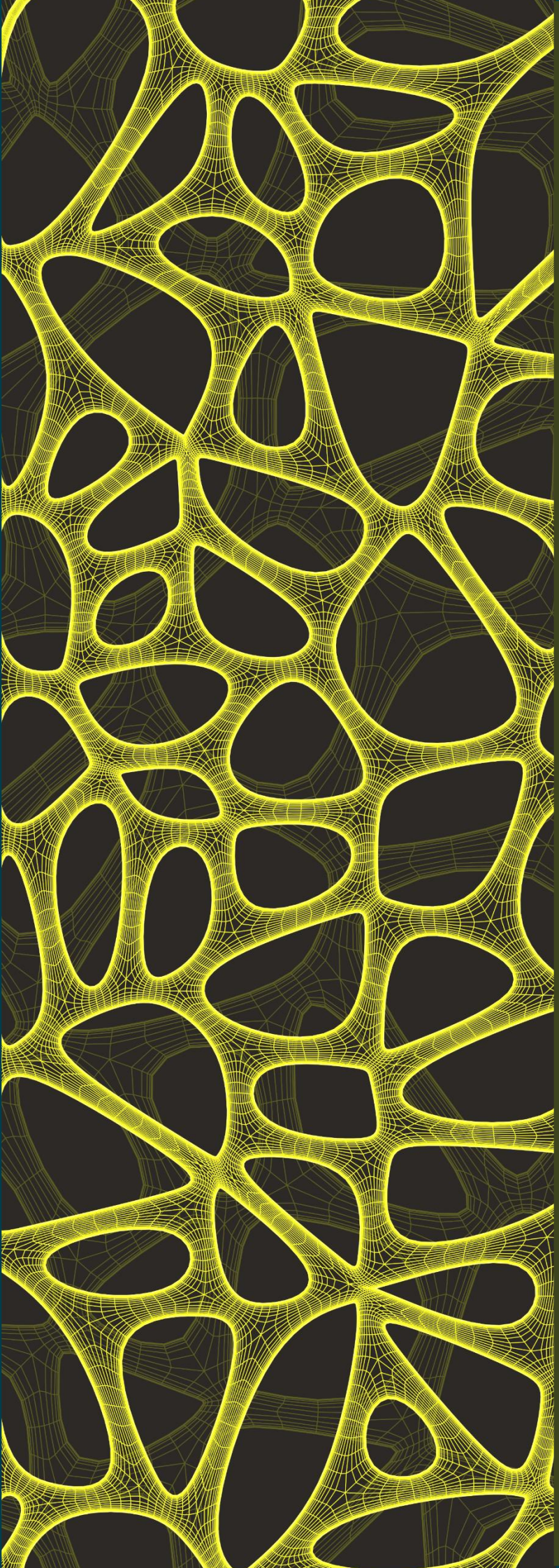


frontier economics

OPEN DATA: THE VALUE AND IMPACT OF EMBL-EBI DATA RESOURCES

A report prepared for EMBL-EBI

2026

EMBL-EBI

Contents

Abstract

This report evaluates the value and socio-economic impact of EMBL-EBI's open biological data resources in 2025 (Frontier Economics, 2026). It updates and extends earlier economic assessments undertaken by Charles Beagrie Ltd in 2015 (published in 2016)[1] and 2020 (published in 2021)[2] using a comparable survey-based modelling framework to enable like-for-like comparisons over the course of a decade. This approach has been improved through the revision of survey questions, development of an updated approach to estimating returns on enabled research and development, and the addition of three case study assessments of downstream impacts.

The impact modelling identifies four headline value indicators. First, users' aggregate willingness to pay for continued access to EMBL-EBI resources is estimated at £1.7bn per year. Second, reported efficiency gains average 11 hours per week per user, equivalent to £11.8bn per year in productivity benefits. Third, the value of the time users spend working with EMBL-EBI resources is estimated at £5.8bn per year. Fourth, EMBL-EBI resources are estimated to enable around £6.3bn per year in returns on research and development, implying a net present value of £14bn to £44bn over a 20 to 30 year horizon. **Across these indicators, benefit-cost ratios are high, with benefits substantially exceeding the costs of maintaining and providing the resources under conservative assumptions.**

These impacts are measured by estimating the value of EMBL-EBI data resources and tools for direct users, who access them directly via EMBL-EBI. It excludes 'secondary' use, whereby users rely on other data resources and tools that are built partially or wholly on data provided by EMBL-EBI. Our estimates therefore represent a lower bound of reach and value.

In 2025, EMBL-EBI's data resources were estimated to serve around 520,000 unique direct users. This represents an increase of 163% from the estimated user base of 198,000 at the time of the 2016 economic assessment. A global user survey with over 2,500 respondents indicates that EMBL-EBI resources are integral to life sciences work: 71% of respondents report their work would be impossible or would require significant additional time and effort without access.

Three case studies illustrate how EMBL-EBI open data resources generate downstream impacts beyond the direct user community through 'secondary' use. These include enabling greater usage of AI-based protein structure predictions across a wider range of scientific fields, supporting the delivery of genomic medicine in healthcare systems, and enabling public and private sector innovation through the reuse of bioimaging and structural biology resources. Crucially, the evidence from the case studies shows that EMBL-EBI's value does not lie solely in the data themselves, but also in the curation, standardisation, integration and open access model. With interviewees highlighting EMBL-EBI's role as a central body that develops

[1] See Beagrie (2016), "The Value and Impact of the European Bioinformatics Institute."

[2] See Beagrie (2021), "Data-driven discovery: The value and impact of EMBL-EBI managed data resources."

standards and coordinates with the open science community, this suggests that the value generated by EMBL-EBI data resources is worth more than the sum of their parts.

Overall, the evidence strongly supports the conclusion that investment in EMBL-EBI data resources represents excellent value for money and plays a critical role in sustaining and advancing the global life sciences research ecosystem.

Executive Summary

Overview of EMBL-EBI

The European Molecular Biology Laboratory (EMBL) is Europe's only intergovernmental life sciences laboratory. Founded in 1974, it now operates six sites across Europe: Barcelona, Grenoble, Hamburg, Heidelberg, Hinxton, and Rome. EMBL's European Bioinformatics Institute (EMBL-EBI), established in 1994 in Hinxton near Cambridge in the UK, is a global leader in the storage, analysis and dissemination of high-quality biological datasets.

In support of its mission "*To freely provide data and bioinformatics services to the scientific community in ways that promote scientific progress*", EMBL-EBI hosts a globally accessible set of around 30 open biological data resources.[3] These support research across genetics, genomics, proteins, bioimaging, chemistry and more, enabling the publication of over 120,000 scientific publications in 2025.[4] One of EMBL-EBI's core principles is to deliver its open data solutions based on a collaborative approach. In particular, most of EMBL-EBI's major data resources are developed through partnerships with organisations and communities from around the world. For clarity and brevity, this report refers to them as "EMBL-EBI data resources"; however, it is important to recognise that the vast majority are the product of collaborative, community-led efforts.

EMBL is supported by over 30 member states, associate member states, and prospect member states.[5] Collectively, they therefore form the largest stream of EMBL-EBI funding. Major funders include UK Research and Innovation, as well as the European Commission, the US National Institutes of Health, and Wellcome. Sustained capital investment from the UK Government has helped EMBL-EBI develop a robust technical infrastructure to efficiently manage the growth and diversity of biological data held in its public data resources. Investment by funders and collaborator contributions enable EMBL-EBI to maintain its comprehensive suite of open data resources for the life sciences.

Purpose of this study

EMBL-EBI has commissioned Frontier Economics to evaluate the value and socio-economic impacts arising from investment in EMBL-EBI's mission to provide freely available data and bioinformatics services.[6] We (Frontier Economics) examine how, and to what extent, these data resources improve research productivity, enable research and development, and

[3] See "EMBL-EBI's Highlights 2025 Report", page 7.

[4] See "EMBL-EBI's Highlights 2025 Report", pages 4-6.

[5] The EMBL member states are Austria, Belgium, Croatia, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, the Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland, and the United Kingdom. Australia is an associate member state, and Serbia and Bulgaria are prospect members.

[6] The assessment does not cover the remaining four Missions, which are outside the scope of this study.

generate wider socio-economic impacts. By quantifying their value and impact, we also aim to strengthen the evidence base for the broader value of open data for the life sciences research community.

Evaluating the value and socio-economic impacts of EMBL-EBI data resources is inherently complex. Because these resources are freely accessible, with no charge to users, their value cannot be observed through market prices and must instead be inferred indirectly. The breadth of resources provided, each serving different user communities, further reinforces the need to identify and articulate the distinct value generated by individual data resources. This is particularly important given that EMBL-EBI data resources are used both by direct users, who access them via EMBL-EBI, and through 'secondary use', whereby users rely on other data resources and tools that are built partially or wholly on data provided by EMBL-EBI. A similar complexity applies to costs, where many EMBL-EBI data resources and tools are the product of longstanding, collaborative, community-led contributions.

Recognising this, we take a conservative approach to valuation. We develop a detailed logic model setting out how the inputs, activities and outputs EMBL-EBI provides are hypothesised to lead to short-term outcomes, long-term outcomes and broader economic impacts. To assess whether these outcomes and impacts have been realised in reality, we apply a robust mixed-methods framework, based on:

- A global user survey with over 2,500 respondents;
- Applying a four-stage method that converts internal EMBL-EBI web-log data into unique users;
- Impact modelling to estimate user-level and societal value; and
- Three case study assessments of the downstream impacts of EMBL-EBI data resources.

Our methodology is designed to allow comparison with earlier assessments published in 2016[7] and 2021[8] by Charles Beagrie Ltd (Beagrie). This latest assessment reflects major developments since 2021, including post-pandemic changes in research practice and the emergence of new AI-enabled resources such as the AlphaFold Database. It also builds upon Beagrie's original methodology, reflecting research best practice and the application of new tools (such as Large Language Models).[9]

[7] See Beagrie (2016), "The Value and Impact of the European Bioinformatics Institute."

[8] See Beagrie (2021), "Data-driven discovery: The value and impact of EMBL-EBI managed data resources."

[9] All values in this study are provided in Great British Pounds (GBP). Global average exchange rates for 2024 were used to convert other currencies to GBP. In line with HM Treasury's Green Book guidance, all values, including those cited from the previous Beagrie (2016, 2021) studies, are reported in 2025 real terms. This is to enable a fair comparison over time.

Evidence on the usage, reach and immediate outcomes of EMBL-EBI data resources

We conservatively estimate the short-term outcomes of providing EMBL-EBI data resources across direct usage, reach and the work users undertake. The findings are summarised in Figure 1 below.

Figure 1 Usage, reach and immediate outcomes of EMBL-EBI data resources

Source: Frontier Economics

Growing scale of usage

Internal EMBL-EBI data suggests that 41m unique IP addresses accessed EMBL-EBI's online portal in 2024. We adjust this figure to estimate the number of yearly users through a four-stage approach. This approach considers that 1) unique IP addresses systematically overstate the true number of individual visitors, 2) there are other methods of accessing EMBL-EBI data resources, 3) there are cases where multiple users sit behind a single IP address, and 4) there has been an increase in hybrid working since the pandemic. This approach is described in Annex B in further detail.

Applying these adjustments, we estimate that EMBL-EBI data resources are accessed directly by around 520,000 unique users per year. This represents continued growth relative to previous years (which was estimated at between 450,000 and 500,000 in 2021), and an increase of 163% from the estimated user base of 198,000 at the time of the 2016 economic assessment. It is likely to be a lower bound as (i) we estimate the user base using conservative assumptions, and (ii) it does not capture secondary usage (i.e., users of data resources and tools that are built partially or wholly on data provided by EMBL-EBI).

The findings suggest that secondary usage is substantial and increasing. Over one-third of survey respondents reported curating secondary resources using EMBL-EBI data, including free public databases, local institutional resources and subscription-based tools. Survey respondents indicated that these curated resources are themselves used by hundreds, and in

some cases thousands of additional users, indicating that the total reach of EMBL-EBI data (including secondary resources) is significantly larger than the direct usage estimates above suggest.

Future work should seek to develop a robust estimate of the secondary user base and the value that arises from it to fully capture the value of EMBL-EBI data resources, as this has not been possible to reflect in the four headline indicators discussed below. For direct users, future work should also seek to more accurately account for accesses made through modern access routes.

Global and cross-sector reach

EMBL-EBI data resources serve a global research community, with survey respondents from Europe, Asia, North America, South America, Africa and Oceania. While academia remains the largest user group, there is increasing use by industry and hospitals (rising from 10% of all respondents in the 2016 study to 20% of all respondents in this study), reflecting the growing importance of open biological data in clinical contexts and commercial settings.

Importance to users' work and study

The survey evidence shows that EMBL-EBI data resources are critical to users' work, with 71% of survey respondents reporting that their work or study would be impossible or would require significant additional time and effort without these data resources.

Further, survey respondents reported that EMBL-EBI resources contribute to a variety of critical areas of research practice:

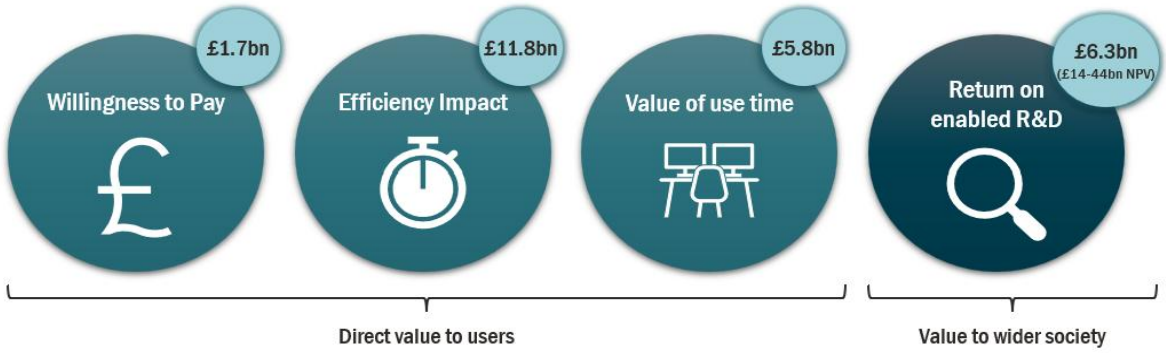
- **Interdisciplinary research:** 78% of respondents reported that EMBL-EBI data resources and tools contributed to collaboration on research across disciplines;
- **Reproducibility:** 72% of respondents reported that EMBL-EBI data resources and tools contributed to validating the reproducibility of data outputs;
- **Funder or open science requirements**: 64% of respondents reported that EMBL-EBI data resources and tools contributed to complying with funder or publisher open science requirements;
- **Citation of outputs:** 60% of respondents reported that EMBL-EBI data resources and tools contributed to providing or receiving credit or citation for data outputs; and
- **Machine learning and AI:** 42% of respondents reported that EMBL-EBI data resources and tools contributed to training and/or evaluating machine-learning or other AI models.

The evidence suggests that users view EMBL-EBI resources as essential infrastructure for modern life sciences research.

Impact modelling of the value of EMBL-EBI data resources

The impact modelling conservatively estimates the overall value of EMBL-EBI in terms of its direct impacts on users and wider socio-economic impacts. Four headline indicators are used, each capturing a different dimension of impact. The findings are summarised in Figure 2 below.

Figure 2 Annual value of EMBL-EBI data resources



Source: Frontier Economics

Willingness to pay

Users' stated willingness to pay (WTP) for access to EMBL-EBI data resources is estimated at £1.7bn per year. This represents a direct expression of the value users place on the benefits they receive.

This figure has increased in real terms since 2021 (from £1.5bn), driven primarily by growth in the user base. It is conservative, as many users report that their stated willingness to pay is constrained by personal or institutional budgets rather than by the true value of the resources.

Efficiency impact

Users report saving an average of 11 hours per week as a result of having access to EMBL-EBI data resources (up from 9 hours per week in 2021). When scaled across the estimated user base, this translates into an annual productivity benefit of approximately £11.8bn per year; a 44% real-terms increase from 2021 (and a 188% real-terms increase from 2016). This highlights the material impact on user productivity from having access to EMBL-EBI resources.

Value of time spent using EMBL-EBI resources

The value of the time users spent working with EMBL-EBI data resources is estimated at £5.8bn per year. This provides an assessment of economic value as users must, as a minimum, value the resources equal to the value of their time they spend accessing and/or

using them. Otherwise, it would not be worth their while and they would spend their time on more valuable activities elsewhere.

Overall, the value of this indicator has declined from around £6.7bn in 2021 in real terms as a result of a decline in the average time spent with EMBL-EBI data (but remains materially above the 2016 value of £3.2bn in real terms). There are multiple explanations that might explain this trend:

- More efficient usage and access methods leading to a reduction in time spent on tasks using EMBL-EBI data;
- A decrease in the value derived from EMBL-EBI data resources;
- An increased use of secondary resources built partially or wholly on EMBL-EBI provided data, which are not considered in the survey results on time spent; and/or
- A reduction in online-based research since the pandemic due to a move back to lab-based working.

Further research is recommended to understand what is driving the reduction in the time spent with EMBL-EBI data. Given this uncertainty, the value of use time indicator represents an imperfect proxy of value and must be considered in the context of the other value indicators.

Returns on research and development enabled

The returns on R&D enabled by EMBL-EBI data resources capture the wider societal impacts from users having access to these resources. Combining evidence on efficiency gains, value of time spent using data and the proportion of research that would otherwise not occur, we estimate that EMBL-EBI data resources enable around £6.3bn per year in returns on research and development. This represents sustained growth from £2.1bn in 2016 and £4.9bn in 2021.

When projected over the lifetime of the knowledge created, the £6.3bn figure equates to a net present value (NPV) of between £14bn and £44bn, depending on assumptions around depreciation and the public or private nature of the research.

Case study modelling assessment of downstream impacts

Three case studies examine specific applications of EMBL-EBI data resources and the downstream benefits arising from their use. They complement the high-level impact modelling by providing a more granular, bottom-up perspective on the impacts of EMBL-EBI's data resources. They also provide examples of secondary usage (which was unable to be captured in the impact modelling above), where EMBL-EBI data was used to develop tools and databases that have their own, separate user bases.

Open data case study: the AlphaFold Database and its impact on research

Protein folding is fundamental to understanding biological processes, and accurate structural predictions can accelerate advances in health research and drug discovery. This is because

many diseases involve proteins folding incorrectly. Developed by Google DeepMind in 2021, AlphaFold 2 is an artificial intelligence-based system capable of predicting a protein's three-dimensional structure from its amino acid sequence. Google DeepMind trained the AlphaFold 2 algorithm on publicly-available data, including data from EMBL-EBI data resources.

Prior to AlphaFold 2's release, scientists had experimentally determined the structures of around 180,000 proteins. AlphaFold 2 revolutionised the field by enabling accurate predictions for over 200 million proteins, dramatically expanding the scope of structural biology. While there were systems before AlphaFold, their accuracy and speed limited their impact.

EMBL-EBI partnered with Google DeepMind to create the AlphaFold Database, which curates and integrates these predictions into an accessible, open resource. By doing so, it removes the need for researchers to have the expertise to run the algorithm, integrate this themselves, and have access to high-performance computing resources. Instead, scientists can directly explore and apply AlphaFold 2's predictions within their research workflows.

The direct impact of the AlphaFold 2 algorithm has been estimated previously and recognised in 2024 with the Nobel Prize for Chemistry.[10] This case study, however, tests two hypotheses pertaining to the impact of the AlphaFold Database, compared to a counterfactual world in which only the AlphaFold 2 algorithm exists:

- **Distributional effect**: the creation of the AlphaFold Database has widened access to the algorithm's predictions to a more diverse array of research fields; and
- **Scale effect**: the creation of the AlphaFold Database has led to greater use of the algorithm's predictions, resulting in more research being conducted.

More generally, this case study provides evidence as to the downstream impacts arising from making new AI-system-derived outputs easily accessible.

Using bibliometric data from OpenAlex, we assess the degree of topic overlap and diversity of papers mentioning the PDB, the AlphaFold Database and the AlphaFold 2 algorithm. Overall, we find evidence in support of both hypotheses. Specifically:

- Research topics of papers mentioning the AlphaFold Database are both different to and more diverse than research mentioning the AlphaFold 2 algorithm alone; and
- Since Q2 2024, the total number of new papers mentioning the AlphaFold Database has continued to rise, while mentions of PDB and the AlphaFold 2 algorithm have remained broadly flat.

While the evidence supports the hypothesis that the AlphaFold Database has had a scale and distributional effect on research, it is not possible to make causal statements. Nevertheless,

[10] https://deepmind.google/blog/alphafold-five-years-of-impact/

the results demonstrate how open data infrastructure can likely amplify the impact of major scientific breakthroughs by making them more widely accessible.

Figure 3 Open data: The AlphaFold Database and its impact on research

Source: Frontier Economics

Data resources and standards case study: their role in European genomic medicine programmes

Genomic sequencing is a technique used to determine the complete or partial DNA sequence of an individual's genome, allowing detailed analysis of genetic variation. By examining either the whole genome or specific genomic regions, this approach can detect disease-causing variants that are not identifiable through conventional diagnostic methods. As a result, genomic sequencing is becoming an increasingly central tool for the diagnosis of rare diseases and other genetic diseases, such as cancer.

EMBL-EBI provides access to a range of data resources, tools and data standards that are relevant to national genomic medicine programmes. This case study tests the hypothesis that these data resources, tools and standards are used by, and have a positive impact on, the genomic sequencing pipelines of national genomics programmes. In particular, it examines the downstream impacts arising from EMBL-EBI's role in (i) making genomic data resources and tools freely available, and (ii) developing and promoting harmonised global standards.

