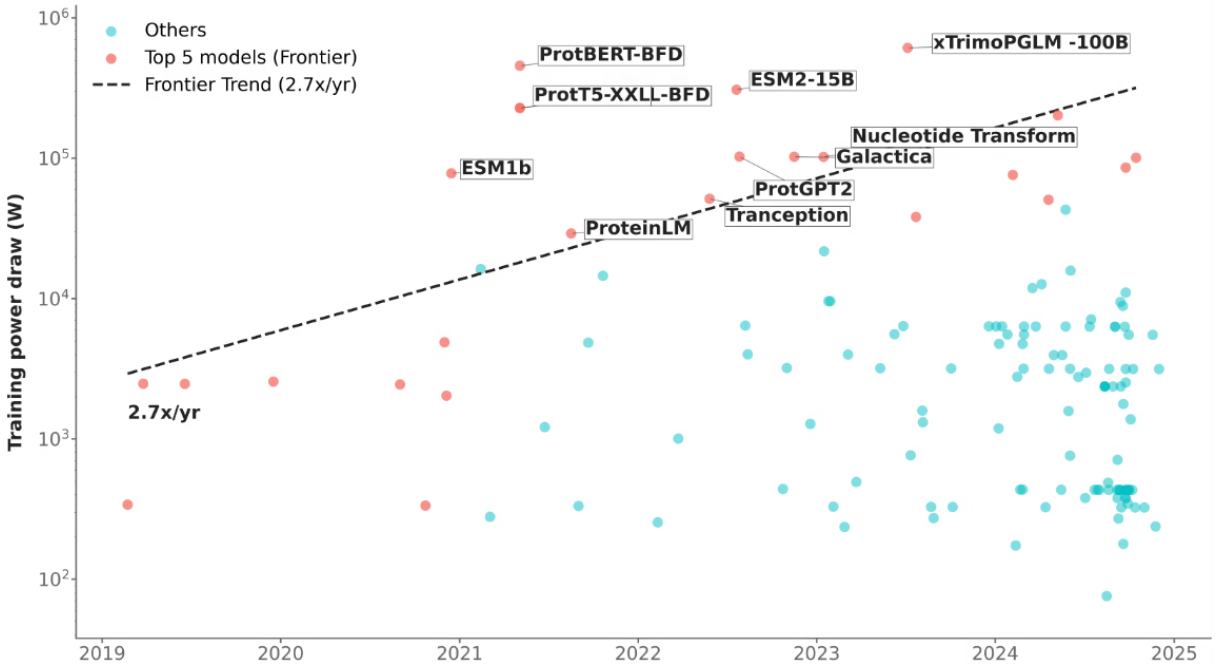
Understanding what drives this growing energy footprint is therefore a relevant question for both environmental and economic policy. A correlation analysis (**Figure 16**) of 236 models in the Epoch AI dataset reveals that training power draw is very strongly correlated with hardware quantity (r=0.95) and strongly correlated with training time (r=0.79), suggesting that models consuming more electricity are typically those running on a larger number of hardware components and for longer durations.

In contrast, training compute (FLOPs) shows only a weak correlation with power draw, and the number of epochs[71] displays low or no correlation with energy consumption (r=0.23 and r=-0.05 respectively). Number of model parameters and training dataset size exhibit moderate positive correlations with energy consumption (r=0.52 and r=0.47 respectively). These findings suggest that

[71] In deep learning, an epoch is a hyperparameter that controls how many times the model processes the full dataset to adjust its weights and reduce errors, with multiple epochs (usually 10-100 or more) often necessary to learn complex patterns (LeCun, Bengio, and Hinton, 2015).

infrastructure configuration and runtime duration are more decisive drivers of environmental impact and operational cost than model scale measured in FLOPs alone, implying that **optimising hardware allocation and reducing training duration** may be **more effective efficiency strategies than** focusing exclusively on **model size**.

Figure 23. Frontier analysis of the Training power draw (W) (141 models).



Source: JRC's analysis.

5. Scientific Barriers and Associated Risks for Real-World Adoption of Biology AI Models

Translating biological AI models from research settings into real-world application requires more than technical performance: it demands scientific robustness, transparent reporting, reproducibility and clear characterisation of uncertainty, representative data, scalable infrastructure, and credible maturity assessment. This section examines where the field currently falls short across these dimensions.

5.1. Training Data Fragmentation

The analysis covered in Section 4.1 identified 75 datasets in the Epoch AI metadata. Their diversity reflects the breadth of biological AI development, but also its fragmentation: there is no single standard corpus, and training pipelines draw on repositories that differ substantially in scale, recency, and maintenance status.

Large, widely adopted repositories such as UniProt variants, PDB, ChEMBL, the AlphaFold Protein Structure Database, RefSeq, UniRef50, and the European Nucleotide Archive are updated frequently, dominate the upper tiers of data volume, and are used repeatedly across many model training workflows (**Figure 24**). At the other end of the spectrum, smaller benchmark datasets including SIDER (Kuhn et al., 2016), BACE (Subramanian et al., 2016), NeoDTI (Wan et al., 2019), PSICOV150 (Jones et al., 2012), and DB5.5 (Vreven et al., 2015) were last updated between 2014 and 2019 and are used by far fewer models, primarily for evaluation rather than training. A middle tier of specialised datasets, such as CrossDocked2020 (Francoeur et al., 2020), ATLAS (Vander Meersche et al., 2024), OpenCell (Cho et al., 2022), and SARS-CoV-2 genome dataset (Zvyagin et al., 2022), has typically been updated within the last three to five years and complements the larger training repositories.

Assessment of maintenance status confirms that large, actively curated datasets cluster in the low-risk zone for obsolescence (**Figure 25**). Several small datasets have not been updated for multiple years; while their role in training is limited, their continued use beyond their original benchmarking purpose raises concerns about representativeness. **Figure 26** further illustrates the uneven maturity of biological data infrastructures: protein datasets dominate across all time periods and remain actively updated, while RNA, imaging, and multi-modal categories remain comparatively sparse, with many empty cells across earlier time periods.

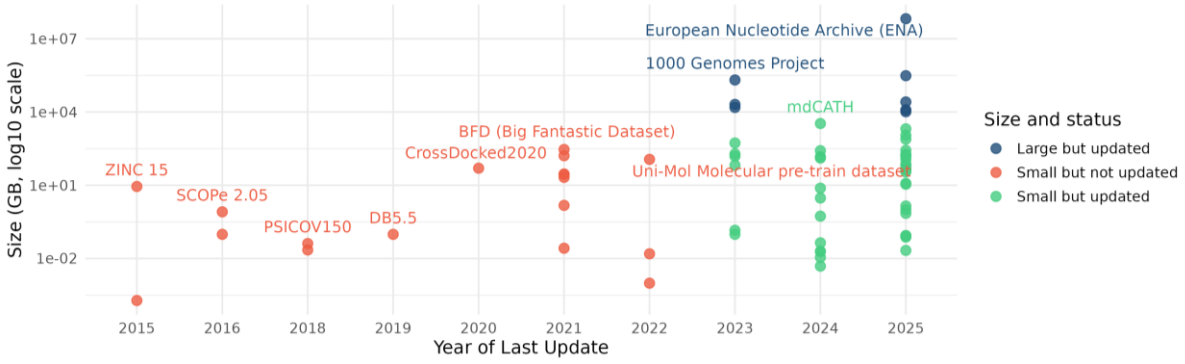
Data currency does not emerge as a major structural concern for dominant foundation model regimes: more than 150 model-dataset pairings show a zero-year lag between dataset update and model publication, and only a small number of cases exhibit lags of one to five years (**Figure 27**, **Table 23**).

The more pressing risk for real-world adoption is fragmentation. Contemporary biological AI relies heavily on a small number of large protein and genome repositories, and the absence of shared, standardised data infrastructure makes comparability across models difficult. Communities working in less data-rich domains, including RNA biology, imaging, and multi-modal biology, face compounding disadvantages that limit both model development and the prospects for deployment in those areas.

Figure 24. Datasets by last update, size and usage

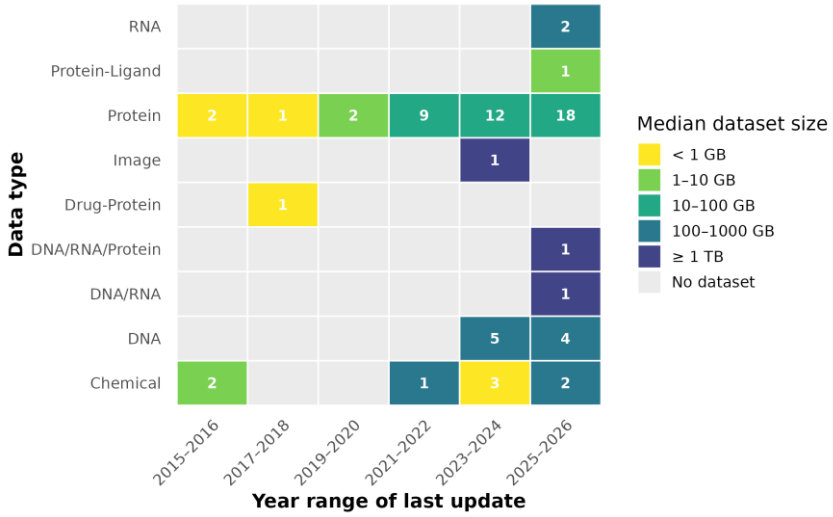
Source: JRC's analysis.

Figure 25. Maintenance status of 75 biological training datasets. Labelled points report the largest, most recently updated (per year) dataset.



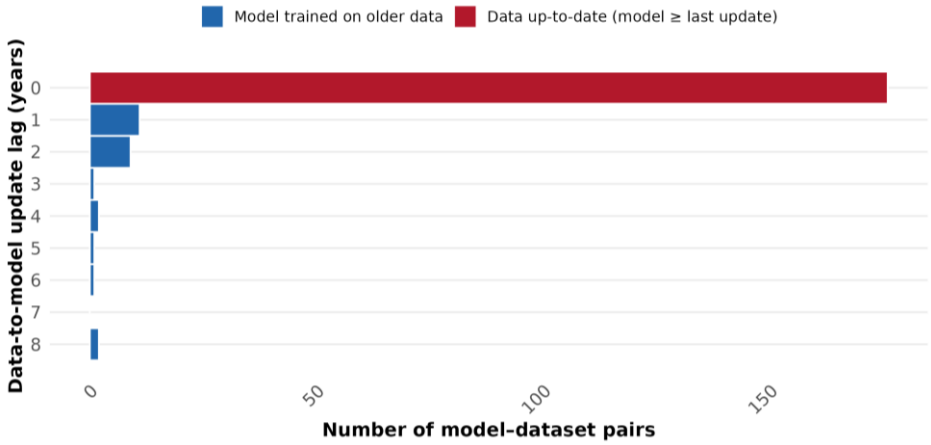
Source: JRC's analysis.

Figure 26. Heatmap of biological training datasets by data type and period of last update. Each cell shows the number of datasets last updated within a given two-year period, grouped by biological data type. Analysis based on 75 datasets.



Source: JRC's analysis.

Figure 27. Lag between dataset last update and model publication, based on 134 and 75 model-dataset pairings retrieved from the Epoch AI dataset. In the figure, model ≥ last update refers to models trained with up-to-date data.



Source: JRC's analysis.

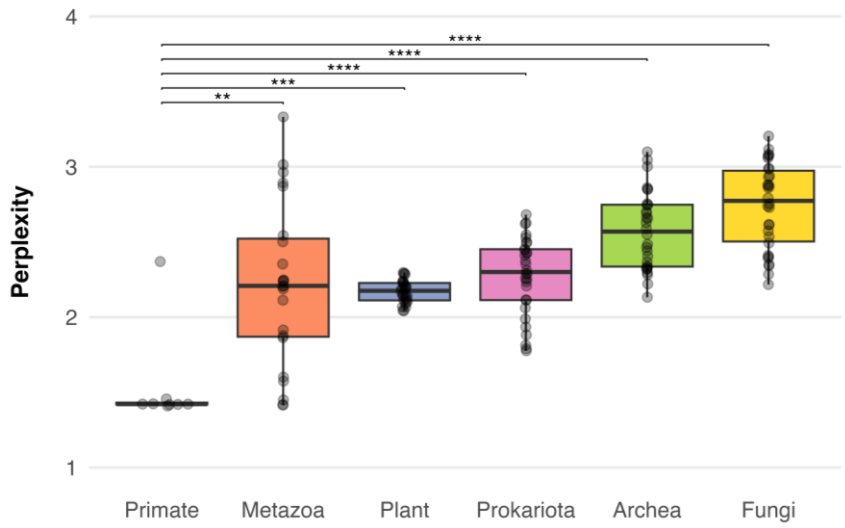
5.2. Training Data Representational Gaps and Generalisation to Novel Contexts

Sequencing artefacts, incomplete functional annotations, and uneven sampling across species, populations, and genomic regions propagate errors into downstream predictions and constrain what models can learn. Whilst architectural choices and data pre-processing strategies may partially mitigate these biases, they are unlikely to fully compensate for fundamental gaps in the underlying training data.

In **genomics**, most benchmark datasets are concentrated on the human genome and a small number of model organisms. This representational imbalance was investigated directly through an experiment using Evo 2, employing *perplexity*[72] as a proxy for training data coverage. Perplexity scores were computed for more than 8 million sequences spanning 150 species across the full tree of life (**Figure 28**).

[72] Perplexity reflects how confidently the model assigns probability to a given sequence; lower values indicate that the sequence is more predictable under the model's learned distribution.

Figure 28. Perplexity distributions across different species and taxa for model Evo2 7B. Wilcoxon tests results indicate statistical significance within all comparisons (P ≤ 0.01).



Source: JRC's analysis.

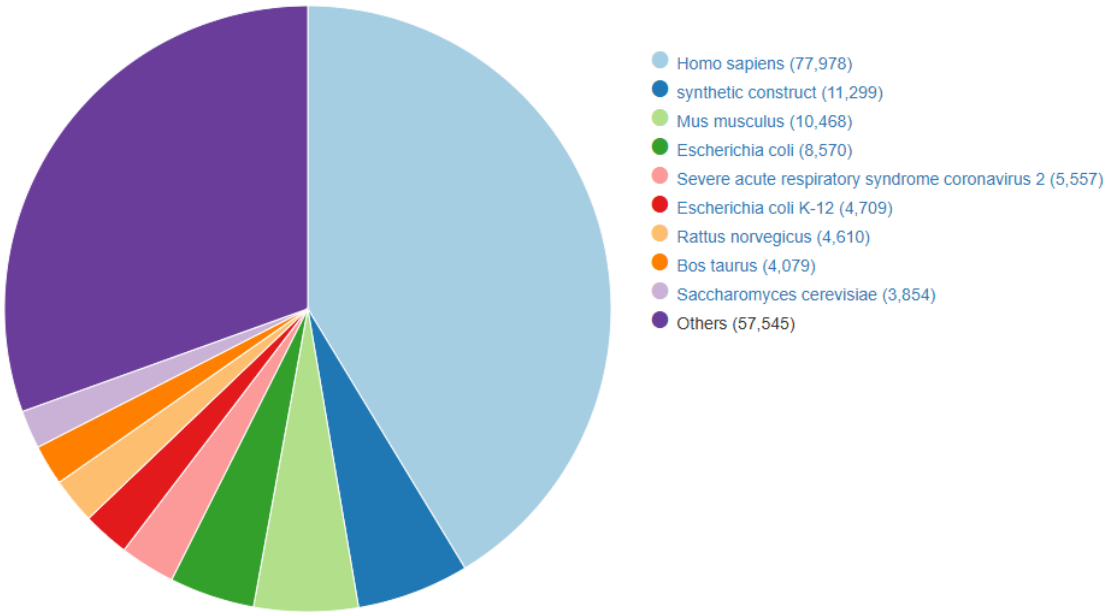
The results revealed a clear gradient in perplexity values across taxa: sequences from clades such as primates, other metazoans, and plants present lower perplexity scores than sequences from fungi and archaea. The observed gradient is therefore consistent with uneven distributional coverage across the tree of life, though it may also partly reflect genuine differences in sequence complexity and homogeneity between clades. Groups with more similar sequences amongst themselves may be intrinsically more predictable, independently of their representation in the training data. Whether and to what extent differences in sequence predictability translate into differences in downstream task performance remains an open empirical question. Nonetheless, the pattern raises the possibility of uneven capability across taxonomic groups, with potential consequences for applications in environmental genomics, agriculture, and infectious disease research. More broadly, different genomic tasks benefit from different evolutionary scopes, and including large amounts of repetitive or intergenic sequence may reduce downstream utility by encouraging models to memorise low-information content rather than learning biologically meaningful patterns (Veiner and Supek, 2026).

For **single-cell** models, training corpora assembled from heterogeneous public repositories frequently suffer from batch effects, variable quality, and inconsistent cell-type annotation. Contrary to the assumption that larger datasets always yield better models, ablation studies suggest diminishing returns beyond a certain corpus size, with curated smaller datasets able to match or exceed noisily aggregated larger ones (Nadig et al., 2025; DenAdel et al., 2025). Models trained on one dataset or tissue type also transfer poorly to others, and the absence of standardised preprocessing and evaluation compounds this by making it difficult to identify where generalisation fails (Qiu et al., 2025; Wenteler et al., 2024).

In **protein** modelling, the PDB provides a clear example of dataset bias, as it is heavily biased towards soluble, globular, well-ordered proteins; membrane proteins are substantially under-represented, as are intrinsically disordered regions, for which the systematic under-representation in training corpora demonstrably reduces model accuracy (Heinzinger et al., 2024). Additionally, the

database currently includes nearly 80,000 entries for *Homo sapiens*, around 10,000 for *Mus musculus*, and only ~5,000 for *Escherichia coli* (**Figure 29**). Such imbalances shape model priors and can lead to uneven performance across biological contexts. Consequently, rapid advances in well-characterised protein spaces may coexist with reduced generalisability in under-sampled domains, slowing progress in therapeutics, microbiome science, plant biology, and non-canonical biomolecular modalities.

Figure 29. Protein Data Bank (PDB) data distribution by scientific name of source organism.



Source: PDB website[73] as of 08/12/2025.

In **drug discovery**, public datasets are skewed towards well-characterised targets and chemical series, leaving large regions of chemical and biological space underrepresented; multi-omics integration introduces further complications from heterogeneous formats and inconsistent annotation standards. Multi-omics integration introduces further complications from heterogeneous formats and inconsistent annotation standards. Models trained on small molecules also generalise poorly to peptides, oligonucleotides, and biologics, where ADMET properties and computational frameworks differ substantially (Tanihata and Iwata, 2026).

5.3. Protein-Centric Bias

As illustrated in Section 3 and Annex 4, protein models dominate the current biological AI landscape. A major reason for this imbalance lies in the maturity of protein data infrastructure. Sequence and structure repositories such as UniProt and the PDB have been curated over decades, yielding large-scale, experimentally validated datasets of a quality and scale that other biological domains cannot yet match. Community benchmarking initiatives such as the Critical Assessment of Protein Structure

[73] https://www.rcsb.org/stats/explore/scientific_name_of_source_organism

Prediction (CASP)[74] further sustain a continuous experimental framework for evaluating and comparing predictive methods. Together, these long-standing investments provide a uniquely rich foundation for training biological AI systems, and the field has responded accordingly.

This concentration of effort has produced genuine scientific advances. Protein structure prediction, function annotation, and increasingly protein design, have reached levels of performance that would have seemed implausible a decade ago. However, the dominance of protein-centric models also reflects and reinforces existing data asymmetries. Biological domains that lack equivalent curation infrastructure (e.g. RNA biology, chromatin organisation, cell signalling, host-pathogen interactions) attract comparatively fewer modelling efforts, not necessarily because the underlying biology is less important, but because the training data needed to build competitive models is thinner and less standardised.

This matters beyond any individual domain. The ambition of much current biological AI research is to move towards multi-modal foundation models capable of integrating information across molecular scales: from genome to transcriptome, proteome, metabolome, and cellular phenotype. Achieving this requires balanced data availability and modelling investment across all these layers. A field disproportionately shaped by protein data risks encoding protein-centric assumptions into the architecture and evaluation practices of the next generation of biological AI systems, making it harder to build or fairly assess models that operate across modalities.

5.4. Oversimplification of Biological Complexity

Biological data presents intrinsic complexity that existing model architectures handle imperfectly. In **genomics**, sequences can span billions of base pairs, yet biologically important signals, such as the interaction between an enhancer and its target gene, can operate across hundreds of kilobases of intervening sequence. Standard transformer architectures, whose computational cost scales quadratically with sequence length, struggle to accommodate these ranges within a single context window. Architectural innovations such as state-space models and hybrid designs have partially addressed this constraint but often introduce trade-offs in training efficiency, resolution, or the ability to capture short-range sequence motifs alongside long-range dependencies (Veiner and Supek, 2026).

In **single-cell** biology, genes within a cell have no natural ordering. To apply transformer architectures, an artificial ordering must be imposed, typically by ranking genes by expression level or discretising expression into bins. Several studies have found only limited benefit from these ordering schemes, with performance often comparable to approaches that use normalised expression values directly, raising questions about how much of the transformer's inductive bias is being exploited (Baek, Song, and Lee, 2025).

For **proteins**, most current structure prediction models, including AlphaFold2 and AlphaFold3, produce a single static structure for a given protein sequence. Real proteins are not rigid objects: they fluctuate, transition between multiple conformations, and change shape upon binding to partners or ligands. This dynamical behaviour is central to how proteins carry out their functions, particularly for enzymes, membrane transporters, and signalling proteins that act as molecular

[74] https://predictioncenter.org/

switches. AlphaFold2 is known to perform poorly on fold-switching proteins and on intrinsically disordered proteins, which lack a stable three-dimensional fold under normal conditions (Chakravarty et al., 2024). AlphaFold3 improves upon this in some respects but continues to struggle with highly dynamic systems. Efforts to extend prediction tools to generate ensembles of conformational states, rather than a single structure, are an active research frontier (Jing et al., 2024).

In **drug discovery**, diffusion-based structure prediction models including AlphaFold3 and Boltz-1 have been shown to produce stereochemical errors in modelled ligands, specifically errors in the three-dimensional arrangement of atoms around chiral centres and in bond lengths and angles, which can be consequential for virtual screening and binding affinity calculations (Ishitani and Moriwaki, 2025).

Addressing these representational limitations is a prerequisite for any meaningful progress toward more integrative approaches to biological AI. The concept of **generalist biological artificial intelligence (GBAI)** envisions systems capable of reasoning jointly across molecular modalities (Rao et al., 2026), from DNA and RNA to proteins and cellular systems, but without models that can faithfully capture the complexity of each individual domain, such ambitions remain a distant objective. The limitations described above are among the foundational barriers to that trajectory. The risks are not only technical: models that oversimplify biological complexity may perform well on benchmarks while failing in precisely the contexts, dynamic, heterogeneous, clinically relevant, where reliable predictions matter most.

5.5. Interpretability

Extracting biologically meaningful mechanistic insights from learned representations remains non-trivial across all domains, hindering both regulatory acceptance and the generation of testable scientific hypotheses. In genomics, attention visualisation provides an incomplete picture of model computations, and state-space models lack established interpretability frameworks entirely (Veiner and Supek, 2026); hybrid models combining predictive and generative objectives introduce further complexity. Single-cell foundation models face the same constraint, limiting uptake in clinical contexts (Xie et al., 2025). In drug discovery, the interpretability gap has direct regulatory consequences: many high-performing models cannot explain their predictions to regulatory agencies, limiting acceptance in regulated development contexts (Jiménez-Luna, Grisoni, and Schneider, 2020).

Regulatory frameworks are beginning to address this. The US FDA published draft guidance in 2025[75] on the use of AI to support regulatory decision-making for drug and biological products, informed by over 500 prior submissions with AI components, and the European Medicines Agency is developing parallel guidance[76]. However, model interpretability remains a feature that can create trust and thereby support broader clinical and regulatory integration that the field has not yet met.

[75] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

[76] https://www.ema.europa.eu/en/news/ema-fda-set-common-principles-ai-medicine-development-0

5.6. Openness and Reproducibility

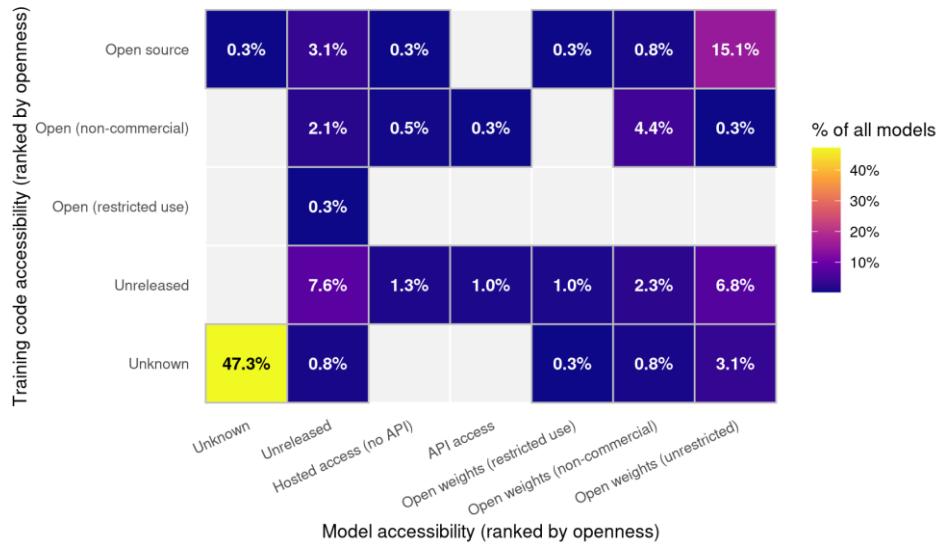
Limited openness in publicly available biological AI models is a tangible barrier to real-world uptake. When model weights and training code are inaccessible, independent validation becomes difficult, regulatory scrutiny is harder to conduct, and clinical or policy adoption is effectively contingent on trusting a black box. Reproducibility is not merely a scientific norm; in high-stakes biological applications, it is a precondition for responsible deployment. In the context of this report, "openness" refers to the degree to which AI4Bio models and their underlying components (model weights and training code) are accessible for inspection, reuse, modification, or deployment. Epoch AI classifies openness based the accessibility of model weights (e.g., open, API-only, unreleased) and training code (e.g., open-source, partially open, unreleased). A full description of this term is provided in **Table 20** (Annex 2).

In AI4Bio, openness should therefore be understood as a layered concept rather than as a binary distinction between open and closed models. Openness may concern different components of a model and its development process, including metadata, model cards, training protocols, benchmarks, evaluation results, source code, model weights, datasets and access through application programming interfaces. These layers do not raise the same scientific, legal or security considerations. While transparent documentation, reproducible protocols and well-described benchmarks are important for scientific scrutiny and responsible reuse, unrestricted access to certain datasets, weights or capabilities may be inappropriate where sensitive health data, intellectual property, biosafety or dual-use concerns are involved. A proportionate approach to openness would therefore support transparency, reproducibility and accountability without assuming that all components of AI4Bio systems must necessarily be released under the same access conditions.

The picture that emerges from the Epoch AI dataset (n=383) is one of **moderate and uneven openness** (see **Figure 30**). Almost half of all models (47.3%, approximately 180 models) provide no disclosure for either model weights or training code. Only 15.1% (approximately 57 models) achieve the most open combination of unrestricted weights and open-source training code. Fully closed systems account for 7.6%, but the more consequential finding is the scale of the unknown category: for nearly half the field, it is simply not possible to determine how accessible a model is. This opacity limits the ability of downstream users, regulators, and the broader scientific community to assess, audit, or build upon existing work.

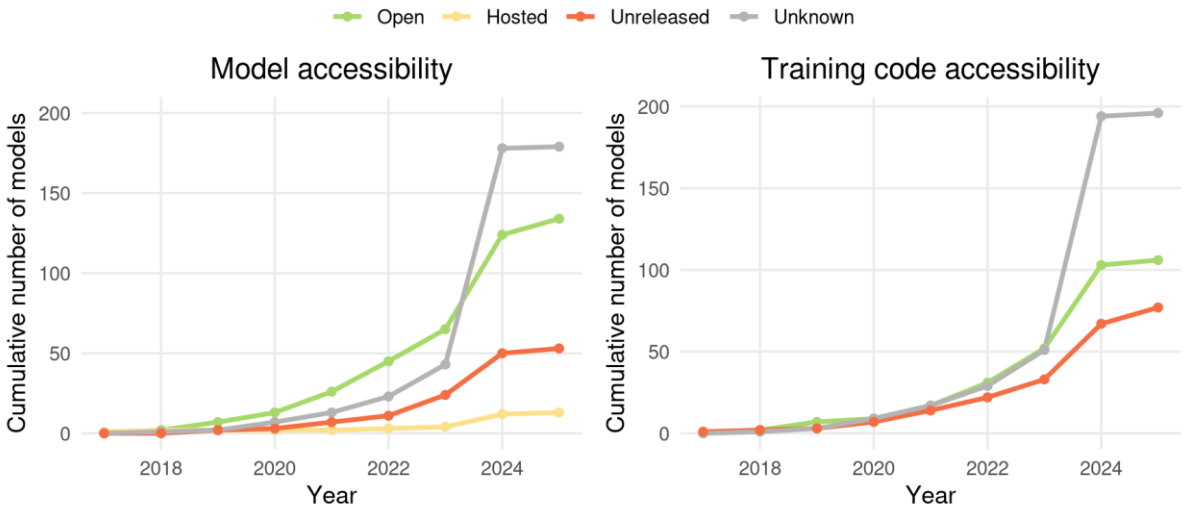
Openness has increased over time, particularly from 2022 onwards, with the cumulative number of models releasing weights or training code growing steadily through 2023 and 2024 (**Figure 31**). However, the volume of undisclosed cases has grown in parallel, suggesting that transparency has not kept pace with the overall expansion of the field.

Figure 30. Model and training code accessibility of 380 biology AI models.



Source: JRC's analysis using Epoch AI dataset.

Figure 31. Temporal distribution of model and training code accessibility of 380 biology AI models.



Source: JRC's analysis using Epoch AI dataset.

5.7. The Maturity Gap: From Benchmark Performance to Deployment Readiness

Biological AI models such as AlphaFold, Evo 2 and AlphaGenome occupy an ambiguous position: publicly available, widely celebrated, and technically impressive, yet lacking the integrated assessment needed to determine whether they are ready for real-world application. This gap reflects the absence of established frameworks that bridge technical development, regulatory requirements, and operational deployment; frameworks that define what readiness means, how it should be measured, and who should determine it.