Through interviews with bioinformaticians at (i) Genomics England, and (ii) The Department for Genomic Medicine at Rigshospitalet, which is a diagnostics department embedded in Copenhagen University Hospital, and has served as the bioinformatics and sequencing facility for the National Genome Centre in Denmark, we find that:

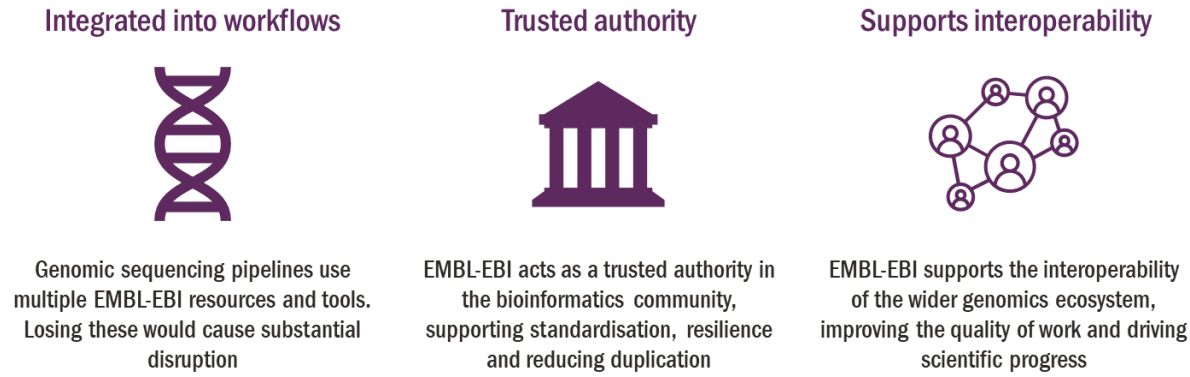
- **Genomic sequencing pipelines use multiple EMBL-EBI resources and tools. Losing these would cause substantial disruption**: EMBL-EBI resources, such as Ensembl, Ensembl VEP, and DECIPHER, are implemented at various stages of the clinical genomics pipeline, including for variant annotation, variant effect prediction, and downstream interpretation. If EMBL-EBI ceased to operate these tools, this would have significant, negative impacts in terms of immediately requiring genomics programmes to

re-design their internal infrastructures, therefore reducing the number of diagnoses made and increasing costs;

- **EMBL-EBI acts as a trusted authority in the bioinformatics community, supporting standardisation, resilience and reducing duplication**: Through the scale of its resources and tools, EMBL-EBI has established itself as a trusted authority in the genomics ecosystem. By working together with the scientific community and other standard-setting organisations such as the Global Alliance for Genomics and Health, EMBL-EBI helps to set sector-wide data standards. This improves the efficiency and reliability with which bioinformaticians can interact with data from different sources across the world. EMBL-EBI also supports resilience throughout the genomics ecosystem, as well as reducing the time and costs spent on duplication of efforts among organisations; and
- **EMBL-EBI supports the interoperability of the wider genomics ecosystem, improving the quality of work and driving scientific progress:** By connecting data resources and tools throughout the field (for example, through MANE**,** which is a collaborative project to ensure that genes and transcripts are described consistently across genome browsers, databases, and clinical tools), as well as collaborating with other life science institutions, the quality of work produced, as well as wider scientific progress, are improved.

Through these channels, the value provided by EMBL-EBI's data resources, tools and standards is greater than the individual value of each resource.

Figure 4 Data resources and standards: their role in European genomic medicine programmes



Source: Frontier Economics

Data reuse case study: supporting public and private sector innovation through bioimaging and structural biology resources

Bioimaging encompasses techniques used to visualise biological structures and processes, from molecules to whole organisms. Structural biology focuses on determining the three-dimensional (3D) structure of macromolecules and their assemblies to understand how their

form underpins their function. Together, these fields help connect molecular structure with biological behaviour. Beyond advancing fundamental science, they support applications such as tracking disease progression, assessing responses to therapies, enabling medical diagnosis and the development of drugs, AI models and biomaterials.

EMBL-EBI hosts several open data resources supporting bioimaging and structural biology research, including EMPIAR, EMDB and PDB (via PDBe). This case study tests the hypothesis that EMBL-EBI's bioimaging and structural biology open data resource ecosystem supports both public and private sector data reuse and innovation through the development of new tools.

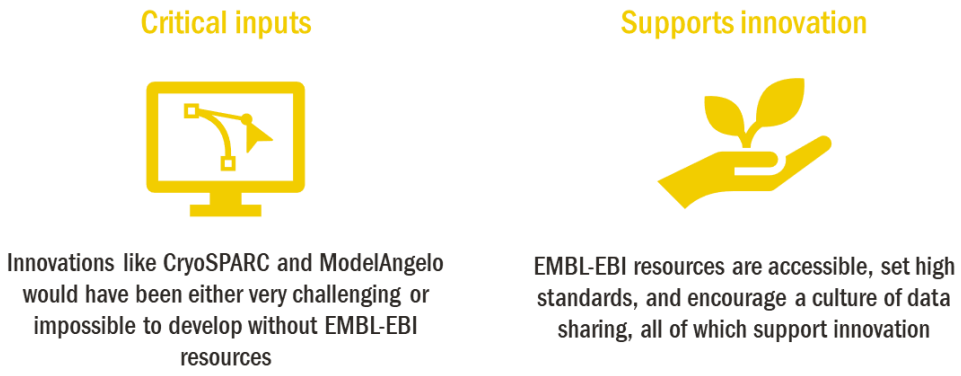
To test this hypothesis, we qualitatively explore the contribution of EMBL-EBI resources to the development of two innovative tools: CryoSPARC (private sector) and ModelAngelo (public sector). These tools accelerate discovery by making complex image processing, reconstruction and interpretation more efficient and more widely usable. This has positive downstream impacts on scientific understanding of living systems across many applications, including drug and vaccine development in human medicine.

Through interviews with the developers of CryoSPARC and ModelAngelo, we find that:

- **Innovations like CryoSPARC and ModelAngelo would have been either very challenging or impossible to develop without EMBL-EBI resources:** The developer of CryoSPARC described that EMBL-EBI resources, particularly EMPIAR, were critical to its development by providing open, high-quality datasets that removed data access bottlenecks. The developer of ModelAngelo stated that EMDB and PDB were essential to the development and training of the tool, with no viable alternatives. EMBL-EBI resources remain critical for ongoing improvements across both tools; and
- **EMBL-EBI resources are accessible, set high standards, and encourage a culture of data sharing, all of which support innovation:** The developer of CryoSPARC emphasised that EMBL-EBI's data standards and open-sharing culture, through repositories such as EMDB and EMPIAR, establish a trusted level of quality; ensure consistent and transparent data deposition; reduce data validation burdens; and remove data-access frictions, thereby accelerating innovation. The developer of ModelAngelo highlighted that EMBL-EBI's well-structured, thoroughly curated and highly accessible open data infrastructure, across resources like PDB and EMDB, substantially lowers barriers to innovation.

The evidence suggests that EMBL-EBI data resources play a key role in enabling the development of tools such as CryoSPARC and ModelAngelo. It therefore supports the realisation of the downstream benefits associated with these tools highlighted above.

Figure 5 Data reuse: supporting public and private sector innovation through bioimaging and structural biology resources



Source: Frontier Economics

Conclusion

Taken together, the evidence suggests that EMBL-EBI data resources deliver substantial value to users and to society, with this value continuing to grow over the past decade. Across all value indicators, we find benefit-cost ratios to be high. Even under conservative assumptions, the benefits of EMBL-EBI data resources substantially exceed the costs of maintaining and providing them.

At the user level, EMBL-EBI resources:

- Are essential to a large and growing number of life sciences researchers;
- Support researchers globally and across academic, industry and healthcare settings;
- Enable work that would otherwise not be possible;
- Save significant amounts of researcher time; and
- Support interdisciplinary research, reproducibility and the development of machine learning and AI systems.

At the societal level, they:

- Enable significant returns to research and development; and
- Realise downstream impacts. For example, widening access to and increasing the volume of science, supporting healthcare systems, and enabling innovation.

Crucially, the evidence from the case studies shows that EMBL-EBI's value does not lie solely in the data themselves, but also in the curation, standardisation, integration and open access model. With interviewees highlighting EMBL-EBI's role as a central body that develops standards and coordinates with the open science community, this suggests that the value generated by EMBL-EBI data resources is worth more than the sum of their parts.

Overall, the evidence strongly supports the conclusion that investment in EMBL-EBI data resources represents excellent value for money and plays a critical role in sustaining and advancing the global life sciences research ecosystem.

1 Introduction

1.1 Overview of EMBL-EBI

The European Molecular Biology Laboratory (EMBL) is Europe's only intergovernmental life sciences laboratory. Founded in 1974, it now operates six sites across Europe: Barcelona, Grenoble, Hamburg, Heidelberg, Hinxton, and Rome.[11] EMBL's European Bioinformatics Institute (EMBL-EBI), established in 1994 in Hinxton near Cambridge in the UK, is a global leader in the storage, analysis and dissemination of large biological datasets. EMBL-EBI conducts its own computational biology research, delivers training programmes to support researchers, and offers partnerships to foster collaboration with industry. As of 2025, EMBL-EBI employed 646 full-time equivalent staff from 65 countries.

Through its collaborations with institutions across the globe, EMBL-EBI seeks to achieve five missions:[12]

1. To freely provide data and bioinformatics services to the scientific community in ways that promote scientific progress;
2. To contribute to the advancement of biology through investigator-driven research in bioinformatics;
3. To provide bioinformatics training to scientists at all levels;
4. To disseminate cutting-edge technologies to industry; and
5. To support, as an ELIXIR Node, the coordination of biomolecular data provision in Europe.

In support of the first mission, EMBL-EBI hosts a globally accessible set of around 30 open biological data resources. These support research across genetics, genomics, proteins, bioimaging, chemistry and more, enabling the publication of over 120,000 scientific publications in 2025.[13]

One of EMBL-EBI's core principles is to deliver its open data solutions based on a collaborative approach. In particular, all of EMBL-EBI's major data resources are developed through partnerships with organisations and communities from around the world. These involve a wide range of partners, including research institutes, universities, industry organisations and scientific communities. They follow various models, including joint grants, consortium agreements, direct partnerships, joint management of data resources and federated models. This collaborative approach brings together expertise from across the

[11] EMBL's six sites are based in Barcelona, Grenoble, Hamburg, Heidelberg, Hinxton (EMBL-EBI), and Rome.

[12] See "EMBL-EBI's Highlights 2025 Report," page 3.

[13] See "EMBL-EBI's Highlights 2025 Report," pages 4-6.

scientific community. For example, in 2025, EMBL-EBI held 173 active grants, of which 146 were collaborative grants with researchers from 860 institutes in 58 countries.[14]

EMBL is supported by over 30 member states, associate member states, and prospect member states.[15] Major funders include UK Research and Innovation, as well as the European Commission, the US National Institutes of Health, and Wellcome. Sustained capital investment from the UK Government has helped EMBL-EBI develop a robust technical infrastructure to efficiently manage the growth and diversity of biological data held in its public data resources. Investment by funders and collaborator contributions enable EMBL-EBI to maintain its comprehensive suite of open data resources for the life sciences.

1.2 Purpose of this study

EMBL-EBI has commissioned Frontier Economics to evaluate the value and socio-economic impacts arising from investment in EMBL-EBI's mission to provide freely available data and bioinformatics services.[16] We (Frontier Economics) examine how, and to what extent, these data resources improve research productivity, enable research and development, and generate wider socio-economic impacts. By quantifying their value and impact, we also aim to strengthen the evidence base for the broader value of open data for the life sciences research community.

Evaluating the value and socio-economic impacts of EMBL-EBI data resources is inherently complex. Because these resources are freely accessible, with no charge to users, their value cannot be observed through market prices and must instead be inferred indirectly. The breadth of resources provided, each serving different user communities, further reinforces the need to identify and articulate the distinct value generated by individual data resources. This is particularly important given that EMBL-EBI data resources are used both by direct users, who access them via EMBL-EBI, and through 'secondary use', whereby users rely on other data resources and tools that are built partially or wholly on data provided by EMBL-EBI. A similar complexity applies to costs, where many EMBL-EBI data resources and tools are the product of longstanding, collaborative, community-led contributions.

Recognising this, we take a conservative approach to valuation. We apply a mixed-methods framework, drawing on both qualitative and quantitative evidence across:

- **A user survey:** This assesses the impact and value of EMBL-EBI data resources based on user experiences. From these user-level findings, we then model the wider impacts and value experienced by the overall direct EMBL-EBI user population, as well as wider

[14] See "EMBL-EBI's Highlights 2025 Report," page 5.

[15] The EMBL member states are Austria, Belgium, Croatia, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, the Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland, and the United Kingdom. Australia is an associate member state, and Serbia and Bulgaria are prospect members.

[16] The assessment does not cover the remaining four Missions, which are outside the scope of this study.

societal benefits. As secondary usage was unable to be captured in the impact modelling, all three case studies (see below) contain examples of where EMBL-EBI data was used to develop tools and databases that have their own, separate user bases.

- **Three case studies:** These case studies evaluate the downstream impacts of EMBL-EBI data resources for three specific use cases. They complement the higher-level survey modelling by focusing on specific examples of how access to these data resources has led to socio-economic impacts.

Throughout this report, we refer to "EMBL-EBI data resources" for brevity. As described above, these resources are developed and sustained through close collaboration with a diverse range of organisations and communities worldwide. This terminology does not detract from the collective effort and shared contributions underpinning their development.

1.3 Link to previous studies

Our study builds on previous economic assessments published by Charles Beagrie Ltd (Beagrie) in 2016 and 2021.[17, 18] We take a similar approach to the user survey modelling, enabling like-for-like comparisons to be made over time. We have however built upon the previous methodology and updated assumptions where new evidence has become available. The key changes include:

- **Additional and updated user survey questions:** we make changes to the survey questions to improve the robustness of the results, taking into account feedback from previous studies. Additional questions are asked to provide further insights into user perceptions of particular aspects of EMBL-EBI data resources;
- **Revised approach to capturing the return on R&D enabled by EMBL-EBI data resources**: this new approach incorporates the impact of improved research efficiency on the overall returns to R&D;
- **New large language model (LLM) analysis:** we apply new, innovative LLM methods to uncover themes and insights from all open text survey responses, allowing us to extract further insight from these responses; and
- **New rigorous case study analysis:** we evaluate the downstream economic impacts arising from three use cases of EMBL-EBI data resources.

It is important to re-evaluate the impact and value of EMBL-EBI data resources following the previous Beagrie studies (2016 and 2021) to capture major developments that have affected the life sciences research community since then. These include:

- **The return to a "new-normal" after the COVID-19 pandemic:** the Beagrie (2021) study took place in the context of the COVID-19 pandemic. Therefore, the return to the "new-

[17] See Beagrie (2016), "The Value and Impact of the European Bioinformatics Institute."

[18] See Beagrie (2021), "Data-driven discovery: The value and impact of EMBL-EBI managed data resources."

normal" might have affected the way that life sciences researchers interact with online resources and tools since 2021.

- **The introduction of new technologies**: in 2021 Google DeepMind developed AlphaFold 2, which is an artificial intelligence (AI) system capable of predicting a protein's three-dimensional structure from its amino acid sequence.[19] The introduction of the AlphaFold 2 algorithm and the AlphaFold Database, along with other new technologies, might have influenced researchers' valuation of open data resources and tools (given some of these systems, like the AlphaFold 2 algorithm, were trained using open data).
- **Developments in user behaviour and expectations shaped by technological change**: the increasing use of AI for data-intensive tasks might have affected the perceived value and impact of research based on open data resources.

1.4 Structure of the report

The rest of this report is structured as follows:

- **Section 2 – Methodology for measuring the value and impact of EMBL-EBI data resources:** includes an overview of our overarching research questions, a framework for answering those questions set out by a logic model, and an outline of the methodology used for assessing the value and impact of EMBL-EBI data resources;
- **Section 3 – Evidence on the usage, reach and immediate outcomes of EMBL-EBI data resources:** reports evidence from the user survey around the usage, reach and immediate outcomes of EMBL-EBI data resources;
- **Section 4 – Impact modelling of the value of EMBL-EBI data resources:** reports the findings from the impact modelling of the overall value of EMBL-EBI data resources;
- **Section 5 – Open data case study: the AlphaFold Database and its impact on research:** outlines the methodology and findings from the first case study, covering the AlphaFold Database and its impact on research;
- **Section 6 – Data resources and standards case study: their role in European genomic medicine programmes:** outlines the methodology and findings from the second case study, covering the role of data resources and standards in European genomic programmes.
- **Section 7 – Data reuse case study: supporting public and private sector innovation through bioimaging and structural biology resources:** outlines the methodology and findings from the third case study, covering the impact of bioimaging and structural resources on public and private sector reuse and innovation.
- **Section 8 – Conclusions:** summarises the findings presented throughout the report and the overall conclusions.

[19] A more detailed description of AlphaFold 2 and the implications for EMBL-EBI data resources is included in Section 2.3.

2 Methodology for measuring the value and impact of EMBL-EBI data resources

2.1 Overview of research questions

This impact assessment aims to establish the socio-economic impacts arising from investment in EMBL-EBI data resources. This includes both direct impacts for users of EMBL-EBI data resources, as well as the wider societal benefits from that use. The methodological approach outlined in this section is designed to address the following high-level research questions:

1. To what extent do EMBL-EBI data resources support users to assimilate and reuse public data/knowledge?
2. What impact does the use of EMBL-EBI data resources have on users' research productivity and efficiency?
3. To what extent do the effects of EMBL-EBI data resources on research productivity and efficiency contribute to wider socio-economic impacts?

To evaluate these research questions, we develop a logic model for EMBL-EBI data resources. This visually sets out the pathways through which outcomes and impacts are expected to arise as a result of researchers having access to EMBL-EBI data resources. This model also provides the framework for our economic analysis.[20] The methodological approach outlined in the rest of this section has therefore been designed to test the impacts and outcomes described in the logic model.

2.2 Logic model for EMBL-EBI data resources

The logic model developed for EMBL-EBI data resources is set out in Figure 6. It starts by setting out the links between primary inputs (e.g., community data contributions and funding), the activities EMBL-EBI undertakes (e.g., curating and annotating entries in databases, developing data deposition tools and engaging the community), and the outputs EMBL-EBI provides (e.g., databases and tools). From this base, it visually shows the process through which these outputs are hypothesised to lead to short-term outcomes, long-term outcomes and broader economic impacts. Using the available evidence we then assess whether these outcomes and impacts have been realised in reality, addressing the three overarching research questions set out in Section 2.1 above.

The types of outcomes and impacts tested include:

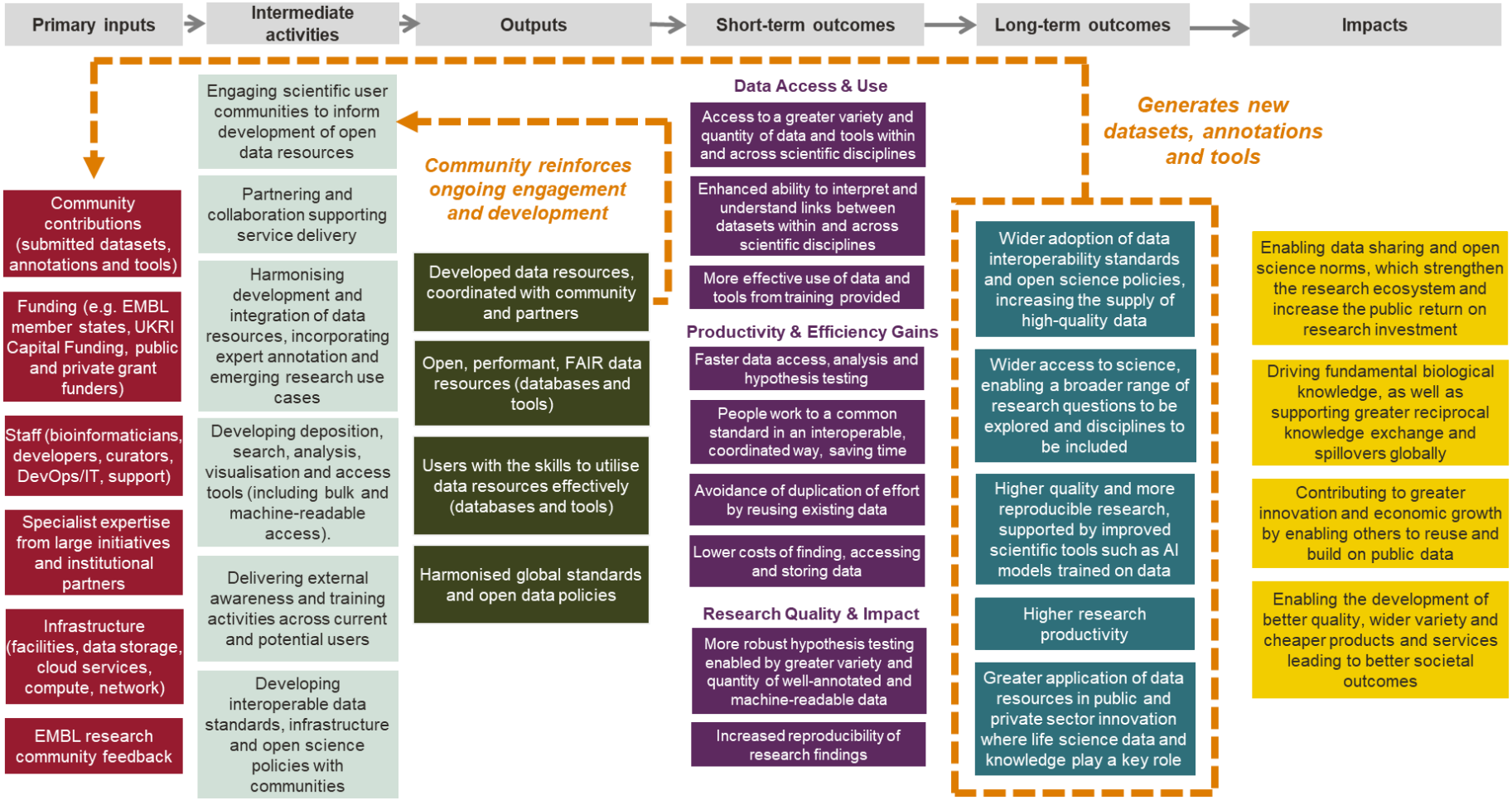
- **Short-term outcomes:** these capture the potential benefits users experience on a day-to-day basis from having access to EMBL-EBI data resources. They include:

[20] A logic model is a core requirement of the HM Treasury's Magenta Book.

- Improved data access and reuse;
- Gains in productivity and efficiency; and
- Increases in the quality, variety and quantity of research produced.

- **Long-term outcomes:** these capture the potential benefits experienced by the wider life sciences community arising from the cumulative effects of the short-term outcomes. They include:
  - Increased quality and availability of data;
  - Development and application of new scientific tools;
  - Improved research productivity; and
  - Increased innovation in the public and private sector.
- **Impacts**: these are the possible wider socio-economic impacts driven by the long-term outcomes, as well as the spillovers to wider society. They include:
  - Increased innovation and returns on research investment from improvements in the rate of data sharing, knowledge exchange and connectivity throughout the research ecosystem; and
  - Improved societal outcomes from the development of cheaper, higher quality and a wider variety of products and services.