This section examines the resulting **maturity paradox**[77] through four interconnected issues (**Figure 32**):

1. Current incentive structures reward public visibility and benchmark recognition over rigorous real-world validation.
2. The benchmarks that confer such recognition measure narrow technical performance rather than clinical or operational impact.
3. Applying existing assessment frameworks places these models squarely at research-stage maturity, with no established pathway yet confirming readiness for real-world deployment.
4. Despite these limitations, models retain the capacity to generate real-world impact, via routes that are not yet well governed, ranging from operational integration into research or industrial pipelines, to direct commercialisation of specific clinical applications, to potential biological risks due to misuse (**Box 6**).

This section focuses primarily on clinical readiness[78], though the structural dynamics described here can apply equally to other deployment contexts, including industrial biotechnology or environmental monitoring.

[77] The *maturity paradox* is a novel conceptual framing within this field; the figure is intended to demonstrate its explanatory value in practice rather than to provide definitive classifications, and establishes a basis for further empirical elaboration and development. As such, it also serves as a tool for structuring the conversation between science and policy on biological AI readiness.

[78] Clinical readiness is used here in a broad sense encompassing three distinct deployment contexts: (i) integration into drug discovery and translational research pipelines; (ii) use as a tool within clinical trials, for example to support patient stratification, endpoint biomarker identification, or trial design; and (iii) potential deployment as regulated clinical tools such as AI-assisted genomic diagnostics or variant interpretation software. AI models such as AlphaFold, Evo 2, and AlphaGenome are not currently classified as medical devices, though downstream applications built on these models may progressively fall within regulated frameworks as the field matures.

Box 6. Misuse potential of AI tools used in biology

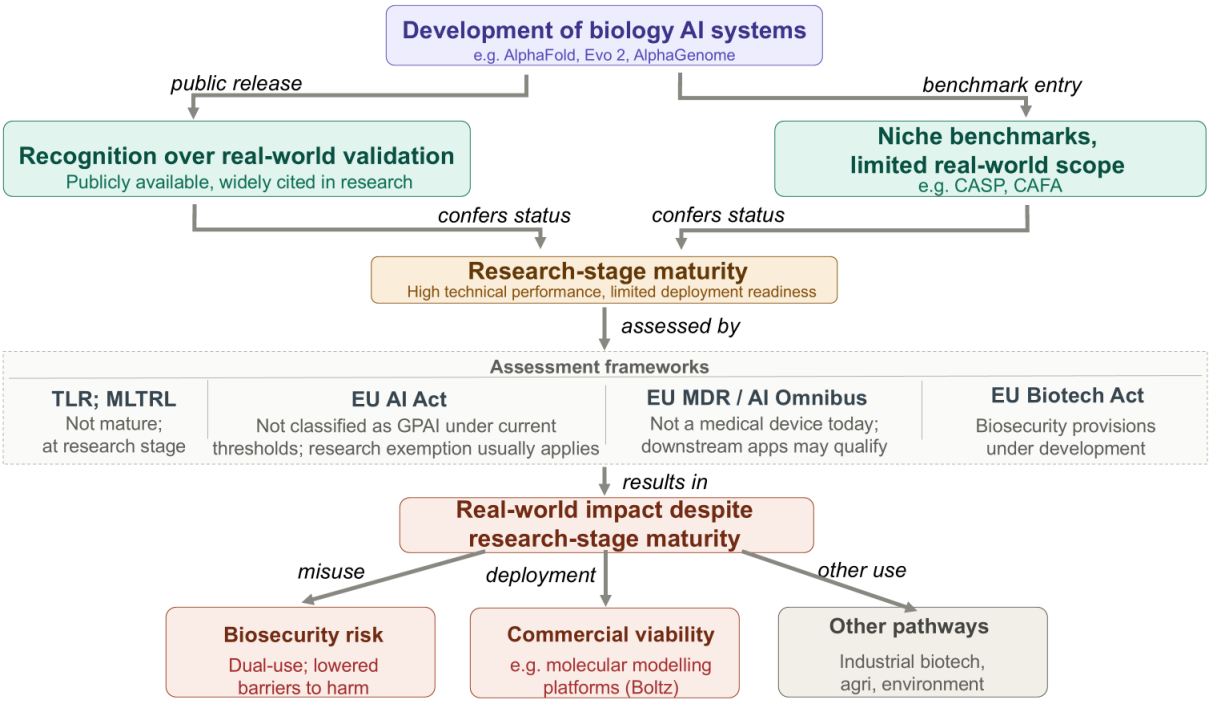
Misuse in the context of biological AI refers to the use of biological foundation models to cause harm in ways that are qualitatively different from pre-AI biosecurity risks. Three dimensions are particularly relevant.

First, AI lowers the barrier to entry for harmful applications which maliciously exploit the large scale generative capabilities of biological foundation models to synthetize potentially harmful biological sequences.AI-driven applications in synthetic biology, protein design, and genetic engineering have expanded the potential for misuse, enabling actors with limited expertise to pursue sophisticated biological capabilities (Wheeler, 2025).

Second, dual-use risks are structural rather than incidental. The same tools that accelerate drug discovery and novel protein design can, if misused, lower the barriers to developing biological weapons with unprecedented efficiency, due to the impressive scale of production and precision (Wheeler, 2025).

Third, existing oversight frameworks are not designed for this risk profile. Risk assessments by biological tool developers are currently outside the scope of the EU AI Act (CFG, 2025), creating a governance gap that the EU is only beginning to address through instruments such as the proposed European Biotech Act (European Commission, 2025), which establishes a framework for preventing the misuse of biotechnology products of concern, including provisions for screening and reporting suspicious transactions[79].

Figure 32. Maturity paradox of biological AI models identifying three main issues. Classifications reflect the authors' expert assessment based on the frameworks described in this report and should be interpreted as indicative rather than definitive[74].



Source: JRC's analysis.

[79] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52025PC1022(01)

5.7.1. Incentive Structures Favour Visibility Over Validation

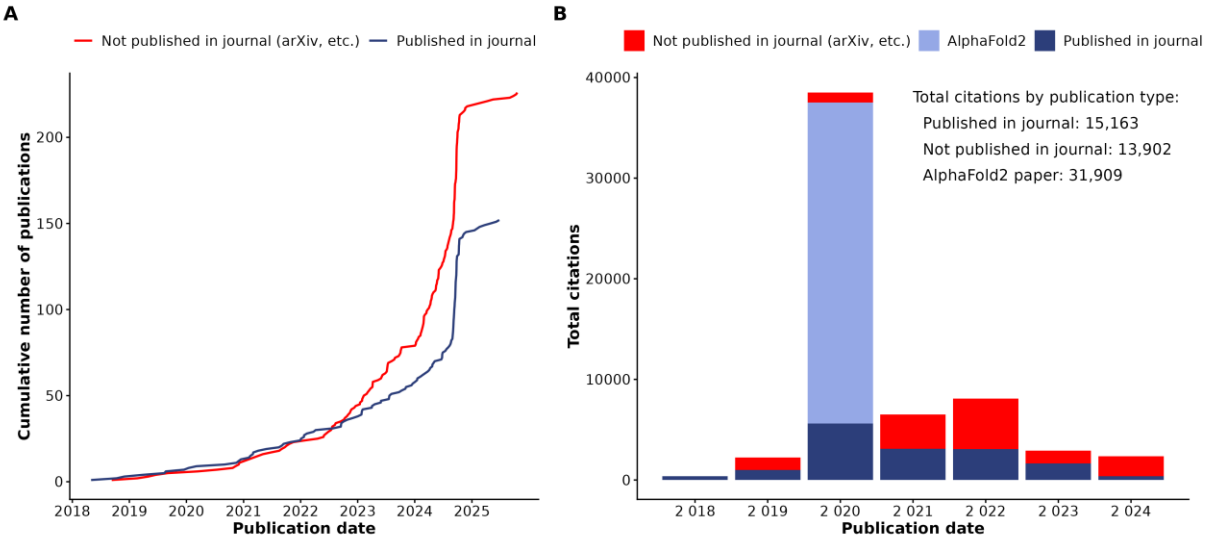
Many biological AI models originate in academic settings and are then refined in biotech startups. However, only a small fraction meets the evidence threshold required for real-world deployment. One explanation is timing: this generation of large, flexible models is still relatively recent, and the field is actively working to define *readiness* beyond benchmark performance, especially under distribution shift, workflow constraints, and post-deployment monitoring expectations (Markowetz, 2024). A structural explanation, however, is that **current incentive structures reward visibility and benchmark performance over long-term clinical validation**.

Academic groups and companies alike push to release models publicly through community hubs such as HuggingFace[80] (Riva et al., 2024), in ways that reflect the visibility and influence of the organisations behind them (Sample, 2026). The annual rate of published preprints on arXiv and similar platforms has consistently outpaced that of peer-reviewed journal articles (**Figure 33**, A), indicating that the dominant mode of dissemination favours speed over formal validation. Citation behaviour reinforces this pattern: researchers reference published and unpublished work with roughly equal frequency (**Figure 33**, B), consistent with a *publish-or-perish* culture, in which accessibility and novelty matter more than peer-review status. The single exception is the period around 2021, where a spike in journal citations is attributable to the AlphaFold model (Jumper et al., 2021b), whose impact was sufficient to distort the overall citation landscape. Beyond that outlier, near-parity between peer-reviewed and preprint citations suggests the community does not systematically reward formal validation when deciding what to build upon.

Early democratisation can theoretically accelerate maturity through independent replication and testing on diverse experimental conditions, allowing faster discovery of failure modes that would otherwise remain hidden. Nevertheless, clinical maturity is determined by intended use, risk profile, clinical evidence, and operational governance, not by popularity or accessibility (Marcinkevičs et al., 2025). This distinction is crucial to understand why current assessment frameworks systematically fail to capture true readiness.

[80] https://huggingface.co/

Figure 33. Publication trends and citation impact of biological AI models. (A) Cumulative number of biological AI models published since 2018, stratified by publication type. (B) Total citations per year, stratified by publication type. AlphaFold2 is highlighted separately given its outsized citation impact. Analysis based on 383 models from the Epoch AI dataset.



Source: JRC's analysis.

5.7.2. Benchmark Performance Does Not Reflect Deployment Readiness

Progress in computational sciences is often driven by benchmarks and challenge-based evaluations. A benchmark typically consists of a shared dataset, a well-defined prediction task, and a fixed evaluation metric used to rank models on a held-out test set. In biological research, this paradigm has been widely adopted across genomics, proteomics, and clinical domains, with challenge platforms organised either by research communities (CASP[81], CAFA[82], DREAM[83], CAGI[84]), or by broader competition platforms such as Kaggle[85] (**Figure 34**). These competitions enable systematic comparison of methods, promote reproducibility, and accelerate technical iteration

[81] https://predictioncenter.org/

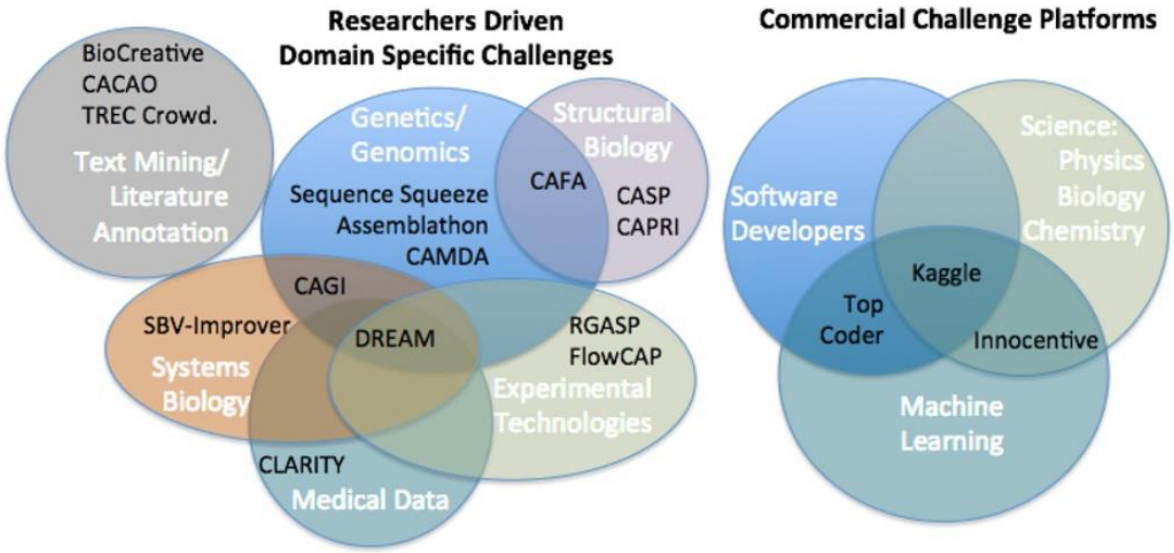
[82] https://biofunctionprediction.org/cafa/

[83] https://dreamchallenges.org/

[84] https://genomeinterpretation.org/challenges.html

[85] https://www.kaggle.com/

Figure 34. Non-exhaustive subset of challenge platforms and organisations in the biomedicine space (left), and some platforms that organise challenges in a variety of areas of science and technology (right).



Source: (Stolovitzky et al., 2023)

However, their design introduces important limitations. Because tasks are narrowly specified and success is tied to fixed metrics, models are often optimised for benchmark performance rather than biological insight or clinical relevance. Improvements in prediction accuracy or reconstruction error do not necessarily translate into mechanistic understanding or actionable outcomes in real-world settings.

This limitation is particularly visible in the uneven distribution of benchmarks across biological domains. Protein-related tasks dominate, largely because they offer well-defined ground truth and standardised metrics, with landmark examples including CASP for structure prediction and CAFA for function prediction, and notable milestones such as AlphaFold's success in CASP14 (Jumper et al., 2021a). Genomic AI benchmarks remain less standardised, reflecting both the relative novelty of large-scale genomic models and the difficulty of defining consistent ground truth across diverse biological contexts.

Beyond domain imbalance, the proliferation of benchmarks has produced fragmentation: multiple challenge initiatives address similar scientific questions using different datasets and protocols, so the same model may perform substantially differently depending on the benchmark chosen. This has given rise to concerns about "**benchmark fatigue**"[86], the continuous introduction of new benchmarks without consolidation, prompting initiatives such as the Benchmarking, Evaluation, and

[86] The term "benchmark fatigue" is used in the literature to describe the growing burden of evaluating an ever-increasing number of methods. Cao et al. (2025a) document this trend empirically, noting that the median number of methods evaluated in benchmark papers increased from 4 to 13 over a nine-year period within the single-cell field, with 40-50% of papers evaluating more methods in their published versions than in their preprint counterparts, reflecting the compounding pressure on evaluation practices across computational biology.

Assessment Consortium for Science (BEACON)[87] to develop shared evaluation infrastructure for areas including drug discovery.

Three further technical phenomena compound these limitations. Saturation occurs when top-performing models cluster near ceiling scores, rendering the benchmark unable to discriminate between methods (Cheng et al., 2025; Akhtar et al., 2026). Obsolescence arises when evaluation tasks fall behind evolving biological standards. For example, CASP itself progressively retired categories such as contact prediction after AlphaFold2 rendered them redundant, replacing them with more challenging targets such as RNA structures and protein ensembles[88]. Data contamination refers to overlap between training and test sets that inflates apparent performance; in biological AI specifically, protein language models pre-trained on large sequence repositories have been shown to produce unrealistically optimistic scores for proteins closely related to pre-training data, with performance dropping substantially for more divergent sequences (Bernett et al., 2024; Hermann et al., 2024).

Finally, the governance of benchmarks raises questions of its own. As some initiatives lose public funding, maintenance may shift towards private stakeholders, sometimes including organisations developing competing models, introducing concerns about impartiality and long-term sustainability in the absence of coordinated oversight (Trang, 2025).

Even well-designed benchmarks, then, **do not fully address the challenge of real-world translation. Strong benchmark performance reflects technical readiness, but not regulatory maturity or deployment readiness: models must also satisfy requirements for validation in clinical workflows, regulatory compliance, interpretability, and safety**.

5.7.3. Fragmented Frameworks Lock Models at Research-Stage Maturity

Strong benchmark results do not imply that a model is ready for deployment in research, industrial, or clinical settings. Across existing readiness, deployment, and regulatory frameworks, biological AI models are typically positioned at a research-stage or nascent level of maturity. Not because relevant frameworks are absent, but because **the current landscape is fragmented**. Different frameworks assess different dimensions of maturity (technical development, evidential robustness, risk management, product safety, regulatory compliance), and no single framework has become the reference for assessing biological AI models across the full pathway from development to deployment.

From a technical perspective, Technology Readiness Levels, and their machine learning adaptation, Machine Learning Technology Readiness Levels (MLTRL) (Lavin et al., 2022), conceptualise progression from early research to fully integrated systems, emphasising that system-level readiness depends on the maturity of all components, including data pipelines, validation protocols, and deployment infrastructure. From an evidential perspective, guidance such as the World Health Organization framework for AI-based medical devices highlights the need for progression from retrospective benchmarking to prospective validation, external evaluation, and post-deployment

[87] Introducing BEACON: The Benchmarking, Evaluation, and Assessment Consortium for Science - Conscience

[88] https://predictioncenter.org/casp15/

monitoring[89]. From a development and quality perspective, Good Machine Learning Practice principles emphasise data quality, transparency, human oversight, and lifecycle management[90].

Box 7. Defining Research-Stage (Nascent) Maturity in Biological AI Models

Biological AI models are often described as being at a research-stage or nascent level of maturity. This classification does not primarily reflect their technical performance, but rather their position within existing readiness, evaluation, and regulatory frameworks.

Drawing on complementary perspectives from technical readiness (e.g. MLTRL), evidential evaluation (e.g. WHO guidance), and development and quality principles (e.g. Good Machine Learning Practice), a model can be considered at a nascent maturity level when it exhibits most of the following characteristics:

- **Benchmark-bound evaluation**: performance is demonstrated primarily on curated benchmark datasets or challenge tasks;
- **Retrospective validation**: evaluation relies on pre-existing datasets rather than prospective or real-time studies;
- **Limited external validation**: performance has not been consistently demonstrated across independent laboratories, institutions, or populations;
- **Lack of workflow integration**: the model is not embedded within real-world research, clinical, or operational decision-making processes;
- **Absence of lifecycle governance**: there are no established mechanisms for monitoring, updating, version control, or managing performance over time;

Undefined intended use and regulatory pathway: the model is not aligned with a clearly specified application context that would trigger formal regulatory assessment. Within this framework, many contemporary biological AI models can be described as benchmark-mature but deployment-immature, highlighting the gap between technical capability and real-world readiness

This distinction further illustrates that evidence generation, independent validation and operational deployment represent separate stages requiring different forms of assessment. This gap between technical capability and deployment readiness is further reflected in existing regulatory frameworks. In the US, the FDA assesses maturity through evidence of safety and effectiveness within a defined clinical context, quality systems, human factors evaluation, and post-market surveillance[91], not through TRLs. The FDA's Software as Medical Device (SaMD) framework provides a regulatory pathway only when a system is intended for medical use, while complementary initiatives such as the National Institute of Standards and Technology (NIST)'s AI Risk Management Framework and the AI Bill of Rights, offer broader but non-binding guidance for implementing biological AI models in real-world healthcare systems (Wells et al., 2025). Additionally, the FDA's drug centre (CDER) has launched an AI council to coordinate AI use and oversight in drug development contexts[92].

A similar separation is observed in the EU. TRLs are widely embedded in research funding and innovation policy to characterise developmental stage. However, clinical deployment of AI systems

[89] who.int/publications/i/item/9789240038462

[90] https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles

[91] https://www.fda.gov/media/154985/download

[92] https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/artificial-intelligence-drug-developmen

is governed separately under the Medical Device Regulation (MDR) which applies only when a system is placed on the market with a defined medical purpose. When the MDR applies, conformity assessment and CE marking[93] evaluate compliance with safety, performance, and quality-management requirements. Importantly, CE marking does not correspond to a specific TRL level, nor does it automatically demonstrate clinical utility. The European Medicine Agency (EMA) has published a reflection paper on the use of AI across the medicinal product lifecycle, advocating a risk-based approach to development and deployment, but does not provide a structured framework for assessing the maturity of biological AI models specifically[94].

The introduction of the Regulation (EU) 2024/1689 (AI Act) (2024) further illustrates this gap. While the Act establishes obligations for certain high-risk AI systems and general-purpose models, it does not fully capture domain-specific risks associated with biological applications that fall outside its scope. Even where biological AI models fall within scope, the research exemption under Article 2(6) means that models used for scientific research and development remain largely outside its practical reach.

In practice, **fragmented frameworks remove the structural incentives that would otherwise drive biological AI models towards deployment readiness,** acting as a barrier to the maturity gate (i.e. the threshold a model must cross to be considered ready for real-world deployment). Regulatory oversight is typically triggered only when a system is marketed for a specific medical purpose, while research-stage models continue to evolve outside formal deployment pathways. In the absence of a clear regulatory signal, developers face little external pressure to bridge the gap between technical performance and the evidential, operational, and governance foundations that real-world deployment requires. Where developers do seek to cross that threshold, regulatory fragmentation across jurisdictions and use-case categories compounds the burden, making the pathway to deployment harder to navigate and easier to abandon in favour of markets with clearer frameworks, reinforcing a pattern where models are benchmark-mature but deployment-immature.

Emerging policy initiatives begin to address this gap. The Biotech Act (European Commission, 2025) recognises AI's potential role in accelerating clinical trials and drug development, proposes an expanded EMA role in evaluating AI systems across the lifecycle of medicinal products (Article 31), and promotes **regulatory sandboxes** as controlled environments for testing novel biotechnological approaches under supervision. These mechanisms may provide a pathway for transitioning AI models from research-stage maturity towards deployment readiness.

5.7.4. Real-World Consequences Without Adequate Oversight

The absence of reference maturity assessment frameworks does not prevent biological AI models from generating real-world consequences. This gap is particularly visible in two domains: biosecurity, and commercial and clinical deployment.

[93] https://single-market-economy.ec.europa.eu/single-market/goods/ce-marking_en

[94] https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle-scientific-guideline

5.7.4.1. Biosecurity risks

Biosecurity refers to the set of measures intended to protect human, animal and environmental health against biological threats, whether these arise from naturally occurring pathogens, accidental release, or the deliberate misuse of biological agents and biotechnologies. The proposed European Biotech Act (European Commission, 2025) places this concept at the core of the EU's life sciences agenda, recognising that rapid biotech innovation necessitates safeguards against misuse. This framing explicitly encompasses accidental release, dual-use research, and technology diffusion, not only malicious actors.

Biological AI models have rapidly emerged as a biosecurity concern in their own right, capable of generating outputs that pose direct physical threats (e.g. designs for toxins, synthetic compounds, or viral sequences realisable in a laboratory) or indirect enabling risks (e.g. detailed guidance on weaponisation or circumvention of biosafety measures). These risks are already demonstrated empirically: a model analysed in this report, Evo 2, have proven vulnerable to jailbreaking, generating outputs that closely mimic dangerous human pathogens (Zhang et al., 2025b) , while earlier work showed that a *de novo* molecule generator could be repurposed to design highly bioactive and toxic molecules (Urbina et al., 2022).

The misuse dilemma is structural: the same models that could revolutionise drug discovery or biomaterial design can, if misused, lower barriers to biological harm. Unlike natural language, where harmful content can be flagged through keyword analysis, biological sequences and chemical compounds pose a distinct challenge. Their complexity means that even trained biohazard experts may struggle to identify dangerous outputs without specialised tools, often the very AI models that generated them. Furthermore, the level of accuracy these technologies achieve in generating complex biological compounds, once requiring years of expert labour, can now be replicated in seconds using widely accessible GPU processors. When combined with autonomous laboratory systems, this lowers the technical barrier to biological engineering for actors with minimal specialised training, representing a paradigm shift in both scientific opportunity and biosecurity risk. The same characteristics that enable beneficial scientific applications may therefore require proportionate governance, monitoring and evidence-based oversight mechanisms.

5.7.4.2. Commercial viability and operational deployment

Biological AI models are increasingly being positioned for commercial applications, particularly in drug development and clinical decision support. Biomedicine has become the dominant frontier for investment by leading AI companies, with industry giants such as Google (through DeepMind and Isomorphic Labs), OpenAI (in collaboration with Retro Biosciences[95]) and Meta[96] committing to numerous activities in the field. Several of these companies have also published internal frameworks addressing biological risk, though these typically sit within broader CBRN[97] or general-purpose AI safety policies rather than being specific to biological AI models. For instance, OpenAI's

[95] https://openai.com/index/accelerating-life-sciences-research-with-retro-biosciences/

[96] Meta's foundational protein language model work (ESM) gave rise to EvolutionaryScale, now a leading AI biology company

[97] CBRN: chemical, biological, radiological and nuclear

Preparedness Framework (2025)[98,99] and Google DeepMind's Frontier Safety Framework (Version 3, September 2025)[100] addresses CBRN risks alongside cyber, machine learning R&D, and harmful manipulation. Together, these signal a shift towards self-regulation ahead of formal policy interventions.

While these steps may appear proactive, such efforts warrant caution. **A structural tension exists when the same actors develop, deploy and set the safety standards for their own technologies**, particularly where independent verification is limited. Recent analysis of OpenAI's Preparedness Framework, for instance, finds that the document requests evaluation of only a small minority of identified AI risks, encourages deployment of systems with 'medium' capabilities for enabling severe harm, and allows the company's CEO to bypass internal safety recommendations (Coggins et al., 2025). Adding to this concern is the limited transparently around critical risk assessments, including findings on model vulnerabilities, which are not consistently shared with policymakers or independent oversight bodies (Orben and Matias, 2025). Taken together, these observations suggest that such measures may constitute baseline compliance necessary to commercialise high-risk products.

[98] https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

[99] These framework cover frontier general-purpose AI systems. The OpenAI's biology-specific model "GPT-4b micro", trained on protein sequences for protein engineering, is not, to our knowledge, evaluated under the Preparedness Framework.

[100] https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf

6. Global Landscape and EU Positioning in Biological AI and Implications for Real-World Adoption

The real-world adoption of biological AI models depends not only on overcoming the scientific and technical barriers described above, but also on the broader structural conditions under which these models are developed and deployed. This chapter examines five such dimensions: the governance and accessibility of training data, the geographic and institutional distribution of model development, the hardware and software infrastructure required to build and run biological AI systems, the collaboration patterns that shape knowledge flows across countries and organisation types, and the funding landscape that sustains this ecosystem. Taken together, these dimensions determine where capabilities are concentrated, where dependencies are forming, and what the implications are for EU positioning and the prospects for real-world translation[101].

6.1. Training Data Management

The availability of data for AI development, the simplification of data rules, and the strengthening of data flows have emerged as a pressing policy priority, as evident in the European Commission's recent initiatives, including the Data Union Strategy[102]. Data underpin the development of AI models in biology, as well. The quality, diversity and accessibility of biological datasets strongly influence model performance, reproducibility and, ultimately, scientific impact.

The governance of biological data raises important legal and ethical considerations; for example, the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits recognises the sovereign rights of nations over their biological resources and the strategic value of genetic data (2011).

Biological AI models are increasingly trained on datasets that are global in ambition but uneven in origin. While AI4Bio research aspires to universal applicability, from protein structure prediction to pathogen surveillance, the underlying data infrastructures remain geographically concentrated and institutionally bounded. Understanding **where datasets are hosted, how collaboration occurs, and under what conditions data are shared** is therefore essential for assessing both the scientific robustness and the governance implications of AI-driven biological research.

Figure 35 shows a strong concentration of training datasets hosted in the United States (US) and the European Union (EU), with these two regions accounting for the clear majority of identified datasets. This pattern reflects long-standing investments in biomedical research infrastructures, large-scale funding programmes, and mature data repositories. The United Kingdom (UK), Switzerland and China form a second tier, while contributions from Japan are present but more limited in scale. Other regions, including Russia, Norway, India, Canada, and Australia, are only marginally represented.

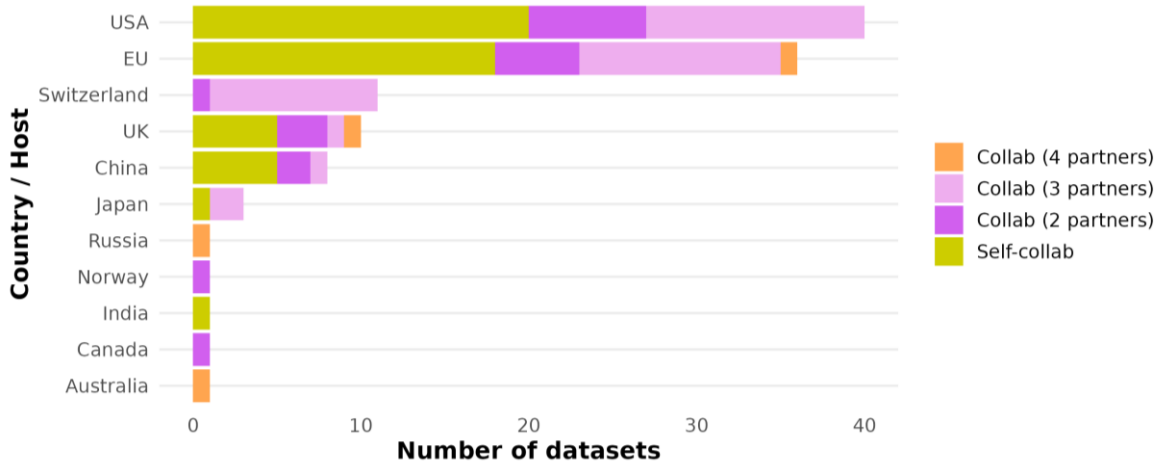
[101] Disclaimer: The analyses presented in this chapter are based exclusively on the Epoch AI dataset. This dataset captures models that have been publicly disclosed, primarily through academic publications and open reporting. As a result, some actor types may be systematically underrepresented for competitive or proprietary reasons. Nonetheless, given the dataset's breadth, the analyses are considered to provide a comprehensive and representative picture of the publicly visible landscape of biological AI model development, offering a solid empirical basis for the findings and comparisons presented in this chapter.

[102] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52025DC0835

Beyond overall volume, the figure also highlights differences in collaboration patterns. In both the US and the EU, a substantial share of datasets is developed through international partnerships, particularly two- and three-country collaborations, alongside a large base of single-country ("self") datasets[103]. Switzerland shows exclusive reliance on collaborative project. UK and China combine a noticeable number of self-hosted datasets with a growing participation in multi-country collaborations. By contrast, countries with smaller overall representation tend to appear either primarily in collaboration or in a limited number of single-country projects.