Figure 6 Logic model for assessing the impact of EMBL-EBI data resources

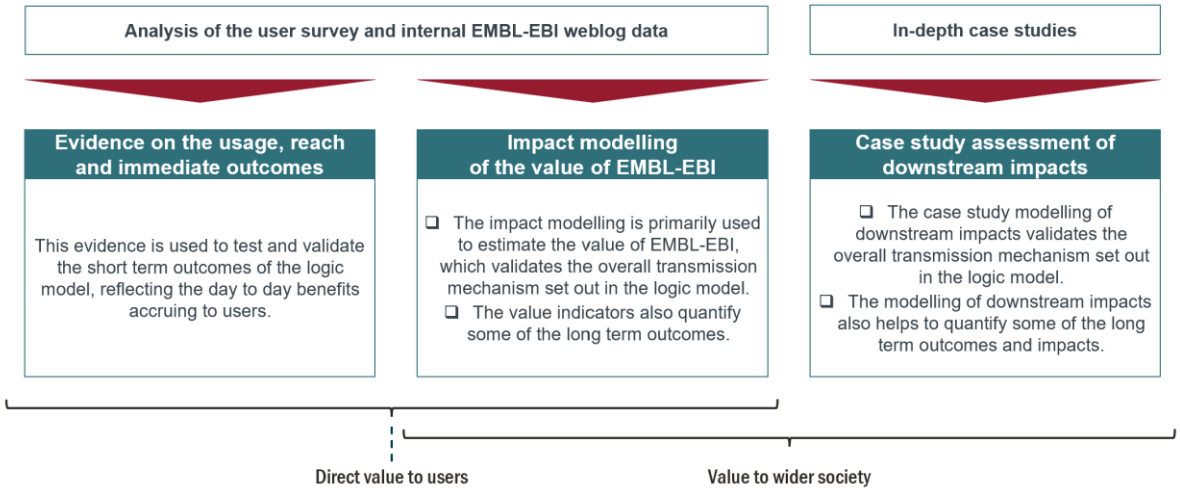


Source: Frontier Economics

2.3 Analytical approach

To assess the outcomes and impacts hypothesised in the logic model, we apply a mixed methods approach. This utilises a survey of EMBL-EBI users (the user survey), user analytics data from internal EMBL-EBI web-logs, and qualitative and quantitative data across three case studies. These sources are applied across three sets of analysis, which are detailed in Figure 7 below. We elaborate on the estimation methodology for each of these analyses in the following sub-sections.

Figure 7 Overview of the methodological approach



Source: Frontier Economics

2.3.1 Evidence on the usage, reach and immediate outcomes of EMBL-EBI data resources

We draw insights around the short-term outcomes of EMBL-EBI data resources based on a series of user survey questions. These insights allow us to measure:

- The usage of EMBL-EBI data resources, quantifying EMBL-EBI's contribution to improvements in worldwide data access and showing how users interact with the service;
- The reach of EMBL-EBI data resources, testing the extent to which EMBL-EBI facilitates widespread data access across a variety of groups and disciplines; and
- Users' views on the immediate outcomes for their research of having access to EMBL-EBI data resources, indicating their importance in facilitating research tasks; the extent to which they enable research that could not have otherwise occurred; and the time saved by users as a result of having access to these resources.

The findings are discussed in Section 3 of the report.

2.3.2 Impact modelling of the overall value of EMBL-EBI data resources

Four headline value indicators quantifying the overall value of EMBL-EBI data resources are estimated in this report, including:

1. Willingness to pay for EMBL-EBI data resources;
2. Self-reported efficiency impact from users having access to EMBL-EBI data resources;
3. The value of time users spend interacting with EMBL-EBI data resources; and
4. The return on R&D enabled by accessing EMBL-EBI data resources.

Indicators 1, 2 and 3 quantify the direct benefits experienced by users. Indicator 4 quantifies the value accruing to society from the research time saved and additional enabled from having access to EMBL-EBI data resources.

To calculate these four indicators, the following four-step approach is taken:

- **Step 1: Estimate the user base.** We apply a four-stage method that converts internal EMBL-EBI web-log data into unique users, conservatively estimating the total EMBL-EBI user population. The resulting user base estimate is described in Section 3.
- **Step 2: Analyse the user survey**. We calculate typical per-user values for each indicator. For example, we estimate the typical reported hourly saving per user from having access to EMBL-EBI data resources for the efficiency indicator. In calculating these values, we remove extreme or unrealistic responses.[21] We also uncover themes and insights from the open text survey responses using LLMs so that responses that might be subject to uncertainty or guesswork are identified. Annex C provides further information on our approach to using LLMs.
- **Step 3: Scale up the user survey results.** We multiply the per-user values in step 2 by the estimated EMBL-EBI user base in step 1 to calculate the total value of the benefits (for each headline indicator) for the whole EMBL-EBI user population.
- **Step 4: Calculate benefit-cost ratios (BCRs)**. We divide the total estimated benefits in step 3 (one for each indicator) by the total cost of maintaining and making available the EMBL-EBI data resources. The resulting benefit-cost ratios provide a standardised metric to assess whether the expected benefits from the provision of EMBL-EBI data resources outweigh the costs.

The four value indicators and user base estimates are calculated on a conservative basis,[22] and in a way that is largely consistent with the previous Beagrie (2016, 2021) approach. As explained in the introduction, we have built upon the previous methodology and updated

[21] See Annex A for further details.

[22] As described in Section 3, the findings throughout this report are conservative, in part, as the user base estimate does not include secondary users (i.e., users accessing the data indirectly via secondary curated resources). In this sense, the findings in this report are likely to be more conservative than the previous Beagrie (2016, 2021) analyses, as secondary usage has increased.

assumptions where new evidence has become available. Annex B details how the user base and the four indicators are calculated, and where changes have been made relative to the previous Beagrie (2016, 2021) studies. The final estimates for each indicator are presented in Section 4.

2.3.3 Case study assessment of downstream impacts

Three case studies are selected to examine specific applications of EMBL-EBI data resources and the downstream benefits arising from their use. These case studies complement the high-level impact modelling by providing a more granular, bottom-up perspective on the impacts of EMBL-EBI's data resources. They also provide examples of secondary usage (which was unable to be captured in the impact modelling above), where EMBL-EBI data was used to develop tools and databases that have their own, separate user bases.

An initial longlist of potential case studies was shortlisted based on three criteria:

- **Additionality:** Is there a clear pathway from EMBL-EBI data resources to one (or several) economic impacts? Have these impacts been assessed before?
- **Feasibility:** Is there readily available primary and/or secondary evidence that can be used to support the case study?
- **Robustness:** Is the evidence robust enough to support the claims being made?

Three case studies were then selected from the shortlist on the basis that each tests a different part of the logic model. This iterative process was performed in close consultation with EMBL-EBI. The final set of case studies chosen were:

- **Case Study 1:** Open data: the AlphaFold Database and its impact on research (see Section 5);
- **Case Study 2:** Data resources and standards: their role in European genomic medicine programmes (see Section 6); and
- **Case Study 3:** Data reuse: supporting public and private sector innovation through bioimaging and structural biology resources (see Section 7).

For each case study, we start by developing hypotheses covering the relevant outcomes and impact pathways from the logic model. We then develop a tailored methodology to test those hypotheses and (where possible) quantify the impacts. Sections 5, 6 and 7 each outline the hypotheses to be tested, methodology applied and results for each case study.

2.4 User survey

Overall, 2,563 responses to the user survey were received between June 2025 and July 2025. This represents a fall from the 5,662 responses in Beagrie (2021) and 4,509 responses in Beagrie (2016). Although it is not possible to determine the cause for this fall, a possible explanation includes response fatigue, where the same survey panel had become tired of

responding to repeated surveys. Alternatively, it is possible that there were fewer responses as our survey took place during summer, compared to previous surveys which took place earlier in the year. However, 2,563 responses remains sufficient for our analysis.

To ensure results could be compared over time, the user survey asked many of the same or similar questions to Beagrie (2016, 2021). However, in some areas we have revised questions to reduce the risk of bias, removed questions that are no longer relevant and added new ones to gain further insight on certain impacts and outcomes. The full survey questionnaire, responses received and approaches to cleaning the data are set out in Annex A.

We employ a range of techniques to address the inherent limitations of survey data. While these limitations cannot be eliminated entirely, we have taken steps to minimise the risk of bias in line with best practice. For example:

- **Survey fatigue:** this occurs when respondents become disengaged throughout the course of responding to surveys. This can lead to lower response rates or careless answers. As a result, survey fatigue can bias the findings and reduce the reliability and validity of the evidence. To minimise this risk, we have:
  - Designed the survey in line with best practices, only asking essential questions relevant to this research so as to minimise total expected completion time.
  - Employed a range of techniques on a question-by-question basis to exclude extreme or unrealistic responses.
- **Cognitive estimation bias:** where respondents have been asked to report a value that is conceptually challenging to estimate (such as the average time that they spent finding and obtaining data on EMBL-EBI), results may be subject to cognitive estimation bias. To minimise this risk, we have:
  - Revised some questions to improve question clarity, based on respondent feedback in the previous Beagrie reports (2016, 2021). Revisions to the survey questions are outlined in Annex A .
  - Used the median instead of the mean for questions with greater uncertainty, as the median is less sensitive to the influence of extreme estimates or outliers.
  - Applied large language model (LLM) techniques to identify key themes from users' open text responses to the survey. This helps isolate questions where users have expressed uncertainty or guesswork in their responses. We explain the LLM analysis methodology further in Annex C .

Another limitation of using survey data that is difficult to avoid is selection bias. This is the risk that, because respondents volunteered to fill in the survey, they might not be representative of all users of EMBL-EBI data resources.[23] While it has not been possible to weight the survey findings to correct for this, the geographic distribution of survey respondents (outlined in Section 3) is similar to the geographic distribution of EMBL-EBI users based on internal EMBL-

[23] For example, if respondents were those who use the data resources and/or value them the most, then this would bias the results.

EBI web-log data.[24] This increases our confidence that the findings from the user survey are broadly representative of the global population of EMBL-EBI users.

2.5 Reporting approach

All values in this study are provided in Great British Pounds (GBP). Global average exchange rates for 2024 were used to convert other currencies to GBP.[25] In line with HM Treasury's Green Book guidance,[26] all values, including those cited from the previous Beagrie (2016, 2021) studies, are reported in 2025 real terms. This is to enable a fair comparison over time.[27]

[24] Both the user survey and EMBL-EBI's internal web-log data suggest that between 56% to 59% of EMBL-EBI's users are based in the top 5 countries in terms either (i) survey responses, or (ii) unique IP addresses respectively. Those five countries are the USA, China, India, the UK and Germany.

[25] We convert monetary values reported throughout the survey responses into GBP using data from the World Bank. Currencies are converted based on the latest full year of data (i.e. 2024). For exchange rates that do not have available data for 2024, we use the next full year of data.

[26] HM Treasury (2026), The Green Book.

[27] All values are shown in 2025 real terms based on an adjustment using the latest UK GDP deflators.

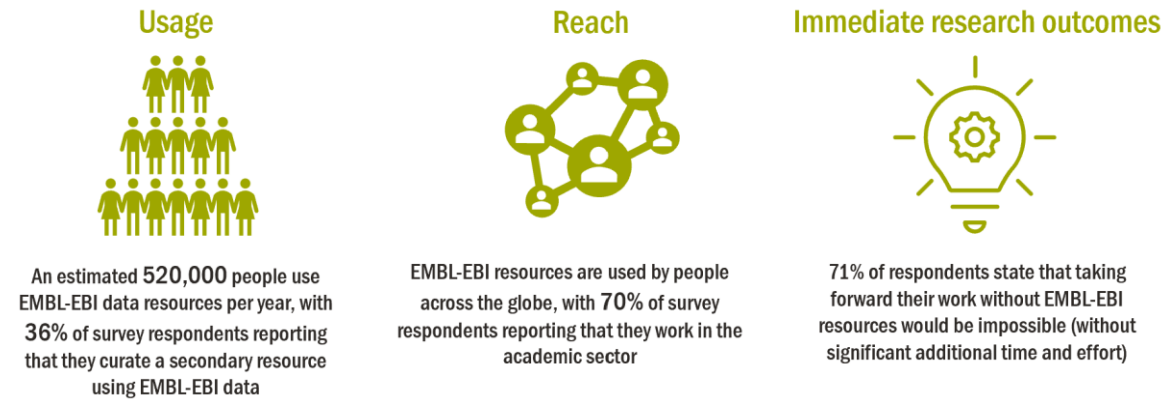
3 Evidence on the usage, reach and immediate outcomes of EMBL-EBI data resources

We analyse the user survey and internal EMBL-EBI web-log data to conservatively estimate the short-term outcomes of providing EMBL-EBI data resources across three domains:

- **Usage:** we estimate the number of direct and secondary users of EMBL-EBI data resources and examine how this has changed over time;
- **Reach:** we analyse user demographics across location, sector, job role and type of work to estimate the reach of EMBL-EBI data resources, how this has changed over time, and the diversity of the research user base across sectors; and
- **Immediate outcomes:** we report user views on how access to EMBL-EBI data resources impacts their work.

Figure 8 summarises the key findings from this analysis, which are set out in more detail below.

Figure 8 Usage, reach and immediate outcomes of EMBL-EBI data resources



Source: Frontier Economics

3.1 Usage of EMBL-EBI data resources

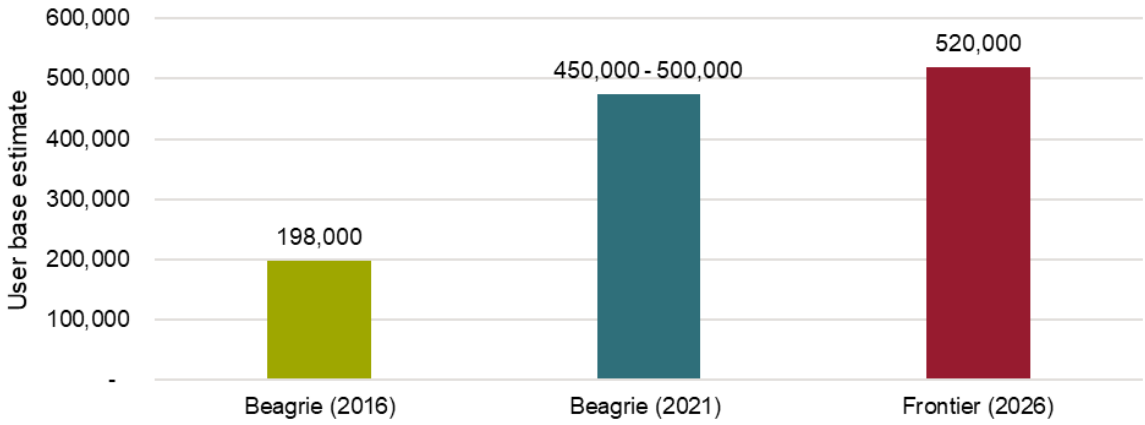
We estimate the contribution of EMBL-EBI data resources to improvements in worldwide data access by quantifying the size of EMBL-EBI's direct user population. We also provide an indication of secondary usage, whereby users rely on other data resources and tools that are built partially or wholly on data provided by EMBL-EBI.

Internal EMBL-EBI data suggests that 41m unique IP addresses accessed EMBL-EBI's online portal in 2024. As explained in Annex B we adopt an approach consistent with previous Beagrie analyses, adjusting this figure to estimate the number of yearly users through a four-stage approach. This approach considers that 1) unique IP addresses systematically overstate the true number of individual visitors, 2) there are other methods of accessing EMBL-EBI data

resources, 3) there are cases where multiple users sit behind a single IP address, and 4) there has been an increase in hybrid working since the pandemic. Applying these adjustments, Figure 9 shows that the estimated direct user base of EMBL-EBI data resources has grown since 2016 and 2021, increasing to around 520,000 users per year.

More recent evidence on the relationship between IP addresses and unique users (see Annex B for further discussion) suggests that the true user base may be up to twice the size of the estimates reported here. The approach taken in this report is therefore highly conservative. Nevertheless, the available evidence indicates a sustained improvement in the reach and use of EMBL-EBI data resources over time.

Figure 9 Estimated user base



Source: User analytics based on adjusted EMBL-EBI annual web-log data

Although the user estimates above give an indication of the number of direct users, they do not capture secondary usage. To assess the extent of secondary usage, we asked survey respondents if they curated a resource using EMBL-EBI data resources, and if so, to provide an estimate of how many people use that resource.

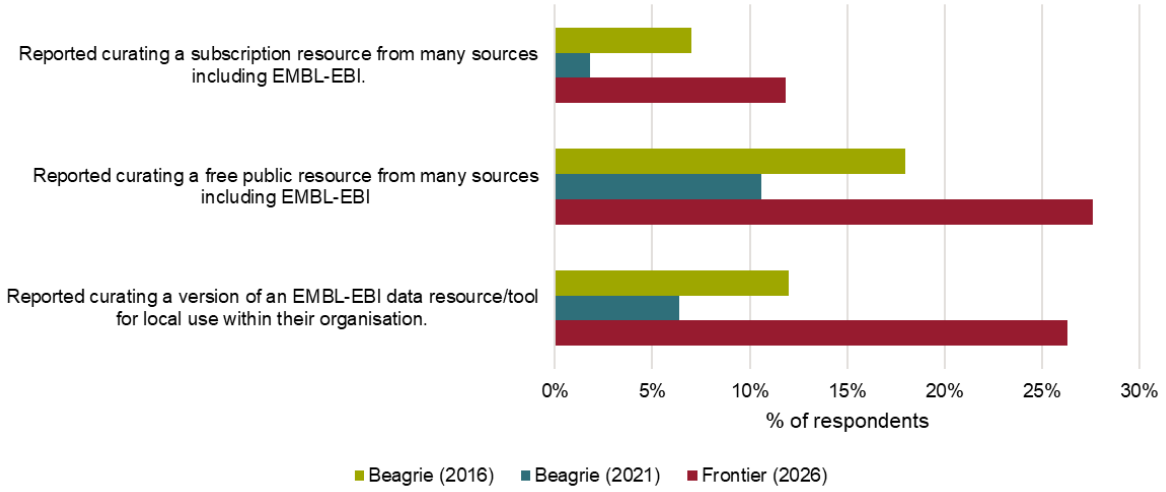
Overall, 36% of users reported curating a secondary resource using EMBL-EBI data. If applied to the estimated direct EMBL-EBI user population of 520,000, this would suggest 187,000 users are curating a secondary resource.

Figure 10 shows this split by the type of resource curated. It reveals that 12% of user survey respondents curated a subscription resource using EMBL-EBI data, 28% curated a free public data resource using EMBL-EBI data, and 26% curated a version for local use within their organisation using EMBL-EBI data. These figures have all increased since the previous Beagrie (2016, 2021) studies.

In addition, the majority of respondents curating resources reported that EMBL-EBI data resources had a moderate, significant or major impact on their secondary resources. In particular, 61% of those curating a subscription resource, 74% of those curating a free public

data resource, and 68% of those curating a version for local use reported a moderate, significant or major impact from having access to EMBL-EBI data.[28]

Figure 10 Share of respondents reporting the curation of secondary resources using EMBL-EBI data resources



Source: *2025 EMBL-EBI User Survey; Beagrie (2016, 2021)*

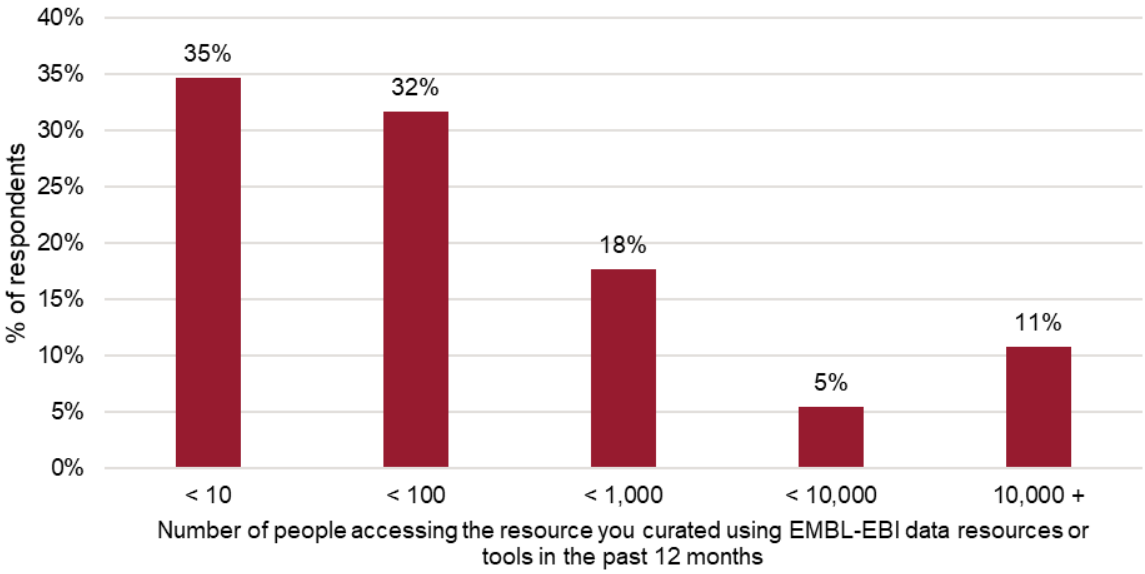
Note: *Question 36 of the survey asked respondents "Which of the following are part of your role (if any), and what is the impact of EMBL-EBI data resources or tools on those curated data resources?" (N = 1,190)*

Figure 11 shows that, of the survey respondents who reported curating a secondary resource, 11% estimate that their resource is accessed by over 10,000 people annually, and 5% estimate that their resource is accessed by between 1,000 and 10,000 people annually.

It is worth noting that the above is just one example of secondary use (curating data resources which are then used by others). Other examples include users of open source tools built using EMBL-EBI resources and the training of AI systems on these resources. These examples of secondary use are explored qualitatively in the case studies in later chapters.

[28] See Annex A for further details.

Figure 11 Number of people accessing curated data resources or tools based on EMBL-EBI data resources



Source: *2025 EMBL-EBI User Survey*

Note: *Question 37 of the survey asked respondents "In the past 12 months, approximately how many people accessed the resource you are curating using EMBL-EBI data resources or tools?." (N = 1,139)*

Taken together, the evidence on secondary usage indicates that the direct user base shown in Figure 9 represents a lower bound on overall usage, as secondary users are not captured. As a result, total use of EMBL-EBI data resources is likely to be substantially higher than the direct user figures reported in this study. Moreover, as a growing share of users curate secondary resources, secondary usage is likely to account for an increasing proportion of overall use in future years.

Based on available information from the survey, it is not possible to produce a robust estimate of the size of the secondary user base. This is because in some cases the same users might be accessing multiple secondary resources, which would risk double counting. Future reporting should therefore seek to develop a robust estimate of the secondary user base to more fully capture the value and impact of EMBL-EBI data resources.