Figure 36 provides a more detailed view of collaboration patterns between countries. The diagonal cells show datasets developed within a single country, while the off-diagonal cells show datasets produced through international collaboration. The heatmap makes clear that **the US and the EU are not only the largest individual hosts, but also the most frequent collaboration partners, with strong co-occurrence across multiple countries, suggesting that they play a central role in cross-border dataset development**. Switzerland, although contributing a smaller total number of datasets, appears prominently in collaborative links, particularly with the EU and the US. Overall, the figure highlights that dataset production is concentrated in a few key regions that are also highly interconnected through international partnerships.

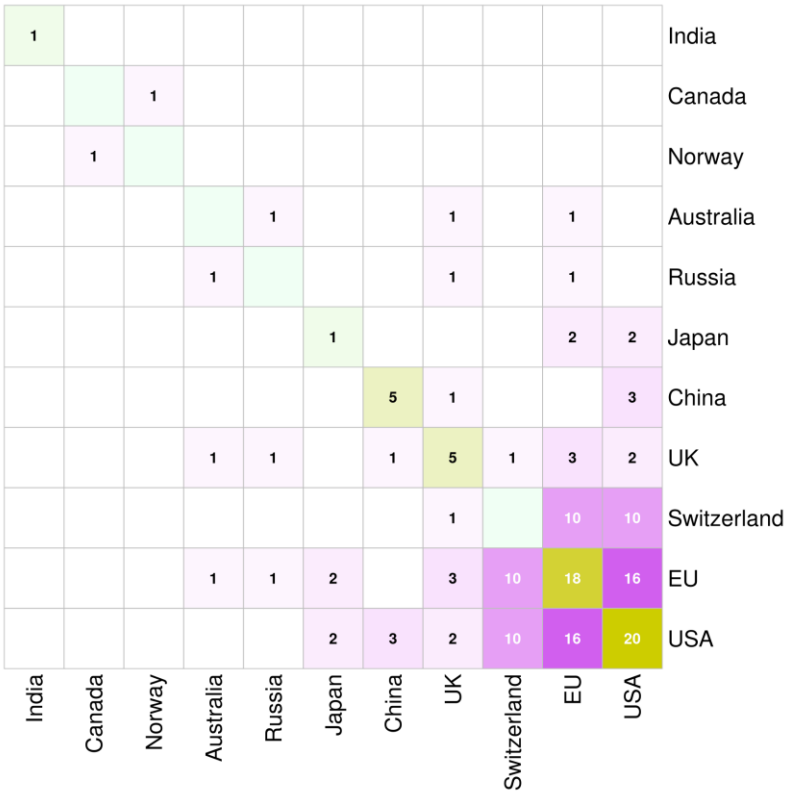
Figure 35. Number of training datasets by host country.



Source: JRC's analysis.

[103] Collaboration categories are assigned at the dataset level. A dataset with institutions from Germany, France, and Italy, for example, counts as a single intra-EU collaboration. Datasets from inter-governmental organisations such as EMBL may appear under multiple country entries, reflecting each national affiliation represented in the figure. For EMBL, the dataset appears under both EU and UK entries, with a collaboration label of 2 partners.

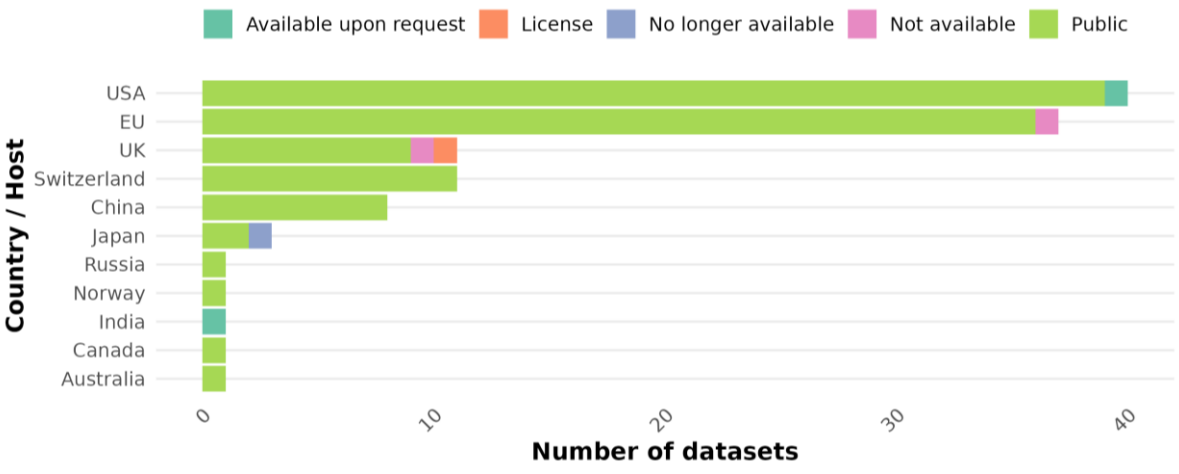
Figure 36. Country collaboration in AI-for-biology datasets. Diagonal cells (yellow) indicate single-country (self) collaborations, and off-diagonal cells (pink) indicate multi-country collaborations.



Source: JRC's analysis.

Figure 37 illustrates the accessibility status of the datasets used for training biological models. The majority are publicly accessible, underscoring the continued importance of open science norms in biology. Three datasets (BRD (Basecamp Research Data), Cambridge Structural Dataset and R-SIM) are subject to access restrictions, including registration requirements, licensing conditions, or limited availability through institutional agreements. Only one dataset (PMD4k) is no longer accessible, reflecting issues of sustainability, governance transitions, or repository shutdowns.

Figure 37. Openness of 75 training datasets.



Source: JRC's analysis.

The observed patterns of concentration, collaboration, and openness intersect directly with broader questions of data governance and ethics in AI-enabled biology.

The most immediate governance concern is geopolitical control over a strategically significant resource. **The concentration of training datasets in a small number of regions means that decisions made by a handful of institutions or governments can determine which data remain accessible to the global research community**. Recent episodes of database shutdowns driven by political priorities, or of asymmetric data accumulation by actors offering limited reciprocity, illustrate that biological data infrastructure is not insulated from geopolitical dynamics[104]. For European policy, this raises a direct question: whether current patterns of dataset hosting and collaboration are consistent with strategic autonomy objectives, and whether collaborative mega-datasets are being built under governance arrangements that protect European interests. The geographic distribution shown in **Figure 35**, particularly when read alongside dataset size, is therefore relevant not only as a descriptive finding but as an indicator of where structural dependencies may be forming.

A related concern arises from the relationship between publicly funded data infrastructures and private model development. **Most training datasets identified in this analysis are publicly accessible**, often the product of sustained public investment in open science. **Private actors can and do build highly capable, and in some cases commercially restricted, models on this public foundation**. This raises a legitimate policy question: *under what conditions should access to publicly funded biological data translate into obligations regarding model openness, licensing, or benefit sharing?* The governance of data access modalities, rather than data availability per se, is where the most consequential policy choices are likely to arise. The UK Biobank model, which conditions access on research use agreements while remaining open to international applicants, offers one reference point for how these tensions can be managed (Sudlow et al., 2015).

Open access data accelerates innovation in public health (Ritoré et al., 2024)**.** However, the same openness that drives scientific progress also raises questions about dual-use risk. The dual-use potential of biological data is not new, and openness has historically been defended on the grounds that the benefits to public health, pandemic preparedness, and basic science far outweigh the risks of misuse by actors who could often obtain the same information through other means (Imperiale and Casadevall, 2015).

What has changed is not the existence of dual-use risk but the capability of models to extract actionable inference from biological sequence data at scale from certain categories, like pathogen data, (Bloomfield et al., 2026). In the context of the 50th anniversary of the Asilomar Conference, more than 100 international researchers issued a public appeal urging safeguards against the misuse of advanced AI systems in harmful biological applications (Bromberg et al., 2025), proposing a tiered system of data access levels. Building on this initiative, Bloomfield et al. (2026) argue for tailored access controls on a narrow subset of high-risk datasets, rather than broad restrictions on biological data as a whole. This perspective is particularly relevant for AI4Bio, where training data can directly shape model capabilities related to pathogen design, virulence prediction, or immune

[104] https://www.science.org/content/article/researchers-china-and-five-other-countries-concern-barred-nih-databases

evasion. The Evo 2 case, however, illustrates the limits of data restriction as a primary safeguard: capable models can be developed (Zhang et al., 2025b) even when specific datasets are withheld[105], suggesting that access controls on training data are a necessary but insufficient instrument, and that model-level governance deserves equal attention.

National and regional frameworks seek to balance openness with sovereignty, ensuring that data sharing aligns with legal, ethical, and security requirements (National Academies of Sciences et al., 2024). European initiatives such as the European Health Data Space[106] exemplify this approach, aiming to enable secondary data use while maintaining strong governance safeguards.

Finally, governance frameworks that introduce stricter access controls, while justified for security or privacy reasons, risk disproportionately affecting researchers in lower-resource settings unless accompanied by capacity-building measures and trusted research environments that enable secure, remote participation. Bloomfield et al. (2026) explicitly caution that biosecurity-motivated controls "*should not exclude researchers from less wealthy institutions or countries*", a consideration that is directly relevant to European policy choices.

Taken together, these findings suggest that data governance is not merely a compliance issue for AI4Bio, but a structural determinant of who can develop, validate, and benefit from biological AI models. As regulatory frameworks such as GDPR, the proposed AI Act, and emerging biotechnology policies continue to evolve, aligning openness, security, and equity will remain a central challenge for the real-world deployment of biological AI across the European and global research and innovation ecosystem.

6.2. Model Development

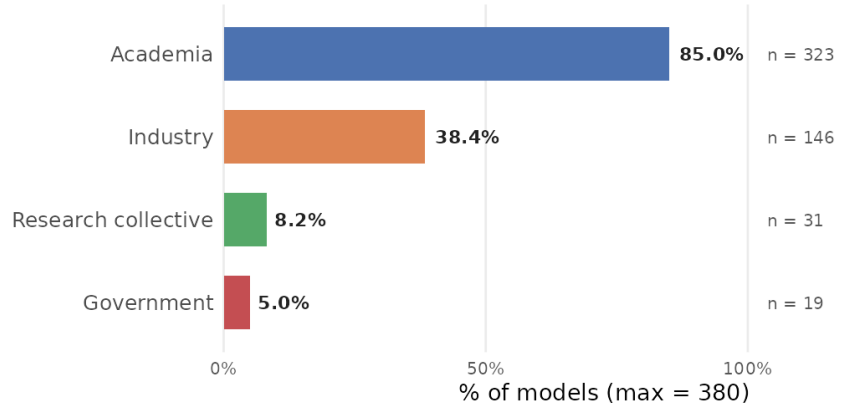
Academic organisations[107] dominate the AI4Bio public landscape (**Figure 38**). They are involved in 85% of models, reflecting the field's strong links to publicly funded research and university-based science. Industry participates in 38% of models, indicating growing commercial engagement but a still secondary role compared with academia. Research collectives and government bodies appear far less frequently, contributing to 8% and 5% of models respectively. Over time, academic participation has grown rapidly since 2021 and accounts for the largest increase in model development (**Figure 39**). Industry involvement has also expanded steadily, particularly from 2022 onwards, although at a lower level. By contrast, research collectives and government actors remain limited throughout the period, with only modest growth.

[105] https://developer.nvidia.com/blog/understanding-the-language-of-lifes-biomolecules-across-evolution-at-a-new-scale-with-evo-2/

[106] https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en

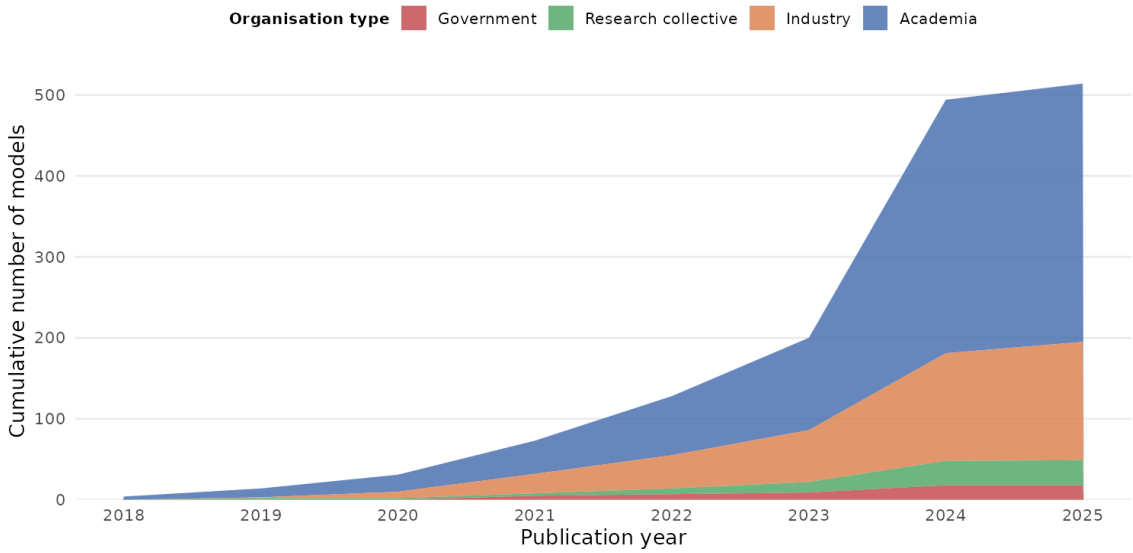
[107] Organisation types follow the classification used by Epoch AI. Four categories are distinguished: academia, industry, research collectives, and government. See **Table 20** for further details on organisation types.

Figure 38. Distribution of organisation type (n=383 models). A model is counted once per category even if the type appears multiple times in its entry.



Source: JRC's analysis using Epoch AI dataset.

Figure 39. Cumulative evolution of organisation type over time. Each model is counted once per category (n=383 models).

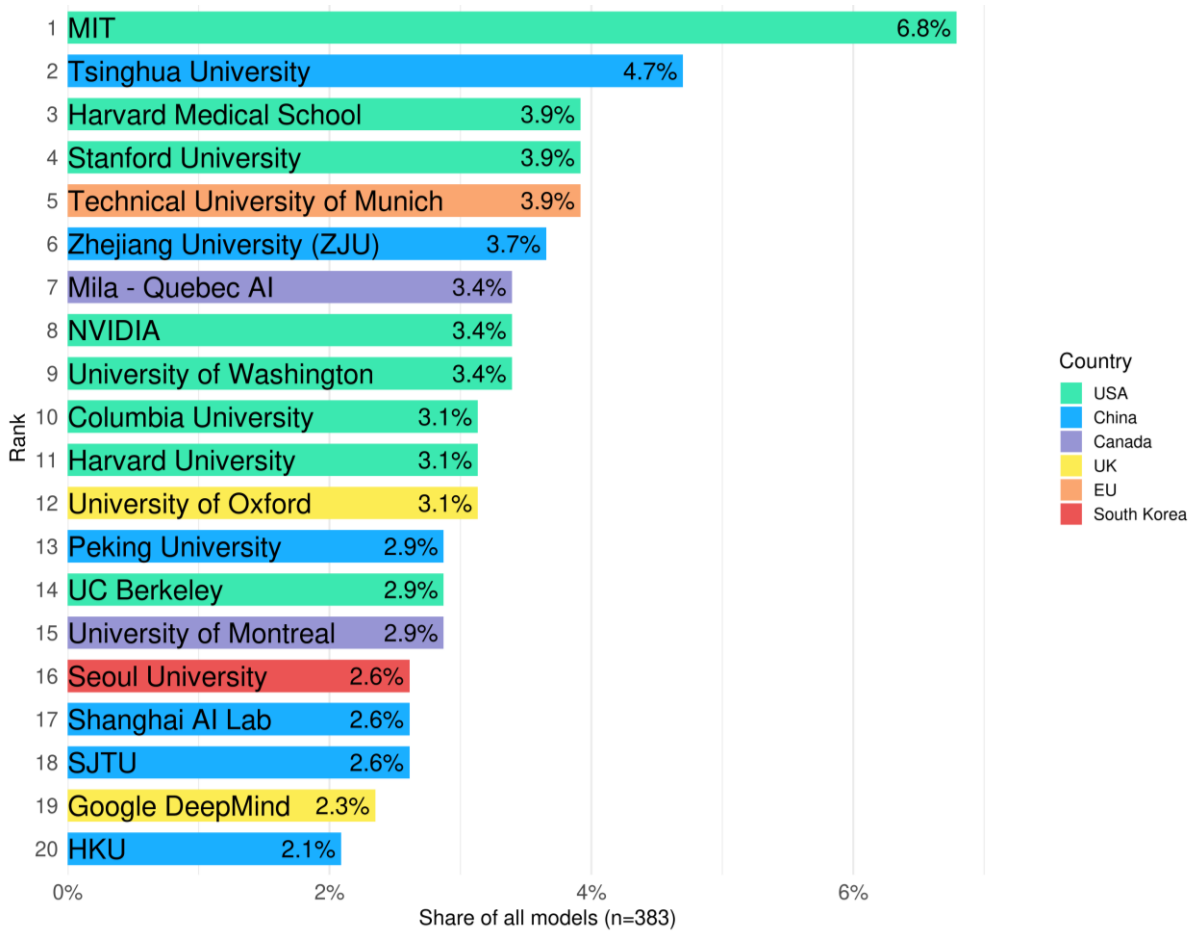


Source: JRC's analysis using Epoch AI dataset.

The **top 20 organisations involved in AI4Bio model development** are dominated by major universities and public research institutions, alongside a small number of leading industry labs (**Figure 40**). The Massachusetts Institute of Technology (MIT) (US) ranks first, contributing to 6.8% of all models, followed by Tsinghua University (4.7%) (CN) and several major US universities, including Harvard Medical School, Stanford University, and Columbia University. Chinese institutions also feature prominently, with Tsinghua University, Zhejiang University, Peking University, Shanghai AI Lab, Shanghai Jiao Tong University (SJTU) and Hong Kong University (HKU) among the top contributors, reflecting China's strong academic investment in AI and biological research. European representation is more limited. The Technical University of Munich is the only EU-based organisation in the top five, while the University of Oxford is the only UK representative. A limited number of industry

actors, including NVIDIA (US) and Google DeepMind (UK), appear in the top tier, indicating that while commercial players are active, model development remains centred in academia.

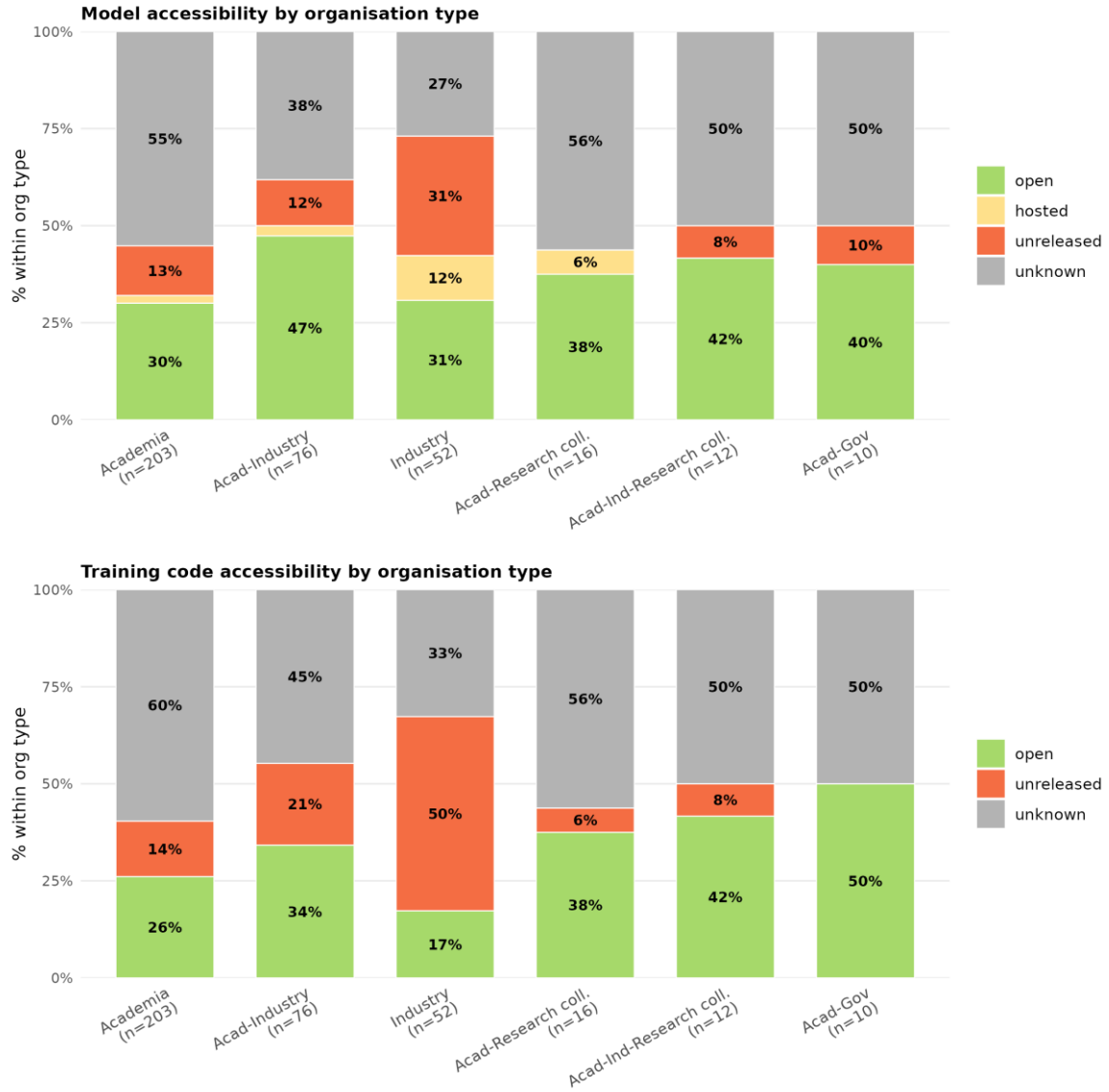
Figure 40. Top 20 organisations contributing to 383 biology AI models.



Source: JRC's analysis.

Openness is structurally uneven across organisation types (**Figure 41**). For model weights, the share of openly accessible models ranges from 30% in academia to 47% in academic-industry collaborations, with most other categories falling between 38% and 42%. Training-code availability follows a broadly similar pattern, ranging from 26% to 50% across non-industry organisation types, and tending to be somewhat higher in academic-industry collaborations. Industry-only models are a clear outlier: only 17% release training code, and 31% of model weights are unreleased, against 8% to 13% elsewhere. Unreleased training code accounts for 50% of industry models, while most other organisation types remain below 21%. The large share of unknown disclosure cases across all organisation types, ranging from 38% to 56% for model accessibility and 45% to 60% for training code, limits interpretation and points to the absence of consistent reporting standards. Notably, industry stands out positively on this dimension: unknown cases are lower for industry models (27% for model accessibility and 33% for training code), suggesting that access conditions are more consistently reported for commercial models than for other categories.

Figure 41. Model and training code accessibility by organisation type (n = 372 biology AI models). Organisation types with less than 10 models are excluded.



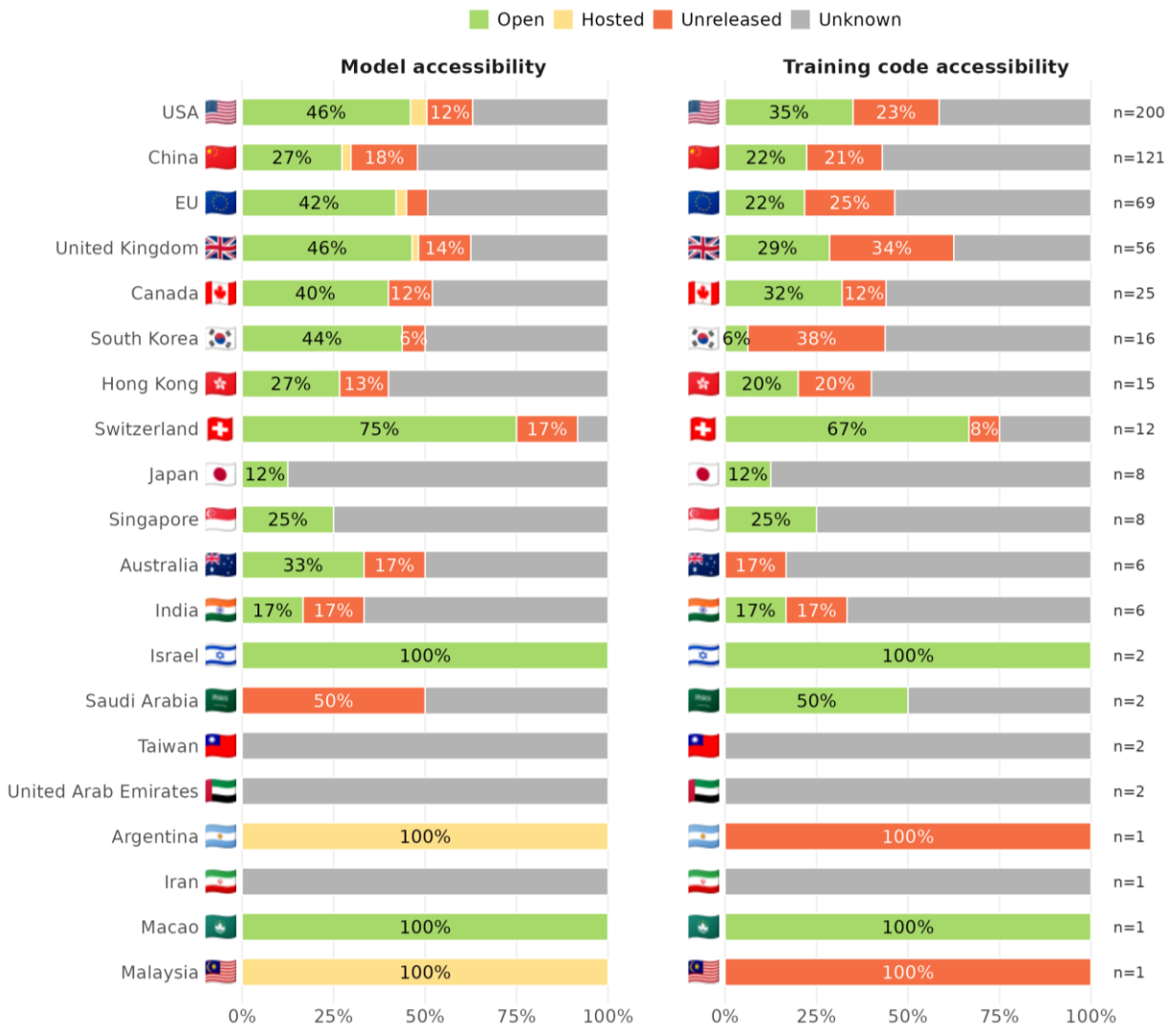
Source: JRC's analysis using Epoch AI dataset.

At country-level[108], a small group of countries leads AI4Bio model development, but openness varies considerably (**Figure 42**). The US is the largest contributor with 200 models, showing relatively high model openness at around 46%, though training-code availability is lower at around 35% and a large share remains undisclosed. China, with 121 models, shows a consistent pattern of lower openness, with roughly a quarter of models openly accessible and similar levels of unreleased

[108] Country counts are based on a full counting approach, whereby each model is attributed to every country associated with it, so the figures reflect the openness characteristics of each country's models and should not be read as a measure of relative contribution (see Annex 8). Country attribution follows author affiliations as recorded in the Epoch AI dataset; for industry-affiliated models, this reflects the institutional location of the contributing authors (see Annex 2).

models and unknown disclosure. The EU (69 models) and the UK (56 models) display broadly similar patterns to each other: around 40% to 46% of models are openly accessible, but training-code release falls to roughly 22% to 29%, with substantial shares of unknown or unreleased code. Switzerland, despite contributing only 12 models, stands out as an exception, with roughly three quarters of models openly accessible and around two thirds releasing training code.

Figure 42. Openness contribution by country (Top 20, EU grouped, n= 381 models). Labels of percentages are shown for segments >5%.



Source: JRC's analysis.

Among mid-ranking contributors, Canada (25 models) and South Korea (16 models) show relatively high model openness at around 40% to 44%, yet training-code availability drops markedly in both cases, particularly in South Korea where unreleased code is more common. Hong Kong (15 models) shows moderate openness in both dimensions. Japan and Singapore (8 models each) have relatively limited openness overall. Beyond these, results for smaller contributors should be interpreted cautiously given very small model counts, which tend to produce skewed or single-category distributions.

In summary, the AI4Bio model development landscape is characterised by a concentration of expertise within a narrow group of institutions, uneven openness across organisation types and countries, and a consistent gap between model release and training-code transparency. The US and China dominate in volume; the EU maintains comparable openness levels but contributes fewer models and has limited representation among the top developing organisations. Industry actors, while increasingly active, release training code at markedly lower rates than academic or collaborative settings, and a large share of disclosure status remains unknown across all categories.

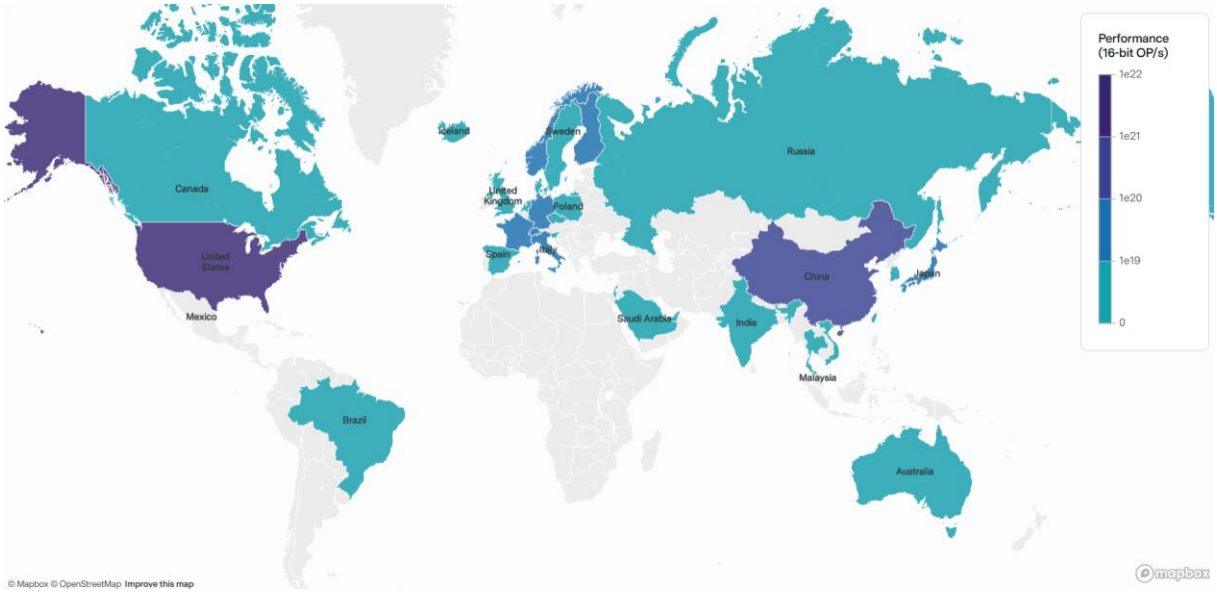
These patterns have direct consequences for the real-world adoption of biological AI models in Europe. Dependence on externally developed, often closed, frontier models creates barriers to independent validation, which is a practical prerequisite for deployment in regulated contexts such as clinical diagnostics or environmental monitoring. For the EU, strengthening domestic model development capacity and establishing clear expectations around openness for publicly funded research are therefore not merely strategic preferences but conditions for translation into real-world applications.