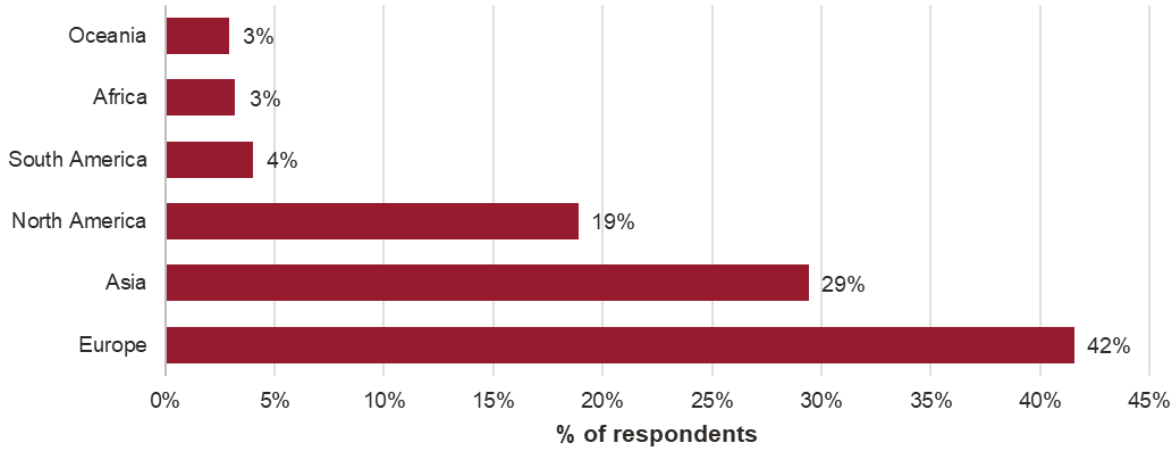
3.2 Reach of EMBL-EBI data resources

The demographic characteristics of user survey respondents provide detailed insight into the different types of users engaging with EMBL-EBI data resources. Compared with internal web-log data, the survey enables a deeper understanding of user types and behaviours, while the two data sources together support the cross-validation of findings.

Starting with geographic location, Figure 12 shows that the majority of respondents study or work in either Europe (42%), Asia (29%) or North America (19%). This is broadly similar

compared to the 2021 survey. It is also consistent with the user demographics indicated by internal web-log data, providing confidence that the findings from the user survey are broadly representative of the global population of EMBL-EBI users.

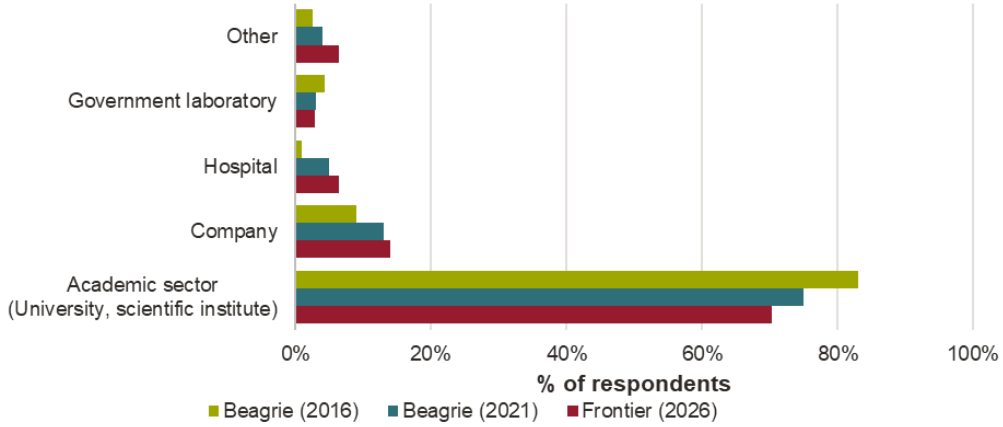
Figure 12 Continents in which user survey respondents work



Source: *2025 EMBL-EBI User Survey*
Note: *Question 1 of the survey asked respondents to report the "country in which you work or study." (N = 2,563)*

Figure 13 shows that the EMBL-EBI user base is becoming increasingly diversified across a range of sectors. While the majority of user survey respondents (70%) work within the academic sector, this share has decreased since the previous Beagrie (2016, 2021) studies, falling from 83% in 2016 and 75% in 2021. The remaining 2025 survey respondents reported working for companies (14%), hospitals (6%), government laboratories (3%) or other sectors (6%).

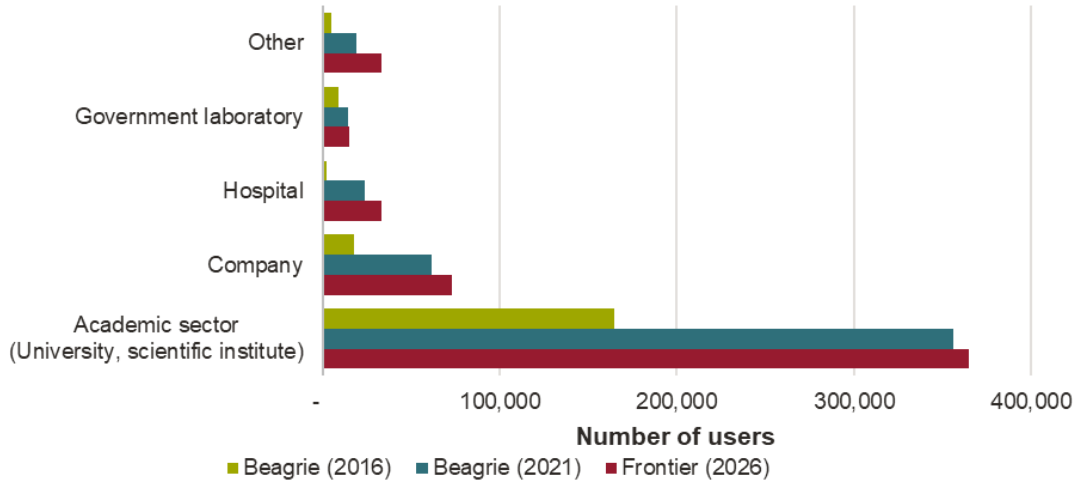
Figure 13 Main affiliation and sector of user survey respondents



Source: *2025 EMBL-EBI User Survey*
Note: *Question 2 of the survey asked respondents to report their "main affiliation and sector." (N = 2,563)*

Although the share of academic sector respondents has decreased, given that the estimated direct user population has increased over that period, in absolute terms the number of academic users has continued to increase (from around 164,000 in 2016 to 365,000 in this report). This is shown in Figure 14 below, which applies the shares in Figure 13 to the user base estimates in Figure 9.

Figure 14 Estimated number of direct users by main affiliation and sector



Source: 2025 EMBL-EBI User Survey; Beagrie (2016, 2021)

Of the user survey respondents affiliated with the academic sector, postgraduate students (24%), research assistants / post-doctorates (16%) and research fellows / associates (12%) were the most commonly listed job roles.[29] Of the user survey respondents affiliated with the non-academic sector, researchers (31%), directors (10%) and bioinformaticians (10%) were the most commonly listed roles.[30] These trends remain broadly similar with those reported by the previous Beagrie (2021) survey.

Analysing the types of work conducted, the survey reveals that there has been a slight shift toward wet laboratory work and away from dry working with computers since the previous Beagrie (2021) study.[31] For example, 42% of respondents to the 2025 survey reported mostly dry working with computers, compared to 48% of respondents to the 2021 survey (and 42% in 2016). In addition, 43% of respondents to the 2025 survey reported mostly wet laboratory working, compared to 37% of the respondents to the 2021 survey (and 48% in 2016). Otherwise, the distribution of respondents between non-scientific (but related to science or research), hospital or clinical practice or other roles has remained similar over time.

[29] Question 3 of the survey asked academic respondents "What is your main role within this affiliation?" (N = 1,772).

[30] Question 4 of the survey asked non-academic respondents "What is your main role within this affiliation?" (N = 735).

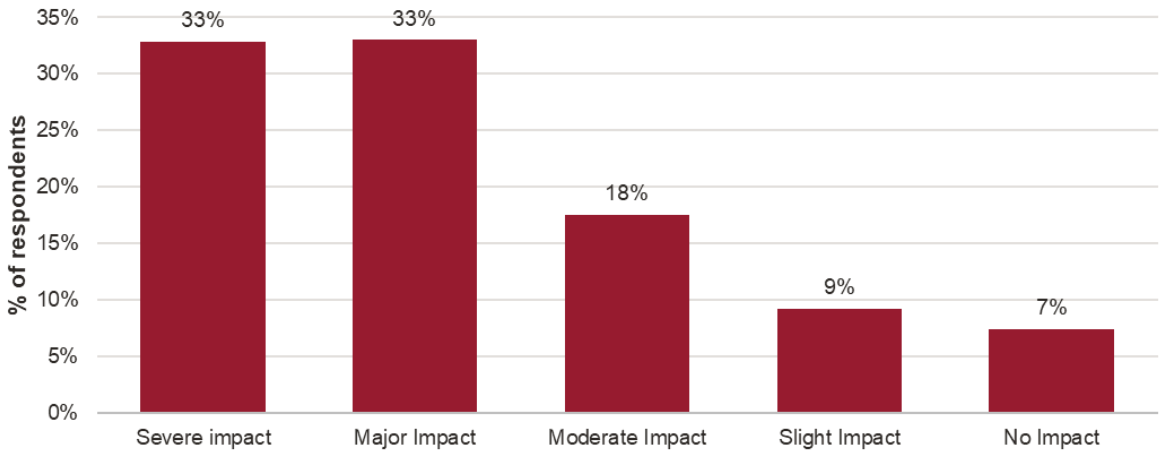
[31] Question 5 of the survey asked respondents "Which of the following most closely describes the nature of your work or study?" (N = 2,443).

3.3 User views on the immediate outcomes of having access to EMBL-EBI data resources

In the survey, we asked users for their views on the role of EMBL-EBI data resources in facilitating research tasks; the extent to which they enable research that could not have otherwise occurred; and the time saved as a result of having access to these resources. This qualitatively reveals the value users place on these resources. In addition, insights from the user survey around their most frequently used resources reveal how users interact with EMBL-EBI's service.

Figure 15 shows that 66% of respondents report that there would be a severe or major impact on their work or study if they lost access to the EMBL-EBI data resources or tools that they use. This compares to 18% of respondents reporting a moderate impact and only 16% of users reporting a slight or no impact. This highlights the key role these resources play in user work or study. While there has been a slight decrease since the previous Beagrie (2021) study (when 69% of respondents reported a severe or major impact), the stated impact is still high, and has increased materially since 2016 (when 55% of respondents reported a severe or major impact).

Figure 15 Stated impact of not having access to EMBL-EBI data resources



Source: *2025 EMBL-EBI User Survey; Beagrie (2016, 2021)*

Note: *Question 18 of the survey asked respondents "What would be the impact on your work or study if you could not access the EMBL-EBI data resources or tools you currently use?" (N = 2,030)*

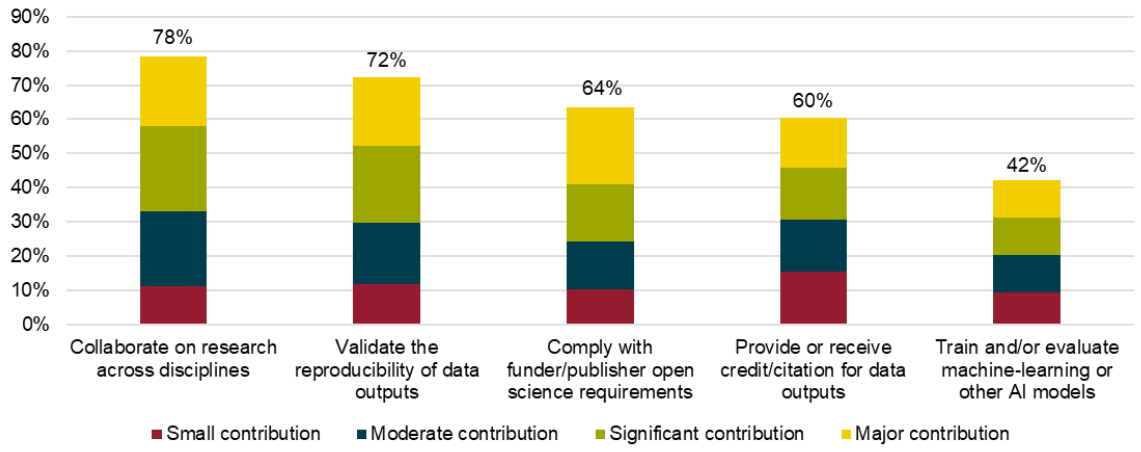
Figure 16 shows that EMBL-EBI data resources contribute across a range of different areas:

- **Interdisciplinary research:** 78% of respondents reported that EMBL-EBI data resources and tools contributed to collaboration on research across disciplines;
- **Reproducibility:** 72% of respondents reported that EMBL-EBI data resources and tools contributed to validating the reproducibility of data outputs;

- **Funder or open science requirements**: 64% of respondents reported that EMBL-EBI data resources and tools contributed to complying with funder or publisher open science requirements;
- **Citation of outputs:** 60% of respondents reported that EMBL-EBI data resources and tools contributed to providing or receiving credit or citation for data outputs; and
- **Machine learning and AI:** 42% of respondents reported that EMBL-EBI data resources and tools contributed to training and/or evaluating machine-learning or other AI models.

Respondents also reported whether EMBL-EBI's contribution to each area was small, moderate, significant or major. Figure 16 shows that respondents were fairly evenly distributed between these categories, but with the majority of respondents reporting a significant or major contribution from EMBL-EBI. Further details are provided in Annex A .

Figure 16 Impact of EMBL-EBI data resources on different areas



Source: *2025 EMBL-EBI User Survey*

Note: *Question 42 of the survey asked respondents "What impact, if any, has access to EMBL-EBI data resources and tools had on your ability to…" (N = 1,072). The figure reports the share of survey respondents who reported that EMBL-EBI data resources contribute to each area. The remainder of respondents either reported "no contribution" or that the question was "not applicable" to them.*

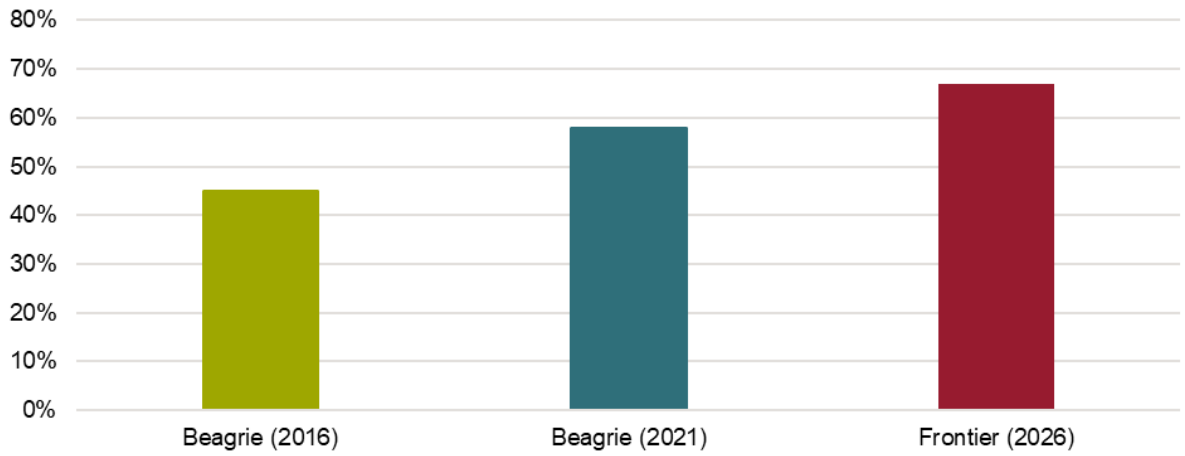
To understand what might be driving the findings in Figure 15, the survey asked users two questions, including (i) if they no longer had access to their most frequently used EMBL-EBI data resource or tool, would they be able to find and obtain equivalent data or tools from another source;[32] and similarly (ii) would they be able to re-create the same or similar dataset or tool themselves.[33]

[32] Question 23 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to find and obtain equivalent data or tools from another source?" (N = 1,888)

[33] Question 26 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to re-create the same or similar dataset or tool yourself?" (N = 1,799)

Survey responses show that EMBL-EBI data resources enable research that would otherwise be impossible to access or recreate. Overall, 69% of respondents reported that they would not be able to (or without significant additional time and effort) find or obtain equivalent data, with 81% reporting that they would not be able to re-create the same or a similar dataset. Combining these two questions, Figure 17 shows the share of respondents reporting that it would not be possible to either (i) find or obtain equivalent data or (ii) recreate the same or similar most frequently used EMBL-EBI data resource. This has consistently increased, rising from 45% in 2016 to 67% based on the latest survey. This highlights how EMBL-EBI data resources increasingly facilitate access to data.

Figure 17 The share of respondents who could neither obtain elsewhere or re-create their EMBL-EBI data resources



Source: *EMBL-EBI User Survey; Beagrie (2016, 2021)*

Note: *The figure combines the findings from Question 23 and Question 26 of the 2025 user survey.*
Question 23 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to find and obtain equivalent data or tools from another source?" (N = 1,888)
Question 26 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to re-create the same or similar dataset or tool yourself?" (N = 1,799)

To understand the importance of EMBL-EBI data resources and tools, the survey then asked whether respondents would be able to take forward their work or study if they did not have access to EMBL-EBI data resources or tools. Figure 18 shows how 71% of respondents stated that it would either be impossible or would require significant additional time and effort for them to take forward their work without EMBL-EBI data resources. This compares to only 11% reporting that they could proceed with their work with similar time and effort. This highlights how EMBL-EBI data resources facilitate a significant volume of research that otherwise would not have been possible.

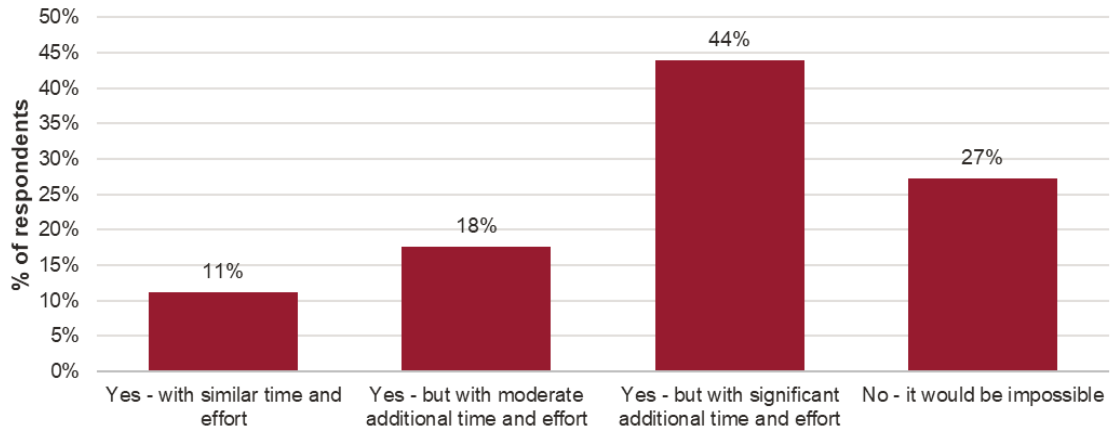
This finding is echoed in the survey free-text responses, with users reporting:

> *"Without these tools, in-depth bioinformatic analysis would be impossible"*

"UniProt is an extremely rich and well-maintained repository for protein information. Performing my work at the same level of quality will be impossible without UniProt."

"A significant amount of research would have to be completed to find several other sources that have the same creditability and accuracy as Ensembl."

Figure 18 Ability to take forward work or study without the use of EMBL-EBI data resources or tools



Source: *2025 EMBL-EBI User Survey*

Note: *Question 19 of the survey asked respondents "Would you be able to take forward your work or study without the EMBL-EBI data resources or tools you currently use?" (N = 2,015)*

The results from the user survey also demonstrate the scale of productivity gains felt by users from having access to EMBL-EBI data resources. When asked about the time taken to find and obtain their most frequently used EMBL-EBI resource, respondents reported a median time of 5 minutes, indicating the efficiency with which users can navigate EMBL-EBI resources.[34] By contrast, when asked about the expected time it would take to replicate an equivalent resource elsewhere, respondents reported much longer time periods, suggesting a median time of 8 hours to find or obtain an equivalent dataset or 150 hours to re-create the same or a similar dataset where this was deemed possible.[35,36] Future studies may also wish to explore how the median access times for EMBL-EBI data and substitute resources differs between user types (i.e., frequent versus infrequent users of EMBL-EBI data resources), thereby determining whether and how the efficiency gains vary between users.