6.3. Infrastructure Capacity

At global scale, computational infrastructure for AI development is highly concentrated in a small number of countries (**Figure 43**)[109]. The US leads by a substantial margin, with aggregate GPU cluster performance reaching the order of 10^{22} 16-bit OP/s, followed by China and, at a lower level, a cluster of countries including the UK, Sweden, Japan, and several EU member states such as Italy, Spain, and Poland. Canada, Brazil, Australia, India, Saudi Arabia, and Malaysia show measurable but more limited capacity. Most countries have no recorded presence in the dataset, reflecting both genuine infrastructure gaps and uneven reporting coverage. This concentration has direct implications for real-world adoption of biological AI: access to frontier training infrastructure determines not only who can develop large-scale models, but also who can fine-tune, validate, and adapt them for specific application contexts.

[109] Disclaimer: The Epoch AI dataset covers an estimated 10–20% of existing global aggregate GPU cluster performance as of March 2025. Planned systems are subject to changes and inherently lower confidence. While coverage varies across companies, sectors, and hardware types due to uneven reporting, we believe the overall distribution remains broadly representative. Future country shares may change dramatically as exponential growth continues. Chinese systems are anonymized and their specifications are rounded.

Figure 43. Global distribution of GPU cluster performance by country (16-bit OP/s). Colour intensity reflects aggregate GPU cluster performance on a logarithmic scale. Grey indicates no data.



Source: Epoch AI GPU Clusters dataset (view: map, tab: country), available at https://epoch.ai/data/gpu-clusters.

Turning to Europe, an analysis of 70 GPU centres[110] reveals that **many European countries have sufficient computational capacity[111] to train the most demanding biological AI models** currently documented (**Table 17**, **Figure 44**). France (e.g. Sesterce Synapse Phase 2), Germany (e.g. Jupiter, Jülich), Italy (e.g. Eni HPC6), UK (e.g. University of Bristol Isambard-AI), Switzerland (e.g. Alps Supercomputer Phase 2), Finland (e.g. EuroHPC LUMI), Sweden (e.g. Sweden 4k H100 Cluster), Norway (e.g. NexGen Cloud Hyperstack AQ Compute Supercomputer), the Netherlands (e.g. Microsoft Azure Pioneer-WEU), Iceland (e.g. Nebius ISEG2), Spain (e.g. Mare Nostrum 5), and Denmark (e.g. Novo Nordisk Gefion) all report facilities with several thousand GPUs, sufficient to train the most computationally intensive AI4Bio models on record, such as ProtBERT-BFD, which requires 1,024 GPUs.

That said, capacity is unevenly distributed across Europe. Switzerland, Finland, Norway, and Italy report the highest median GPU counts per centre, while Poland, the Netherlands, Luxembourg, Czechia, and Slovenia have substantially lower capacity. Within-country variation is also considerable: Germany, for instance, has a mean of nearly 3,000 GPUs per centre but a median of 740, reflecting the outsized contribution of a small number of large facilities.

Beyond existing GPU infrastructure, Europe is actively expanding its AI compute capacity through the EuroHPC AI Factories initiative, a network of 19 AI-optimised hubs established across Europe as of late 2025[112]. Several of these explicitly target health and life sciences as priority sectors,

[110] https://epoch.ai/data/gpu-clusters

[111] Capacity figures reflect aggregate GPU cluster performance regardless of ownership. Several facilities listed are privately owned (e.g. Eni HPC6) and would not necessarily be available for publicly driven research initiatives. Publicly accessible infrastructure, including EuroHPC joint undertaking systems, represents the more reliable baseline for assessing capacity available to the broader research community.

[112] https://www.eurohpc-ju.europa.eu/ai-factories_en

including the Spanish 1HealthAI factory (focused on One Health, genomics, and personalised medicine), the Dutch NLAIF (with emphasis on sensitive health data), and the Polish Gaia AI Factory (targeting healthcare among other sectors).

This development is relevant for AI4Bio because AI Factories can be framed not only as an expansion of European AI compute capacity, but also as dynamic ecosystems fostering innovation, collaboration and AI development. In line with the Commission's description, AI Factories bring together computing power, data and talent, and link supercomputing centres, universities, SMEs, industry and financial actors. In the context of AI4Bio, this ecosystem dimension is particularly relevant: biological AI models require high-performance computing resources, but their development, refinement and integration into key applications also depend on access to curated and interoperable datasets, specialised scientific expertise, support services, secure experimentation environments and links with users in research, industry, healthcare and public authorities.

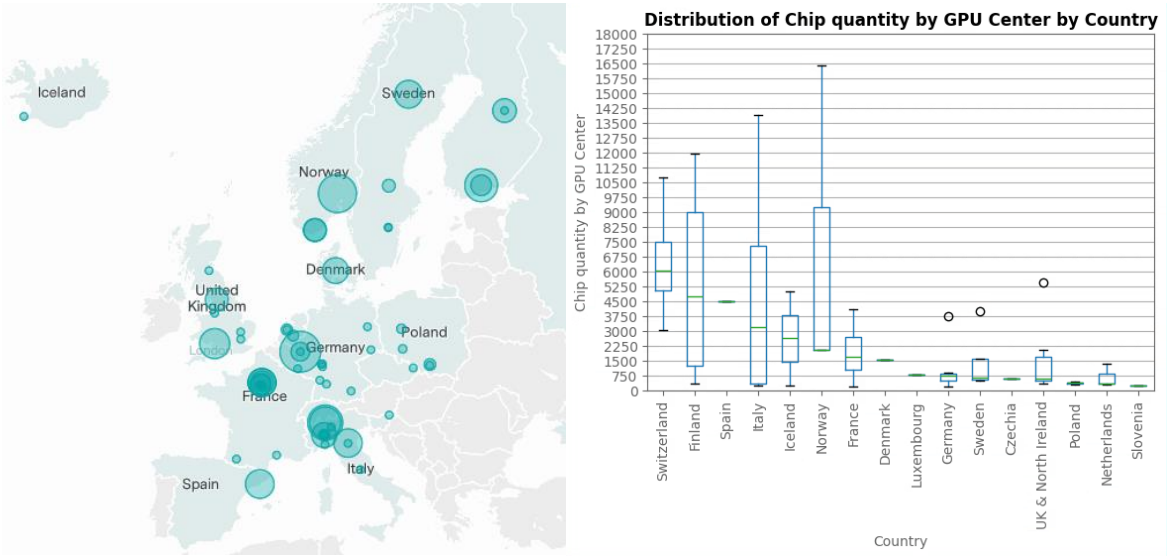
The added value of AI Factories for AI4Bio may therefore lie in helping to connect compute capacity with the broader conditions needed for trustworthy AI development and uptake. This is particularly important for start-ups, SMEs, universities and research teams that may not have all the in-house capacity required to combine AI engineering, biological expertise, access to data, validation resources and sector-specific knowledge. In this sense, AI Factories could support users moving from model training and fine-tuning towards testing, benchmarking and integration into relevant applications. This may be particularly relevant where AI Factories are networked with other European AI-support initiatives, including Testing and Experimentation Facilities, which offer dedicated resources for testing AI solutions, and the network of European Digital Innovation Hubs.

Table 17. Descriptives of GPU centres in Europe by country.

Country	Total N centres	Mean number of GPU per centre	SD	Median	Max number of GPU per centre
France	14	1899	1335	1682	4096
Germany	11	2954	6895	740	23536
Italy	9	5104	5448	3170	13888
UK	6	1568	2004	572	5448
Switzerland	4	6474	3199	6052	10752
Finland	4	5438	5478	4760	11912
Sweden	4	1444	1708	648	4000
Poland	4	369	58	366	440
Norway	3	6827	8277	2048	16384
Netherlands	3	651	574	352	1312
Iceland	2	2620	3355	2620	4992
Spain	1	4480	-	4480	4480
Denmark	1	1528	-	1528	1528
Luxembourg	1	800	-	800	800
Czechia	1	576	-	576	576
Slovenia	1	240	-	240	240

Source: JRC's analysis.

Figure 44. Locations of GPU centres in Europe and distribution of chip quantity by GPU centre by country (sorted by median).



Source: Epoch AI and JRC's analysis based on Epoch AI data[113].

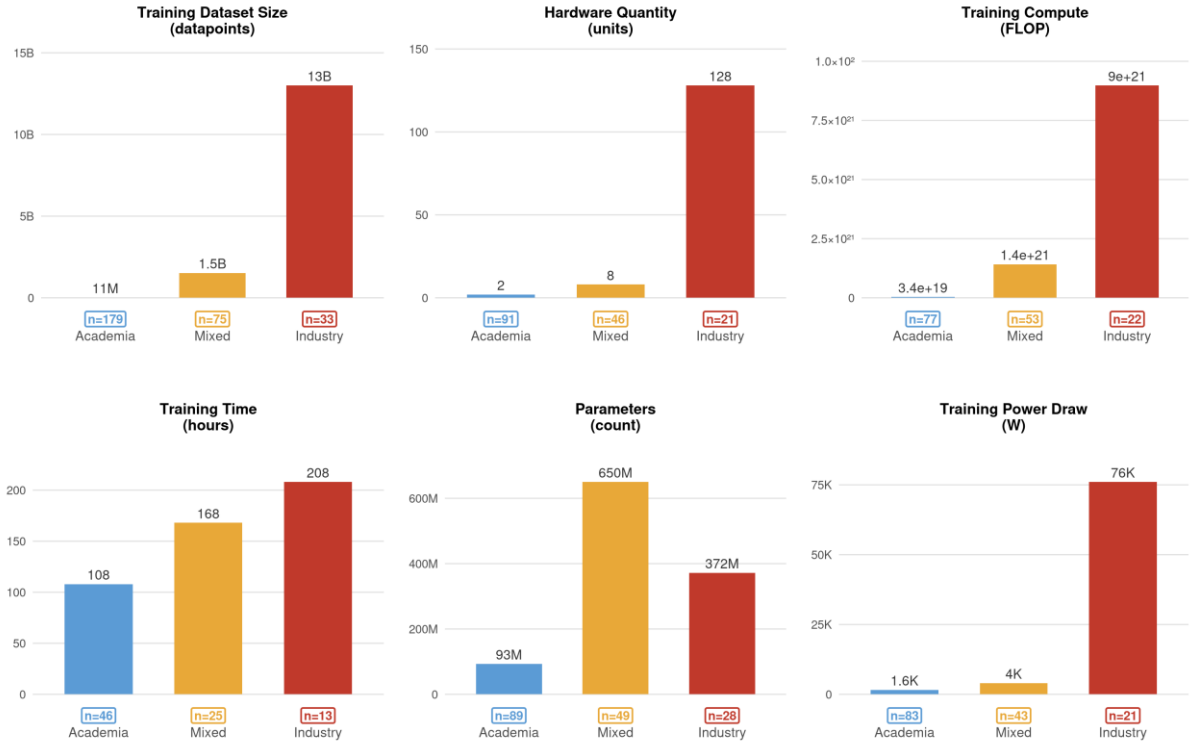
Infrastructure intensity differs systematically by organisation type. **Industry-only models mobilise substantially greater infrastructure than academia-only models across every resource dimension where the comparison is meaningful** (**Figure 45**). The median[114] industry model is trained on 13 billion datapoints, compared with 11 million for the median academic model, a difference of more than three orders of magnitude. Industry models also use more hardware (median 128 units), substantially higher training compute (9.0×10^{21} vs. 3.5×10^{19} FLOP) and considerably greater training power draw (76.0 kW vs. 1.6 kW). Two metrics depart from this pattern. Median training time is longest for industry-only models (208 hours), yet mixed affiliation[115] models are not far behind (168 hours), both well above academia-only models (108 hours). Median parameter count is highest for mixed models (650 million), exceeding both industry-only (372 million) and academia-only models (93 million); for this variable in particular, the combination of academic expertise in model architecture and industry computing capacity appears to confer an advantage that neither sector achieves independently.

[113] https://epoch.ai/data/gpu-clusters

[114] The underlying distributions contain extreme outliers, particularly at the upper end, which would distort mean comparisons. Medians more reliably represent the central tendency of the resource profiles being compared.

[115] Models with mixed affiliations, those produced by collaborations spanning industry and academia, were assigned to a separate Mixed category for the organisational comparison.

Figure 45. Evolution of infrastructure per organisation type (academia vs industry)[116].



Source: JRC's analysis.

These infrastructure patterns have direct implications for real-world adoption. The resource asymmetry between industry and academia means that the models best positioned for deployment in demanding applications, such as clinical genomics or large-scale drug screening, are predominantly developed in settings with limited openness and restricted access. Academic and public research institutions, which account for the majority of model development, often lack the infrastructure to train or fine-tune frontier-scale models independently, creating a dependency on industry-developed systems for translation into applied contexts.

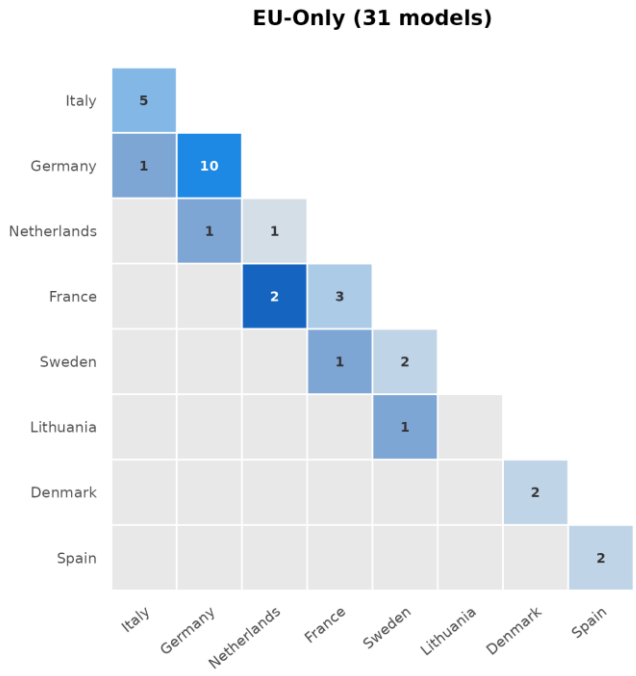
6.4. Collaboration Patterns

The geographic distribution of collaborations reveals a marked contrast between limited intra-EU cooperation and the broader international network in which EU actors participate. **Figure 46** and **Figure 47** show pairwise collaboration patterns based on organisational affiliations (see Annex 8 for more details), with diagonal cells indicating solo-country models and off-diagonal cells indicating cross-country collaborations.

[116] Data provided from the Epoch AI dataset and corresponding to models from the following countries: Armenia, Australia, Austria, Belgium, Brazil, Canada, China, Croatia, Denmark, France, Germany, Hong Kong, India, Iran (Islamic Republic of), Ireland, Israel, Italy, Japan, South Korea, Lithuania, Luxembourg, Macao, Netherlands, New Zealand, Norway, Poland, Russia, Saudi Arabia, Singapore, Spain, Sweden, Switzerland, Taiwan, United Arab Emirates, UK, United States of America and Vietnam.

Within the EU-only subset (n=31 models), collaboration is concentrated in a small number of Member States and remains relatively modest in scale (**Figure 46**). Germany is the main hub, with 10 solo-country models, as well as collaborations with Italy (1) and the Netherlands (1). Italy has a notable domestic presence with 5 single-country models. France contributes 3 single-country models and collaborates with the Netherlands (2) and Sweden (1). Other Member States appear only sporadically, including Denmark and Spain, each with 2 single-country models, and Lithuania, with 1 single-country model. Overall, this analysis shows model development activity concentrated in few EU countries with limited cross-border collaboration within the EU.

Figure 46. Pairwise collaboration between EU countries in the development of biology AI models, based on organisational affiliations. Diagonal values indicate solo-country models. Off-diagonals corresponds to number of pairwise collaborations.



Source: JRC's analysis.

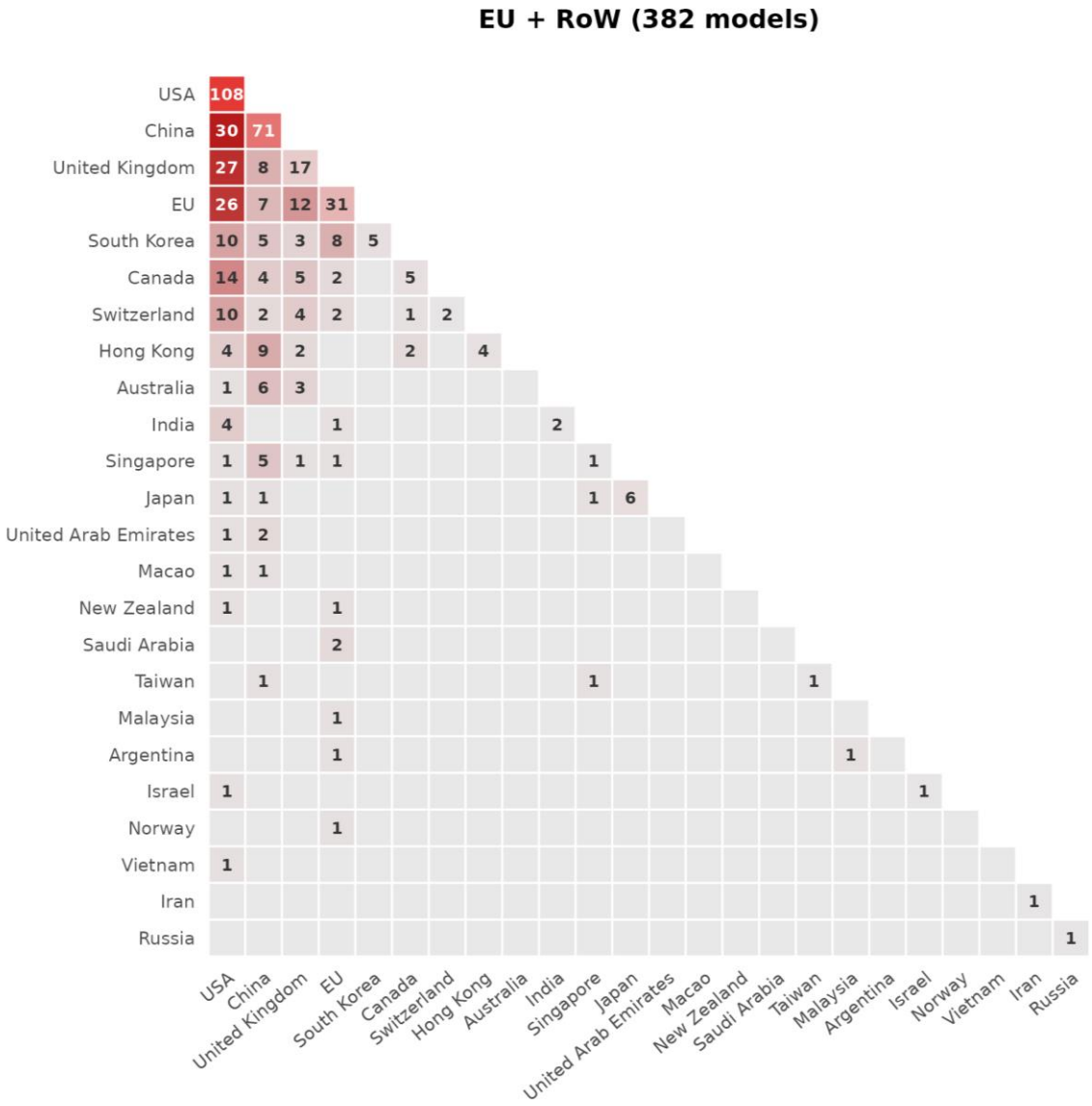
The broader EU and rest-of-world (RoW) network (n=382 models) is far more extensive (**Figure 47**). The US is the single largest hub, with 108 solo-country models, followed by China (71) and the EU (31). The largest bilateral links are between the US and China (30), the US and the UK (27), and the US and the EU (26). The EU also collaborates substantially with the UK (12) and China (7), but these links remain smaller than those involving the US. Other countries, such as South Korea, Canada and Switzerland, play secondary but still visible roles in the international network.

Two observations follow from this comparison:

1. **Intra-EU collaboration is much more limited than the EU's external collaboration with major global partners**, especially the US, the UK and China.

2. **The EU is present as an international actor but does not occupy the same central position as the US in the global collaboration network**. This is consistent with broader evidence on fragmentation in the European research and innovation system: as (Balland, Pierre-Alexandre et al., 2025) argue, European hubs tend to be less interconnected than their US counterparts, particularly in complex technology domains. Europe therefore has a visible role in global AI4Bio research, but one shaped more by connections to leading non-EU countries than by a dense internal collaboration base.

Figure 47. Pairwise collaboration countries in the development of biology AI models, based on organisational affiliations. Diagonal values indicate solo-country models. Off-diagonals corresponds to number of pairwise collaborations.

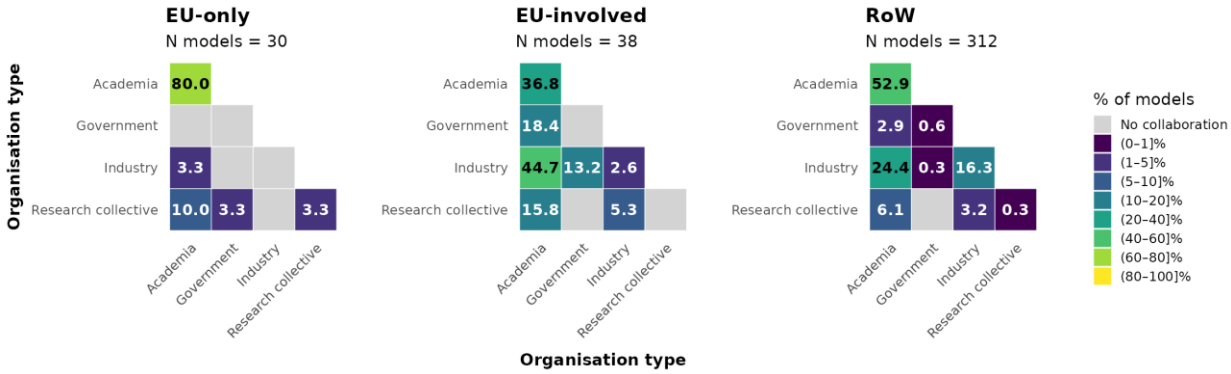


Source: JRC's analysis.

Collaboration patterns also differ considerably by organisation type across EU-only, EU-involved[117], and RoW models (**Figure 48**).

- **EU-only models (30 models)**: academia-only collaborations dominate at 80%. Other types are infrequent: academia-research collective represents 10%, while academia-industry, government-research collective, and industry-research collective each account for around 3%, indicating limited cross-sector engagement.
- **EU-involved models (38 models)**: the picture is more diversified, with academia-industry collaboration as the most common type at 45%, followed by academia-only at 37% and government-only at 18%. Academia-research collective (16%) and government-industry (13%) collaborations also feature more prominently than in the EU-only group.
- **RoW models (312 models)**: academia-only collaboration dominates at 53%, while academia-industry is present but less prominent at 24%, and industry-only models are more frequent at 16%.

Figure 48. Collaboration patterns between organisation types.



Source: JRC's analysis.

In summary, **academic institutions are the backbone of AI4Bio collaboration across all groups**. However, organisational diversity increases considerably once partnerships extend beyond the EU, with academia-industry collaboration emerging as a defining feature of internationally connected EU research. For real-world adoption, this matters: the translation of biological AI models into applied settings typically requires precisely the kind of cross-sector engagement that is currently more common in internationally connected collaborations than in EU-only activity.

6.5. Funding

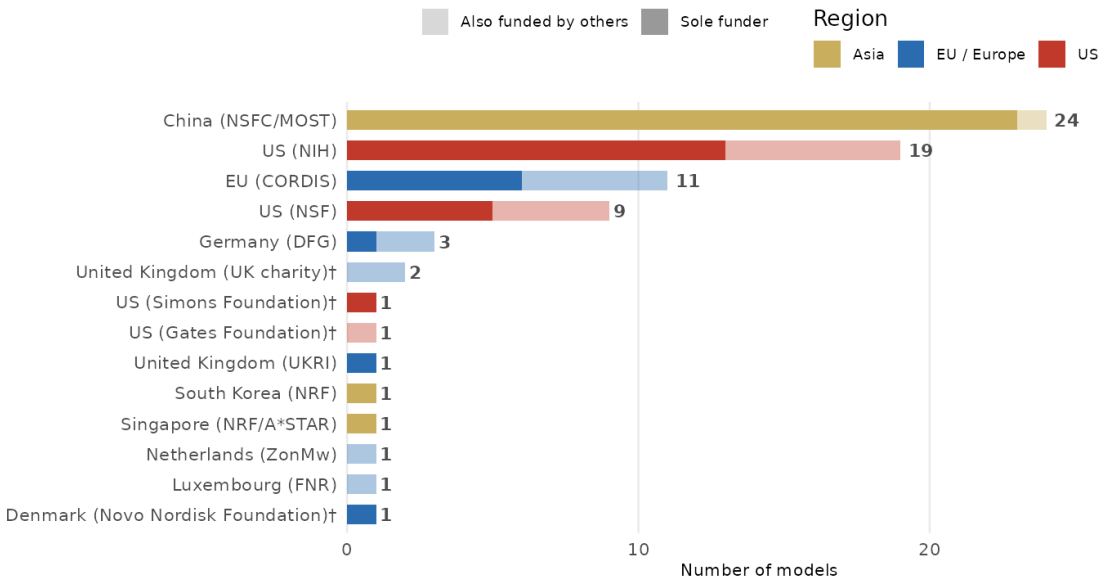
Funding is a primary lever through which governments and companies shape the direction, pace, and openness of AI research. In the context of biological AI, funding decisions determine not only

[117] EU-involved models refer exclusively to collaborations involving both EU and non-EU actors; EU-only models are excluded from this category.

what gets built but also who can access the resulting models, under what conditions results are shared, and whether the capabilities developed translate into applications relevant to public health, clinical practice, or environmental monitoring. Understanding the funding landscape is therefore essential for assessing the prospects for real-world adoption and the EU's positioning within it, though the available data allow only a partial view, as detailed in Annex 9.

Of the 383 Epoch AI models, 64 (around 17%) had at least one funding source recorded, and quantitative amounts were available for 28. The China's National Natural Science Foundation (NSFC/MOST) and the NIH and are the two most frequent funders, supporting 24 and 19 models respectively (**Figure 49**). European funding programmes identified through CORDIS are linked to 11 models, placing them third among the sources recorded. The range of funders represented in the dataset, spanning US federal agencies, Chinese national bodies, and European sources, including EU-level programmes recorded in CORDIS as well as national funders[118], reflects the increasingly international character of AI research.

Figure 49. Number of AI models (n=64) supported by each funding source. The faded segments of each bar indicate models that are co-funded with at least one other organisation. Symbol "†" denotes philanthropic or private foundation.



Source: JRC's analysis.

Eleven models are linked to two or more funding sources, forming nine unique agency pairings (**Table 18**). The most common co-funding arrangement is NIH–NSF (four models), consistent with the practice of US federal agencies jointly supporting translational research. EU funding appears in four of the nine pairings, in combination with NIH, FNR Luxembourg (Luxembourg National Research Fund), UK Charity (The Wellcome Trust and Cancer Research UK), DFG (German Research Foundation), and ZonMw (The Netherlands Organisation for Health Research and Development), reflecting the consortium structure typical of Horizon Europe and H2020 projects. While this

[118] Germany (DFG), United Kingdom (UK charity and UKRI), Netherlands (ZonMw), Luxebourg (FNR Luxenbourg), Denmark (Novo Nordisk Foundation).

indicates that European programmes do support internationally collaborative work, the overall number of co-funded models is too small to draw firm conclusions.

Table 18. Co-funding pairs: models jointly supported by two or more agencies.

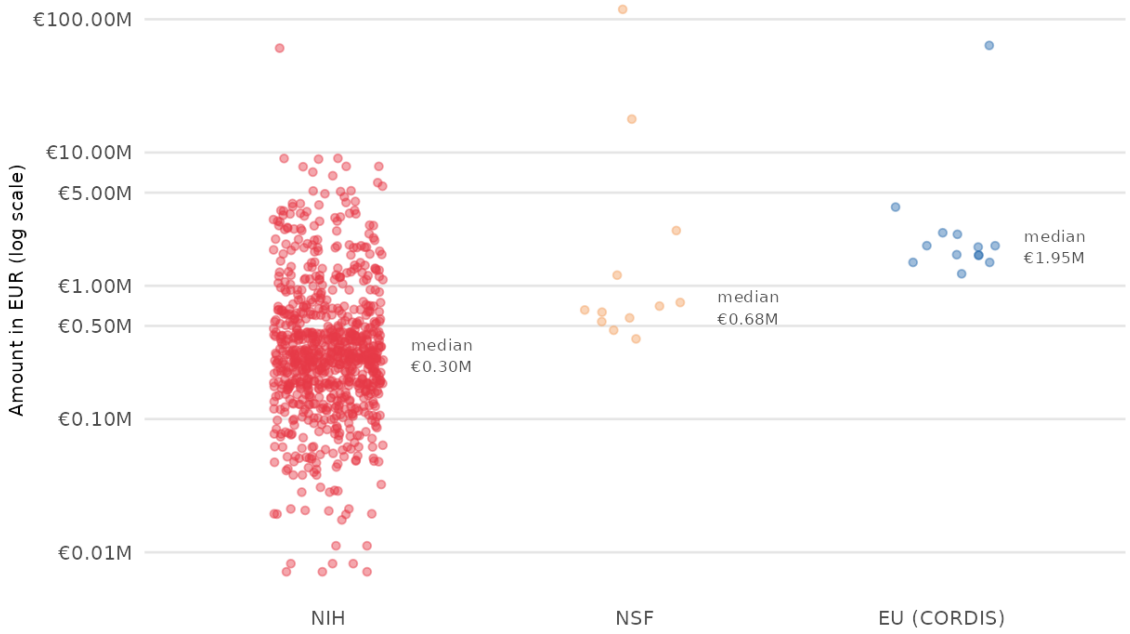
Funder A	Funder B	Models	Model names
NIH	NSF	4	AntiFormer; MMAPLE; RNA language models predict mutations that improve RNA function; TripletRes
China (NSFC/MOST)	NIH	1	TripletRes
China (NSFC/MOST)	NSF	1	TripletRes
DFG	EU	1	SPOT
DFG	UK Charity	1	AlphaLink
EU	FNR Luxembourg	1	SO3LR
EU	NIH	1	MPNNsol
EU	UK Charity	1	DMPFold
EU	ZonMw	1	Jaeger
Gates Foundation	NIH	1	ProteinGenerator

Source: JRC's analysis.

The 28 models with quantitative funding data, the median NIH grant is €0.30 M, compared with €0.68 M for NSF and €1.95 M for projects recorded in CORDIS (**Figure 46)**. The EU sample comprises primarily ERC investigator grants (9 of 12 projects), alongside two Research and Innovation Actions and one MSCA network, and the EU values are not directly comparable to NIH and NSF awards given differences in funding instruments. In aggregate, US federal agencies account for approximately €744 M of matched funding (NIH: €599 M across 19 models; NSF: €145 M across 9 models), while CORDIS projects contribute €88 M across 11 models (**Figure 47**). On a per-model basis, the median CORDIS project budget exceeds the NIH median by a factor of approximately six, yet the aggregate total remains considerably lower owing to the smaller number of matched models.

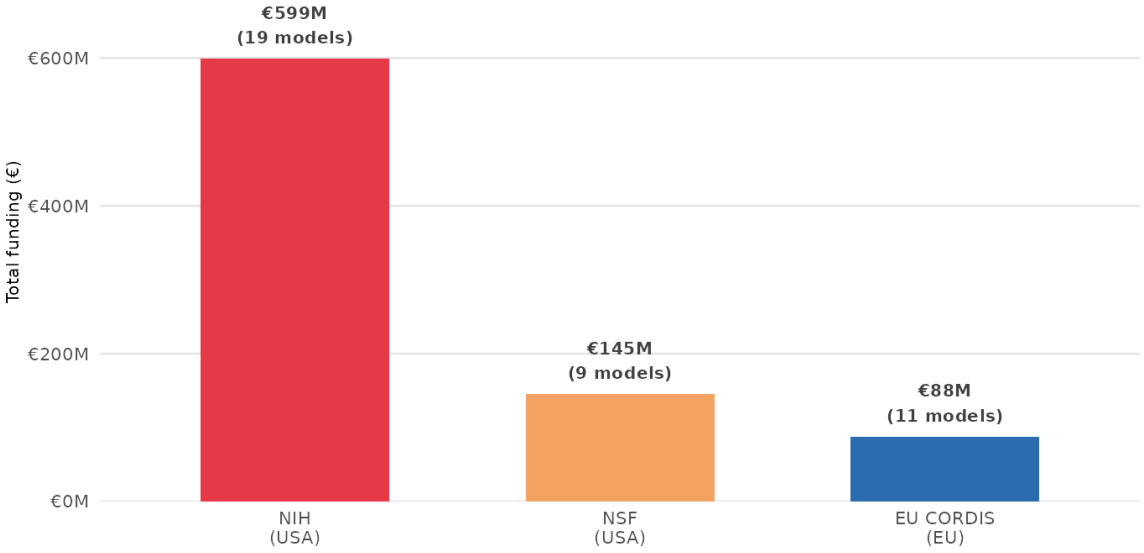
These findings offer a partial and indicative picture only. The 17% funding coverage means that no strong conclusions can be drawn about the relative investment levels of different actors. What the data do suggest, tentatively, is that European public funding supports a visible but modest share of the biological AI models for which any funding trace exists, and that this share is predominantly channelled through investigator-led research rather than large-scale coordinated programmes. Whether this funding profile is well matched to the infrastructure, validation, and translation requirements of real-world adoption is a question that better data would be needed to answer properly. **Systematically linking AI research outputs to funding databases at scale, via DOIs, author affiliations, and grant acknowledgements, should be considered a priority for functional monitoring of public investment in this field.**

Figure 50. Grant sizes of NIH and NSF awards vs EU CORDIS project costs (n=28 models).



Source: JRC's analysis.

Figure 51. Total recorded funding flowing into AI research in US vs EU (n=28 models).



Source: JRC's analysis.

7. Policy Considerations

This section draws together the principal challenges identified across the preceding analyses of biological AI models and sets out policy considerations addressed primarily to EU-level decision-making. Each subsection corresponds to a challenge that emerged empirically from the analysis in Chapters 3 to 6: the structural skew of the model landscape towards protein-centric and predictive tasks (Section 7.1); the fragmentation and geographic concentration of training data and data infrastructure (Section 7.2); the EU's external dependencies and limited intra-EU collaboration base (Section 7.3); and the persistent gap between technical performance and deployment readiness (Section 7.4). The policy considerations are grounded in the empirical findings of this report rather than derived from general principles.

7.1. A Protein-Centric and Predictive Model Landscape

The biological AI model landscape surveyed in this report is structurally skewed across three related dimensions. First, transformer-based architectures account for 66% of surveyed systems and dominate across application areas (Section 4.2.2). Second, the majority of models are oriented towards prediction rather than generation or closed-loop experimental integration (Section 4.2.1). Third, the field is protein-centric: amino acid sequence and structure constitute the primary input modality for most state-of-the-art systems, reflecting where training data is richest and benchmarks are most established (Section 5.3). This convergence was accelerated by the landmark success of AlphaFold and its successors in protein structure prediction, which concentrated research investment and talent around a comparatively narrow slice of biological space.

The data confirm that this concentration is not merely a feature of the technology but of the ecosystem around it. Protein datasets dominate across all time periods in the training data analysis and remain the most actively updated category (**Figure 26**). Community benchmarking initiatives maintain a continuous evaluation infrastructure that other domains cannot yet match (Section 5.7.2). Single-cell biology, metabolomics, RNA biology, plant biology, and multi-organism systems remain comparatively underserved, a pattern with direct consequences for EU policy priorities in areas such as antimicrobial resistance, rare disease genomics, ecological monitoring, and the circular bioeconomy.

Broadening the biological and architectural scope of biological AI model development is not merely a matter of scientific completeness: it is a precondition for addressing the full range of life sciences challenges that matter to European society.

Box 8. Considerations for Broadening the Biological and Architectural Scope of Biology AI Models

1. Targeted investment in underrepresented biological modalities. European research funding should explicitly address the imbalance between protein-centric and other biological AI models. Priority areas could include genomics and transcriptomics foundation models, metabolic and multi-omics integration, non-model organisms of agricultural and ecological significance, and multi-modal systems that bridge molecular, cellular and organismal scales. European research funding programmes and public-private partnerships are well placed to anchor such targeted investment.

2. Generative and hybrid model development beyond prediction. The dominance of predictive, transformer-based models reflects current data availability and benchmark incentives, but the next generation of biological AI applications, including *de novo* molecular design, closed-loop experimental automation and whole-cell simulation, requires generative and hybrid architectures. Funding instruments and evaluation frameworks should be adapted to support and assess these more architecturally diverse approaches.

3. EU strategic roadmap for biological AI scope. There is a need for a strategic roadmap identifying the biological AI capabilities needed to address EU priorities in health, food security, climate adaptation and the circular bioeconomy. This roadmap should inform research funding priorities, infrastructure investment and international collaboration strategies, and should be updated on a regular cycle to track the pace of field development.

7.2. Training Data and Data Infrastructure Fragmentation

The heavy reliance on a narrow set of curated, high-volume repositories represents both a strength and a systemic vulnerability. The training data analysis identified 75 datasets in the Epoch AI metadata (Section 5.1), and the most widely used, UniProt, the Protein Data Bank, AlphaFold-predicted structures, RefSeq, and the European Nucleotide Archive, are actively maintained at terabyte scale. Yet this concentration creates fragility: when most state-of-the-art models are trained on the same datasets, their behaviour becomes tightly coupled to biases embedded in those sources. The Protein Data Bank, for instance, contains nearly 80,000 entries for *Homo sapiens* but only approximately 5,000 for *Escherichia coli* (**Figure 29**), systematically privileging well-studied organisms. Perplexity analysis of Evo 2 across more than eight million sequences spanning 150 species confirms that this representational imbalance translates directly into uneven model performance across the tree of life (**Figure 28**, Section 5.2).

A parallel fragmentation challenge concerns dataset governance and geographic distribution. The US and EU account for most identified training datasets, with Switzerland, the UK, and China forming a secondary tier (**Figure 35**, Section 6.1). Outside these regions, contributions remain marginal. Datasets vary considerably in maintenance status (**Figure 24** and **Figure 25**): whilst large foundation repositories are actively updated, some smaller benchmark datasets have not been updated for up to a decade, carrying risk of scientific currency loss when used beyond their original benchmarking purpose. Openness adds a further dimension: the analysis of model accessibility (**Figure 30**, Section 5.6) shows that nearly half of all surveyed models disclose neither model weights nor training code, limiting independent validation and complicating regulatory scrutiny. Private actors can and do build highly capable, commercially restricted models on publicly funded data infrastructure, raising questions about the conditions under which access to publicly funded biological data should translate into obligations regarding openness or benefit sharing.

An emerging concern is the increasing use of AI-generated synthetic data as training material, which risks propagating and amplifying errors through successive model generations and creating distorted biological distributions at scale. Guidelines governing this practice have not kept pace with its spread.

Box 9. Considerations for Enhanced Data Infrastructure

1. Data coordination framework. Existing initiatives (e.g. the Global Biodata Coalition and, within Europe, ELIXIR) provide partial foundations for coordinating biological data infrastructure, but AI-readiness (standardised quality metrics, bias documentation, and machine-actionable access conditions) is not yet a systematic requirement. EU policy should build on these structures by expanding their mandates to encompass AI training data needs, and by requiring EU-funded repositories to adopt standardised dataset descriptions covering content, access conditions, quality metrics and known biases. A centralised registry of available biological datasets, anchored within an existing body such as ELIXIR, would consolidate this function at the European level while remaining interoperable with global counterparts.

2. Targeted data generation initiatives. European research funding programmes could be more strongly aligned with goals, such as (1) expanding biological data in underrepresented biological domains, e.g., membrane proteins, non-model organisms, rare genetic variants or plant biology, and (2) prioritise data generation in regions currently underrepresented in global repositories.

3. Transparent data quality assessment. Establish independent mechanisms for assessing and certifying data quality across repositories. Publish comparative analyses documenting known biases, coverage gaps, artefacts and limitations in major datasets, enabling researchers to make informed choices about data suitability for specific applications.

4. Synthetic data governance. As AI-generated data increasingly enters training pipelines, guidelines distinguishing between legitimate uses (supplementing sparse domains) and problematic practices (replacing real experimental data) should be established. These guidelines should require transparent labelling of synthetic versus experimental data origins in models.

7.3. Strategic Dependencies and International Positioning

The EU's role in global AI4Bio research, whilst visible, remains characterised by external dependencies rather than continental leadership. The analysis of model development (Section 6.2) shows that approximately 200 models in the Epoch AI dataset originate from the US, compared with 69 from the EU collectively (**Figure 42**). The top-20 contributing organisations are dominated by major US universities and Chinese research institutions; the Technical University of Munich is the only EU-based institution in the top tier (**Figure 40**). This positioning reflects historical investment patterns but simultaneously constrains Europe's influence over the technological trajectories and governance frameworks shaping the field.

The collaboration analysis reveals two further structural limitations. First, intra-EU collaboration is substantially lower than EU engagement with non-EU partners: within the EU-only model subset, 80% of models involve academia-only collaborations, with limited cross-sector engagement (**Figure 48**). By contrast, EU-involved models, those developed through international partnerships, show a more diversified picture, with academia-industry collaboration accounting for 45% of cases. This suggests that international partnerships drive the kind of cross-sector engagement that real-world adoption requires, but that this dynamic has not yet generated comparable activity within the EU's own research ecosystem.

Second, the infrastructure intensity analysis (Section 6.3, **Figure 45**) shows that industry models are trained on datasets two orders of magnitude larger than academic models (median 9.0 billion versus 13.9 million datapoints), with substantially more hardware and compute. European GPU infrastructure is, in principle, sufficient to train the most computationally intensive biological AI models on record (**Table 17**), but this capacity is distributed unevenly across Member States and is not yet matched to model development activity at the scale of leading non-EU institutions. The gap the EU faces is one of mobilisation, not of underlying capacity.

Box 10. Considerations for Strengthened European Positioning

The EuroHPC AI Factories initiative, a network of 19 AI-optimised hubs established across Europe as of late 2025, provides relevant general-purpose AI infrastructure but does not directly address the specific needs of AI4Bio development. The following considerations identify actions needed to strengthen European positioning in this domain and collectively motivate the establishment of "AI4Bio Factories".

1. European biological AI foundation model initiative. Development of large-scale biological foundation models should be supported as strategic infrastructure, analogous to high-performance computing initiatives. These should be positioned as public goods accessible to European researchers and companies, thereby reducing dependence on foundation models developed outside the EU.

2. Incentives for intra-EU collaboration and coordinated AI4Bio research programmes. The creation of large-scale collaborative programmes linking institutions across multiple Member States around shared research objectives should be considered under existing and successor EU research funding frameworks. Funding must include recruitment of world-class talent and infrastructure investment comparable to leading non-European centres.

3. Technology transfer and capacity building. European funding programmes should enable academic discoveries in biological AI to transition into industrial applications within Europe. Strengthening EU positioning should demonstrate a coherent innovation ecosystem that connects research, clinical settings, industry and deployment at scale.

7.4. Maturity and Technological Readiness Assessment

Section 3.5.2 identifies a *maturity paradox* at the heart of the AI4Bio development. Whilst several applications have advanced towards mature implementation, the field broadly lacks the integrated frameworks required to assess, govern, and communicate deployment readiness. AlphaFold2 and AlphaFold3 represent the paradigm case: domain-mature by any measure, with more than 200 million structures in an open database and recognised by the 2024 Nobel Prize in Chemistry, yet assessed at TRL 5 to 6 because no regulatory authority has formally accepted their outputs as a certified component of a drug development submission (**Table 7**).

The gap between domain maturity and TRL is a structural feature of the current landscape, as set out in Section 5.7. Current incentive structures reward benchmark performance and public visibility over long-term real-world validation (Section 5.7.1). Addressing this gap will require evaluation approaches that extend beyond benchmark performance and incorporate reproducibility, validation context and evidence quality. Benchmarks themselves are domain-concentrated, subject to saturation, obsolescence, and data contamination, and do not assess the dimensions, including workflow integration, safety monitoring, access controls, and lifecycle management, that deployment readiness requires (Section 5.7.2). The result is a field in which models are benchmark-mature but deployment-immature.

A particular concern arises from the divergence between academic model release practices and the requirements for clinical or industrial deployment. Academic institutions and companies prioritise public release through community hubs, framing this as the democratisation and acceleration of science. However, these releases often occur without accompanying governance infrastructure or pre-deployment biosafety assessments, despite the models being vulnerable to jailbreaking that can generate outputs closely mimicking dangerous human pathogens (Section 5.7.4.1). The absence of reference maturity frameworks means that this gap is neither systematically identified nor managed.

Box 11. Considerations for Effective Maturity and Technological Readiness Assessment

1. Coherent European maturity assessment frameworks. No major jurisdiction currently operates a shared framework for assessing the maturity and deployment readiness of biological AI systems, making this a gap with both European and international dimensions. At the EU level, the Commission should consider establishing an independent scientific advisory group with expertise spanning AI, life sciences and regulatory science, tasked with providing ongoing advice on maturity criteria, biosecurity risks and the appropriateness of available assessment tools. Any framework should be designed to evolve iteratively as the field advances. At the international level, coordination with bodies such as WHO, OECD and ISO would help avoid divergent standards and support mutual recognition of maturity assessments.

2. Clinically and biologically relevant benchmarking standards. Maturity assessment must move beyond algorithmic performance scores and should also encourage transparent reporting of assumptions, limitations and sources of uncertainty where these influence interpretation thereby ensuring that benchmarks capture real-world utility. Evaluation frameworks for biological AI systems should incorporate workflow integration (how seamlessly can the system be embedded in existing research or clinical practice?), safety monitoring (what protocols govern continuous surveillance for anomalous outputs or adverse events?), access controls (who and what has access to the system, and under what conditions?) and lifecycle management (how does the provider address updates, version control and long-term maintenance?). These dimensions are already emerging in healthcare AI guidance and should be formalised for the life sciences context, drawing on the experience of health providers and independent organisations developing FAIR AI guidelines.

3. Taxonomy of model roles and corresponding governance pathways. A regulatory distinction should be established between biological AI systems operating as research and development tools, as Software as a Medical Device (SaMD) under the Medical Device Regulation or In Vitro Diagnostic Regulation (MDR/IVDR), or in hybrid capacities spanning both contexts. Each role carries distinct maturity requirements, liability frameworks and oversight obligations under the EU AI Act, MDR/IVDR, or the proposed Biotech Act advisory mechanisms. Clarifying these pathways would reduce regulatory uncertainty for developers and provide clearer gatekeeping criteria for deployment in high-stakes settings such as clinical trials, drug discovery and genetic intervention.

8. Future Directions and Emerging Trends

Model Development

The consolidation of AI architectures around transformer-based approaches signals that future progress will increasingly focus on architectural specialisation rather than broad paradigm shifts. Despite language-based models continue to dominate, currently representing 66% of surveyed systems, **emerging research points toward more sophisticated multimodal frameworks** that seamlessly combine sequences, structures, images and experimental measurements. For example, the recognition of transformer limitations for extremely long genomic contexts has catalysed innovation in alternative architectures such as HyenaDNA and Evo, which achieve sub-quadratic complexity through implicit convolutions (Nguyen et al., 2024). Future development will prioritise hybrid architectures combining transformer scalability with these efficiency gains, enabling researchers to model lager biological contexts.

The field will likely see continued **evolution toward foundation models capable of simultaneous prediction and generation**, moving beyond the current separation of these modalities into unified systems. This shift reflects recognition that prediction and generation are complementary rather than antagonistic capabilities, both essential for accelerating discovery and design. Models integrating these objectives, such as Evo 2, demonstrate the scientific significance of hybrid approaches, though they remain limited in number at present, and will be positioned to provide richer biological insights whilst supporting more versatile applications across research contexts.

Metadata Gaps and Model Transparency

The results presented in this report are constrained by limited availability and inconsistent reporting of metadata for AI models in the life sciences. The analysis relies heavily on the Epoch AI dataset and manual curation of the literature, reflecting broader challenges in systematically identifying and comparing models. Key information such as training data, model design and evaluation is often incomplete or difficult to access, limiting reproducibility and large-scale analysis. This lack of structured metadata also makes it more difficult to reliably identify patterns and anticipate future trends in model development. Addressing this gap will be important for the next phase of progress. Initiatives such as EMBL-EBI's BioAIrepo[119] provide an early step towards more structured and transparent model sharing. As a pilot repository within the BioStudies database, it enables researchers to share model weights alongside metadata, datasets and validation resources. Wider adoption of such standards could improve discoverability, reduce reliance on manual curation and support more robust and reproducible AI research in biology.

Infrastructure and Computational Capacity

Whilst frontier models continue scaling toward increasing computational intensity, the simultaneous emergence of efficient, resource-constrained alternatives suggests future development will follow complementary rather than convergent paths. Highly capable foundation models will remain concentrated in well-resourced institutions, serving as reference systems, whilst specialised and

[119] https://www.ebi.ac.uk/biostudies/BioAIrepo

efficient models optimised for specific tasks will proliferate across distributed teams. This heterogeneous landscape reflects practical realities: not all research questions require billion-parameter models, and many applications benefit from smaller, more interpretable systems deployable on institutional infrastructure. The analysis in section 4.3 demonstrates that this diversity is already observable, with models spanning orders of magnitude in parameter count and training requirements.

The weak correlation between training compute (FLOPs) and energy consumption, contrasted with stronger relationships for hardware quantity and runtime duration, will guide future optimisation toward pragmatic efficiency. Rather than pursuing ever-larger models, **future work will optimise hardware utilisation, reduce training time through efficient scheduling, and develop models achieving comparable capability with fewer parameters through improved architectural design**. This simultaneously addresses environmental concerns and democratises access by lowering participation barriers for resource-constrained teams.

Applications with Emphasis on Explainability

Interpretability and explainability represent critical frontiers as biological AI systems transition from research tools to decision-support systems in clinical and industrial contexts. Currently, the field exhibits a concerning asymmetry: benchmark performance is comprehensively measured whilst mechanistic interpretability receives limited attention. This gap becomes acute in applications where predictions directly influence decisions such as variant effect assessment in clinical diagnostics, protein designs for therapeutics, or drug candidates selected for synthesis.

Early work demonstrates the potential of attention visualisation and feature attribution methods in biological contexts (Brandes et al., 2022), though systematic application remains limited. Future directions will establish integrated frameworks for mechanistic validation coupling computational predictions with biological interpretation.

Best Practices Toward Maturity and Readiness

Some health providers and independent organisations are beginning to publish guidelines on effectively integrating the latest AI models in the healthcare ecosystem (Wells et al., 2025). While these guidelines are specific to the healthcare sector and may not offer practical tools or methods for monitoring, they do address key areas of concern and could serve as inspiration for life sciences, particularly in drug design, clinical trials, and artificial genomes research. Key elements of those guidelines include:

1. Incorporation **of best practices** from existing frameworks, guidelines, and regulations;
2. Understanding **stakeholder needs**: including patients, providers, operational leaders, and AI developers in order to adopt a comprehensive approach that considers various perspectives and requirements in the healthcare ecosystem;
3. **Multidisciplinary** synthesis: drawing on expertise from different fields will ensure a well-rounded and effective implementation strategy.

These efforts demonstrate that a proactive and transparent approach from both smaller and larger stakeholders, across the public and private sectors, could be crucial in ensuring that biological AI systems meet necessary maturity benchmarks, thereby facilitating their effective integration into healthcare systems.

In addition to technical validation and implementation guidelines, AI literacy and interdisciplinary capacity will be important enabling conditions for the responsible uptake of AI4Bio. Biological AI models operate at the intersection of AI engineering, molecular and cellular biology, clinical expertise, data governance, cybersecurity, ethics, regulatory requirements and biosecurity considerations. As a result, their effective development, assessment and deployment require teams that are able to interpret both the technical behaviour of models and the biological meaning, limitations and potential implications of their outputs.

This skills dimension is particularly relevant because several of the risks identified in this report arise at the interface between benchmark performance and biological interpretation. A model may perform well in a controlled evaluation setting while still being difficult to assess in terms of clinical relevance, experimental validity, uncertainty, safety or misuse potential. Strengthening interdisciplinary AI literacy would therefore support more informed model development, better downstream use, and more robust interaction between researchers, developers, regulators, healthcare professionals and industrial actors.

This is also consistent with the broader approach of the EU AI Act, which requires providers and deployers of AI systems to take measures to ensure a sufficient level of AI literacy among staff and other persons dealing with the operation and use of AI systems on their behalf. In the AI4Bio context, such literacy should be understood not only as general familiarity with AI, but as domain-specific capacity to evaluate model limitations, data provenance, validation context, uncertainty and potential safety or security implications.

Agentic Platforms for Research

A key emerging direction is the increasing automation of biological research workflows. Recent systems such as *CellVoyager* illustrate a shift towards agent-based approaches that can autonomously analyse complex datasets, generate hypotheses and iteratively refine analyses with limited human intervention (Alber et al., 2026). This represents a conceptual extension of existing applications rather than a separate development trajectory: AI is already embedded across multiple stages of the research pipeline, but typically as specialised components within human-driven workflows. The next step is the integration of these components into more autonomous, end-to-end systems capable of navigating the full design–build–test–learn cycle[120] with reduced human supervision.

Early instances of this transition are already visible in the maturity analysis presented in Section 3.5.2.2. Lab-in-loop platforms such as Arc Institute's MULTI-evolve and Cradle Bio's Cradle-1 combine protein language models with active learning and focused experimental testing to compress engineering cycles from months to weeks, with Cradle-1 reporting four- to sevenfold acceleration relative to rational design across antibodies, enzymes and gene-editing tools. These

[120] The design–build–test–learn (DBTL) cycle is the iterative engineering framework at the core of synthetic biology and metabolic engineering. It comprises four phases: design (specifying a target biological system computationally), build (assembling the genetic constructs or biological parts), test (measuring performance experimentally) and learn (using the resulting data to inform the next design iteration). The cycle was formalised in the context of automated biofoundries by Carbonell et al. (2018), who demonstrated that fully automated DBTL pipelines could deliver order-of-magnitude improvements in microbial production yields within a small number of iterations

platforms are not yet fully autonomous, but they exemplify the architectural shift from AI-as-tool to AI-as-coordinator within an experimental loop.

As biological research workflows become more agentic and semi-autonomous, evaluation should increasingly consider not only the performance of individual model components, but also the behaviour of integrated workflows. This includes assessing how different models, tools and experimental steps interact across the design–build–test–learn cycle, as well as identifying appropriate human oversight mechanisms, intervention points and accountability structures. Such a workflow-level perspective would help ensure that increasing automation in biological research remains compatible with scientific reliability, traceability and responsible human control .

Taken together, these developments suggest a **transition from tool-based AI towards semi-autonomous research agents that can orchestrate multiple modelling tasks**, potentially accelerating discovery while reshaping the role of human oversight in biological research. Future evaluation approaches may therefore need to consider the performance of integrated workflows in addition to individual model components.

From Specialised Models to Generalist Biological AI

The trajectories outlined above, multimodal integration, unified prediction and generation, and agentic orchestration of research workflows, converge on a more ambitious horizon: the development of Generalist Biological AI (GBAI). GBAI, touched on earlier in this report in relation to the 'language of life' framing (Section 1) and the limits of current biological representations (Section 5.4), is developed more fully in what follows. Rao et al. (2026) define GBAI as a class of systems capable of modelling the *language of life*, the flow of information from DNA to RNA to proteins and ultimately to cellular function, within a single integrated framework rather than through discrete, task-specific models. Whereas current biological AI systems remain organised around isolated tasks, predicting a protein structure, generating a regulatory sequence, classifying a cell type, GBAI describes systems able to operate across these domains in an integrated manner, performing multiple critical biological tasks within a single framework rather than switching between specialised models.

The most concrete embodiment of this ambition is the virtual cell: a computational model that integrates genomic, transcriptomic, proteomic and metabolic information to represent cellular state and dynamics in silico (Bunne et al., 2024). Rather than answering one question at a time, a virtual cell would, in principle, allow researchers to perturb a gene, introduce a compound, or alter an environmental condition and observe the predicted downstream consequences across regulatory networks, signalling cascades and metabolic flux. This represents a qualitative shift in what biological AI is asked to do: from pattern recognition over fixed inputs to mechanistic simulation of integrated biological systems. Rao et al. (2026) identify the integration of foundation models within virtual cells as a defining frontier for the field, alongside the synergy of language and structural AI and the development of AI agents for autonomous discovery.

Realising this ambition will require advances on several of the fronts discussed in this report simultaneously: foundation models that span modalities rather than specialising in one; training data infrastructures capable of supporting integrative learning across heterogeneous repositories; agentic systems able to coordinate prediction, generation and experimental validation; and interpretability frameworks robust enough to make whole-cell predictions trustworthy in scientific and regulatory contexts. Rao et al. (2026) explicitly identify data, biological complexity, scalability and experimental validation as the principal challenges that must be addressed before GBAI can be realised. Progress will be incremental and uneven, and current efforts remain at an early stage.

Nonetheless, the trend is now clear enough that GBAI, and especially the virtual cell as its most concrete example, can reasonably be seen as the main direction for the next phase of biological AI development.

For EU policy, this trajectory reinforces the arguments developed in this report. A field moving towards integrative, multi-scale models will place even greater demands on data interoperability, infrastructure capacity and harmonised assessment frameworks than the current generation of specialised systems. The maturity gap identified in this work will not be resolved by further benchmark gains alone; it will require governance instruments capable of evaluating systems whose outputs are simulations of biological reality rather than predictions of individual properties.

9. Conclusions

This report has examined the landscape of 480 AI models applied to molecular and cellular biology, drawing on a dataset of 383 systems compiled by Epoch AI, supplemented by metadata from 97 models extracted from the scientific literature. Its aim has been to provide an empirically grounded account of what these models can do, how they are built and resourced, and what their current state of maturity means for policy and governance. Three overarching conclusions emerge from this analysis.

1) Capability is real, but the field is narrower than its public profile suggests.

Biological AI models have advanced substantially. Protein structure prediction, molecular design, and genome editing assistance have reached levels of technical performance that were considered out of reach only a few years ago. Yet the landscape that generates these advances is structurally concentrated: transformer-based architectures account for 66% of surveyed systems, inputs are predominantly protein sequences and structures, and tasks are mostly predictive rather than generative. The "AlphaFold effect", whereby a landmark success concentrates talent, data and investment around a specific area of biological space, has proven a powerful engine of progress, and its influence has extended well beyond protein structure to energise the broader AI4Bio field. At the same time, it reflects a recurring dynamic of concentration rather than a one-off event, with implications for which biological domains receive sustained attention.

The practical consequence is that headline capability claims in biological AI are often accurate for the specific slice of biology where data is richest and benchmarks are most mature, but do not transfer straightforwardly to the domains where European public health, food security, and environmental priorities are concentrated. For science-for-policy purposes, reported capability must therefore be considered alongside evidence quality and relevance, reproducibility and deployment readiness. Policy attention calibrated to benchmark performance rather than to the structure of the underlying evidence base risks misallocating investment and regulatory effort.

2) Scientific maturity and deployment readiness are distinct and conflating them has governance consequences.

The most consequential finding of this report is methodological as much as empirical: domain maturity and technology readiness, though frequently treated as equivalent, measure different things and routinely diverge. A model can be transformative within its biological domain, widely adopted by the research community, and even Nobel Prize-recognised, while remaining at a research-stage Technology Readiness Level (TRL) if it has never been formally qualified for regulated deployment.

This distinction matters for governance because incentive structures in the field currently reward domain maturity, specifically benchmark performance and public visibility, while the infrastructure for assessing deployment readiness, shared frameworks, prospective validation pathways, lifecycle governance, does not yet exist for this class of systems. The result is not that dangerous models are being deployed recklessly, but that the field lacks a unified process and standardized criteria for evaluating deployment readiness. Frontier systems can become publicly available and technically impressive yet occupy an ambiguous position with respect to real-world impact, including biosecurity risk, that is not currently governed by any systematic framework.