[34] Question 21 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. Approximately how long did it take you to find and obtain the EMBL-EBI data resource or tool you were looking for?" (N = 1,075)

[35] Question 24 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. Approximately how long would it take you to find and obtain an equivalent data resource or tool from another source?" (N = 686)

[36] Question 27 of the survey asked respondents "Approximately how much time would it take to collect and recreate the data or tool you used most frequently in the last 12 months?" (N = 337)

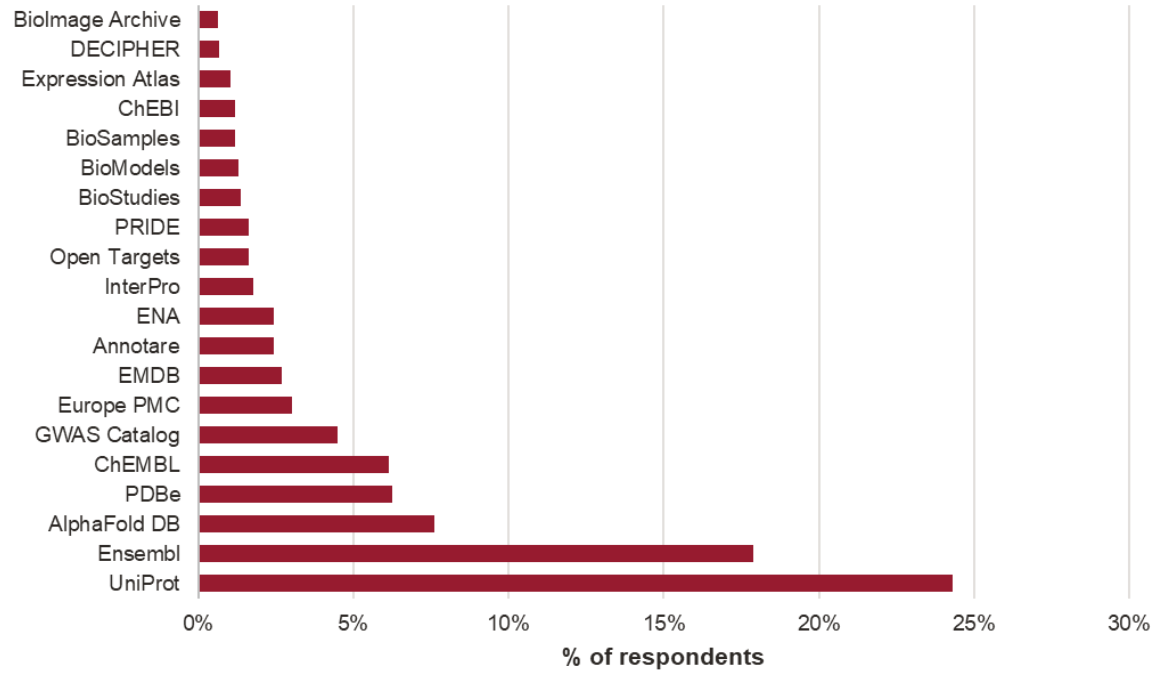
The free-text responses to the survey provide further insight into how EMBL-EBI data resources improve research efficiency and productivity:[37]

> *"Extracting suitable datasets on scale from the literature is theoretically possible, but would require *a lot* of time and set us back by years"*
>
> *"There are alternatives for all of the services we use, but they are much slower, less convenient and more difficult to access for us."*
>
> *"Having all genomes in one place in a standardized format is invaluable"*
>
> *"Without large databases provided by organizations like EMBL-EBI, it would be so time-consuming to pull out and validate information from individual data sources that only a fraction of my research would be possible"*
>
> *"Such well-curated datasets are very hard to find. This is an excellent resource that saves a lot of time."*
>
> *"Data curation is extremely time consuming. I expect it would take months to build resources to cover my own needs, which could then be used [regularly] in house only. EMBL resources have streamlined this process"*
>
> *"With access, I find what I need within about 1h/week across all databases I use; without it, the time required would easily increase to 10 fold"*

Finally, the survey highlights the wide range of EMBL-EBI data resources used by respondents. Figure 19 reports that 24% of respondents indicated UniProt as their most frequently used resource in the last 12 months, followed by Ensembl (18%), AlphaFold DB (8%), PDBe (6%) and ChEMBL (6%). These resources can all be used for a variety of disciplines. For example, UniProt is a comprehensive resource of protein sequences and functional annotation whereas Ensembl provides comprehensive genome data for vertebrates and other species.

[37] Question 19 of the survey asked respondents "Would you be able to take forward your work or study without the EMBL-EBI data resources or tools you currently use?"

Figure 19 Top 20 EMBL-EBI resources that users most frequently used



Source: *2025 EMBL-EBI User Survey*

Note: *Questions 20 of the survey asked respondents "Which EMBL-EBI data resource did you use most frequently in the last 12 months?" (N = 1,917).*

Insights from the open text responses to the user survey corroborate that EMBL-EBI data resources can be used across a range of disciplines and use cases. For example, when asked whether they would be able to take forward their work without EMBL-EBI data resources,[38] the following responses were received:

> *"I completely rely on ChEMBL as my source of bioactivity data for many aspects of the work I carry out as a comp chem consultant."*
>
> *"EMBL-EBI data resources are highly valuable and well-known stable sources for all scientist at all level[s], losing these tools would mean a disaster for the scientific community […] such as computation chemistry and biology, data science, crystallography etc."*
>
> *"Let's use AlphaFoldDB as an example. I have been using it with high schoolers that have never heard of bioinformatics or visualized a protein structure before. […] It makes it easy and relatively straightforward for a novice to access this information and learn from it."*
>
> *"My team and I use EMBL-EBI resources and tools in our bioinformatics training. These tools and resources allow us to provide biological examples, as well as tools and resources for bioinformatics, to our learners."*

[38] Question 19 of the survey asked respondents "Would you be able to take forward your work or study without the EMBL-EBI data resources or tools you currently use?"

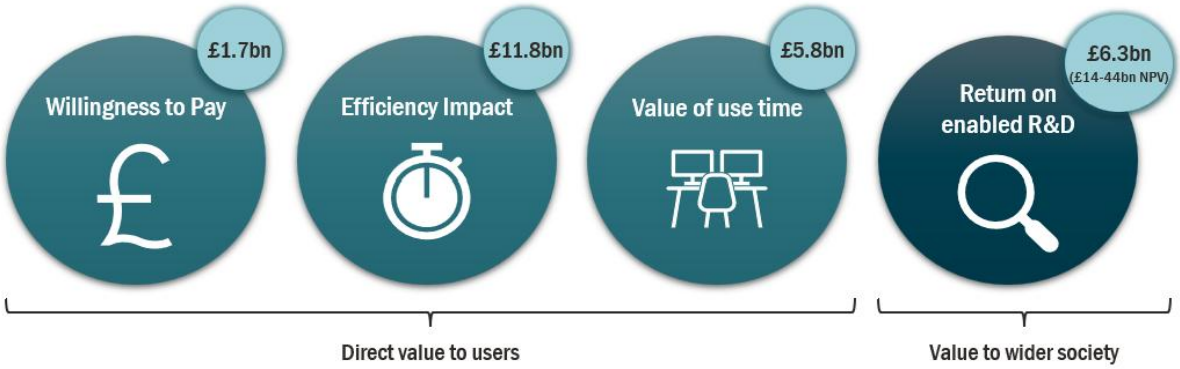
4 Impact modelling of the value of EMBL-EBI data resources

The impact modelling conservatively estimates the overall value of EMBL-EBI in terms of its direct impacts on users and related socio-economic impacts. The findings across the four value indicators are shown in Figure 20. Together, the findings reveal:

- **Users' stated willingness to pay (WTP) for EMBL-EBI data resources has continued to increase, rising from £1.5bn in 2021 to £1.7bn per year in 2025 in real terms**. As WTP directly reflects user valuations of the benefits that they experience from having access to EMBL-EBI data resources, this suggests these benefits have continued to rise. Users did, however, report that budget constraints as limiting their willingness to pay. This likely explains why the total WTP value is smaller compared to the remaining three value figures.
- **The total value of time users reported saving from having access to EMBL-EBI data resources comes to £11.8bn per year, an increase of 44% from 2021 in real terms.** Users reported saving 11 hours per week in the latest survey, up from 9 hours per week in 2021. This highlights the material impact on user productivity of having access to these resources.
- **The total value of time users spent working with EMBL-EBI data resources is estimated as £5.8bn per year.** This value has declined from around £6.7bn in 2021 in real terms as a result of a decline in the average time spent with EMBL-EBI data. There are multiple explanations that might underlie this trend:
  - On the one hand, decreases in time spent could reflect productivity improvements from EMBL-EBI data resources owed to shorter use journeys (as shown in indicator #2). Such a change would lead to a decrease in the value of use time indicator, despite this being driven by an improvement in the service.
  - On the other hand, decreases in time spent could reflect a decrease in the value derived from EMBL-EBI data resources. The latter explanation would however be at odds with the willingness to pay findings in indicator #1 above.
  - Other explanations for the decline in use time might be (i) a substitution towards usage via secondary resources, which are not considered in the survey results on time spent or (ii) a reduction in online-based research since the pandemic due to a move back to lab-based working (possibly explaining the decline since 2021);
  - Given the potential ambiguity, this highlights how the value of use time indicator is an imperfect proxy of value and must be considered in the context of the other value indicators.
- **The return on R&D enabled by having access to EMBL-EBI data resources comes to £6.3bn annually, or between £14bn and £44bn in net present value (NPV) terms over 20 to 30 years**. This has increased from £2.1bn in 2016 and £4.9bn in 2021. This metric reflects the return to R&D arising from the time saved from having access to EMBL-

EBI data resources (indicator #2) plus the total time spent on work that could not otherwise have taken place without EMBL-EBI data resources (a portion of indicator #3).

Figure 20 Annual value of EMBL-EBI data resources



Source: Frontier Economics

Overall, we see that the direct and wider societal value of EMBL-EBI data resources has continued to increase over time across most indicators. The following sub-sections include a more detailed breakdown of these findings, including:

- The trends over time;
- The factors underlying these trends; and
- The BCRs.[39]

4.1 Annual willingness to pay

Users' WTP directly reflects their valuation of the benefits that they experience as a result of having access to EMBL-EBI. It captures in one metric a wide range of potential outcomes such as improvements in individuals' access to data, efficiency and research quality. It is, however, limited by the budgets of participants and/or their organisations.

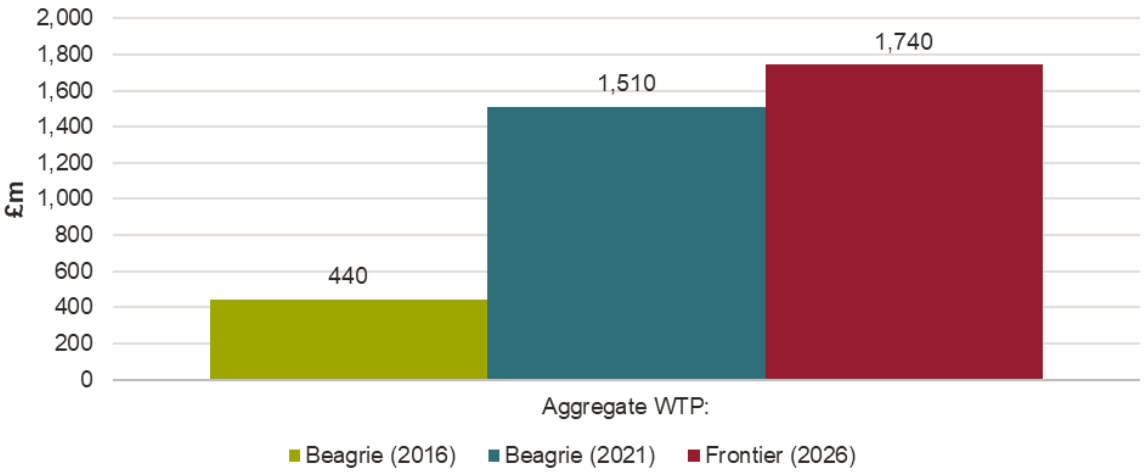
On average, the user survey respondents reported that they would be willing to pay £3,352 per year for access to EMBL-EBI data resources.[40] This results in an aggregate WTP across all EMBL-EBI users of £1.7bn per year. Consequently, the benefits are 16 times the annual cost of maintaining and making available EMBL-EBI data resources (i.e., the BCR is 16).

[39] See Annex B.6 for an overview of how the cost of running EMBL-EBI's open data resources has evolved over time.

[40] The LLM analysis also revealed that that 17% of users who responded to the open text question admitted to guessing, using intuition, or not having a clear rationale for their answer. There is a risk that unreliable estimates resulting from guesswork may introduce bias into our findings. Therefore, on a conservative basis, we include a sensitivity analysis that removes all "guess" responses when estimating the average willingness to pay. This sensitivity is outlined in Annex A but does not have a material impact on results.

Figure 21 shows that users' aggregate willingness to pay has increased since the previous Beagrie (2016, 2021) findings, rising from £1.5bn in 2021 to £1.7bn in 2025 in real terms. This increase in value is driven by both (i) a rise in the real average WTP (from £3,335 in 2021 to £3,352 in 2025), and (ii) an increase in the user base, with the latter effect being the dominant driver. Taken together, this shows how the benefits users derive from accessing EMBL-EBI data resources have continued to rise.

Figure 21 Aggregate annual willingness to pay of EMBL-EBI users



Source: 2025 EMBL-EBI User Survey; Beagrie (2016, 2021)
Note: All values are shown in 2025 real terms based on an adjustment using the latest UK GDP deflators.

Insights from the LLM analysis show that 38% of users who responded to the open text question when estimating their WTP base their answer on what they or their institution could realistically afford. They often referenced limited funding, grant restrictions, or personal income.[41] This finding suggests that average WTP would be even higher if users' individual budget constraints were relaxed. For example, when asked to report how they estimated their WTP, users responded:

> *"It is invaluable, but I don't have my own funding"*
>
> *"Largely based on the funds I have available"*
>
> *"It's based on how much money I would be able to secure from a grant. The value could easily be 100000 DKK [i.e. > £11k] and still willing to pay, as it would save me so many hours of work."*
>
> *"This is how much I would personally pay for my own individual access. My employer would likely be willing to pay more."*

[41] Question 39 of the survey asked respondents "How have you estimated the value you provided in the previous question?" (N = 571).

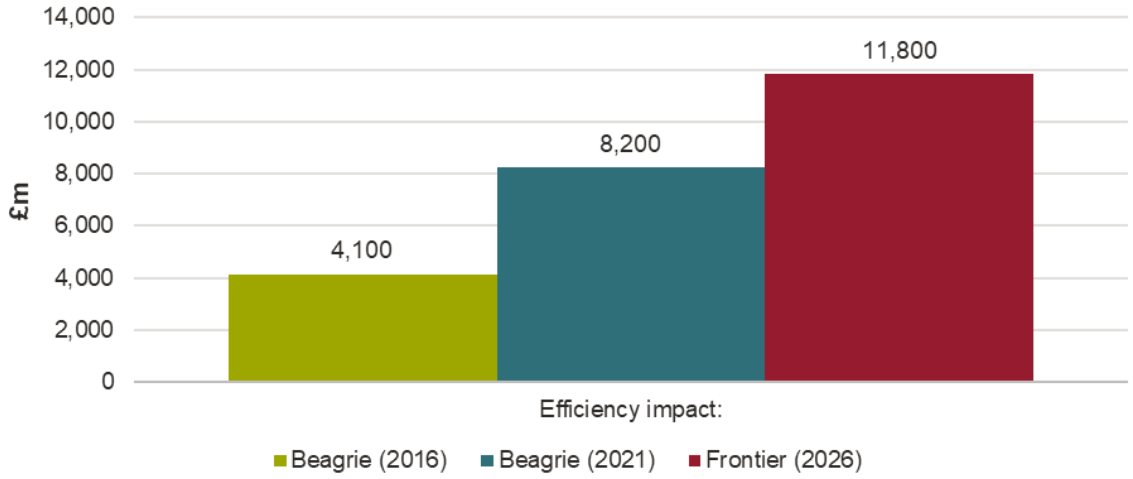
4.2 Annual efficiency impact

The efficiency impact reflects the time saved by users from having access to EMBL-EBI data resources. In other words, it indicates the broader productivity gains from the additional, potential research time researchers now have available from accessing these resources.

On average, the user survey respondents reported that they save 2.2 hours per day (or 11 hours per week) from having access to EMBL-EBI data resources.[42] Applying these productivity savings across the estimated EMBL-EBI user base and applying an assumption around the value of users' time, we estimate the overall efficiency impact from time saved at £11.8bn per year. This is 108 times the annual cost of maintaining and making available EMBL-EBI data resources (i.e., the BCR is 108).

Figure 22 shows that the efficiency impact has increased since the previous Beagrie (2016, 2021) reports, rising from around £4.1bn in 2016 and around £8.2bn in 2021. The continued increase in 2025 is driven by both the increase in the reported time savings (from around 9 hours per week in 2021) as well as the increase in the estimated user base. This highlights how EMBL-EBI data resources have continued to improve user productivity over the past four years.

Figure 22 Aggregate annual efficiency impact



Source: *2025 EMBL-EBI User Survey; Beagrie (2016, 2021)*

Note: *All values are shown in 2025 real terms based on an adjustment using the latest UK GDP deflators. Beagrie (2016, 2021) reported a range of possible values for the efficiency impact as a result of ambiguity in the framing of the relevant survey questions. So that a comparison can be made, we use the midpoint of the ranges reported by Beagrie. See Annex A for further detail.*

[42] Insights from the LLM analysis reveal that 27% of users who responded to the open text question when estimating their average time saving stated they could not estimate, found the question impossible, irrelevant, or refused to answer, often citing the essential nature of the resources or the difficulty of imagining an alternative. In addition, 11% of responses were based on guesswork. There is a risk that unreliable estimates may introduce bias into our findings. Therefore, on a conservative basis, we include a sensitivity analysis that removes these responses. This sensitivity is outlined in Annex A and does not have a material impact on the results.

The average efficiency savings reported by users (11 hours per week) account for over a third of the average time spent conducting research per week (which was 31 hours according to the user survey). This time saving can be invested into other productive research tasks or working fewer hours. Although we cannot know for sure how users reinvest this additional time saving, the results of the user survey suggest that, on average, users spend the same amount of time conducting research each week (31 hours) as they did in 2021 (Beagrie, 2021). Taken together, the findings of (i) an increase in time saved, and (ii) the same overall time spent on research per week, suggest that the additional hours made available by having access to EMBL-EBI data resources is likely realised as more productive research time (as opposed to working fewer hours).

4.3 Annual value of time using EMBL-EBI data resources

This indicator uses the time users spend working or accessing EMBL-EBI data resources (value of use time) to assess their economic value. This is because users must, as a minimum, value the resources equal to the value of their time they spend accessing and/or using them. Otherwise, it would not be worth their while and they would spend their time on more valuable activities elsewhere. We present two ways of calculating this metric, reflecting different approaches to assessing the time users interact with EMBL-EBI data resources. To provide a conservative estimate for the minimum value users attribute to EMBL-EBI, the value of access time captures the time spent finding and obtaining EMBL-EBI data resources. To provide a more comprehensive insight into the time spent using EMBL-EBI data, the value of use time captures the time spent finding, obtaining and analysing those data resources.

4.3.1 Value of access time

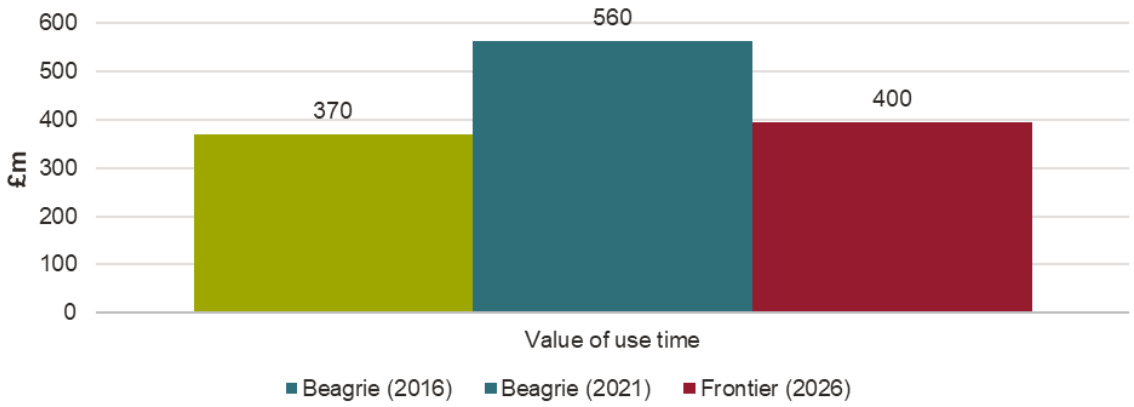
To calculate the value of access time, the median time users spend accessing EMBL-EBI data resource is used (5 minutes).[43] This estimate is combined with the median number of times that survey respondents accessed EMBL-EBI data resources (199 times per year). Applying this use time across the estimated EMBL-EBI user base and applying an assumption around the value of users' time, the value of access time is estimated at around £400m per year. This is 4 times the annual cost of maintaining and making available EMBL-EBI data resources (i.e., the BCR is 4).

Figure 23 shows that the value of access time has increased from around £370m in 2016, but declined from around £560m since 2021. The increase in value since 2016 is primarily driven by the sharp increase in the user base from around 198k to around 520k. However, the decline in value since 2021 is driven by the decline in the median number of annual accesses per

[43] The LLM analysis reveals that 12% of users who responded to the open text question when estimating the length of time it took them to find EMBL-EBI resources expressed confusion, uncertainty, or inability to answer due to not understanding the question or lacking enough experience. Similarly, 6% of responses were based on guesswork. Therefore, on a conservative basis, we include a sensitivity analysis that removes these responses to improve the overall accuracy of the findings. This sensitivity is outlined in Annex A and does not affect the median access time.

user.[44] Meanwhile, the median time spent accessing EMBL-EBI data resources is unchanged from the previous Beagrie (2016, 2021) findings.

Figure 23 Annual value of access time



Source: 2025 EMBL-EBI User Survey; Beagrie (2016, 2021)
Note: All values are shown in 2025 real terms based on an adjustment using the latest UK GDP deflators.

Although it is not possible to determine from the survey why the median number of annual accesses per user has decreased, possible explanations include:

- There has been a reduction in online-based research since the pandemic due to a move back to lab-based working (possibly explaining the decline since 2021);
- Users are better able to retrieve the EMBL-EBI data resources they need in fewer sittings, reflecting an improvement in the service provided;
- Secondary usage has increased, thereby reducing the rate of direct data accesses to EMBL-EBI data resources, and/or
- Users derive less value from EMBL-EBI data resources, therefore reducing the time spent using them.

Further research would be required to understand what is driving this reduction in the median number of annual accesses.

The efficiency of the time taken to access EMBL-EBI data resources is emphasised throughout the open text questions. For example, when asked to estimate how long it took them to find or obtain their most frequently used EMBL-EBI resource, users responded:

> *"It's as easy as a google search"*
>
> *"I use this all the time and it is very fast and easy to use"*

[44] Beagrie (2016) reported that the median number of times that the survey respondents accessed EMBL-EBI data resources was 445 per year. Beagrie (2021) do not report the median number of times that the survey respondents accessed EMBL-EBI data resources. However, we can determine that the value has fallen since 2021 given that the overall use value has declined (as all other parameters are equal).

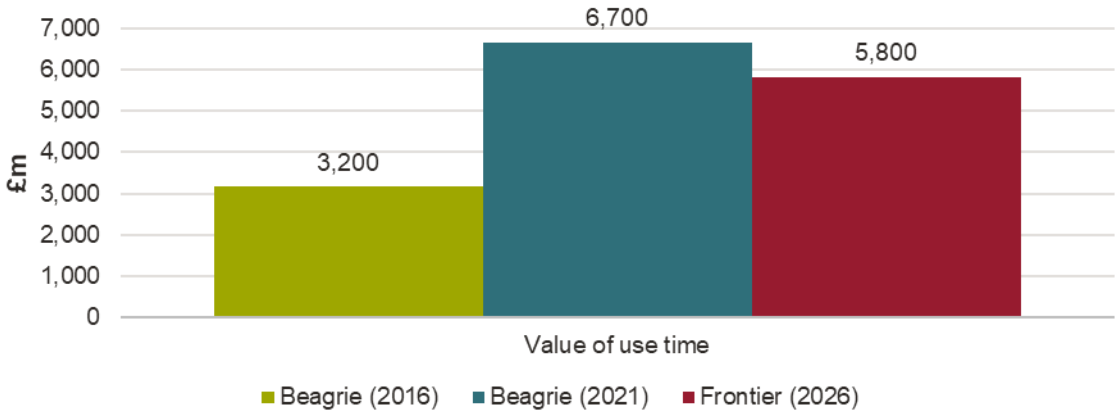
"Easy to search tool, takes minutes"

4.3.2 Value of use time

To calculate the value of use time, the mean time users spend working with EMBL-EBI data resources is used (1.1 hours per user per day, or 5.4 hours per week). Applying this use time across the estimated EMBL-EBI user base and applying an assumption around the value of users' time, the value of use time for the wider user population is around £5.8bn per year. This is 53 times the annual cost of maintaining and making available EMBL-EBI data resources (i.e., the BCR is 53).