Developing coherent EU-level maturity assessment frameworks, ones that integrate domain-specific classification with the TRL logic already embedded in EU research and innovation policy, is a precondition for meaningful regulatory oversight. The EU AI Act (EU) 2024/1689, the Medical Device

Regulation (EU) 2017/745, and the proposed Biotech Act (EU) 2025/0406 collectively create a regulatory environment that will shape how biological AI systems are evaluated and compliant for release, but these frameworks presuppose assessment methods that do not yet fully exist for this class of systems.

3) The infrastructure and governance conditions for EU leadership are not yet in place.

The analysis of training data, infrastructure, collaboration, and funding reveals a field in which enabling conditions are as decisive as algorithms. The EU is not absent: EuroHPC JU pre-exascale and exascale systems provide compute resources sufficient to train even the most infrastructure-intensive biological AI models surveyed. European research institutions are active contributors to global model development, and European public funding supports a visible, if modest, share of the field. The gap between the EU and its main competitors in biological AI is one of mobilisation and coordination, not of underlying capacity.

Two specific coordination limitations emerge from the analysis. First, training data is fragmented across many specialised repositories with heterogeneous standards and limited interoperability, and geographic concentration outside Europe remains pronounced; no coordinating mechanism currently maps this infrastructure or aligns investment to fill coverage gaps. Second, intra-EU collaboration in model development is less developed than the EU's external collaboration with the US, China, and the UK; academic-industry cross-sector partnerships, which are the most common organisational form in internationally connected EU research, remain uncommon in EU-only activity.

Looking ahead

The window for establishing European positioning in AI-enabled life sciences is open. The scientific foundations are strong, the application potential is real and growing, and the EU's strategic ambitions in life sciences and AI are clearly articulated in policy. What is needed is the translation of those ambitions into the specific instruments, including coordinated data governance, targeted infrastructure mobilisation, harmonised maturity frameworks, and structured support for cross-sector collaboration, that the policy considerations set out in Section 7 address in detail.

Significant uncertainties remain. Safety evaluation methods for generative biological AI systems are underdeveloped. The dual-use risk profile of openly released frontier models is not yet systematically assessed. The trajectory of EU developer capacity relative to leading non-EU institutions depends on investment and coordination decisions that are still being made. Progress will be uneven, and the pace of field development means that any framework developed now will need to be revisited regularly. Sustaining European positioning will require treating AI4Bio not only as a scientific opportunity but as a governance challenge: one that demands the same strategic coordination the EU has brought to bear on other deep-technology domains where capability, sovereignty, and public interest intersect.

References

Abramson, J., J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, et al., 'Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3', *Nature*, Vol. 630, No. 8016, June 2024, pp. 493–500.

Ahmed, N., and M. Wahed, 'The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research', arXiv, October 22, 2020.

Akhtar, M., A. Reuel, P. Soni, S. Ahuja, P.S. Ammanamanchi, R. Rawal, V. Zouhar, et al., 'When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation', arXiv, February 18, 2026.

Alamdari, S., N. Thakkar, R. van den Berg, N. Tenenholtz, R. Strome, A.M. Moses, A.X. Lu, N. Fusi, A.P. Amini, and K.K. Yang, 'Protein Generation with Evolutionary Diffusion: Sequence Is All You Need', bioRxiv, November 4, 2024.

Alber, S., B. Chen, E. Sun, A. Isakova, A.J. Wilk, and J. Zou, 'CellVoyager: AI CompBio Agent Generates New Insights by Autonomously Analyzing Biological Data', *Nature Methods*, March 17, 2026, pp. 1–11.

Alley, E.C., G. Khimulya, S. Biswas, M. AlQuraishi, and G.M. Church, 'Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning', *Nature Methods*, Vol. 16, No. 12, December 2019, pp. 1315–1322.

Andrianov, G.V., E. Haroldsen, and J. Karanicolas, 'vScreenML v2.0: Improved Machine Learning Classification for Reducing False Positives in Structure-Based Virtual Screening', *International Journal of Molecular Sciences*, Vol. 25, No. 22, January 2024, p. 12350.

Ashyrmamatov, I., S.J. Gwak, S.-Y. Jin, I. Jun, U.V. Ucak, J.-Y. Lee, and J. Lee, 'A Survey on Large Language Models in Biology and Chemistry', *Experimental & Molecular Medicine*, Vol. 58, No. 4, April 2026, pp. 970–980.

'Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions | FDA', n.d. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/assessing-credibility-computational-modeling-and-simulation-medical-device-submissions.

Avsec, Ž., V. Agarwal, D. Visentin, J.R. Ledsam, A. Grabska-Barwinska, K.R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D.R. Kelley, 'Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions', *Nature Methods*, Vol. 18, No. 10, October 2021, pp. 1196–1203.

Avsec, Ž., N. Latysheva, J. Cheng, G. Novati, K.R. Taylor, T. Ward, C. Bycroft, et al., 'Advancing Regulatory Variant Effect Prediction with AlphaGenome', *Nature*, Vol. 649, No. 8099, January 2026, pp. 1206–1218.

Baek, M., F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G.R. Lee, J. Wang, et al., 'Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network', *Science*, Vol. 373, No. 6557, August 20, 2021, pp. 871–876.

Baek, S., K. Song, and I. Lee, 'Single-Cell Foundation Models: Bringing Artificial Intelligence into Cell Biology', *Experimental & Molecular Medicine*, Vol. 57, No. 10, October 2025, pp. 2169–2181.

Balland, Pierre-Alexandre, Benoit, Florence, Ravet, Julien, and Hobza, Alexandr, *Divided We Fall behind: Why a Fragmented EU Cannot Compete in Complex Technologies*, Publications Office of the European Union, 2025.

Baltrušaitis, T., C. Ahuja, and L.-P. Morency, 'Multimodal Machine Learning: A Survey and Taxonomy', arXiv, August 1, 2017.

Benegas, G., C. Albors, A.J. Aw, C. Ye, and Y.S. Song, 'GPN-MSA: An Alignment-Based DNA Language Model for Genome-Wide Variant Effect Prediction', bioRxiv, April 6, 2024.

Bernett, J., D.B. Blumenthal, D.G. Grimm, F. Haselbeck, R. Joeres, O.V. Kalinina, and M. List, 'Guiding Questions to Avoid Data Leakage in Biological Machine Learning Applications', *Nature Methods*, Vol. 21, No. 8, August 2024, pp. 1444–1453.

Bixby, E., G. Brunner, D. Danciu, R.D. Rosa, N. Deutschmann, C. Ferragu, F. Geiger, et al., 'What Comes after de Novo? Automated Lead Optimization of Proteins with CRADLE-1', bioRxiv, March 12, 2026.

Bloomfield, D., J.R.M. Black, O. Crook, N. Brandes, M.S. Hanke, T.V. Inglesby, A. Cicero, et al., 'Biological Data Governance in an Age of AI', *Science*, Vol. 391, No. 6785, February 5, 2026, pp. 558–561.

Bohacek, R.S., C. McMartin, and W.C. Guida, 'The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective', *Medicinal Research Reviews*, Vol. 16, No. 1, 1996, pp. 3–50.

Bommasani, R., D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, et al., 'On the Opportunities and Risks of Foundation Models', arXiv, July 12, 2022.

Brandes, N., D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, 'ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function', *Bioinformatics (Oxford, England)*, Vol. 38, No. 8, April 12, 2022, pp. 2102–2110.

Brixi, G., M.G. Durrant, J. Ku, M. Naghipourfar, M. Poli, G. Sun, G. Brockman, et al., 'Genome Modelling and Design across All Domains of Life with Evo 2', *Nature*, March 4, 2026, pp. 1–13.

Bromberg, Y., R. Altman, M. Imperiale, E. Horvitz, M. Dus, R. Townshend, V. Yao, et al., '3.1 Artificial Intelligence and the Future of Biotechnology', 2025.

Bunne, C., Y. Roohani, Y. Rosen, A. Gupta, X. Zhang, M. Roed, T. Alexandrov, et al., 'How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities', *Cell*, Vol. 187, No. 25, December 12, 2024, pp. 7045–7063.

Cao, Y., L. Yu, M. Torkel, S. Kim, Y. Lin, P. Yang, T.P. Speed, S. Ghazanfar, and J.Y.H. Yang, 'The Current Landscape and Emerging Challenges of Benchmarking Single-Cell Methods', *Briefings in Bioinformatics*, Vol. 26, No. 5, September 1, 2025a, p. bbaf380.

Cao, Z., L. Zhang, R. Rousseau, and G. Sivertsen, 'Measuring the Relative Intensity of Collaboration with Fractional Counting Methods', *Journal of Informetrics*, Vol. 19, No. 4, November 1, 2025b, p. 101743.

Cao, Z.-J., and G. Gao, 'Multi-Omics Single-Cell Data Integration and Regulatory Inference with Graph-Linked Embedding', *Nature Biotechnology*, Vol. 40, No. 10, October 2022, pp. 1458–1466.

CFG, 'Priorities to Uphold European Biosecurity in 2025', *Centre for Future Generations*, July 10, 2025. https://cfg.eu/european-biosecurity-2025/.

Chakravarty, D., J.W. Schafer, E.A. Chen, J.F. Thole, L.A. Ronish, M. Lee, and L.L. Porter, 'AlphaFold Predictions of Fold-Switched Conformations Are Driven by Structure Memorization', *Nature Communications*, Vol. 15, No. 1, August 24, 2024, p. 7296.

Chen, B., X. Cheng, P. Li, Y. Geng, J. Gong, S. Li, Z. Bei, et al., 'xTrimoPGLM: Unified 100B-Scale Pre-Trained Transformer for Deciphering the Language of Protein', bioRxiv, December 3, 2024.

Chen, J.-Y., J.-F. Wang, Y. Hu, X.-H. Li, Y.-R. Qian, and C.-L. Song, 'Evaluating the Advancements in Protein Language Models for Encoding Strategies in Protein Function Prediction: A Comprehensive Review', *Frontiers in Bioengineering and Biotechnology*, Vol. 13, January 21, 2025.

Chen, Z., W.C. King, A. Hwang, M. Gerstein, and J. Zhang, 'DeepVelo: Single-Cell Transcriptomic Deep Velocity Field Learning with Neural Ordinary Differential Equations', *Science Advances*, Vol. 8, No. 48, November 30, 2022, p. eabq3745.

Cheng, S., L. Zhang, B. Jin, Q. Zhang, X. Lu, M. You, and X. Tian, 'GraphMS: Drug Target Prediction Using Graph Representation Learning with Substructures', *Applied Sciences*, Vol. 11, No. 7, January 2021, p. 3239.

Cheng, Z., S. Wohnig, R. Gupta, S. Alam, T. Abdullahi, J.A. Ribeiro, C. Nielsen-Garcia, et al., 'Benchmarking Is Broken -- Don't Let AI Be Its Own Judge', arXiv, October 15, 2025.

Cho, N.H., K.C. Cheveralls, A.-D. Brunner, K. Kim, A.C. Michaelis, P. Raghavan, H. Kobayashi, et al., 'OpenCell: Endogenous Tagging for the Cartography of Human Cellular Organization', *Science*, Vol. 375, No. 6585, March 11, 2022, p. eabi6983.

Chuai, G., H. Ma, J. Yan, M. Chen, N. Hong, D. Xue, C. Zhou, et al., 'DeepCRISPR: Optimized CRISPR Guide RNA Design by Deep Learning', *Genome Biology*, Vol. 19, No. 1, June 26, 2018, p. 80.

Coggins, S., A.K. Saeri, K.A. Daniell, L.P. Ruster, J. Liu, and J.L. Davis, 'The 2025 OpenAI Preparedness Framework Does Not Guarantee Any AI Risk Mitigation Practices: A Proof-of-Concept for Affordance Analyses of AI Safety Policies', *arXiv.Org*, September 29, 2025. https://arxiv.org/abs/2509.24394v2.

Comajuncosa-Creus, A., G. Jorba, X. Barril, and P. Aloy, 'Comprehensive Detection and Characterization of Human Druggable Pockets through Binding Site Descriptors', *Nature Communications*, Vol. 15, No. 1, September 10, 2024, p. 7917.

Committee on Assessing and Navigating Biosecurity Concerns and Benefits of Artificial Intelligence Use in the Life Sciences, Board on Life Sciences, Division on Earth and Life Studies, Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, Committee on International Security and Arms Control, Policy and Global Affairs, and National Academies of Sciences, Engineering, and Medicine, *The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations*, National Academies Press, Washington, D.C., 2025.

Cui, H., A. Tejada-Lapuerta, M. Brbić, J. Saez-Rodriguez, S. Cristea, H. Goodarzi, M. Lotfollahi, F.J. Theis, and B. Wang, 'Towards Multimodal Foundation Models in Molecular Cell Biology', *Nature*, Vol. 640, No. 8059, April 2025, pp. 623–633.

Cui, H., C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang, 'scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI', *Nature Methods*, Vol. 21, No. 8, August 2024, pp. 1470–1480.

Dalla-Torre, H., L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A.H. Grzywaczewski, F. Oteri, C. Dallago, et al., 'Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics', *Nature Methods*, Vol. 22, No. 2, February 2025, pp. 287–297.

Dauparas, J., I. Anishchenko, N. Bennett, H. Bai, R.J. Ragotte, L.F. Milles, B.I.M. Wicky, et al., 'Robust Deep Learning–Based Protein Sequence Design Using ProteinMPNN', *Science*, Vol. 378, No. 6615, October 7, 2022, pp. 49–56.

DeMeo, B., C. Nesbitt, S.A. Miller, D.B. Burkhardt, I. Lipchina, D. Fu, P. Holderrieth, et al., 'Active Learning Framework Leveraging Transcriptomics Identifies Modulators of Disease Phenotypes', *Science*, Vol. 390, No. 6776, October 23, 2025, p. eadi8577.

DenAdel, A., M. Hughes, A. Thoutam, A. Gupta, A.W. Navia, N. Fusi, S. Raghavan, P.S. Winter, A.P. Amini, and L. Crawford, 'Evaluating the Role of Pre-Training Dataset Size and Diversity on Single-Cell Foundation Model Performance', bioRxiv, November 5, 2025.

Discovery, C., J. Boitreaud, J. Dent, M. McPartlon, J. Meier, V. Reis, A. Rogozhnikov, and K. Wu, 'Chai-1: Decoding the Molecular Interactions of Life', bioRxiv, October 15, 2024.

Du, Z., W. Fu, X. Guo, D. Caragea, and Y. Li, 'FusionESP: Improved Enzyme-Substrate Pair Prediction by Fusing Protein and Chemical Knowledge', bioRxiv, October 14, 2024.

Elnaggar, A., M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, et al., 'ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 10, October 2022, pp. 7112–7127.

European Commission, *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Establishing a Framework of Measures for Strengthening Union's Biotechnology and Biomanufacturing Sectors Particularly in the Area of Health and Amending Regulations (EC) No 178/2002, (EC) No 1394/2007, (EU) No 536/2014, (EU) 2019/6, (EU) 2024/795 and (EU) 2024/1938 (European Biotech Act)*, 2025.

Evans, R., M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, et al., 'Protein Complex Prediction with AlphaFold-Multimer', bioRxiv, March 10, 2022.

Fazzari, M.J., and J.M. Greally, 'Introduction to Epigenomics and Epigenome-Wide Analysis', in H. Bang, X.K. Zhou, H.L. van Epps, and M. Mazumdar (eds), *Statistical Methods in Molecular Biology*, Humana Press, Totowa, NJ, 2010, pp. 243–265.

Ferruz, N., S. Schmidt, and B. Höcker, 'ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design', *Nature Communications*, Vol. 13, No. 1, July 27, 2022, p. 4348.

Fishman, V., Y. Kuratov, A. Shmelev, M. Petrov, D. Penzar, D. Shepelin, N. Chekanov, O. Kardymon, and M. Burtsev, 'GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences', *Nucleic Acids Research*, Vol. 53, No. 2, January 27, 2025, p. gkae1310.

Fleming, J., P. Magana, S. Nair, M. Tsenkov, D. Bertoni, I. Pidruchna, M.Q. Lima Afonso, et al., 'AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities', *Journal of Molecular Biology*, Vol. 437, No. 15, August 1, 2025, p. 168967.

Francoeur, P.G., T. Masuda, J. Sunseri, A. Jia, R.B. Iovanisci, I. Snyder, and D.R. Koes, 'Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design', *Journal of Chemical Information and Modeling*, Vol. 60, No. 9, September 28, 2020, pp. 4200–4215.

Fuchs, J.E., G. Sivertsen, and R. Rousseau, 'Measuring the Relative Intensity of Collaboration within a Network', *Scientometrics*, Vol. 126, No. 10, October 1, 2021, pp. 8673–8682.

Galkin, F., V. Naumov, S. Pushkov, D. Sidorenko, A. Urban, D. Zagirova, K.M. Alawi, et al., 'Precious3GPT: Multimodal Multi-Species Multi-Omics Multi-Tissue Transformer for Aging Research and Drug Discovery', bioRxiv, July 25, 2024.

Gayoso, A., P. Weiler, M. Lotfollahi, D. Klein, J. Hong, A. Streets, F.J. Theis, and N. Yosef, 'Deep Generative Modeling of Transcriptional Dynamics for RNA Velocity Analysis in Single Cells', *Nature Methods*, Vol. 21, No. 1, January 2024, pp. 50–59.

Greener, J.G., S.M. Kandathil, and D.T. Jones, 'Deep Learning Extends de Novo Protein Modelling Coverage of Genomes Using Iteratively Predicted Structural Constraints', *Nature Communications*, Vol. 10, No. 1, September 4, 2019, p. 3977.

Gu, X., A. Aranganathan, and P. Tiwary, 'Empowering AlphaFold2 for Protein Conformation Selective Drug Discovery with AlphaFold2-RAVE', *eLife*, Vol. 13, August 13, 2024.

Guan, J., X. Zhou, Y. Yang, Y. Bao, J. Peng, J. Ma, Q. Liu, L. Wang, and Q. Gu, 'DecompDiff: Diffusion Models with Decomposed Priors for Structure-Based Drug Design', arXiv, February 26, 2024.

Guo, F., R. Guan, Y. Li, Q. Liu, X. Wang, C. Yang, and J. Wang, 'Foundation Models in Bioinformatics', *National Science Review*, Vol. 12, No. 4, March 7, 2025.

Hao, M., J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang, and L. Song, 'Large-Scale Foundation Model on Single-Cell Transcriptomics', *Nature Methods*, Vol. 21, No. 8, August 2024, pp. 1481–1491.

Hayes, T., R. Rao, H. Akin, N.J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, et al., 'Simulating 500 Million Years of Evolution with a Language Model', *Science*, Vol. 387, No. 6736, February 21, 2025, pp. 850–858.

He, F., R. Fei, J.E. Krull, Y. Yu, X. Zhang, X. Wang, H. Cheng, et al., 'Harnessing the Power of Single-Cell Large Language Models with Parameter-Efficient Fine-Tuning Using scPEFT', *Nature Machine Intelligence*, Vol. 8, No. 1, January 2026, pp. 118–133.

Heimberg, G., T. Kuo, D.J. DePianto, O. Salem, T. Heigl, N. Diamant, G. Scalia, et al., 'A Cell Atlas Foundation Model for Scalable Search of Similar Human Cells', *Nature*, Vol. 638, No. 8052, February 2025, pp. 1085–1094.

Heinzinger, M., A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost, 'Modeling Aspects of the Language of Life through Transfer-Learning Protein Sequences', *BMC Bioinformatics*, Vol. 20, No. 1, December 2019, p. 723.

Heinzinger, M., K. Weissenow, J.G. Sanchez, A. Henkel, M. Mirdita, M. Steinegger, and B. Rost, 'Bilingual Language Model for Protein Sequence and Structure', *NAR Genomics and Bioinformatics*, Vol. 6, No. 4, December 1, 2024, p. lqae150.

Hermann, L., T. Fiedler, H.A. Nguyen, M. Nowicka, and J.M. Bartoszewicz, 'Beware of Data Leakage from Protein LLM Pretraining', bioRxiv, July 24, 2024.

Hie, B.L., 'Mapping the Landscape of AI-Enabled Biological Design', *The Age of AI in the Life Sciences (National Academies Press*, 2025.

High Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, Directorate-General for Communications Networks, Content and Technology, 2019.

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de L. Casas, et al., 'Training Compute-Optimal Large Language Models', arXiv, March 29, 2022.

Hou, W., and Z. Ji, 'Assessing GPT-4 for Cell Type Annotation in Single-Cell RNA-Seq Analysis', *Nature Methods*, Vol. 21, No. 8, August 2024, pp. 1462–1465.

Hu, L., H. Qin, Y. Zhang, Y. Lu, P. Qiu, Z. Guo, L. Cao, et al., 'RegFormer: A Single-Cell Foundation Model Powered by Gene Regulatory Hierarchies', *Nature Communications*, May 5, 2026.

Huang, K., P. Chandak, Q. Wang, S. Havaldar, A. Vaid, J. Leskovec, G.N. Nadkarni, B.S. Glicksberg, N. Gehlenborg, and M. Zitnik, 'A Foundation Model for Clinician-Centered Drug Repurposing', *Nature Medicine*, Vol. 30, No. 12, December 2024a, pp. 3601–3613.

Huang, Z., L. Yang, Z. Zhang, X. Zhou, Y. Bao, X. Zheng, Y. Yang, Y. Wang, and W. Yang, 'Binding-Adaptive Diffusion Models for Structure-Based Drug Design', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 11, March 24, 2024b, pp. 12671–12679.

Huang, Z., S. Luo, Z. Wang, Z. Zhang, B. Jiang, Q. Nie, and J. Zhang, 'Deep Learning Linking Mechanistic Models to Single-Cell Transcriptomics Data Reveals Transcriptional Bursting in Response to DNA Damage', *bioRxiv*, January 1, 2024c, p. 2024.07.10.602845.

Imperiale, M.J., and A. Casadevall, 'A New Synthesis for Dual Use Research of Concern', *PLOS Medicine*, Vol. 12, No. 4, April 14, 2015, p. e1001813.

Ingraham, J.B., M. Baranov, Z. Costello, K.W. Barber, W. Wang, A. Ismail, V. Frappier, et al., 'Illuminating Protein Space with a Programmable Generative Model', *Nature*, Vol. 623, No. 7989, November 2023, pp. 1070–1078.

Ishitani, R., and Y. Moriwaki, 'Improving Stereochemical Limitations in Protein–Ligand Complex Structure Prediction', *ACS Omega*, Vol. 10, No. 46, November 25, 2025, pp. 56075–56084.

Ivanenkov, Y.A., D. Polykovskiy, D. Bezrukov, B. Zagribelnyy, V. Aladinskiy, P. Kamya, A. Aliper, F. Ren, and A. Zhavoronkov, 'Chemistry42: An AI-Driven Platform for Molecular Design and Optimization', *Journal of Chemical Information and Modeling*, Vol. 63, No. 3, February 13, 2023, pp. 695–701.

Ji, Y., Z. Zhou, H. Liu, and R.V. Davuluri, 'DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome', *Bioinformatics*, Vol. 37, No. 15, August 9, 2021, pp. 2112–2120.

Jia, Y., B. Gao, J. Tan, J. Zheng, X. Hong, W. Zhu, H. Tan, et al., 'Deep Contrastive Learning Enables Genome-Wide Virtual Screening', *Science*, Vol. 391, No. 6781, January 8, 2026, p. eads9530.

Jiménez-Luna, J., F. Grisoni, and G. Schneider, 'Drug Discovery with Explainable Artificial Intelligence', *Nature Machine Intelligence*, Vol. 2, No. 10, October 2020, pp. 573–584.

Jing, B., H. Stärk, T. Jaakkola, and B. Berger, 'Generative Modeling of Molecular Dynamics Trajectories', arXiv, September 26, 2024.

Jones, D.T., D.W.A. Buchan, D. Cozzetto, and M. Pontil, 'PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments', *Bioinformatics*, Vol. 28, No. 2, January 15, 2012, pp. 184–190.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, et al., 'Applying and Improving AlphaFold at CASP14', *Proteins: Structure, Function, and Bioinformatics*, Vol. 89, No. 12, December 1, 2021a, pp. 1711–1721.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, et al., 'Highly Accurate Protein Structure Prediction with AlphaFold', *Nature*, Vol. 596, No. 7873, August 2021b, pp. 583–589.

Kalfon, J., J. Samaran, G. Peyré, and L. Cantini, 'scPRINT: Pre-Training on 50 Million Cells Allows Robust Gene Network Predictions', *Nature Communications*, Vol. 16, No. 1, April 16, 2025, p. 3607.

Kamimoto, K., B. Stringa, C.M. Hoffmann, K. Jindal, L. Solnica-Krezel, and S.A. Morris, 'Dissecting Cell Identity via Network Inference and in Silico Gene Perturbation', *Nature*, Vol. 614, No. 7949, February 2023, pp. 742–751.

Koh, H.Y., Y. Zheng, M. Yang, R. Arora, G.I. Webb, S. Pan, L. Li, and G.M. Church, 'AI-Driven Protein Design', *Nature Reviews Bioengineering*, Vol. 3, No. 12, December 2025, pp. 1034–1056.

Kommu, S., Y. Wang, Y. Wang, and X. Wang, 'Prediction of Gene Regulatory Connections with Joint Single-Cell Foundation Models and Graph-Based Learning', *Bioinformatics*, Vol. 41, No. Supplement_1, July 1, 2025, pp. i619–i627.

Krishna, R., J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G.R. Lee, et al., 'Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom', *Science*, Vol. 384, No. 6693, March 7, 2024, p. eadl2528.

Kuhn, M., I. Letunic, L.J. Jensen, and P. Bork, 'The SIDER Database of Drugs and Side Effects', *Nucleic Acids Research*, Vol. 44, No. Database issue, January 4, 2016, pp. D1075–D1079.

Lal, A., D. Garfield, T. Biancalani, and G. Eraslan, 'Designing Realistic Regulatory DNA with Autoregressive Language Models', *Genome Research*, Vol. 34, No. 9, October 11, 2024, pp. 1411–1420.

Lavin, A., C.M. Gilligan-Lee, A. Visnjic, S. Ganju, D. Newman, S. Ganguly, D. Lange, et al., 'Technology Readiness Levels for Machine Learning Systems', *Nature Communications*, Vol. 13, No. 1, October 20, 2022, p. 6039.

LeCun, Y., Y. Bengio, and G. Hinton, 'Deep Learning', *Nature*, Vol. 521, No. 7553, May 1, 2015, pp. 436–444.

Leslie, M., 'The Silicon Cell', Vol. 390, Science, October 30, 2025.

Li, Q., Z. Hu, Y. Wang, L. Li, Y. Fan, I. King, G. Jia, S. Wang, L. Song, and Y. Li, 'Progress and Opportunities of Foundation Models in Bioinformatics', *Briefings in Bioinformatics*, Vol. 25, No. 6, September 23, 2024a, p. bbae548.

Li, S., T. Xiao, Y. Lan, C. Wu, Z. Li, R. Liu, Q. Fang, and C. Zhang, 'Artificial Intelligence Revolution in Transcriptomics: From Single Cells to Spatial Atlases', *Advanced Science*, Vol. 13, No. 5, 2026, p. e18949.

Li, S., Z. Wang, Z. Liu, D. Wu, C. Tan, J. Zheng, Y. Huang, and S.Z. Li, 'VQDNA: Unleashing the Power of Vector Quantization for Multi-Species Genomic Sequence Modeling', arXiv, June 2, 2024b.

Li, W., F. Yang, F. Wang, Y. Rong, L. Liu, B. Wu, H. Zhang, and J. Yao, 'scPROTEIN: A Versatile Deep Graph Contrastive Learning Framework for Single-Cell Proteomics Embedding', *Nature Methods*, Vol. 21, No. 4, April 2024c, pp. 623–634.

Lin, Z., H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, et al., 'Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model', *Science (New York, N.Y.)*, Vol. 379, No. 6637, March 17, 2023, pp. 1123–1130.

Linder, J., D. Srivastava, H. Yuan, V. Agarwal, and D.R. Kelley, 'Predicting RNA-Seq Coverage from DNA Sequence as a Unifying Model of Gene Regulation', *Nature Genetics*, Vol. 57, No. 4, April 2025, pp. 949–961.

Liu, O., S. Jaghouar, J. Hagemann, S. Wang, J. Wiemels, J. Kaufman, and W. Neiswanger, 'METAGENE-1: Metagenomic Foundation Model for Pandemic Monitoring', arXiv, January 3, 2025.

Liu, S., W. Nie, C. Wang, J. Lu, Z. Qiao, L. Liu, J. Tang, C. Xiao, and A. Anandkumar, 'Multi-Modal Molecule Structure–Text Model for Text-Based Retrieval and Editing', *Nature Machine Intelligence*, Vol. 5, No. 12, December 2023, pp. 1447–1457.

Loeffler, H.H., J. He, A. Tibo, J.P. Janet, A. Voronov, L.H. Mervin, and O. Engkvist, 'Reinvent 4: Modern AI–Driven Generative Molecule Design', *Journal of Cheminformatics*, Vol. 16, February 21, 2024, p. 20.

Lopez, R., J. Regier, M.B. Cole, M.I. Jordan, and N. Yosef, 'Deep Generative Modeling for Single-Cell Transcriptomics', *Nature Methods*, Vol. 15, No. 12, December 2018, pp. 1053–1058.

Lotfollahi, M., M. Naghipourfar, M.D. Luecken, M. Khajavi, M. Büttner, M. Wagenstetter, Ž. Avsec, et al., 'Mapping Single-Cell Data to Reference Atlases by Transfer Learning', *Nature Biotechnology*, Vol. 40, No. 1, January 2022, pp. 121–130.

Marcinkevičs, R., K. Sokol, A. Paulraj, M.A. Hilbert, V. Rimili, S. Wellmann, C. Knorr, B. Reingruber, J.E. Vogt, and P. Reis Wolfertstetter, 'External Validation of Predictive Models for Diagnosis, Management and Severity of Pediatric Appendicitis', *Frontiers in Pediatrics*, Vol. 13, August 29, 2025.

Marin, F.I., F. Teufel, M. Horlacher, D. Madsen, D. Pultz, O. Winther, and W. Boomsma, 'BEND: BENCHMARKING DNA LANGUAGE MODELS ON BIOLOGICALLY MEANINGFUL TASKS', 2024.

Markowetz, F., 'All Models Are Wrong and Yours Are Useless: Making Clinical Prediction Models Impactful for Patients', *Npj Precision Oncology*, Vol. 8, No. 1, February 28, 2024, p. 54.

Martinez, P.F., G.E. Gomez, and J. Hernández-Orallo, 'AI Watch: Assessing Technology Readiness Levels for Artificial Intelligence', *JRC Publications Repository*, 2020. https://publications.jrc.ec.europa.eu/repository/handle/JRC122014.

Martínez-Plumed, F., F. Caballero, D. Castellano-Falcón, D. Fernández-Llorca, E. Gómez, I. Hupont-Torres, L. Merino, C. Monserrat, European Commission Joint Research Centre, and J. Hernández Orallo, *AI Watch, Revisiting Technology Readiness Levels for Relevant Artificial Intelligence Technologies*, *EUR, 31066*, Publications Office, Luxembourg, 2022.

Masters, M.R., A.H. Mahmoud, and M.A. Lill, 'Investigating Whether Deep Learning Models for Co-Folding Learn the Physics of Protein-Ligand Interactions', *Nature Communications*, Vol. 16, No. 1, October 6, 2025, p. 8854.

Moen, E., D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, 'Deep Learning for Cellular Image Analysis', *Nature Methods*, Vol. 16, No. 12, December 2019, pp. 1233–1246.

Moldwin, A., and A. Shehu, 'Foundation models for AI-enabled biological design', 2025.

Munson, B.P., M. Chen, A. Bogosian, J.F. Kreisberg, K. Licon, R. Abagyan, B.M. Kuenzi, and T. Ideker, 'De Novo Generation of Multi-Target Compounds Using Deep Generative Chemistry', *Nature Communications*, Vol. 15, No. 1, May 6, 2024, p. 3636.

Nadig, A., A. Thoutam, M. Hughes, A. Gupta, A.W. Navia, N. Fusi, S. Raghavan, P.S. Winter, A.P. Amini, and L. Crawford, 'Consequences of Training Data Composition for Deep Learning Models in Single-Cell Biology', bioRxiv, February 24, 2025.

Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilization to the Convention on Biological Diversity: Text and Annex, Secretariat of the Convention on Biological Diversity, United Nations Environmental Programme, 2011.

National Academies of Sciences, E., P. and G. Affairs, D. on E. and L. Studies, I.N. and Cooperation, B. on L. Sciences, T. Tucholski, and A. Johnson, 'Data Governance Principles for Life Science Research Across the Globe', *Engaging Scientists in Central Asia on Life Science Data Governance Principles: Proceedings of a Workshop Series*, National Academies Press (US), 2024.

Nguyen, E., M. Poli, M.G. Durrant, B. Kang, D. Katrekar, D.B. Li, L.J. Bartie, et al., 'Sequence Modeling and Design from Molecular to Genome Scale with Evo', *Science*, Vol. 386, No. 6723, November 15, 2024, p. eado9336.

Nguyen, E., M. Poli, M. Faizi, A. Thomas, C. Birch-Sykes, M. Wornow, A. Patel, et al., 'HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution', *ArXiv*, November 14, 2023, p. arXiv:2306.15794v2.

Nijkamp, E., J.A. Ruffolo, E.N. Weinstein, N. Naik, and A. Madani, 'ProGen2: Exploring the Boundaries of Protein Language Models', *Cell Systems*, Vol. 14, No. 11, November 15, 2023, pp. 968-978.e3.

Notin, P., A. Kollasch, D. Ritter, L. van Niekerk, S. Paul, H. Spinner, N. Rollins, et al., 'ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design', *Advances in Neural Information Processing Systems*, Vol. 36, December 15, 2023, pp. 64331–64379.

OECD, 'Guidance Document on Good In Vitro Method Practices (GIVIMP)', *OECD Series on Testing and Assessment*, Vol. 2018, December 9, 2018.

OECD, 'Synthetic Biology, AI and Automation: A Forward-looking Technology Assessment', *OECD Science, Technology and Industry Policy Papers*, December 3, 2025.

Orben, A., and J.N. Matias, 'Fixing the Science of Digital Technology Harms', *Science*, Vol. 388, No. 6743, April 11, 2025, pp. 152–155.

Pannu, J., D. Bloomfield, R. MacKnight, M.S. Hanke, A. Zhu, G. Gomes, A. Cicero, and T.V. Inglesby, 'Dual-Use Capabilities of Concern of Biological AI Models', *PLOS Computational Biology*, Vol. 21, No. 5, May 8, 2025, p. e1012975.

Patterson, D., J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, 'Carbon Emissions and Large Neural Network Training', arXiv, April 23, 2021.

Pearce, J.D., S.E. Simmonds, G. Mahmoudabadi, L. Krishnan, G. Palla, A.-M. Istrate, A. Tarashansky, et al., 'TranscriptFormer: A Generative Cell Atlas across 1.5 Billion Years of Evolution', *Science*, Vol. 0, No. 0, May 7, 2026, p. eaec8514.

Peng, X., S. Luo, J. Guan, Q. Xie, J. Peng, and J. Ma, 'Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets', arXiv, July 11, 2025.

Penić, R.J., T. Vlašić, R.G. Huber, Y. Wan, and M. Šikić, 'RiNALMo: General-Purpose RNA Language Models Can Generalize Well on Structure Prediction Tasks', November 12, 2024.

Purificato, Erasmo, Bili, Danai, Jungnickel, Robert, Ruiz-Serra, Victoria, Fabiani, Josefina, Abendroth-Dias, Kulani, Fernández-Llorca, David, and Gómez, Emilia, *The Role of Artificial Intelligence in Scientific Research: A Science for Policy, European Perspective*, Publications Office of the European Union, 2025.

Pyeon, M., J. Lee, M. Lee, J. Yun, H. Choi, J. Kim, J. Kim, Y. Hu, J. Jang, and S. Lee, 'EXAONE Path 2.0: Pathology Foundation Model with End-to-End Supervision', arXiv, 2025.

Qiu, P., Q. Chen, H. Qin, S. Fang, Y. Zhang, Y. Zhang, T. Xia, et al., 'BioLLM: A Standardized Framework for Integrating and Benchmarking Single-Cell Foundation Models', *Patterns*, Vol. 6, No. 8, August 8, 2025, p. 101326.

Ramsauer, H., B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, et al., 'Hopfield Networks Is All You Need', arXiv, April 28, 2021.

Rangan, R., R. Feathers, S. Khavnekar, A. Lerer, J.D. Johnston, R. Kelley, M. Obr, A. Kotecha, and E.D. Zhong, 'CryoDRGN-ET: Deep Reconstructing Generative Networks for Visualizing Dynamic Biomolecules inside Cells', *Nature Methods*, Vol. 21, No. 8, August 2024, pp. 1537–1545.

Rao, V.M., S. Zhang, B.S. Plosky, P.D. Hsu, B. Wang, J. Zou, M. Zitnik, E.J. Topol, and P. Rajpurkar, 'Generalist Biological Artificial Intelligence in Modeling the Language of Life', *Nature Biotechnology*, March 20, 2026, pp. 1–16.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance), 13.6.2024.

Repecka, D., V. Jauniskis, L. Karpus, E. Rembeza, I. Rokaitis, J. Zrimec, S. Poviloniene, et al., 'Expanding Functional Protein Sequence Spaces Using Generative Adversarial Networks', *Nature Machine Intelligence*, Vol. 3, No. 4, April 2021, pp. 324–333.

Ritoré, Á., C.M. Jiménez, J.L. González, J.C. Rejón-Parrilla, P. Hervás, E. Toro, C.L. Parra-Calderón, L.A. Celi, I. Túnez, and M.Á. Armengol De La Hoz, 'The Role of Open Access Data in Democratizing Healthcare AI: A Pathway to Research Enhancement, Patient Well-Being and Treatment Equity in Andalusia, Spain', Edited by P.-C. Kuo, *PLOS Digital Health*, Vol. 3, No. 9, September 16, 2024, p. e0000599.

Riva, M., T.L. Parigi, F. Ungaro, and L. Massimino, 'Hugging Face's Impact on Medical Applications of Artificial Intelligence', *Computational and Structural Biotechnology Reports*, Vol. 1, December 1, 2024, p. 100003.

Rives, A., J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, et al., 'Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences', *Proceedings of the National Academy of Sciences*, Vol. 118, No. 15, April 13, 2021, p. e2016239118.

Rosen, Y., Y. Roohani, A. Agrawal, L. Samotorčan, T.S. Consortium, S.R. Quake, and J. Leskovec, 'Universal Cell Embeddings: A Foundation Model for Cell Biology', bioRxiv, April 8, 2026.

Ruffolo, J.A., S. Nayfach, J. Gallagher, A. Bhatnagar, J. Beazer, R. Hussain, J. Russ, et al., 'Design of Highly Functional Genome Editors by Modelling CRISPR–Cas Sequences', *Nature*, Vol. 645, No. 8080, September 2025, pp. 518–525.

Sample, I., 'Google DeepMind Launches AI Tool to Help Identify Genetic Drivers of Disease', *The Guardian*, January 28, 2026, sect. Science.

Schiff, Y., C.-H. Kao, A. Gokaslan, T. Dao, A. Gu, and V. Kuleshov, 'Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling', arXiv, June 5, 2024.

Schneider, V.A., T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P.A. Kitts, T.D. Murphy, et al., 'Evaluation of GRCh38 and de Novo Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly', *Genome Research*, Vol. 27, No. 5, May 2017, pp. 849–864.

Sevilla, J., L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, 'Compute Trends Across Three Eras of Machine Learning', *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

Sgarbossa, D., C. Malbranke, and A.-F. Bitbol, 'ProtMamba: A Homology-Aware but Alignment-Free Protein State Space Model', *Bioinformatics*, Vol. 41, No. 6, June 1, 2025, p. btaf348.

Sterling, T., and J.J. Irwin, 'ZINC 15--Ligand Discovery for Everyone', *Journal of Chemical Information and Modeling*, Vol. 55, No. 11, November 23, 2015, pp. 2324–2337.

Stokes, J.M., K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N.M. Donghia, C.R. MacNair, et al., 'A Deep Learning Approach to Antibiotic Discovery', *Cell*, Vol. 180, No. 4, February 20, 2020, pp. 688-702.e13.

Stolovitzky, G., J. Saez-Rodriguez, J. Bletz, J. Albrecht, G. Andreoletti, J.C. Costello, and P. Boutros, 'AI Competitions and Benchmarks: The Life Cycle of Challenges and Benchmarks', arXiv, December 8, 2023.

Stringer, C., T. Wang, M. Michaelos, and M. Pachitariu, 'Cellpose: A Generalist Algorithm for Cellular Segmentation', *Nature Methods*, Vol. 18, No. 1, January 2021, pp. 100–106.

Strubell, E., A. Ganesh, and A. McCallum, 'Energy and Policy Considerations for Deep Learning in NLP', in A. Korhonen, D. Traum, and L. Màrquez (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650.

Su, J., C. Han, Y. Zhou, J. Shan, X. Zhou, and F. Yuan, 'SaProt: Protein Language Modeling with Structure-Aware Vocabulary', bioRxiv, April 19, 2024.

Subramanian, G., B. Ramsundar, V. Pande, and R.A. Denny, 'Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches', *Journal of Chemical Information and Modeling*, Vol. 56, No. 10, October 24, 2016, pp. 1936–1949.

Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, et al., 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age', *PLOS Medicine*, Vol. 12, No. 3, March 31, 2015, p. e1001779.

Sun, D., W. Gao, H. Hu, and S. Zhou, 'Why 90% of Clinical Drug Development Fails and How to Improve It?', *Acta Pharmaceutica Sinica B*, Vol. 12, No. 7, July 1, 2022, pp. 3049–3062.

Tang, Z., Z. Li, T. Hou, T. Zhang, B. Yang, J. Su, and Q. Song, 'SiGra: Single-Cell Spatial Elucidation through an Image-Augmented Graph Transformer', *Nature Communications*, Vol. 14, No. 1, September 12, 2023, p. 5618.

Tanihata, S., and H. Iwata, 'Next-Generation Artificial Intelligence for ADME Prediction in Drug Discovery: From Small Molecules to Biologics', *Yonago Acta Medica*, Vol. 69, No. 1, February 2026, pp. 1–13.

Tejada-Lapuerta, A., A.C. Schaar, R. Gutgesell, G. Palla, L. Halle, M. Minaeva, L. Vornholz, et al., 'Nicheformer: A Foundation Model for Single-Cell and Spatial Omics', *Nature Methods*, Vol. 22, No. 12, December 2025, pp. 2525–2538.

Theodoris, C.V., L. Xiao, A. Chopra, M.D. Chaffin, Z.R. Al Sayed, M.C. Hill, H. Mantineo, et al., 'Transfer Learning Enables Predictions in Network Biology', *Nature*, Vol. 618, No. 7965, June 2023, pp. 616–624.

Topol, E.J., 'Learning the Language of Life with AI', *Science*, Vol. 387, No. 6733, January 30, 2025, p. eadv4414.

Townshend, R.J.L., S. Eismann, A.M. Watkins, R. Rangan, M. Karelina, R. Das, and R.O. Dror, 'Geometric Deep Learning of RNA Structure', *Science*, Vol. 373, No. 6558, August 27, 2021, pp. 1047–1051.

Tran, V.Q., M. Nemeth, L.J. Bartie, S.S. Chandrasekaran, A. Fanton, H.C. Moon, B.L. Hie, S. Konermann, and P.D. Hsu, 'Rapid Directed Evolution Guided by Protein Language Models and Epistatic Interactions', *Science*, Vol. 0, No. 0, February 19, 2026, p. eaea1820.

Trang, B., 'AlphaFold Developer Google DeepMind to Fund CASP as NIH Funding Falls Short', *STAT*, July 21, 2025. https://www.statnews.com/2025/07/21/casp-new-funding-from-alphafold-developer-google-deepmind-after-nih-grant-runs-out/.

Urbina, F., F. Lentzos, C. Invernizzi, and S. Ekins, 'Dual Use of Artificial Intelligence-Powered Drug Discovery', *Nature Machine Intelligence*, Vol. 4, No. 3, March 2022, pp. 189–191.

Vander Meersche, Y., G. Cretin, A. Gheeraert, J.-C. Gelly, and T. Galochkina, 'ATLAS: Protein Flexibility Description from Atomistic Molecular Dynamics Simulations', *Nucleic Acids Research*, Vol. 52, No. D1, January 5, 2024, pp. D384–D392.

Varadi, M., D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, et al., 'AlphaFold Protein Structure Database in 2024: Providing Structure Coverage for over 214 Million Protein Sequences', *Nucleic Acids Research*, Vol. 52, No. D1, January 5, 2024, pp. D368–D375.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, 'Attention Is All You Need', arXiv, August 2, 2023.

Veiner, M., and F. Supek, 'The DNA Dialect: A Comprehensive Guide to Pretrained Genomic Language Models', *Molecular Systems Biology*, Vol. 22, No. 3, March 1, 2026, pp. 309–332.

Vreven, T., I.H. Moal, A. Vangone, B.G. Pierce, P.L. Kastritis, M. Torchala, R. Chaleil, et al., 'Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2', *Journal of Molecular Biology*, Vol. 427, No. 19, September 25, 2015, pp. 3031–3041.

Wan, F., L. Hong, A. Xiao, T. Jiang, and J. Zeng, 'NeoDTI: Neural Integration of Neighbor Information from a Heterogeneous Network for Discovering New Drug–Target Interactions', *Bioinformatics*, Vol. 35, No. 1, January 1, 2019, pp. 104–111.

Wang, D., C. Zhang, B. Wang, B. Li, Q. Wang, D. Liu, H. Wang, et al., 'Optimized CRISPR Guide RNA Design for Two High-Fidelity Cas9 Variants by Deep Learning', *Nature Communications*, Vol. 10, No. 1, September 19, 2019, p. 4284.

Wang, Y., J. Wang, Z. Cao, and A. Barati Farimani, 'Molecular Contrastive Learning of Representations via Graph Neural Networks', *Nature Machine Intelligence*, Vol. 4, No. 3, March 2022, pp. 279–287.

Watson, J.L., D. Juergens, N.R. Bennett, B.L. Trippe, J. Yim, H.E. Eisenach, W. Ahern, et al., 'De Novo Design of Protein Structure and Function with RFdiffusion', *Nature*, Vol. 620, No. 7976, August 2023, pp. 1089–1100.

Wells, B.J., H.M. Nguyen, A. McWilliams, M. Pallini, A. Bovi, A. Kuzma, J. Kramer, et al., 'A Practical Framework for Appropriate Implementation and Review of Artificial Intelligence (FAIR-AI) in Healthcare', *Npj Digital Medicine*, Vol. 8, No. 1, August 11, 2025, p. 514.

Wenteler, A., M. Occhetta, N. Branson, V. Curean, M. Huebner, W. Dee, W. Connell, et al., 'PertEval-scFM: Benchmarking Single-Cell Foundation Models for Perturbation Effect Prediction', Bioinformatics, October 3, 2024.

Wheeler, N.E., 'Responsible AI in Biotechnology: Balancing Discovery, Innovation and Biosecurity Risks', *Frontiers in Bioengineering and Biotechnology*, Vol. 13, February 5, 2025, p. 1537471.

Wohlwend, J., G. Corso, S. Passaro, N. Getz, M. Reveiz, K. Leidal, W. Swiderski, et al., 'Boltz-1 Democratizing Biomolecular Interaction Modeling', bioRxiv, May 6, 2025.

Wong, C.H., K.W. Siah, and A.W. Lo, 'Estimation of Clinical Trial Success Rates and Related Parameters', *Biostatistics*, Vol. 20, No. 2, April 1, 2019, pp. 273–286.

Wouters, O.J., M. McKee, and J. Luyten, 'Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018', *JAMA*, Vol. 323, No. 9, March 3, 2020, pp. 844–853.

Wu, R., F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, et al., 'High-Resolution de Novo Structure Prediction from Primary Sequence', bioRxiv, July 22, 2022.

Xia, J., C. Zhao, B. Hu, Z. Gao, C. Tan, Y. Liu, S. Li, and S.Z. Li, 'Mole-BERT: Rethinking Pre-Training Graph Neural Networks for Molecules', 2022.

Xie, F., B. Zhao, S. Xu, Z. Wang, J.J. Moon, and L.X. Garmire, 'Overcoming Barriers to the Wide Adoption of Single-Cell Large Language Models in Biomedical Research', *Nature Biotechnology*, Vol. 43, No. 11, November 2025, pp. 1758–1762.

Yang, F., W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, and J. Yao, 'scBERT as a Large-Scale Pretrained Deep Language Model for Cell Type Annotation of Single-Cell RNA-Seq Data', *Nature Machine Intelligence*, Vol. 4, No. 10, October 2022, pp. 852–866.

Yang, L., Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, 'Diffusion Models: A Comprehensive Survey of Methods and Applications', arXiv, September 27, 2025.

Yang, X., G. Liu, G. Feng, D. Bu, P. Wang, J. Jiang, S. Chen, et al., 'GeneCompass: Deciphering Universal Gene Regulatory Mechanisms with a Knowledge-Informed Cross-Species Foundation Model', *Cell Research*, Vol. 34, No. 12, December 2024a, pp. 830–845.

Yang, Z., W. Zhong, Q. Lv, T. Dong, G. Chen, and C.Y.-C. Chen, 'Interaction-Based Inductive Bias in Graph Neural Networks: Enhancing Protein-Ligand Binding Affinity Predictions From 3D Structures', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 12, December 2024b, pp. 8191–8208.

Zablocki, L.I., L.A. Bugnon, M. Gerard, L. Di Persia, G. Stegmayer, and D.H. Milone, 'Comprehensive Benchmarking of Large Language Models for RNA Secondary Structure Prediction', *Briefings in Bioinformatics*, Vol. 26, No. 2, March 1, 2025, p. bbaf137.

Zeng, Y., J. Xie, N. Shangguan, Z. Wei, W. Li, Y. Su, S. Yang, et al., 'CellFM: A Large-Scale Foundation Model Pre-Trained on Transcriptomics of 100 Million Human Cells', *Nature Communications*, Vol. 16, No. 1, May 20, 2025, p. 4679.

Zhang, Q., W. Chen, M. Qin, Y. Wang, Z. Pu, K. Ding, Y. Liu, et al., 'Integrating Protein Language Models and Automatic Biofoundry for Enhanced Protein Evolution', *Nature Communications*, Vol. 16, No. 1, February 11, 2025a, p. 1553.

Zhang, X.-M., L. Liang, L. Liu, and M.-J. Tang, 'Graph Neural Networks and Their Current Applications in Bioinformatics', *Frontiers in Genetics*, Vol. 12, July 29, 2021.

Zhang, Z., M. Wang, and Q. Liu, 'FlexSBDD: Structure-Based Drug Design with Flexible Protein Modeling', 2024.

Zhang, Z., Z. Zhou, R. Jin, L. Cong, and M. Wang, 'GeneBreaker: Jailbreak Attacks against DNA Language Models with Pathogenicity Guidance', arXiv, May 28, 2025b.

Zhong, E.D., T. Bepler, B. Berger, and J.H. Davis, 'CryoDRGN: Reconstruction of Heterogeneous Cryo-EM Structures Using Neural Networks', *Nature Methods*, Vol. 18, No. 2, February 2021, pp. 176–185.

Zhou, Z., Y. Ji, W. Li, P. Dutta, R.V. Davuluri, and H. Liu, 'DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes', 2023.

Zhou, Z., R. Riley, S. Kautsar, W. Wu, R. Egan, S. Hofmeyr, S. Goldhaber-Gordon, et al., 'GenomeOcean: An Efficient Genome Foundation Model Trained on Large-Scale Metagenomic Assemblies', *bioRxiv*, February 5, 2025a, p. 2025.01.30.635558.

Zhou, Z., W. Wu, H. Ho, J. Wang, L. Shi, R.V. Davuluri, Z. Wang, and H. Liu, 'DNABERT-S: Pioneering Species Differentiation with Species-Aware DNA Embeddings', *Bioinformatics*, Vol. 41, No. Supplement_1, July 1, 2025b, pp. i255–i264.

Zvyagin, M., A. Brace, K. Hippe, Y. Deng, B. Zhang, C.O. Bohorquez, A. Clyde, et al., 'GenSLMs: Genome-Scale Language Models Reveal SARS-CoV-2 Evolutionary Dynamics', 2022.

List of abbreviations and definitions

Abbreviations	Definitions
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
AI	Artificial Intelligence
AI4Bio	Artificial Intelligence for Biology
BEACON	Benchmarking and Evaluation Assessment Consortium for Open Networks
BFD	Big Fantastic Database
CAFA	Critical Assessment of Function Annotation
CASP	Critical Assessment of Protein Structure Prediction
CBRN	Chemical, Biological, Radiological, and Nuclear
CCLE	Cancer Cell Line Encyclopedia
CDER	Center for Drug Evaluation and Research
CE	Conformité Européenne
CNNs	Convolutional Neural Networks
CORDIS	Community Research and Development Information Service
CPUs	Central Processing Units
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
EC	European Commission
EMA	European Medicines Agency
ENA	European Nucleotide Archive
EuroHPC JU	European High-Performance Computing Joint Undertaking
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration

Abbreviations	Definitions
FLOP	Floating-Point Operations Per Second
GAN	Generative Adversarial Network
GBAI	Generalist Biological Artificial Intelligence
GDPR	General Data Protection Regulation
GEO	Gene Expression Omnibus
GIVIMP	Good in vitro Method Practices
GO	Gene Ontology
GPT	Generative Pre-trained Transformer
GPUs	Graphics Processing Units
HKU	Hong Kong University
IVDR	In Vitro Diagnostic Regulation
JRC	Joint Research Centre
MDR	Medical Device Regulation
MIT	Massachusetts Institute of Technology
MLTRL	Machine Learning Technology Readiness Levels
MSAs	Multiple Sequence Alignments
MSCA	Marie Skłodowska-Curie Actions
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NSF	National Science Foundation
NSFC	National Natural Science Foundation of China

Abbreviations	Definitions
OECD	Organisation for Economic Co-operation and Development
PCA	Principal Component Analysis
PDB	Protein Data Bank
QSAR	Quantitative Structure-Activity Relationship
Real-World	Real-World Data/Evidence
RoW	Rest of World
RTD	Research and Innovation
SaMD	Software as a Medical Device
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SJTU	Shanghai Jiao Tong University
TPUs	Tensor Processing Units
TRLs	Technology Readiness Levels
VAEs	Variational Autoencoders

List of boxes

List of figures

List of tables

Annexes

Annex 1. Glossary

In recent years, there has been a significant surge in research activities focused on the AI4Bio domain. Despite this growing body of work, numerous studies and publications often employ distinct definitions, which can lead to ambiguity.

To ensure clarity, a set of definitions that are used throughout the report are enumerated, providing a common framework for understanding the ideas presented herein. European Commission documents served as a primary source. Another important source was Epoch AI, as it provides a valuable resource informing about AI models applied to biology that is utilised in various sections of this report.

Table 19: Main definitions relevant to this report.

Concept	Considered Definition	Source
Artificial Intelligence (AI)	Is a family of technologies that when in a system is defined as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments"	Article 3, Regulation (EU) 2024/1689 (AI Act) (2024)
AI model	Is a program that has been trained on a set of data to recognize certain patterns or make certain decisions without further human intervention. This model utilises algorithms, which are step-by-step procedures often expressed in mathematical terms, to process data and achieve specific tasks. The key distinction between algorithm and model lies in the fact that algorithms are the methods used to create the model, whereas the model itself is the result of applying these algorithms to a dataset. In essence, the model makes predictions or decisions after training based on the logic provided by the algorithm.	IBM[121]
General-purpose AI model (GPAI)	Is an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market.	Article 3, Regulation (EU) 2024/1689 (AI Act) (2024)
Foundation Model	Are models (e.g., BERT, DALL-E, GPT-3) trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks.	Center for Research on Foundation Models (Bommasani et al., 2022)

[121] https://www.ibm.com/think/topics/ai-model

Concept	Considered Definition	Source
Specialised model	Is a smaller model that is trained or adapted to excel in specific domains, such as medical diagnosis, legal research or AI contracting, providing expert support quickly and efficiently. Such specialised models and applications are commonly produced by distilling/finetuning their knowledge into more lightweight architectures suitable for targeted, efficient deployment, or by integrating tools like external knowledge bases.	Apply AI Strategy[122]
Biology	Is a division of the natural sciences concerned with the study of life and living organisms.	IATE[123]
Biological data	Is any type of data related to Biology, such as genetic sequences, medical images, or sensor readings from biological systems.	JRC's analysis
Biology-AI Models	Are models that are directly and explicitly trained on biological data, including at least one of the following categories: • Biological sequence data (e.g., DNA/RNA/protein sequences); • Biomolecular structure data (e.g., protein/RNA/complex structures); • Biological property labels for biomolecules (e.g., fitness, pathogenicity, binding, activity); • Cell-level data (e.g., cell types, gene expression levels, spatial omics, microscopy/imaging).	Epoch AI[124,125]
AI for Biology (AI4Bio)	AI technologies specifically developed for application in the field of biology or life sciences related sub-areas.	JRC's analysis

Source: JRC's analysis.

Terminology note. Throughout the report, the terms "biology–AI models", "biology AI models", "biology AI systems", "biology models", "biological AI models", and "biological models" are used interchangeably unless the context explicitly indicates otherwise

[122] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52025DC0723

[123] https://iate.europa.eu/entry/result/45385/en

[124] https://epoch.ai/data/biology-ai-models-documentation#inclusion

[125] Epoch AI's definitions are adopted here for practical purposes, as their taxonomy directly underpins the model selection criteria used in their dataset, which this report draws upon. It should be noted, however, that these definitions are not universally agreed upon and reflect Epoch AI's own classificatory choices

Annex 2. Epoch AI dataset descriptors

The main source of data for the analyses presented in this document is the Epoch AI's Biology AI Models dataset, or just the Epoch AI dataset, a resource released in early 2025 that consolidates information on representative biology AI models and their characteristics.

The Epoch AI dataset includes 383 biology AI models as of 5th January 2026. Some descriptors of the models are base model, batch size, task, epochs, finetune compute (FLOP), hardware quantity, hardware utilisation (MFU), model, training compute (FLOP), training compute estimation method, training dataset, training hardware, training time (hours), training power draw (W), possibly over 1e23 FLOP, model accessibility and training code accessibility. These descriptors enable a fine-grained analysis of the AI-for-biology landscape across tasks, data types, and methods (see **Table 20** for a detailed description).

Table 20. Descriptors of the Epoch AI dataset consulted on 07/11/2025.

Column	Type	Definition
Abstract	Text	Abstract text from the publication associated with the model.
Authors	Text	Comma-separated list of authors. Authors are named in the way that they report their names in their publications, if applicable. For example, Lê Viết Quốc is credited as "Quoc V. Le" in his publications.
Base model	Categorical (single select)	Which base model the model was fine-tuned from, if applicable.
Batch size	Numeric	Batch size used during training.
Citations	Numeric	Number of citations as of last update. Values are collected from Semantic Scholar where available, otherwise manually from Google Scholar.
Confidence	Categorical (single select)	Metadata describing our confidence in the recorded values for Training compute, Parameters, and Training dataset size. This describes confidence for the *most* uncertain of these parameters, where they have a non-empty entry (compute is typically the most uncertain). The confidence statuses are defined as followed: Confident: This is based on primary or reliable sources who are providing specific technical details that allow us to make this estimate using a methodology we are confident is correct. If a lab discloses in a technical paper that a model was trained on 30,000 H100s for 45 days, we would make a confident estimate. Deepseek v3 is in this category. Likely: Similar to confident in terms of methodology, but the inputs to the methodology are not as reliable. For instance, we might base an estimate of how many GPUs were used for a training run on a detailed leak, a less reliable source (i.e., what an engineer discusses on a podcast), the size of a datacenter we are tracking, or a vague but trustworthy source. Grok 3 is in this category, Speculative: Our best guess, based on methodologies or sources we think are informative but don't have confidence in. For instance, we might base an estimate on the compute used to train other models with similar

Column	Type	Definition
		capabilities. Speculative estimates may change substantially as more information comes to light. An example of a speculative estimate would be GPT-4o, where we used benchmark scores to estimate how much compute was required for training. We also provide further statuses: Unknown - we have too little information to even make a speculative estimate. Wrong - we know this estimate is incorrect, and it has been queued for correction Unverified - this estimate has not yet been assessed for confidence.
Country (of organisation)	Categorical (multiple select)	Country/countries associated with the developing organisation(s). Multinational is used to mark organisations associated with multiple countries.
Domain	Categorical (multiple select)	The machine learning domain(s) of application associated with the model. This is fairly high-level, for example "Language" incorporates many different ML tasks. Possible values: Biology, Image generation, Language, Materials science, Mathematics, Medicine, Multimodal, Vision.
Task	Categorical (multiple select)	The fine-grained task(s) that the model is designed to perform. These are specific applications of the model to different problems, and can span multiple domains. Task labels are assigned by following a flowchart[126]. Each applicable branch of the flowchart is followed until a leaf node is reached. If the task is already in the database, the model is tagged with that task. If the task does not yet exist in the database, the model is tagged with that task and the task is added to the flowchart. Examples: Face recognition Visual question answering Tic Tac Toe Weather forecasting
Epochs	Numeric	How many epochs (repetitions of the training dataset) was used to train the model.
Finetune compute (FLOP)	Numeric	Compute used to fine-tune the model, if applicable.
Hardware quantity	Numeric	Indicates the quantity of the hardware used in training, i.e. the number of chips.