Similar to the value of access time, Figure 24 shows that the value of use time has increased from around £3.2bn since 2016. This is driven by the sharp increase in the user base, as the average time spent with EMBL-EBI data has decreased from 6.8 hours per week in 2016. However, the value of use time has declined from around £6.7bn since 2021 as a result of the decline in the average time spent with EMBL-EBI data from around 7.1 hours per week.

Figure 24 Annual value of use time



Source: 2025 EMBL-EBI User Survey; Beagrie (2016, 2021)
Note: All values are shown in 2025 real terms based on an adjustment using the latest UK GDP deflators.

In a similar vein to the value of access time, it is not possible to determine from the survey why the average time spent with EMBL-EBI data per week has decreased. The same explanations as listed in Section 4.3.1 also apply here.

Across both the value of access and use time findings, the decreases in time spent could reflect productivity improvements from EMBL-EBI data resources (as shown in the annual efficiency impact indicator above). Equally, it could reflect a decrease in the value derived from EMBL-EBI data resources. The latter explanation would however be at odds with the willingness to pay indicator above. Overall, this highlights how these indicators are imperfect proxies of value, and must be considered in the context of the other value indicators presented.

4.4 Return on R&D enabled by EMBL-EBI data resources

The returns on R&D enabled by EMBL-EBI data resources (return on enabled R&D) capture the wider societal impacts from researchers having access to EMBL-EBI data resources. It brings together the annual efficiency impact and value of use time indicators, highlighting the potential returns to R&D arising from the research time that otherwise could not have taken place without users having access to EMBL-EBI data resources.

We estimate the annual value of this return by:

- Summing (i) the value of time spent with EMBL-EBI data resources that otherwise would not have occurred (described in Section 4.3 above), and (ii) the efficiency impact (described in Section 4.2);
- Multiplying the result by an average social rate of return to investment in R&D of 40%.[45]

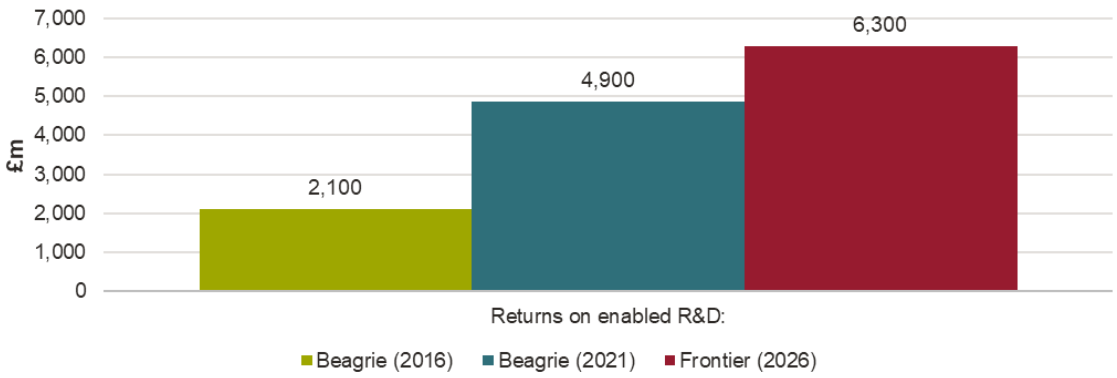
The resulting estimate is that the annualised returns on R&D enabled by EMBL-EBI data resources are around £6.3bn. We assume that the research enabled by EMBL-EBI generates returns over many years, with the exact time profile depending on the share of the R&D that takes place in the public sector versus the private sector.[46] To capture how the NPV of the returns might vary depending on the mix of public versus private sector research enabled by EMBL-EBI, we assess the NPV over a range of 20 to 30 years and consider multiple possible depreciation rates. As a result, we estimate that the NPV of the returns sits within the range of £14bn and £44bn across 20 to 30 years.

Figure 25 shows that the annualised returns on enabled R&D have increased from £2.1bn in 2016 and £4.9bn in 2021 to £6.3bn in 2025. It highlights the significant value of research that has likely been directly enabled as a result of users having access to EMBL-EBI data resources. This increase since 2021 is primarily the result of the increase in the efficiency impact of EMBL-EBI data resources described in Section 4.2 above.

[45] See Annex B for an explanation of the assumptions made in this section.

[46] As described in Annex B.5, in comparison to private sector research, public sector research might deliver benefits for a longer time period and depreciate more slowly if it is aimed at building general purpose knowledge that can be re-used and built upon.

Figure 25 Annual returns on enabled R&D



Source: *2025 EMBL-EBI User Survey; Beagrie (2016, 2021)*

Note: *All values are shown in 2025 real terms based on an adjustment using the latest UK GDP deflators. This indicator was not reported by Beagrie (2016, 2021) and is therefore reconstructed based on the findings around the value of use time, the share of usage that could not have otherwise occurred, and the efficiency impact.*

5 Open data case study: the AlphaFold Database and its impact on research

5.1 Background and hypotheses to test

Protein folding is fundamental to understanding biological processes, and accurate structural predictions can accelerate advances in health research and drug discovery. This is because many diseases involve proteins folding incorrectly. Developed by Google DeepMind in 2021, AlphaFold 2 is an artificial intelligence-based system capable of predicting a protein's three-dimensional structure from its amino acid sequence. Google DeepMind trained the AlphaFold 2 algorithm on publicly-available data, including data from EMBL-EBI data resources.

Prior to AlphaFold 2's release, scientists had experimentally determined the structures of around 180,000 proteins. AlphaFold 2 revolutionised the field by enabling accurate predictions for over 200 million proteins, dramatically expanding the scope of structural biology.[47] While there were systems before AlphaFold 2, their accuracy and speed limited their impact.

To make these predictions widely accessible, EMBL-EBI collaborated with Google DeepMind to create the AlphaFold Database. This database enhances the utility of the algorithm by curating, organising, and integrating its predictions with existing biological resources such as the Protein Data Bank (PDB) and UniProt. By doing so, it removes the need for researchers to have the expertise to run the algorithm, integrate this themselves and have access to high-performance computing resources. Instead, scientists can directly explore and apply AlphaFold 2's predictions within their research workflows.

The direct impact of the AlphaFold 2 algorithm has been estimated previously and recognised in 2024 with the Nobel Prize for Chemistry.[48] This case study, however, tests two hypotheses pertaining to the impact of the AlphaFold Database, compared to a counterfactual world in which only the AlphaFold 2 algorithm exists:

- **Distributional effect**: the creation of the AlphaFold Database has widened access to the algorithm's predictions to a more diverse array of research fields.
- **Scale effect**: the creation of the AlphaFold Database has led to greater use of the algorithm's predictions, resulting in more research being conducted.

More generally, this case study provides evidence as to the downstream impacts arising from making new AI-system-derived outputs easily accessible. In doing so, it assesses whether the long-term outcomes set out in the logic model in Section 2 are likely to be realised.

[47] See AlphaFold Database, "About."

[48] https://deepmind.google/blog/alphafold-five-years-of-impact/

Figure 26 summarises the findings from this case study, which are outlined in more detail in the following sections.

Figure 26 Open data: The AlphaFold Database and its impact on research

Source: Frontier Economics

5.2 Methodology

To test the two hypotheses above, we analyse bibliometric data covering papers that mention either the AlphaFold Database or the Protein Data Bank (PDB).[49] PDB is also a publicly accessible database managed by the wwPDB consortium, which includes EMBL-EBI. The characteristics of these papers were extracted from OpenAlex, an open repository with enriched information covering paper topic; authorship; institution; location published and many other characteristics.[50] By comparing papers that mention the AlphaFold Database to those that mention PDB, we investigate how the AlphaFold Database has influenced research since its release.

PDB was chosen as the baseline for assessing the nature of research conducted pre-AlphaFold 2 and post-AlphaFold 2 for two reasons:

1. It represented the primary source of validated 3D macromolecular structures for proteins and nucleic acids before the release of the AlphaFold Database in 2021.
2. PDB is also a publicly accessible database managed by EMBL-EBI.

Its similarity therefore makes it a suitable baseline to assess how research has evolved since the AlphaFold Database was released.

We hypothesised that the papers mentioning the AlphaFold Database may overlap with those referencing the AlphaFold 2 algorithm. To isolate the effect of the database from the algorithm,

[49] Data listing the DOIs of papers mentioning either the Protein Data Bank (PDB) or the AlphaFold Database from January 2018 to July 2025 is used. This data was captured by EMBL-EBI, using text mining approaches.

[50] See OpenAlex, here

we also extracted information from OpenAlex for papers citing the AlphaFold 2 algorithm.[51] We then compare papers that mention only the AlphaFold Database with papers that mention only the algorithm, to assess the degree of overlap and difference.

The following sub-sections set out the quantitative methods used to test these hypotheses as well as the results.

5.3 Hypothesis 1: distributional effect

Do papers mentioning the AlphaFold Database cover different topics to those mentioning the algorithm and PDB?

To establish a baseline, the topics of papers mentioning PDB before the AlphaFold Database was released (in 2021) are compared with topics of papers mentioning PDB after the AlphaFold Database was released. This first step shows how topics of research have evolved over time and their degree of difference. We then compare the baseline degree of topic overlap with the AlphaFold Database, to reveal how different or not the AlphaFold Database topics are to pre-2021 PDB topics.

To do this, we calculate the Hellinger Distance (HD).[52] The HD measures the difference between the distributions of research topics for papers published using the different resources. It lies between 0 and 1, with 0 indicating the papers mentioning each resource cover the exact same topics, and 1 that the topics do not overlap at all.

Figure 27 presents the HD between pre-2021 PDB topics and post-2021 PDB topics in the baseline, yielding a score of 0.33. It also shows the HD between the AlphaFold Database and pre-2021 PDB topics, revealing a score of 0.44. As the HD is higher than in the baseline (0.44>0.33) and is highly statistically significant at the 95% level, this indicates that there is less topic overlap between research mentioning the AlphaFold Database and research mentioning PDB before its release, than in the baseline.

We then compare the topics researched by papers mentioning the AlphaFold Database to those mentioning PDB since AlphaFold 2's release. Figure 27 shows that the topics continue to be different to those mentioning PDB since AlphaFold 2's release, with the difference even higher than before 2021 (as the 0.47 score is high, relative to the differences cited above). In other words, topics covered by papers citing the AlphaFold Database are even more different to research mentioning PDB since AlphaFold 2's release, when compared to the baseline.

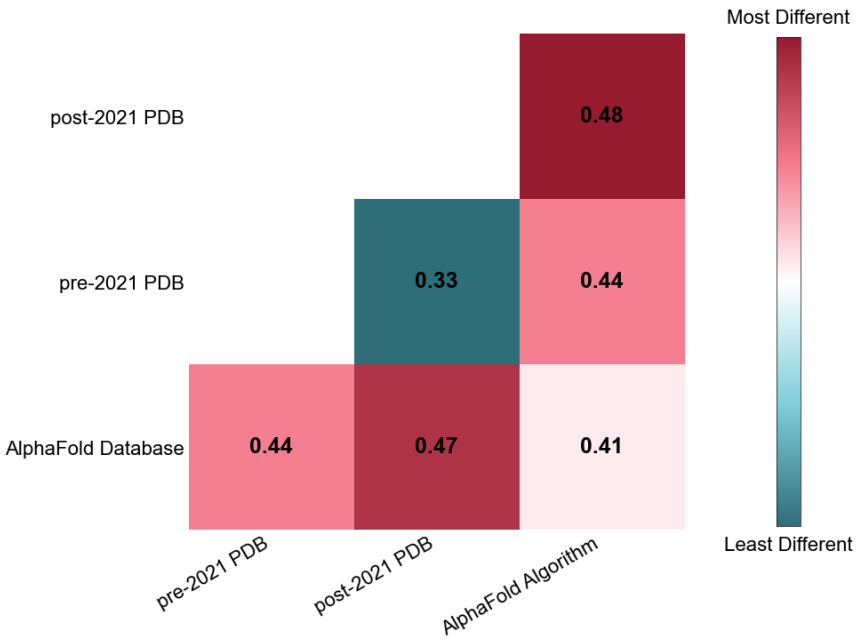
[51] We used papers citing the original AlphaFold 2 paper and/or ColabFold paper. See Jumper, Evans, Pritzel *et al.* (2021), "Highly accurate protein structure prediction with AlphaFold"; and Mirdita, Schütze, Moriwaki *et al.* (2022), "ColabFold: making protein folding accessible to all."

[52] The Hellinger Distance (and Shannon Index, discussed later on) are both distribution-level statistical measures, meaning they help summarise the properties of an entire distribution. They are particularly suited to the highly rich and detailed topic information provided by OpenAlex. Further detail on both metrics and our sampling approach are contained in Annex D .

To distinguish the impact of the AlphaFold Database from the algorithm, we assess the topic overlap between the two and compare this to the baseline. As shown in Figure 27, the lowest HD for both the database and the algorithm is between each other (0.41). This is unsurprising given their shared origin, as both are AlphaFold 2 predictions in different forms. However, the value of 0.41 is still statistically significantly higher than the HD between pre- and post-2021 PDB (0.33).[53] This suggests that AlphaFold Database paper topics are more different to AlphaFold 2 algorithm paper topics, relative to how topics of research have evolved over time in the baseline. This indicates that the database is not proxying for the algorithm, but actually covering different research topics.

Taken together, the evidence suggests that the AlphaFold Database has had a distributional effect on the topics of research conducted. By hosting AlphaFold 2 predictions openly in the form of the AlphaFold Database, the evidence suggests protein structure data is being used across a wider range of research topics, compared to a counterfactual world in which only the AlphaFold 2 algorithm exists.

Figure 27 Matrix of all Hellinger Distances between the AlphaFold Database, the AlphaFold 2 algorithm, and PDB topics



Source: Frontier Economics

Do papers mentioning the AlphaFold database cover a more diverse range of topics than those mentioning the algorithm alone?

As the AlphaFold Database curates, organises, and integrates the algorithm's predictions, the topics researched by papers mentioning the AlphaFold Database are expected to be more

[53] This difference is significant at the 95% confidence level.

diverse than those mentioning the AlphaFold 2 algorithm. This is because researchers do not need to have the expertise to run the algorithm and have access to high-performance computing resources, which users of the algorithm would require.

To measure topic diversity, we use a version of the Shannon Index (SI) called the Number of Effective Categories[54] (N_{eff}). N_{eff} measures both the concentration and number of unique categories, where a category is a topic or published location. A higher value for N_{eff} indicates that the papers mentioning each tool are more diverse than papers with a lower value for N_{eff}. To ensure a fair comparison, we take a representative sample of papers so there are an equivalent number of papers used to calculate the N_{eff} for each tool. See Annex D for further information.

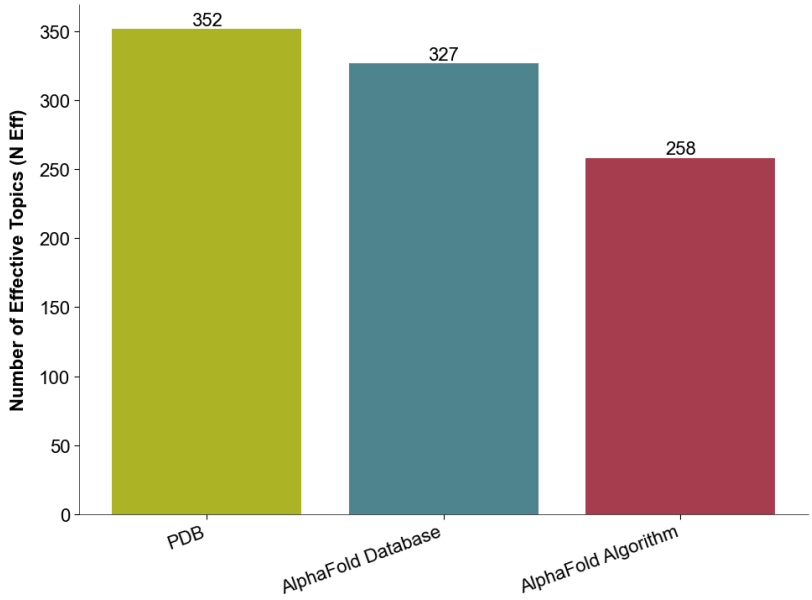
Figure 28 shows the Effective Number of Topics (N_{eff}) for PDB post-2021, the AlphaFold Database and the algorithm. The AlphaFold Database has a statistically significantly higher N_{eff} value (327) than the AlphaFold 2 algorithm (258).[55] The diversity of research topics is therefore higher for papers mentioning the database than the algorithm, suggesting that the database has had a distributional impact on the topics of research conducted.

We also see that PDB has a statistically significantly higher N_{eff} value (352) than the AlphaFold Database (327). While the results in the previous subsection suggest PDB post-2021 and AlphaFold Database papers cover different topics to each other (relative to the baseline), the diversity of topics within each database is higher in PDB than the AlphaFold Database. While this may reflect the maturity of PDB, given it has been available for a much longer period than the AlphaFold Database, it is not possible to say for certain.

[54] Where categories are research topics, publish locations, or author nationalities depending on the variable of interest.

[55] This difference is significant at the 95% level.

Figure 28 The Effective Number of Topics researched using the AlphaFold Database, the AlphaFold 2 algorithm, and PDB



Source: Frontier Economics

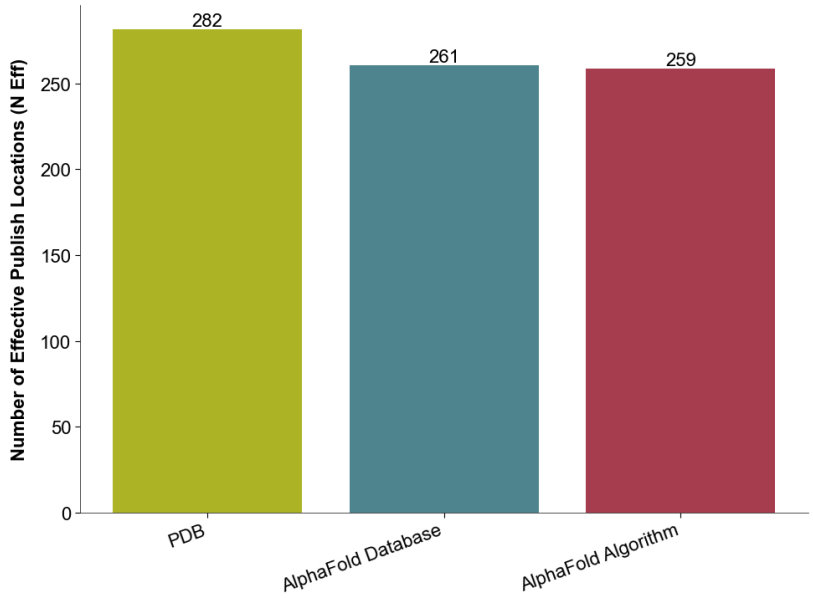
Are papers mentioning the AlphaFold Database published in a wider range of locations than the algorithm?

For similar reasons to topic diversity, papers mentioning the AlphaFold Database are expected to be published in a more diverse array of journals than research mentioning the AlphaFold 2 algorithm. This is because if the database makes it easier for researchers from more fields to access AlphaFold 2 predictions, then the final published locations of their research should also reflect a broader range of fields.

Figure 29 shows the N_{eff} for published location for PDB post-2021, the AlphaFold Database and the algorithm. The diversity of published locations is similar between the AlphaFold Database (261) and algorithm (259). This, and the greater diversity of published locations covered by the more established PDB (282), seem to be at least partly driven by the recency of both AlphaFold 2 tools, which manifests in both having a high proportion of papers published in pre-print journals. So while the diversity of topics is higher in the AlphaFold Database than the algorithm (as per the previous section), the diversity of published locations is similar.

This result may reflect the fact that the algorithm and database are still relatively new. If these pre-prints are then published in topic-specific journals in future, this result may change. For now, only the topic-based diversity measure supports the distributional hypothesis.

Figure 29 The Effective Number of Published Locations for the AlphaFold Database, the AlphaFold 2 algorithm, and PDB



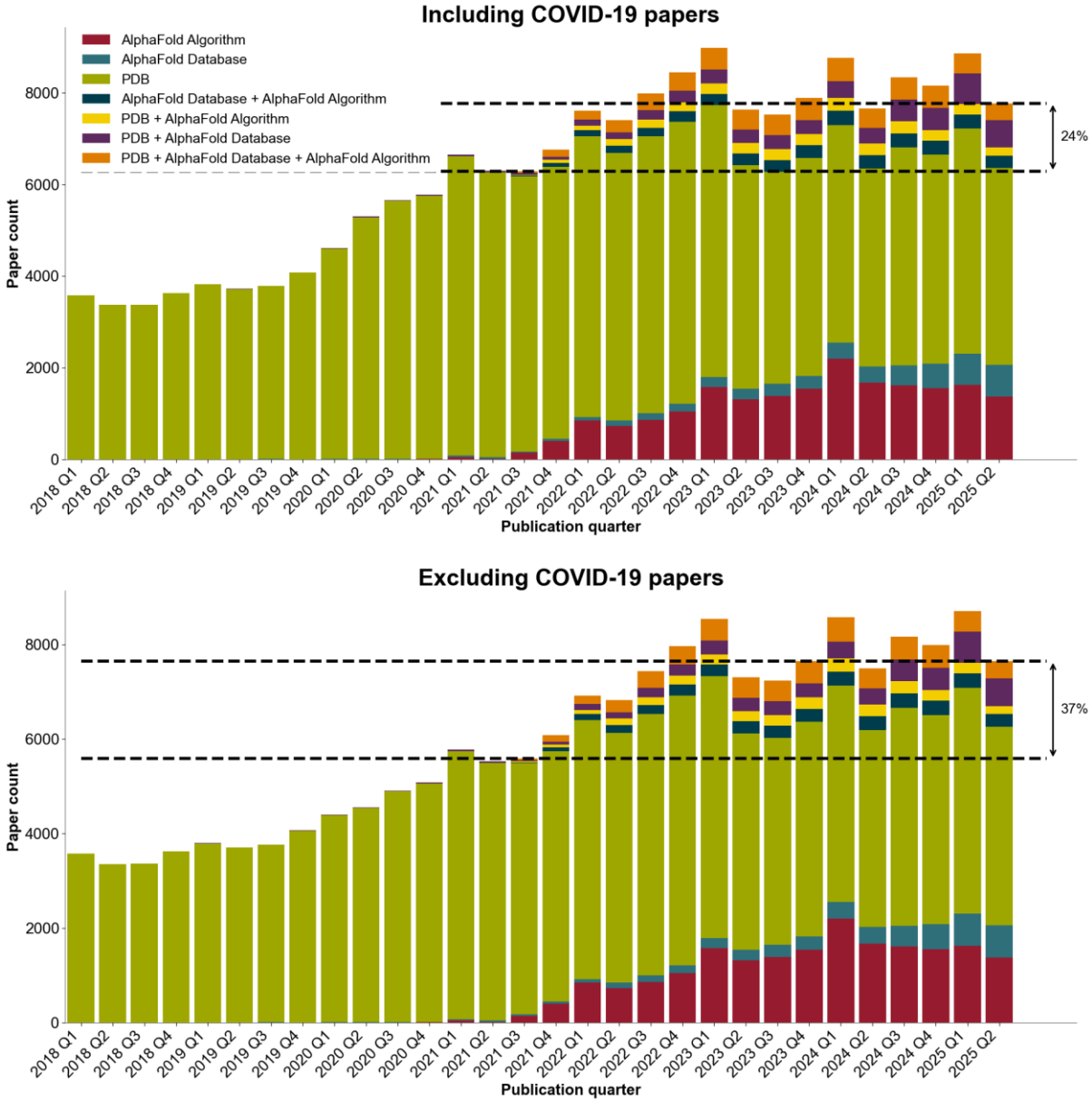
Source: Frontier Economics

5.4 Hypothesis 2: scale effect

To test whether the release of the AlphaFold Database has increased the scale of research conducted, we plot the number of new papers published each month for PDB, the AlphaFold Database and the AlphaFold 2 algorithm. We also indicate where papers mention more than one tool (e.g., AlphaFold Database and PDB).

As shown in the top chart of Figure 30, since the release of the AlphaFold Database in July 2021, the total quantity of research papers mentioning one or more of these resources has increased by 24% (see black dotted lines). Papers that only mention the AlphaFold Database constituted 9% of all papers published in Q2 2025 in this set, and a further 16% in the same quarter mention AlphaFold Database in combination with other resources. This proportion seems to be growing since Q2 2024. Simultaneously, the quantity of new papers that only mention the AlphaFold 2 algorithm seems to be falling over the same period, potentially reflecting either a shift in researchers' usage from the algorithm to the database and / or a move to AlphaFold 3 or AlphaFold-Multimer.

Figure 30 Quantity of papers published by quarter and research tool combination



Source: Frontier Economics

As the COVID-19 pandemic may have led to a temporary increase in the volume of protein research between 2019-2022, we exclude papers tagged with the topic "SARS-CoV-2 and COVID-19 Research" from the dataset. If this were the case, this would make the volume of pre-AlphaFold 2 research appear higher than it would have otherwise have been, without the pandemic. This would obfuscate the effect of AlphaFold 2 on the scale of research.

The bottom graph shows that, with COVID-19 topics removed, the total quantity of research papers since the release of the AlphaFold Database has increased by 37% (see black dotted lines). This is materially higher than the 24% increase shown before.

Finally, the quantity of papers mentioning PDB (either individually or in combination with other resources) has actually decreased by 1% since the release of the AlphaFold Database (excluding the COVID-19 topics), making up 70% of total papers published in Q2 2025. Over the same period, the number of papers that mentioned the AlphaFold Database (either individually or in combination with other resources) has grown to 25% of total papers published in Q2 2025. This suggests that the AlphaFold Database is not a substitute for PDB. While it is not possible without using more causal methods to say exactly what proportion of these papers would still have been produced had the database not existed, it is clear that the AlphaFold Database is supporting the generation of net new research, and playing an increasing role over time.

5.5 Conclusions

AlphaFold 2 represents a major breakthrough in the use of AI in protein research. The results of this case study indicate that the release of the AlphaFold Database has widened access to the algorithm's predictions, leading to a more diverse array of research fields publishing papers mentioning the database. They also suggest that the database has led to greater use of the algorithm's predictions, resulting in more research being conducted. Taken together, the results demonstrate how open data infrastructure can amplify the impact of major scientific breakthroughs by making them more widely accessible.

While the evidence supports the hypothesis that the AlphaFold Database has had a scale and distributional effect on research, it is not possible to make causal statements. To do so would require a suitable counterfactual to be constructed and econometric techniques for causal inference applied. These methods were outside of the scope of this case study. Future research could seek to apply these methods, isolating the impact of the AlphaFold Database from other factors shaping research trends over the same period.

6 Data resources and standards case study: their role in European genomic medicine programmes

6.1 Background and hypotheses to test

Genomic sequencing is a technique used to determine the complete or partial DNA sequence of an individual's genome, allowing detailed analysis of genetic variation. By examining either the whole genome or specific genomic regions, this approach can detect disease-causing variants that are not identifiable through conventional diagnostic methods. As a result, genomic sequencing is becoming an increasingly central tool for the diagnosis of rare diseases and other genetic diseases, such as cancer.

EMBL-EBI provides access to a range of data resources, tools and data standards that are relevant to national genomic medicine programmes. These include:

- **Ensembl:** This public and open data resource provides access to genomes, annotations, tools and methods. Its goal is to enable genomic science by providing high-quality, integrated and consistent annotation on all cellular genomes, providing the coordinates and biological context needed to interpret where variants fall in the genome.[56]
- **Ensembl Variant Effect Predictor (VEP):** This tool annotates and predicts the effect of individual variants on gene transcripts, protein sequences, and regulatory regions. It reports reference data to facilitate variant prioritisation and interpretation, including the frequency of observed variants within human populations.[57]
- **The Matched Annotation from the NCBI and EMBL-EBI (MANE) Standards:** MANE is a collaborative project to ensure that genes and transcripts are described consistently across genome browsers, databases, and clinical tools. For almost every gene in the human genome, the project has agreed on a single "preferred" transcript that is identical in both the EMBL-EBI-led GENCODE gene set and its counterpart RefSeq gene set, provided by the US institution The National Center for Biotechnology Information (NCBI). This transcript, called MANE Select, is intended to be the default for displaying genes and reporting variants in clinical sequencing. A second category of transcripts - MANE Plus Clinical - have been curated to include additional transcripts which contain clinically relevant variants. MANE builds on earlier work to create stable, trusted reference sequences for clinical variant reporting.[58]
- **DECIPHER:** This tool is used by the clinical community to share and compare phenotypic and genotypic data to support the diagnosis of rare genetic disorders. It enables

[56] See Ensembl (2026), "Homepage."

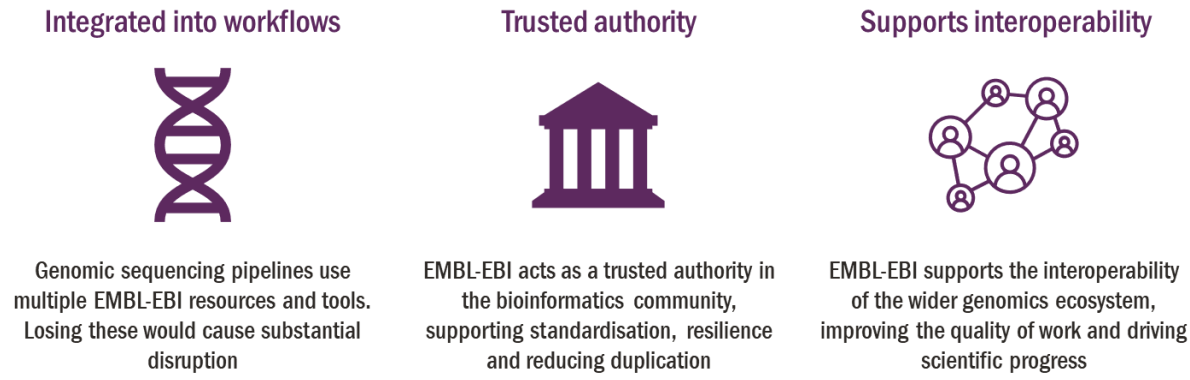
[57] See Ensembl (2026), "Ensembl VEP."

[58] EMBL-EBI (2026), "Genome Interpretation."

comparison of variants from genome or exome sequencing across patients, helping identify pathogenic variants through genotype–phenotype matching.[59]

This case study tests the hypothesis that EMBL-EBI data resources, tools and standards like the above are used by, and have a positive impact on, the genomic sequencing pipelines of national genomics programmes. In particular, it examines the downstream impacts arising from EMBL-EBI's role in (i) making genomic data resources and tools freely available, and (ii) developing and promoting harmonised global standards. Through this analysis, the case study assesses whether the impacts set out in the logic model in Section 2 are likely to be realised (e.g., relating to the development of better quality, wider variety and cheaper products and services leading to better societal outcomes).

Figure 31 Data resources and standards: their role in European genomic medicine programmes



Source: Frontier Economics

6.2 Methodology

To qualitatively assess EMBL-EBI's contribution to national genomics programmes, we conducted two 1-hour long interviews with a total of four bioinformaticians at the following genomics institutions:

- **Genomics England:** owned by the UK Department of Health and Social Care, Genomics England partners with the UK's National Health Service (NHS) to provide whole genomic sequencing diagnostics, as well as driving genomic research and innovation in healthcare. Through their use of genomics, they seek to improve the accuracy of diagnostics and the effectiveness of treatment throughout the country.[60]
- **The Department for Genomic Medicine at Rigshospitalet:** a specialist clinical and research unit within Rigshospitalet, which is Denmark's largest hospital. The department is embedded in the hospital's clinical structure and works closely with multiple

[59] Decipher (2026), "Homepage."

[60] Genomics England (2026), "Homepage."

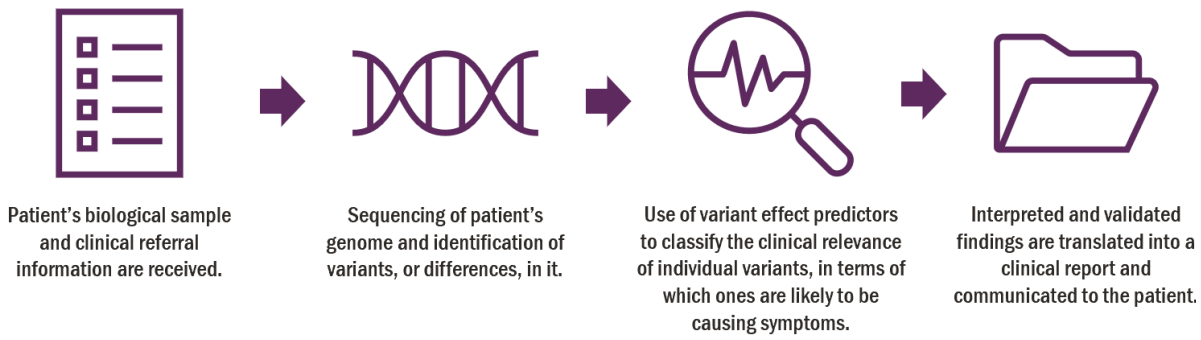
departments, such as rare disease services, cancer care, and specialised diagnostics, to apply genomic sequencing in patient care. It also plays a key role in the wider research field, collaborating with academic and industrial partners to advance the wider field. The Department for Genomic Medicine is the bioinformatics and sequencing facility for the National Genome Centre in Denmark.[61]

The interviews focused on assessing the current impact of EMBL-EBI's data resources, tools, and standards on the two institutions' genomics sequencing workflows, compared to a counterfactual world where EMBL-EBI data resources, tools and standards no longer existed. An interview guide was developed and applied to ensure consistency in approach across the two interviews.

6.3 Overview of how EMBL-EBI data resources are used in genomics sequencing workflows

The primary workflow we explored in the interviews was the clinical genomics sequencing pipeline for patients with rare diseases. Figure 32 summarises this process, showing how patients' biological samples are analysed using genomic sequencing, how variants identified across the genome are annotated and classified according to their clinical relevance, and how the results are interpreted to inform a diagnosis, where possible. By facilitating earlier and more accurate diagnoses for rare diseases, this process can lead to better, more personalised care pathways, avoid unnecessary diagnostic investigations and improve patient health from treatments (where available).|

Figure 32 Overview of clinical genomic sequencing pipelines



Source: Frontier Economics

Both Genomics England and Rigshospitalet also operate research functions beyond their clinical pipeline. For example, Genomics England performs secondary diagnostic discovery analyses, which seek to further investigate individual cases wherever a diagnosis is not able to be made during the process outlined above. These analyses are often undertaken by a separate team, using a new set of tools and methods to validate and expand upon the primary

[61] Rigshospitalet (2026), "Department of Genomic Medicine."

analysis. Rigshospitalet also conducts external research, focusing on implementing the latest methodologies to keep Next-Generation Sequencing (NGS) platforms at their cutting edge.[62]

EMBL-EBI data resources, tools and standards were reported as being central to both the clinical and research functions. For example, Ensembl plays a key role in underpinning variant annotation at Genomics England, and supports a large share of data-driven research at Rigshospitalet. Ensembl VEP is used by both organisations at different points in the workflow:

- **Genomics England:** it is mainly employed to cross-check and validate internal variant effect predictions or support secondary analyses.
- **Rigshospitalet:** it is embedded in the variant annotation process, as well as being used to support wider research.

The MANE standards are used consistently across both organisations to provide matched, equivalent transcripts between GENCODE and RefSeq, supporting a standardised approach to variant interpretation. Finally, although DECIPHER is not used directly for variant annotation or variant effect prediction, Rigshospitalet reported that it is used to determine the commonality of a variant amongst the general population, and Genomics England reported that it is a key tool used downstream in NHS labs. Further detail on how EMBL-EBI resources, tools and standards are used in each organisation is provided in Table 1.

Table 1 Contribution of EMBL-EBI data resources, tools, and standards to national genomics programme

EMBL-EBI data resource, tool or standards	Genomics England	Rigshospitalet
Ensembl data	■ The Ensembl database is used as an input for Genomics England's internal software that they use for variant annotation and variant effect prediction. ■ Genomics England rely heavily on the external references from the main Ensembl database, which are used to link together the inputs from a variety of different sources.	■ The Ensembl database underpins the vast majority of data-driven research undertaken at Rigshospitalet, outside of the clinical pipeline. ■ Ensembl is not typically used for variant annotation on the clinical side, as the majority of the literature referred to builds on RefSeq.
Ensembl VEP	■ Although Genomics England use their own internal software for the variant effect prediction and classification process, Ensembl VEP is often used to cross-check and validate internal predictions.	■ Ensembl VEP sits within the early stages of the pipeline, supporting variant annotation. ■ Ensembl VEP also supports the wider research functions.

[62] Rigshospitalet (2026), "Department of Genomic Medicine."

	■ Ensembl VEP is used outside of the main clinical genomics pipeline in cases where no diagnosis is found from the primary clinical pipeline, providing a new perspective in secondary analyses.	
MANE standards	■ The MANE standards are used to connect matched transcripts from Ensembl and RefSeq throughout the clinical pipeline. ■ The MANE transcripts are particularly important in The Generation Study (which screens newborns) to maintain high specificity.	■ The MANE transcripts are always used for interpretation to ensure that the most clinically relevant transcript is selected. ■ By using the MANE transcripts, this ensures that the internal team always refer to the same transcript.
DECIPHER	■ DECIPHER is used for variant interpretation in NHS labs downstream of the Genomics England pipeline.	■ DECIPHER is used on an ad hoc basis, especially to determine the commonality of a variant amongst the general population.

Source: Frontier Economics : findings from interviews with bioinformaticians.

Beyond the resources summarised in Table 1, other EMBL-EBI resources also play key roles in genomic sequencing pipelines. For example, Genomics England reported that their wider teams rely on inputs from EMBL-EBI's Gene2Phenotype dataset; UniProt data; and the Expression Atlas. Furthermore, Rigshospitalet reported that, between their clinical pipeline and their research operations, they use almost all of EMBL-EBI's data resources, either directly or indirectly through secondary resources.

6.4 The impact of EMBL-EBI resources on national genomics programmes

Throughout the interviews with Genomics England and Rigshospitalet, we identified three key ways in which EMBL-EBI positively supports the genomic sequencing pipeline for national genomics programmes:

- Genomic sequencing pipelines use multiple EMBL-EBI resources and tools. Losing these would cause substantial disruption;
- EMBL-EBI acts as a trusted authority in the bioinformatics community, supporting standardisation, resilience and reducing duplication; and
- EMBL-EBI supports the interoperability of the wider genomics ecosystem, improving the quality of work and driving scientific progress.

6.4.1 Genomic sequencing pipelines use multiple EMBL-EBI resources and tools. Losing these would cause substantial disruption.

The interviews revealed that EMBL-EBI's data resources and tools are deeply embedded at multiple stages of the clinical genomic sequencing pipelines at both Genomics England and Rigshospitalet. In particular, as described above, resources such as Ensembl and Ensembl VEP support variant annotation; classification; and the validation of findings based on

alternative tools. Tools such as DECIPHER are also used further downstream to support variant interpretation.

These resources are not used in isolation; rather, they are tightly interwoven throughout internal pipelines at both institutions. In particular, interviewees from Genomics England highlighted that their internal variant database, which is used for variant annotation and for variant effect prediction, has been built around the Ensembl database since its creation. Meanwhile, interviewees at Rigshospitalet highlighted that EMBL-EBI resources are deeply integrated into their clinical pipeline, which relies on 15 different EMBL-EBI data resources, either directly or indirectly through secondary resources. The interviewees at Rigshospitalet also emphasised that a wide range of commercial resources and services are derived using EMBL-EBI data or code, demonstrating the extent to which EMBL-EBI's data is integrated throughout the wider genomics field.

When asked to consider a counterfactual scenario in which EMBL-EBI ceased to exist, interviewees reported that, given the centrality of EMBL-EBI data resources and tools to their pipelines, this would have a materially negative impact. In particular, interviewees raised the following potential impacts on their workflows:

- **Major re-engineering of existing infrastructures:** the loss of the Ensembl database would result in both institutions having to re-engineer their existing infrastructures around new data resources and tools. Whilst these infrastructures could continue to use offline copies of the Ensembl database as inputs in the short term, in the longer term, switching to alternatives from NCBI would involve revalidating pipelines, rewriting code, retraining staff, and re-establishing clinical confidence in outputs. Such restructuring would be highly costly in both time and financial budget, with Genomics England suggesting that this process could take *"two or three years"*. In particular, interviewees highlighted that there would be a significant opportunity cost in terms of bioinformaticians' time that would be directed towards re-engineering internal infrastructures instead of contributing to the diagnostic pipeline. This would have an impact on internal productivity and the ability to deliver outputs responsively.
- **Reduction in the number of diagnoses made:** Even where alternative databases exist (for example, RefSeq as an alternative to GENCODE gene sets), interviewees at Rigshospitalet reported that the number of diagnoses made would fall in a scenario where the available data and annotations were halved. The reduction in diagnostic yield was highlighted as an issue particularly for rare diseases, where the marginal benefits of having additional data and annotations available are higher than for common diseases. Genomics England reported that their diagnostic yield would particularly suffer amongst The Generation Study, which is highly dependent on the MANE transcripts.[63] Whilst Genomics England were of the view that, in the long term, their diagnostic yield for rare

[63] The Generation Study, delivered by Genomics England, aims to recruit 100,000 newborn babies and screen them for over 200 genetic conditions. This is to "facilitate the earlier identification of rare genetic conditions in babies, to gather genomic data for wider research purposes and to explore the risks and benefits of storing an individual's genome over their lifetime."

diseases would be the same if they were to switch to NCBI alternatives, they highlighted that, in the short term, it would take longer for workers to achieve the same diagnostic yield due to resources being diverted towards establishing a new pipeline (as described above).

- **High cost of substitution towards commercial alternatives:** Moving away from EMBL-EBI resources towards commercial tools would be significantly more expensive. Commercial solutions were described as costly to license and often less transparent, with lower ability to internally validate outputs. In contrast, interviewees highlighted how EMBL-EBI resources provide open, community-validated standards that support reproducibility.

In particular, in relation to the hypothetical scenario where EMBL-EBI ceased to exist, interviewees reported:

> *"[There would be a] catastrophic impact in terms of resources that we will need to divert to re-engineer our systems."*
>
> *"It would have a massive impact on productivity internally and a massive impact on our ability to deliver features responsively."*
>
> *"If we imagine a world where we'd only have half the annotation that we have today, then we would have twice the problems."*

6.4.2 EMBL-EBI acts as a trusted authority in the bioinformatics community, supporting standardisation, resilience and reducing duplication

Interviewees consistently described EMBL-EBI as being a central authority within the global bioinformatics community, playing a critical role in setting shared standards that enable genomic data to be used reliably at scale. A key example is the MANE transcript standard, which connects inputs from across the genomics landscape to enhance interoperability. Interviewees underlined that EMBL-EBI's role as a trusted authority stems not only from its individual resources, but from the scale of its overall contribution. By contrast, smaller institutions, even if technically capable in specific areas, would lack the legitimacy, scale, and reach to establish international standards in the same way. Therefore, by curating and maintaining a wide range of data resources and tools under a trusted institutional umbrella, EMBL-EBI has established itself as a standard-setter, delivering value that is greater than the sum of its parts.

Furthermore, interviewees reported the value of having multiple institutions (such as NCBI) in the genomics field, instead of just one, as this provides resilience to the wider community. In particular, interviewees described experiences where one service was "*not responding from one day to another*," during which time they were able to rely on EMBL-EBI's resources as an easy substitute.

Interviewees also highlighted the importance of having a trusted authority to deliver centralised and openly accessible data resources, as smaller institutions would not be able to do so. When asked about a counterfactual scenario where EMBL-EBI ceased to exist, it was described as *"out of the question"* that an organisation such as Genomics England could replicate the

EMBL-EBI resources they use, given their differences in scale and remit. Attempting to recreate equivalent databases, standards, update cycles and governance mechanisms internally would be infeasible. Therefore, EMBL-EBI delivers significant value by reducing the requirement for duplication of effort across organisations. For example, we heard that:

> *"If you don't have an authority in the field, then things can maybe be solved anyway, but it just takes so much more time."*

Finally, interviewees emphasised that EMBL-EBI's authority and branding provide substantial downstream benefits for researchers and clinicians. Given that EMBL-EBI resources are widely recognised as high-quality and community-validated, users have deep confidence in their outputs. This trust reduces the need for extensive local quality assurance or revalidation, saving time and effort within clinical and research pipelines. In this way, EMBL-EBI's role as an industry authority underpins efficiency, reliability and confidence across the bioinformatics ecosystem. In particular, interviews highlighted that:

> *"They do things so we can trust that whenever we use it, we don't have to benchmark [it]"*

6.4.3 EMBL-EBI supports the interoperability of the wider genomics ecosystem, improving the quality of work and driving scientific progress

EMBL-EBI data resources, tools and standards play a crucial role in enabling interoperability across the wider open science and genomics ecosystem**.** Interviewees from both institutions highlighted the MANE transcript standard as the primary example of this, through the creation of a mapping between Ensembl and RefSeq transcripts such that the two identifiers can be used together. In doing so, MANE acts as a common anchor point that allows different parts of the genomics ecosystem to connect reliably. This allows users to compare the outputs from either source, or switch easily between the two depending on the setting. As a result, the improved interoperability across the field enhances the efficiency and accuracy of the genomic pipeline where transcripts from multiple sources are used.