[126] https://www.mermaidchart.com/raw/2e755f16-5525-46cc-8d03-6c091e18a3ae?theme=light&version=v0.1&format=svg

Column	Type	Definition
Hardware utilization (MFU)	Numeric	Number between 0.00 and 1.00 indicating the hardware utilization ratio, i.e. utilized FLOPs / theoretical maximum FLOPs. This value reflects utilization based on computations successfully applied to model training, and does not include computations performed by the hardware which do not ultimately affect the model.
Link	URL	Link(s) to best-choice sources documenting a model. This should preferentially be a journal or conference paper, preprint, or technical report. If these are not available, the links should point to other supporting evidence, such as an announcement post, a news article, or similar.
Model	Text	The name of the model. This should be unique within the database, and should be the best-known name for a given model. This column must be filled in, and is used as the primary key for indexing entries in the dataset.
Notability criteria	Categorical (multiple select)	The criteria met by the model which qualify it for notability. To be notable, a model must meet at least one criterion. Possible values are highly cited, large training cost, significant use, state of the art, or historical significance. These are discussed further in Inclusion[127].
Organization	Categorical (multiple select)	Organization(s) who created the model. Organizations may have multiple different names, but we aim to standardize organization names where they refer to the same organization. Therefore, organizations are periodically reviewed in Airtable and standardized to the most common name for them. For example, “University of California, Berkeley” and “Berkeley” have been changed to “UC Berkeley”. Note that some organizations have similar names but genuinely are different organizations, for example Google Brain versus Google versus Google DeepMind.
Organization categorization	Categorical (multiple select)	Categorization of the organization(s), automatically populated from the Organization entry. Models are categorized as “Industry” if their authors are affiliated with private sector organizations, “Academia” if the authors are affiliated with universities or academic institutions, or “Industry - Academia Collaboration” when at least 30% of the authors are from each. Possible values: Industry, Research Collective, Academia, Industry - Academia Collaboration (Industry leaning), Industry - Academia Collaboration (Academia leaning), Non-profit
Parameters	Numeric	Number of learnable parameters in the model. For neural networks, these are the weights and biases. Further information is provided in Estimation[128].

[127] https://epoch.ai/data/biology-ai-models-documentation#inclusion

[128] https://epoch.ai/data/biology-ai-models-documentation#estimation

Column	Type	Definition
Publication date	Date	The publication, announcement, or release date of the model, in YYYY-MM-DD format. If the year and month are known but the day is unknown, the day is filled in as YYYY-MM-15. If the year is known but the month and day are unknown, the month and day are filled in as YYYY-07-01.
Reference	Text	The literature reference for the model, such as the title of the journal or conference paper, academic preprint, or technical report.
Training compute (FLOP)	Numeric	Quantity of compute used to train the model, in FLOP. This is the total training compute for a given model, i.e. pretrain + finetune. It should be filled in here when directly reported, or calculated via GPU-hours or backpropagation gradient updates. Further guidance is provided in Estimation[124].
Training compute estimation method	Categorical (multiple select)	Indicates how the quantity of training compute was found or estimated. Options include: "Reported", when the developers report how much training compute was used. "Operation counting", when the parameters, training data, and/or architecture are known, and these are used to estimate the number of operations performed while training. "Hardware", when the training time and/or hardware type are known, allowing an estimate based on the rate of computation. "Cost", when the training cost is known and the compute is estimated through the hardware usage budget "Benchmarks", when the model's training compute was estimated using its benchmark performance.
Training dataset	Categorical (multiple select)	Standard datasets are often used, and can be selected as multiple choice options. If a custom, unreleased dataset is used, it is set as "Unspecified Unreleased". Where this entry is empty, it has not yet been entered for a given model.
Training hardware	Categorical (multiple select)	Type of training hardware used. Entries are cross-referenced against Epoch AI's database of ML training hardware.
Training time (hours)	Numeric	Training time of the model, if reported. This refers to the time elapsed during the training process, *not* the number of chip-hours. For example, if a model were trained with 10 GPUs for 1 hour, the training time would be 1 hour. Includes the duration of all training phases conducted to develop the model, such as pre-training, post-training, RL, SFT, etc. If the model is fine-tuned from a previously published model, then that base model's training time is *not* included.

Column	Type	Definition
Training power draw (W)	Numeric	Power draw of the hardware used to train the model, in watts. Calculated as hardware quantity times processor TDP times datacenter PUE times server overhead. More details are provided in Estimating power draw[129].
Frontier model	Boolean	Indicates whether a model was within the frontier, defined as models that were in the top 10 by training compute as of their release date.
Possibly over 1e23 FLOP	Boolean	Indicates whether a model was (or may have been) trained with at least 10^{23} floating-point operations, which qualifies it for inclusion in the large-scale models dataset.
Model accessibility	Categorical (multiple select)	The accessibility of the model in terms of whether the model weights can be downloaded or, if the model weights are not accessible, whether the model can be used in an API or product.
Training code accessibility	Categorical (single select)	Denotes how the model can be accessed and used by the public. "Open weights (unrestricted)", "Open weights (restricted use)" and "Open weights (non-commercial)" all mean that the model weights are downloadable by the public, but with different restrictions on use. "API access" means the model can only be interacted with via an application programming interface, and possibly also a hosted service. "Hosted access (no API)" means the model can only be interacted with via a hosted service. "Unreleased" means there is no way for the public to access the model.
Notes fields, e.g. "Training compute notes"	Text	Metadata documenting the reasoning and/or evidence for a given column, e.g. training compute or dataset size. This is particularly important to note in cases where such information isn't obvious. This field is unstructured text.

Source: Epoch AI.

Annex 3. Literature sources

In addition to the Epoch AI dataset, several external literature reviews (see **Table 21**) were consulted to complement and contextualise the findings. These sources were chosen based on relevance, scientific impact, and alignment with the report's focus, ensuring a comprehensive understanding of both foundational work and the latest developments within the AI4Bio domain.

For each selected literature source, structured tables were collected and automatically homogenised to ensure consistency across sources. This process enabled the extraction of information for 130 additional models, of which 97 are not included in the Epoch AI dataset. The resulting consolidated tables categorise models according to key characteristics, including biological input type, task type, model architecture, and input modality class. Together, these harmonised literature-derived annotations provide an enriched and more comprehensive view of the AI4Bio landscape and serve as a complementary basis for the analyses presented in the following sections.

[129] https://epoch.ai/data/ai-models-documentation#estimating-power-draw

When interpreting the resulting synthesis, it is important to recognise potential sources of uncertainty arising from heterogeneous reporting practices, publication bias, differing benchmark definitions and incomplete disclosure of industrial systems.

Table 21. Summary of selected literature sources on AI models in biology and their rationale for inclusion.

Title	Reference	Justification for Selection
Progress and Opportunities of Foundation Models in Bioinformatics	(Li et al., 2024a)	Comprehensive review that outlines the development and application of foundation models in bioinformatics. It discusses how models are adapted to biological sequence analysis (e.g., DNA, RNA, proteins), structure prediction, and function annotation — and provides comparisons of architectures and challenges like data noise and interpretability.
Foundation Models in Bioinformatics	(Guo et al., 2025)	Detailed review of foundation models in bioinformatics, categorising model types (language, vision, graph, multimodal) and surveying their use across genomics, transcriptomics, proteomics, drug discovery, and single-cell analysis. Emphasises both biological data modalities and computational frameworks for selecting models.
Mapping the Landscape of AI-Enabled Biological Design	(Hie, 2025)	Overview of AI models for biological design - including generative, predictive, and foundation models - with examples in protein engineering and modality integration (e.g., genomic to phenotypic). Discusses datasets and future potential, offering context for how models map biological distributions and generate outputs.
Foundation Models for AI-Enabled Biological Design	(Moldwin and Shehu, 2025)	Preprint survey that taxonomises contemporary foundation models used in biological design covering protein engineering, small molecule design, and genomic sequence design. Highlights challenges such as controllability and multimodal integration and addresses architectural considerations for biological sequence modelling.
Single-cell foundation models: bringing artificial intelligence into cell biology	(Baek, Song, and Lee, 2025)	Detailed review of single cell foundation models, with description of model architectures, training paradigms and current application areas. Includes description of popular models and provides an outlook for future directions of downstream applications of single cell foundation models.

Source: JRC's analysis.

Annex 4. Downstream Task Coverage Across Biology AI Models

This section presents an overview of model capabilities using the *Task* field as defined in the original Epoch AI dataset (see **Table 20** for further details). The analysis relies exclusively on the task labels assigned by Epoch AI, which were intended to describe the primary objective or application area of each model. The results are presented in **Figure 52**.

Tasks labels across 383 models in the dataset highlights a strong concentration of biological AI development around protein-related applications. Tasks such as "*protein or nucleotide language*

models (pLM/nLM)" (n=95) dominate the dataset followed by "*proteins*" (n=62), "*protein folding prediction"* (n=49), "*protein design"* (n=45), and "*protein generation"* (n=41), indicating that much of the recent modelling effort has focused on learning representations of protein sequences and structures. This concentration reflects both the central role of proteins in molecular biology and the suitability of data-driven approaches for sequence- and structure-based problems.

The "*drug discovery*" task also appears prominently (n=42), reflecting the growing use of AI methods for the drug research & development (R&D) pipeline, from antibiotic discovery (Stokes et al., 2020) to virtual screening and lead optimisation (Jia et al., 2026). These models often build on protein representations while extending their application to small-molecule design, target identification, and prediction of drug–target interactions, underscoring the versatility of protein-centric AI approaches in end-to-end therapeutic development.

Beyond protein-centric tasks, a smaller but diverse set of applications can be observed, including RNA-related predictions, molecular property prediction, and image-based tasks such as cell segmentation and cryo-electron microscopy reconstruction. These categories appear with substantially lower frequencies, suggesting either a more limited adoption of AI methods in these areas or differences in how such models are captured and labelled in the dataset. The long tail of infrequently occurring task labels points to a wide but unevenly represented range of biological applications.

Figure 52. Number of biological AI models associated with each task label as reported in the Epoch AI dataset (n=383). Colour indicates the primary biological molecule or domain addressed by each task. Task assignment follows Epoch AI's original labelling; molecule-type classification is based on manual assessment.

Source: JRC's analysis.

The dominance of protein-related tasks is expected given the strength of its experimental and data ecosystem (see Section 4.1). This higher number of protein AI models illustrates how data availability and quality drive AI development. Protein models currently benefit from well-established experimental training data sources and data-sharing frameworks, whereas other data types, such as DNA, RNA, or cellular datasets, pose different technical challenges. Recognising these differences can guide the targeted allocation of resources, ensuring that areas with higher complexity receive appropriate support to advance their integration into AI-driven biological R&D.

While the task annotations provided in the original Epoch AI dataset are informative and offer a useful high-level overview of where modelling activity is concentrated, they also present limitations for more detailed interpretation. The *Task* field was designed to capture the primary objective or application area of each model, yet in practice the labels vary in specificity and consistency, ranging from clearly defined biological objectives (e.g. *protein folding* or *RNA structure prediction*) to broader or more ambiguous categories that may conflate task, modality, or model type. As a result, some annotations provide limited insight into the actual functional role of a model, and similar models may be labelled in slightly different ways.

Annex 5. Training Datasets and Databases

Within the Epoch AI catalogue, information on training data is reported for 134 of the 383 biological AI systems included, and the 75 distinct datasets and databases identified correspond only to this subset of models and to the data that are explicitly documented. Each of the 75 datasets/databases were manually annotated with key information such as size, year of first release, most recent update, biological scope and access conditions. This process involved harmonising naming conventions, removing duplicates and clarifying cases where repositories contain overlapping or closely related sub-collections. The results are provided in **Table 22**.

It is important to note the limitations of the available metadata. Neither the 134 models nor the 75 datasets should be interpreted as exhaustive of the AI4Bio landscape. Despite these limitations, this curated set provides a meaningful overview of the current AI4Bio training data landscape and enables an analysis of the main biological domains, data modalities, dataset scale over time, and geographic and access patterns.

Table 22. Manually curated Table with training datasets and databases reported in the Epoch AI dataset.

Training dataset	Derived from	Year founded	First Release	Last Update	Size	Availability	Year of size calc.	Country / Host	Data category	Data type	Type
AlphaFold database (AFDB)	-	2021	2022	2025	25.2TB	Public	2024	EU	Protein	Predicted protein structures	Database
1000 Genomes Project	-	2008	2019	2023	200TB	Public	2015	UK, China, US	DNA	Human genetic variation; population-scale sequencing data	Database
ATLAS	PDB	2022	2022	2023	100MB	Public	2017	EU	Protein	Protein molecular dynamics simulations	Database
BACE	PDB	2016	2016	2016	100MB		2016	US	Protein	Drug discovery	Dataset

BFD (Big Fantastic Dataset)	UniprotKb+Metaclust+SRC	2021	2021	2021	292GB	Public	2021	US, EU	Protein	Sequence profile	Database
BFD30	BFD	2021	2021	2021	21GB	Public	2023	US, EU	Protein	Sequence profile	Dataset
BindingDB	-	2000	2000	2025	1.42GB	Public	2025	US	Protein-Ligand	Protein-ligand binding affinities	Database
BRD (Basecamp Research Data)	-	2000	2024	2024	NA	Available upon request	2025	US	Protein	Protein sequences	Dataset
BRENDA	-	1987	2000	2025	77MB	Public	2025	EU	Protein	Enzyme information	Database
BV-BRC	-	2022	2022	2025	10TB	Public	2024	US	DNA	Bacterial and viral genomics	Database
Cambridge Structural Dataset	-	NA	2016	2025	NA	License	NA	UK	Chemical	Small molecule crystal structures	Database
CATH	PDB	1995	1997	2025	1GB	Public	2025	UK	Protein	Protein domain classification	Database
ChEMBL	-	2009	2009	2025	35GB	Public	2024	EU	Chemical	Bioactive molecules with drug-like properties	Database
ConSurf10k	ConSurf-DB	2021	2021	2021	27MB	Public	2021	EU	Protein	Conservation analysis	Dataset

CrossDocked2020	PDB	2020	2020	2020	50GB	Public	2020	US	Protein	Protein-ligand docking poses	Dataset
DB5.5	PDB	2015	2015	2019	100MB	Public	2019	US	Protein	Protein-protein docking benchmark	Dataset
DeepLoc	UniprotKb	2017	2017	2024	20MB	Public	2024	EU	Protein	Subcellular localization	Dataset
DIPS	PDB	2018	2018	2021	28GB	Public	2021	US	Protein	Protein-protein interactions	Dataset
DMS39	-	2022	2022	2022	16MB	Public	2022	EU	Protein	Deep mutational scanning	Dataset
DMS4	-	2022	2022	2022	1BM	Public	2022	EU	Protein	Deep mutational scanning	Dataset
Eff10k	SNAP2	2022	2022	2022	NA	Not available	NA	EU	Protein	Deep mutational scanning	Dataset
ESM3 Dataset	UniRef, MGnify90	2022	2022	2023	15TB	Public	2023	US	Protein	Protein sequences for ESM3 training	Dataset
European Nucleotide Archive (ENA)	-	1980	1980	2025	63PB	Public	2024	EU	DNA/RNA	Nucleotide sequences	Database
Galactica Corpus	-	2022	2022	2022	106 billion tokens	Not available	2022	UK	Text	Scientific literature	Dataset

GuacaMol	CHEMBL	2024	2024	2024	45MB	Public	2024	UK,Switzerland	Chemical	Molecular generation benchmark	Dataset
Human Protein Atlas (HPA)	-	2003	2005	2025	300TB	Public	2025	EU	Protein	Protein expression and localization	Database
Human Reference Genome (GRCh38/hg38)	-	2000	2013	2024	3GB	Public	2024	US	DNA	Human reference genome	Database
Immune Epitope Database (IEDB)	-	2004	2004	2024	556MB	Public	2024	US	Protein	Immune epitopes	Database
JASPAR 2024	-	2004	2004	2024	5MB	Public	2024	Norway, Canada	DNA	Transcription factor binding sites	Database
Ligand Discovery at CeMM	-	2024	2024	2024	20MB	Public	2024	EU	Chemical	Chemical probes and screening	Dataset
mdCATH	CATH	2024	2024	2024	3.3TB	Public	2024	EU	Protein	Molecular dynamics	Dataset
Metaclust	UniProt, JGI,NCBI,OM-RGC	2017	2017	2021	162GB	Public	2021	EU	Protein	Protein sequence clusters	Dataset

MGnify	ENA,NCBI	2017	2017	2025	74GB	Public	2025	UK,EU	Protein	Metagenomic sequences	Database
MGnify	ENA,NCBI	2018	2017	2025	173GB	Public	2025	UK,EU	DNA	Metagenomic sequences	Database
MIDAS	UHGG,GTDB	2017	2016	2023	539GB	Public	2023	US	Protein	Metagenomic sequences	Database
MISATO	PDB	2023	2023	2023	193GB	Public	2023	EU	Protein	Molecular dynamics trajectories	Dataset
Mol instructions	PubChem, USPTO, UniProtKB, BC4CHEMD, ChemProt, BC5CDR, MoleculeNet, MMLU, PubMedQA, MedMCQA, Agency for Toxic Substances and Disease Registry (ATSDR), FDA Pharm Classes, LiverTox, Drug Database, Clinicalinfo.hiv.gov, Drug Bank, ChEBI, Yeast Metabolome Database (YMDB), LOTUS - the natural	2023	2024	2024	261GB	Public	2024	China	Chemical	Instruction tuning for molecules	Dataset

	products occurrence database, GlyCosmos Glycoscience Portal, CAMEO Chemicals, E. coli Metabolome Database (ECMDB)										
NeoDTI	-	2018	2018	2018	23MB	Public	2018	China,US	Drug-Protein	Drug-target interactions	Dataset
Observed Antibody Space (OAS) database	SRA	2018	2018	2025	1.1TB	Public	2024	UK	Protein	Antibody sequences	Database
OpenCell	-	2021	2021	2023	20TB	Public	2023	US,EU	Image	Protein localization	Database
OpenGenome	NCBI,ENA	2024	2024	2024	147GB	Public	2024	US	DNA	Genomic sequences	Dataset
OpenGenome 2	NCBI,ENA	2025	2025	2025	185GB	Public	2025	US	DNA	Genomic sequences	Dataset
PDB (Protein Data Bank)		1971	1971	2025	270GB	Public	2024	US, EU, Japan	Protein	Experimentally determined protein structures	Database

PDB25	PDB	-	-	2025	NA	Public	2025	US, EU, Japan	Protein	Experimentally determined protein structures	Dataset
PDBbind	PDB	2004	2004	2024	7.6GB	Public	2024	China	Protein	Protein-ligand binding affinities	Database
Pfam	Uniprot, PDB	1997	1997	2024	130GB	Public	2025	EU, US	Protein	Protein families	Database
PMD4k	PMD	1999	2022	2022	NA	No longer available	NA	Japan	Protein	Deep mutational scanning	Dataset
ProteinGym	-	2023	2023	2025	11GB	Public	2025	UK,US	Protein	Deep mutational scanning benchmark	Dataset
ProteinKG25	Swiss-Prot, Gene Ontology	2022	2023	2023	147MB	Public	2023	China	Protein	Knowledge graph	Dataset
PSICOV150	PDB	2014	2014	2014	5MB	Public	2014	UK	Protein	Contact prediction benchmark	Dataset
PubChem	-	2004	2004	2025	150GB	Public	2025	US	Chemical	Small molecules and bioactivity	Database
RBPSuite	POSTAR3	2020	2020	2025	722MB	Public	2025	China	Protein	RNA-binding proteins	
RefSeq	GenBank	1989	1989	2025	2TB	Public	2025	US	DNA/RNA/Protein	Reference sequences	Database
Rfam 14	-	2003	2003	2025	772GB	Public	2025	EU	RNA	RNA families	Database

RNAcentral	Rfam	2014	2014	2025	43GB	Public	2025	EU	RNA	Non-coding RNA	Database
R-SIM	-	2022	2022	2025	NA	Available upon request	NA	India	RNA	RNA-small molecule affinities	Dataset
SARS-CoV-2 genome dataset	BV-BRC	2020	2022	2023	155GB	Public	2023	Multiple	DNA	Viral genomes	Dataset
SCOPe 2.05	SCOP	2014	2015	2016	836MB	Public	2021	US	Protein	Protein structural relationships	Database
SCOPe 2.07	SCOP	2014	2018	2021	1.5GB	Public	2021	US	Protein	Protein structural relationships	Database
SIDER	-	2010	2010	2015	200kB	Public	2015	EU	Chemical	Drug side effects	Database
SwissProt	-	1986	1986	2025	89MB	Public	2025	Switzerland, EU, US	Protein	Manually curated protein sequences	Database
Therapeutic Data Commons (TDC)	Multiple	2021	2021	2025	15.9 million data points	Public	2021	US	Multiple	Therapeutic data	Database
trRosetta	PDB	2020	2020	2021	29GB	Public	2021	US,China	Protein	Protein structure prediction training	Dataset
Uni-Mol Molecular	PDB, ZINC, CHEMBL	2022	2022	2022	115GB	Public	2022	China	Chemical	Molecular conformations	Dataset

pre-train dataset											
UniParc	UniProtKB	2002	2002	2025	12TB	Public	2025	Switzerland, EU, US	Protein	Protein sequence archive	Database
Uniprot	-	2002	2002	2025	48GB	Public	2025	Switzerland, EU, US	Protein	Protein sequences and annotations	
UniProtKB	-	2002	2002	2025	48GB	Public	2025	Switzerland, EU, US	Protein	Protein sequences and annotations	Database
UniProtKB/ Swiss-Prot	-	1986	1986	2025	89MB	Public	2025	Switzerland, EU, US	Protein	Manually curated protein sequences	Database
UniRef100	UniProtKB	2007	2007	2025	116GB	Public	2025	Switzerland, EU, US	Protein	100% identity clusters	Database
UniRef30 (FKA UniClust30)	UniProtKB	2016	2016	2023	66GB	Public	2023	Switzerland, EU, US	Protein	30% identity clusters	Database
UniRef50	UniProtKB	2007	2007	2025	12GB	Public	2025	Switzerland, EU, US	Protein	50% identity clusters	Database
UniRef90	UniProtKB	2007	2007	2025	65GB	Public	2025	Switzerland, EU, US	Protein	90% identity clusters	Database
Uniref9B	UniRef	2024	2024	2024	11.4MB	Public	2024	Switzerland, EU, US	Protein	Protein sequences	Dataset
VDJdb	-	2017	2017	2025	22MB	Public	2025	EU,UK,Russia, Australia	Protein	T-cell receptor sequences	Database

ZINC 15	-	2005	2005	2015	8.9GB	Public	2022	US	Chemical	Purchasable “drug-like” compounds	Database

Annex 6. Infrastructure analysis

The analysis draws on the Epoch AI dataset, applying a filter to exclude models developed prior to 2017. This results in a final sample of 368 models. In addition, a separate Epoch AI dataset containing metadata on more than 500 GPU clusters[130] is used to characterise large-scale hardware facilities supporting AI model training and inference.

The analysis is primarily descriptive. Summary statistics, including the mean, standard deviation, and maximum, are reported for key infrastructure-related metrics.

Additionally, the analysis characterises the technological requirements frontier, operationally defined as the Top 5 models per year with highest resource consumption. This approach identifies the leading edge of the field's resource-intensive trajectory, focusing on experiments that define the limit of what is technically and logistically feasible. For this, the dataset is cleaned by removing records with missing publication dates or missing metric values. Observations from the year 2025 were excluded to avoid partial-year bias. Due to the presence of extreme values, a logarithmic transformation was applied to the y axis. The trajectory of the resource frontier was modelled using a Linear Regression which was fitted exclusively to the frontier data points (Top 5 per year), with the ordinal date of publication serving as the independent variable. Yearly growth estimation is provided by calculating the ratio between the line extremities and normalising it over the time delta. The resulting value (e.g., 2.5x/yr) represents the factor by which the technological frontier expands every year.

The frontier of infrastructure requirements is examined across several variables, including training dataset size, hardware quantity, training compute (FLOPs), training time, and training power draw (W) (see **Table 20**). For each variable, the analysis identifies the model with the highest observed resource allocation, serving as a reference point for the upper bound of resource consumption currently observed in the field.

Descriptive statistics are also used to assess the characteristics and geographic distribution of GPU clusters in Europe, providing additional context on regional computational capacity. Finally, correlation analyses are conducted to explore relationships among infrastructure-related variables, with particular attention to identifying factors associated with energy consumption during model training.

Annex 7. Model-Dataset Update Lag

Update lag is defined here as the time elapsed between a dataset's most recent update and the publication year of the model that used it. The analysis covers 134 model-dataset pairings from the literature-curated dataset and 75 from the Epoch AI dataset. As discussed in Section 5.1, the large majority of pairings show a zero-year lag, indicating that most models were trained on data that was current at the time of publication. **Table 23** lists the subset of cases where a lag of two or more years was recorded.

Datasets such as PSICOV150, SIDER, BACE, and CrossDocked2020 were used with update gaps ranging from three to five years, but these are compact, task-focused resources designed primarily

[130] https://epoch.ai/data/gpu-clusters

for benchmarking or protein–ligand interaction evaluation. Their use in models such as Hopfield Networks (Ramsauer et al., 2021), DMPFold (Greener, Kandathil, and Jones, 2019), BindDM (Huang et al., 2024b), and GraphMS (Cheng et al., 2021) reflects targeted fine-tuning or evaluation choices rather than dependence on outdated biological baselines.

The two longest lags in the table, ZINC 15 used by MegaMolBART (6 years) and MoLFormer-XL (8 years), and SCOPe 2.05 used by AtomFlow (8 years), warrant brief comment. ZINC 15 is a large compound library not subject to the same continuous curation cycle as biological sequence databases; its chemical content does not become biologically outdated in the same way as, for example, a genomic or structural repository, and both models use it primarily as a molecular pre-training corpus. SCOPe 2.05 is a structural classification database whose taxonomy is deliberately stable across versions; AtomFlow's use of an older version reflects a choice to train on a fixed classification scheme rather than an oversight. Neither case represents a substantive data currency concern for the conclusions drawn in the main text.

Table 23. Update lag between models and datasets release dates.

Model	Publication date	Dataset-to-model update lag (years)
ATLAS	Boltz-2	2
BACE	Hopfield Networks (2020)	4
CrossDocked2020	BindDM	4
DB5.5	EquiDock	2
DIPS	DiffDock-PP	2
ESM3 Dataset	Odyssey 102B	2
MIDAS	Boltz-2	2
MISATO	Boltz-2	2
NeoDTI	GraphMS	3
SCOPe 2.05	DeepConPred2	2
SCOPe 2.05	AtomFlow	8
SCOPe 2.07	Genie-SCOPe (bio)	2
SIDER	Hopfield Networks (2020)	5
trRosetta	AlphaLink	2
ZINC 15	MegaMolBART	6
ZINC 15	MoLFormer-XL	8

Source: JRC's analysis based on the Epoch AI dataset.

Annex 8. Country and Organisation Type Counting Method

For country-level analysis, this report applies *full counting*, meaning that each country associated with a model is counted once, regardless of how many other countries are also involved. This choice reflects the aim of capturing *participation* rather than estimating *relative contribution*.

The Epoch AI dataset does not provide explicit links between authors, organisations, and their countries, which makes more refined approaches, such as measures of relative collaboration intensity (Fuchs, Sivertsen, and Rousseau, 2021) or fractional counting methods that attribute proportional credit across contributors (Cao et al., 2025b), not feasible. Full counting also enhances the interpretability. However, it has clear limitations: it may overrepresent countries frequently involved in multinational collaborations and does not adjust for unequal levels of contribution within a model.

For organisation type analysis, the same full counting logic applies to the distribution of model counts across categories: each model is counted once per organisation type present in its entry, after deduplication. A model developed by two academic institutions is therefore counted once under academia; a model developed jointly by an academic institution and an industry partner is counted once under academia and once under industry. This approach captures the participation of each organisational category without inflating counts from repeated affiliations within the same category.

Where the analysis moves from descriptive counts to comparisons of behaviour across organisation types, models with mixed affiliations spanning academia and industry are assigned to a separate *Mixed* category rather than counted in both constituent categories. This prevents double-counting from distorting comparisons and allows the Mixed group to be examined as a distinct organisational form.

Annex 9. Funding data retrieval

Funding data were extracted for the 383 Epoch AI models via DOI lookup and keyword matching against major public funding databases. Of these, 64 models had at least one funding source recorded, and quantitative amounts were available for 28: 19 matched to US National Institutes of Health (NIH) grants, 9 to US National Science Foundation (NSF) grants, and 11 to projects recorded in the EU's Community Research and Development Information Service (CORDIS).

NIH and NSF amounts were retrieved in USD and converted to EUR at an approximate 2019 to 2025 average rate of 0.90 EUR/USD; CORDIS amounts were used directly in EUR. The conversion rate of 0.90 EUR/USD represents the arithmetic mean of annual average exchange rates from 2019 to 2024. Annual averages were sourced from the U.S. Internal Revenue Service Yearly Average Currency Exchange Rates (2021-2024: 0.846 EUR/USD, 0.951 EUR/USD, 0.924 EUR/USD, 0.924 EUR/USD respectively) and Exchange-Rates.org (2019: 0.893 EUR/USD; 2020: 0.877 EUR/USD). Sources: Internal Revenue Service, "Yearly Average Currency Exchange Rates," accessed March 11, 2026, https://www.irs.gov/individuals/international-taxpayers/yearly-average-currency-exchange-rates; Exchange-Rates.org, "USD to EUR Exchange Rate History," accessed March 11, 2026, https://www.exchange-rates.org/exchange-rate-history/usd-eur

These results should be interpreted as indicative rather than exhaustive: only ~17% of models carry any funder information, and quantitative amounts are available for a smaller subset still. Matching was performed via DOI lookup and keyword search against funding databases; models without DOIs or with ambiguous grant identifiers could not be matched. Furthermore, NIH and NSF figures reflect individual researcher awards, whereas EU CORDIS entries correspond to total project budgets for European funding programmes and may span multiple institutions and work packages. These two types of amounts are therefore not directly comparable.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/write-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

Scan the QR code to visit:

Joint Research Centre
https://joint-research-centre.ec.europa.eu

Publications Office
of the European Union