Interviewees also emphasised the mutually reinforcing relationship between EMBL-EBI and NCBI**.** Rather than being competing substitutes, the two infrastructures are complementary (i.e., each containing distinct annotations) and have many interdependencies. In addition, the presence of a credible peer encourages competition, pushing both organisations to maintain high standards, innovate, and respond to user needs. Consequently, interviewees consistently noted that, in a counterfactual scenario where EMBL-EBI resources no longer existed, both (i) NCBI resources, and (ii) the wider variant annotation process would likely be less effective, and progress across the field would slow down. In particular, when discussing this counterfactual scenario, interviewees noted that:

> *"If there was no Ensembl, […] NCBI wouldn't be of the quality they are."*
>
> *"In five years, you would see that the progression has stalled"*

Finally, interviewees highlighted that both Genomics England and Rigshospitalet actively use EMBL-EBI resources as a means of validating the outputs of alternative tools. EMBL-EBI's open datasets provide a neutral benchmark against which results can be compared. This highlights the importance of the interoperability of resources to support data validation and the wider confidence of outputs.

6.5 Conclusions

Genomics sequencing is a crucial process supporting the diagnosis of rare diseases and the ultimate treatment of patients, which itself provides significant socio-economic benefits to society. EMBL-EBI plays a critical role in supporting national genomics programmes by providing open data resources and tools that are key to genomic sequencing pipelines. These resources are implemented at various stages of the pipeline, including for variant annotation, variant effect prediction, and downstream interpretation. If EMBL-EBI ceased to exist, this would have significant, negative impacts in terms of immediately requiring genomics programmes to re-design their internal infrastructures, therefore reducing the number of diagnoses made and increasing costs.

Through the scale of its resources and tools, EMBL-EBI has established itself as a trusted authority in the genomics ecosystem. As a result, EMBL-EBI is able to set sector-wide standards, improving the efficiency and reliability with which bioinformaticians can interact and draw together inputs from different sources. Through this role, EMBL-EBI also supports resilience throughout the genomics ecosystem, as well as reducing the time and costs spent on duplication of efforts among organisations. Furthermore, EMBL-EBI plays a central role in supporting the interoperability across the wider genomics scientific community, without which the quality of work would likely be weakened and scientific progress might slow. Through these channels, the value provided by EMBL-EBI's data resources, tools and standards are far greater than the incremental value of each resource.

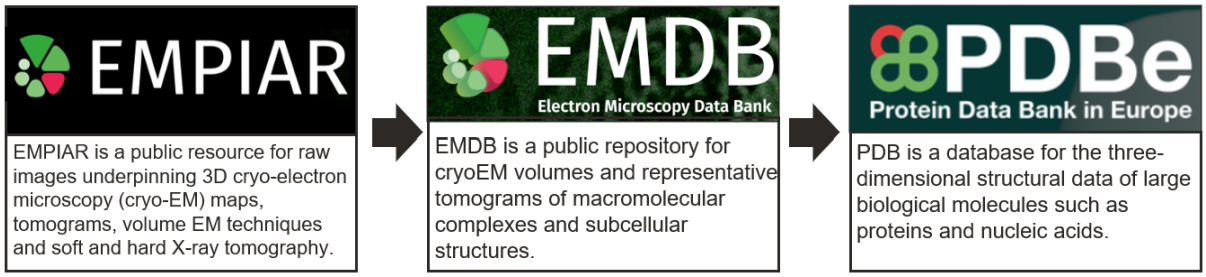
7 Data reuse case study: supporting public and private sector innovation through bioimaging and structural biology resources

7.1 Background and hypotheses to test

Bioimaging encompasses techniques used to visualise biological structures and processes, from molecules to whole organisms. Structural biology focuses on determining the three-dimensional (3D) structure of macromolecules and their assemblies to understand how their form underpins their function. Together, these fields help connect molecular structure with biological role. Beyond advancing fundamental science, they support applications such as tracking disease progression, assessing responses to therapies, enabling medical diagnosis and the development of drugs, AI models and biomaterials. As such, bioimaging and structural biology are cornerstones of modern biomedical research.

EMBL-EBI hosts several open data resources supporting bioimaging and structural biology research. For this case study, we focus on EMPIAR, EMDB and PDB (via PDBe). These resources are linked: EMPIAR archives raw electron microscopy (EM) image datasets that underpin reconstructed 3D EM volumes, and where atomic models are built for those volumes, the corresponding coordinates are deposited in the PDB and cross-referenced to the relevant EMDB entries.

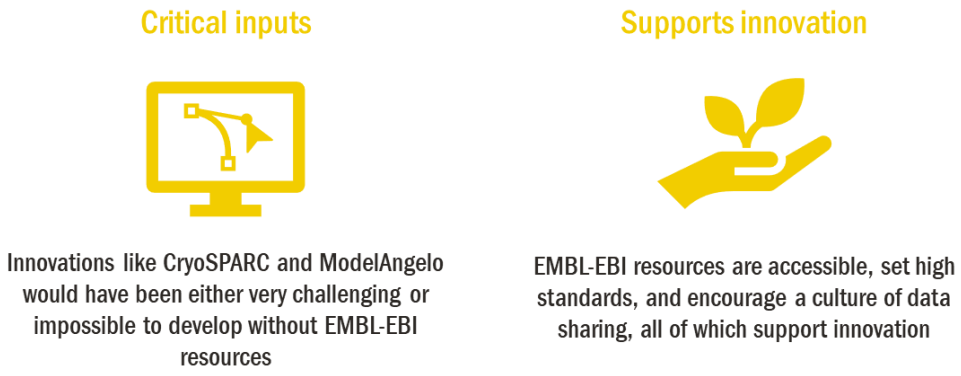
Figure 33 Pipeline of EMBL-EBI open data resources in scope for this case study



Source: Frontier Economics

This case study tests the hypothesis that the bioimaging and structural biology open data resource ecosystem (EMPIAR, EMDB and PDB) supports both public and private sector data reuse and innovation through the development of new tools. This is compared to a counterfactual world in which none of these resources are available. Through this analysis, the case study assesses whether the impacts set out in the logic model in Section 2 are likely to be realised (e.g., relating to EMBL-EBI's contribution to greater innovation and economic growth by enabling others to reuse and build on public data).

Figure 34 Data reuse: supporting public and private sector innovation through bioimaging and structural biology resources



Source: Frontier Economics

7.2 Methodology

To test this hypothesis, we qualitatively explore the contribution of EMBL-EBI resources to the development of two innovative tools developed in the public and private sectors:

- **CryoSPARC:** developed by Canadian Firm Structura Biotechnology (Structura), this software supports rapid, automated processing of cryo-electron microscopy (cryo-EM) image data (the type archived in EMPIAR) to reconstruct high-resolution 3D EM volumes of biological molecules (the type archived in EMDB). By automating major steps in this process, it reduces turnaround time and the level of specialist intervention typically required by researchers.
- **ModelAngelo:** developed by researchers at the MRC Laboratory of Molecular Biology in Cambridge, this machine-learning tool supports automated atomic model building (models suitable for deposition in PDB) from high-resolution 3D EM volumes (the type archived in EMDB). By automating this process, it reduces the need for manual model building, also saving researchers time.

CryoSPARC and ModelAngelo are therefore examples of tools that generate data at different stages of the cryo-EM structure determination pipeline, which are deposited to the different EMBL-EBI resources shown in Figure 33, where these data become available for reuse. To qualitatively assess EMBL-EBI's contribution to the development of these tools, we conducted a total of two interviews with (i) one of the developers of CryoSPARC, and (ii) one of the developers of ModelAngelo. An interview guide was developed and applied to ensure consistency in approach across the two interviews.

7.3 CryoSPARC: EMBL-EBI's role in private sector innovation

This sub-section summarises the key findings on whether and how EMBL-EBI data resources contributed to the development of CryoSPARC.

7.3.1 The creation of CryoSPARC would have been challenging and much slower without EMBL-EBI data resources

When asked whether EMBL-EBI data resources contributed to the development of CryoSPARC, the interviewee emphasised how the launch of EMPIAR in particular was crucial in enabling the widespread data access required to develop the algorithms underpinning the tool. This is because prior to the creation of EMPIAR, there were "*few labs that could collect high quality, high resolution Cryo-EM data*," meaning that limited data availability posed a significant constraint on innovation. This meant the development of CryoSPARC was initially dependent on collaborations with a small number of academic laboratories to obtain suitable experimental datasets. By providing a novel open repository of cryo-EM data, EMPIAR quickly became the "*dominant source of test data and real experimental data on which to develop methods."* This removed the bottleneck of having to individually negotiate the provision of data with specific laboratories, accelerating the testing and development of the tool.

While not directly involved in the development of the algorithm, the interviewee noted that EMDB serves as the "*final home*" for the processed 3D density maps generated using tools such as CryoSPARC. As a result, the quality and resolution standards set by EMDB were an important consideration in designing the outputs from the tool. The interviewee also noted that PDB plays an important role in methods development. Although PDB sits further downstream in the pipeline, access to repositories of solved atomic structures and associated validation metrics is valuable for benchmarking, diagnostics, and identifying data-processing issues that could otherwise propagate into downstream analyses.

Taken together, the interviewee stated that without EMBL-EBI data resources:

> *"It would be much, much more challenging, and it would be a lot slower [to develop CryoSPARC]."*
>
> *"All of Cryo-EM methods development would have been a lot slower and a lot less far along than it is if these resources didn't exist."*

The interviewee also confirmed that future updates and developments to the CryoSPARC tool will continue to rely on EMBL-EBI data resources (particularly EMPIAR) to develop new methods. This demonstrates the importance of continuing to provide and update EMBL-EBI data resources.

7.3.2 The standards and data sharing ecosystem created by EMBL-EBI plays a significant role in supporting innovation

The interviewee also highlighted how data standards and the culture of sharing fostered by EMBL-EBI creates an environment that supports innovation. This echoes a similar finding from case study 2 in Section 6.

For example, EMDB hosts the processed 3D EM volumes generated using CryoSPARC and other image refinement and 3D reconstruction tools. The resolution and validation standards required to deposit into it were viewed as establishing a "*quality bar*" across computational structural biology. This ensures that researchers do not "*cut corners*" and that consistently high data standards are maintained across the field. Similarly, EMPIAR's structured deposition requirements, including detailed data collection parameters such as electron voltage and magnification, were noted as ensuring that datasets are uploaded in a consistent, transparent format. This standardisation improves data quality and strengthens the reliability of downstream analyses. Therefore, by acting as a "*centralised, trusted single authority*" that defines and enforces standards, EMBL-EBI facilitates the supply of trustworthy data which reduces the time users of that data (such as Structura) spend on validation, accelerating the development of tools like CryoSPARC.

More broadly, EMBL-EBI's open data resources were noted as having fostered a "*culture of [data] sharing*" that accelerates knowledge exchange and innovation. This was viewed as far less common prior to the launch of EMPIAR. As outlined above, the development of CryoSPARC initially relied on bilateral arrangements with academic laboratories to access cryo-EM data, which took significant time to both a) sustain relationships, and b) agree whether data used to develop the tool could be published. By opening access across the scientific community and establishing a norm in which data are "*safe and fair to use*", EMPIAR significantly reduced these frictions and enabled the more rapid development of CryoSPARC.

Although CryoSPARC is chargeable to private companies, it is free for non-profit academic use. The interviewee highlighted that the public-private model aims to use revenue from commercial licences to fund ongoing maintenance and further development of the software for the wider research community. In 2025, CryoSPARC reported around 6,500 active installations across academic institutions and laboratories worldwide. Comparable figures for use by private organisations were not available. Overall, the 6,500 active installations figure provides an example of secondary usage arising from tools developed using EMBL-EBI resources.

7.4 ModelAngelo: EMBL-EBI's role in public sector innovation

This sub-section summarises the key findings on whether and how EMBL-EBI data resources contributed to the development of ModelAngelo.

7.4.1 ModelAngelo could not exist without EMBL-EBI open data resources

The interviewee highlighted how two EMBL-EBI data resources, EMDB and PDB, were foundational to the development and training of ModelAngelo. The developers trained the tool using paired cryo-EM density maps from EMDB alongside their corresponding atomic models and protein sequences deposited in the PDB. The interviewee emphasised that the tool would not be possible without access to these open data resources and identified no alternative datasets that could have supported the innovation. In particular, they stated:

> *"Without [EMBL-]EBI, there would be no ModelAngelo."*

Looking ahead, the interviewee expected that ongoing improvements to ModelAngelo would continue to rely on EMDB and PDB for training and validation. This highlights EMBL-EBI's role in enabling and sustaining innovations in the life sciences community.

7.4.2 EMBL-EBI open data resources are built in an accessible way that facilitates innovation

The interviewee emphasised that EMBL-EBI's open data resources are designed to be highly accessible and usable, significantly lowering the barriers to innovating. By ensuring that data are well-structured, "thorough" in terms of their content, and straightforward to integrate into new tools, EMBL-EBI enables researchers to build and validate methods more efficiently. For example, the integration of Q-Scores,[64] a metric developed by academic researchers, into EMBL-EBI databases allows users to assess the fit of individual amino acids within cryo-EM reconstruction maps more easily and consistently. Embedding such validation metrics directly within the data infrastructure consolidates a broad range of information in one place, and therefore improves research quality while also reducing the time and technical effort required for analysis. The interviewee also highlighted that other innovations, such as Blush regularisation,[65] have been supported by EMBL-EBI's accessible and well-curated data resources. This demonstrates how the accessibility of EMBL-EBI's infrastructure directly facilitates and accelerates innovation across structural biology and bioinformatics. In particular, the interviewee cited that:

> *"Structural biology is quite a lucky field because very early on PDB was structured extremely thoroughly. […] When electron microscopy came along as a new technique and EMDB was set up, the same rigour was used for that database too. For machine learning, that has been a treasure trove."*

> *"Model Angelo, Blush [regularisation], […] would not have been possible without such well-maintained, annotated databases.*

[64] A Q-score is a quantitative metric that measures how well an individual atom or residue in an atomic model is supported by the surrounding cryo-EM density, providing an indicator of local model quality and resolvability.

[65] Blush regularisation is a machine-learning–based method used in cryo-electron microscopy (cryo-EM) image processing to improve the quality of reconstructed 3D maps.

7.5 Conclusion

Bioimaging and structural biology play critical roles in modern life science research. By providing access to high-quality, accessible data resources, EMBL-EBI enables the development of tools such as CryoSPARC and ModelAngelo. These tools then accelerate discovery in these fields by making complex image processing, reconstruction and interpretation more efficient and more widely usable. This has positive downstream impacts on scientific understanding of living systems across many applications, including drug and vaccine development in human medicine. Without EMBL-EBI's open data resources, the evidence suggests innovations like CryoSPARC and ModelAngelo would have been either very challenging or impossible to develop, negating these downstream benefits.

More generally, the evidence also supports a virtuous cycle where the tools accelerate the generation of outputs that are deposited back into the same resources they were developed using. In other words, the provision of EMBL-EBI data resources helps drive innovation and data reuse, with the innovation leading to further deposition of data, enhancing the breadth and depth of data available for all.

8 Conclusions

8.1 Summary of findings

This study has taken a mixed methods approach to assess the overall value and socio-economic impact of EMBL-EBI's data resources on users and wider society. We employed evidence from (i) a survey of EMBL-EBI users, (ii) user analytics data from internal EMBL-EBI web-logs, and (iii) qualitative and quantitative data across three case studies. These sources were applied across three sets of analysis, which sought to validate the short-term outcomes, long-term outcomes and downstream impacts of EMBL-EBI's data resources.

Taken together, the evidence presented throughout this report suggests that EMBL-EBI data resources deliver substantial value to users and to society, with this value continuing to grow over the past decade. Across all value indicators, we find benefit-cost ratios to be high. Even under conservative assumptions, the benefits of EMBL-EBI data resources substantially exceed the costs of maintaining and providing them.

At the user level, EMBL-EBI resources:

- Are essential to a large and growing number of life sciences researchers;
- Support researchers globally and across academic, industry and healthcare settings;
- Enable work that would otherwise not be possible;
- Save significant amounts of researcher time; and
- Support interdisciplinary research, reproducibility and the development of machine learning and AI systems.

At the societal level, they:

- Enable significant returns to research and development; and
- Realise downstream impacts. For example, widening access to and increasing the volume of science, supporting healthcare systems, and enabling innovation.

Crucially, the evidence from the case studies shows that EMBL-EBI's value does not lie solely in the data themselves, but also in the curation, standardisation, integration and open access model. With interviewees highlighting EMBL-EBI's role as a central body that develops standards and coordinates with the open science community, this suggests that the value generated by EMBL-EBI data resources is worth more than the sum of their parts.

Overall, the evidence strongly supports the conclusion that investment in EMBL-EBI data resources represents excellent value for money and plays a critical role in sustaining and advancing the global life sciences research ecosystem.

8.2 Long-term trends and considerations for future work

This impact assessment builds on earlier economic assessments published by Charles Beagrie Ltd in 2016 and 2021. The repeated nature of the study over a 10-year period has enabled the tracking of changes in the value of EMBL-EBI open data resources over time. These changes likely reflect how the resources themselves have evolved, but also major developments that have affected the life sciences research community. These include:

- **The return to a "new-normal" after the COVID-19 pandemic:** this may have affected the value of time indicator presented in this report, with a return to lab-based working potentially explaining the decline in time spent working with EMBL-EBI data resources since 2021;
- **The introduction of new technologies**: Case Study 1 demonstrated the distributional and scale impacts on research from the development of the AlphaFold Database, with Case Study 3 highlighting the key role EMBL-EBI data resources play in supporting public and private sector innovation; and
- **Developments in user behaviour and expectations shaped by technological change**: 42% of survey respondents reported that EMBL-EBI data resources contributed to training and/or evaluating machine-learning or other AI models.

It will be important to continue monitoring these and emerging developments, and to assess how they affect the value and impact of EMBL-EBI's data resources.

As discussed throughout this report, a crucial element of our study involves accurately estimating the direct user base. We have highlighted the following challenges around tracking the user base, regarding the ability to measure (i) accesses through modern access routes, and (ii) secondary usage. Future research should consider how to comprehensively capture all usage of EMBL-EBI data resources to successfully capture their overall value.

- **Accesses through modern access routes:** users increasingly rely on modern access routes that are not captured by EMBL-EBI's internal web-log data, making it increasingly difficult to accurately measure the user base. For example, cloud providers (behind which multiple users may sit) appear as a single IP in EMBL-EBI's usage data. Whilst we provide an indicative adjustment for this in our user base estimates, future work should more robustly assess the number of users accessing data via these new routes.
- **Secondary usage:** although the user base estimated in this study gives an indication of the number of direct users, it does not capture secondary usage via resources curated based on EMBL-EBI data, users of open source tools built using EMBL-EBI resources and the training of AI systems on these resources. Based on the information available, it has not been possible to produce a robust estimate of the size of the secondary user base. However, the evidence presented throughout this report demonstrates that secondary usage is substantial and increasing. Therefore, future work should seek to measure the number of secondary users of EMBL-EBI data resources and their value to comprehensively capture the overall impact of EMBL-EBI data resources.

Annex A – Summary of the user survey

This annex sets out a granular breakdown of the user survey findings on a question-by-question basis, including:

- A summary of users' responses;
- A mapping of the updates made to the survey questionnaire compared to the previous Beagrie (2016, 2021) user surveys;
- An overview of the approaches used to remove extreme or unreliable responses; and
- An overview of the insights from the LLM analysis covering free-text responses highlighting uncertainty or guesswork.

A.1 Questions about the respondent

Q1: Country in which you work or study (N = 2,563)

Respondents were asked to report the country in which they work or study. A summary of users' responses is included in Section 3. In particular, the top five countries reported by respondents are the USA (16%), China (14%), the UK (12%), Germany (7%) and India (7%).

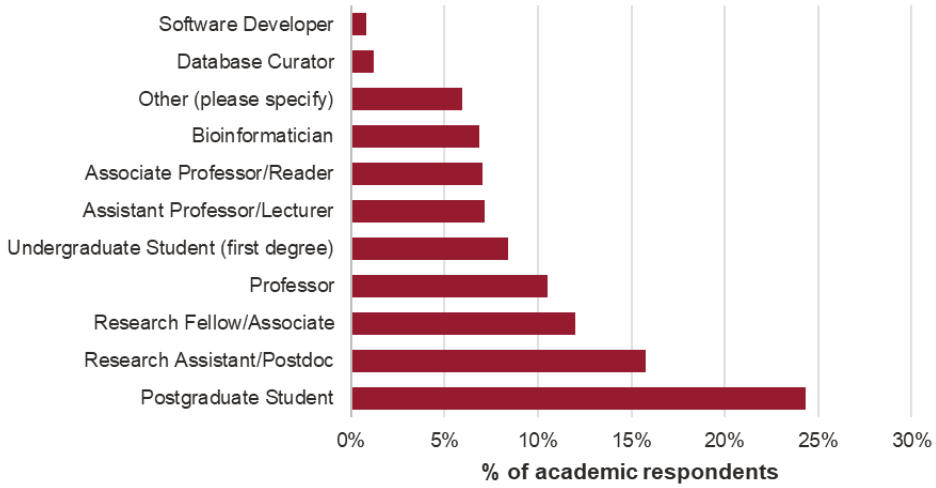
Q2: Your main affiliation or sector (N = 2,563)

Respondents were asked to report their main affiliation or sector. A summary of users' responses is included in Section 3.

Q3 & Q4: What is your main role within this affiliation? (N = 2,507)

Figure 35 shows that, of the user survey respondents affiliated with the academic sector, postgraduate students (24%), research assistants / post-doctorates (16%) and research fellows associates (12%) were the most commonly listed job roles. Since the Beagrie (2021) survey, the distribution of survey respondents between job roles remains broadly similar.

Figure 35 Main role within the academic sector

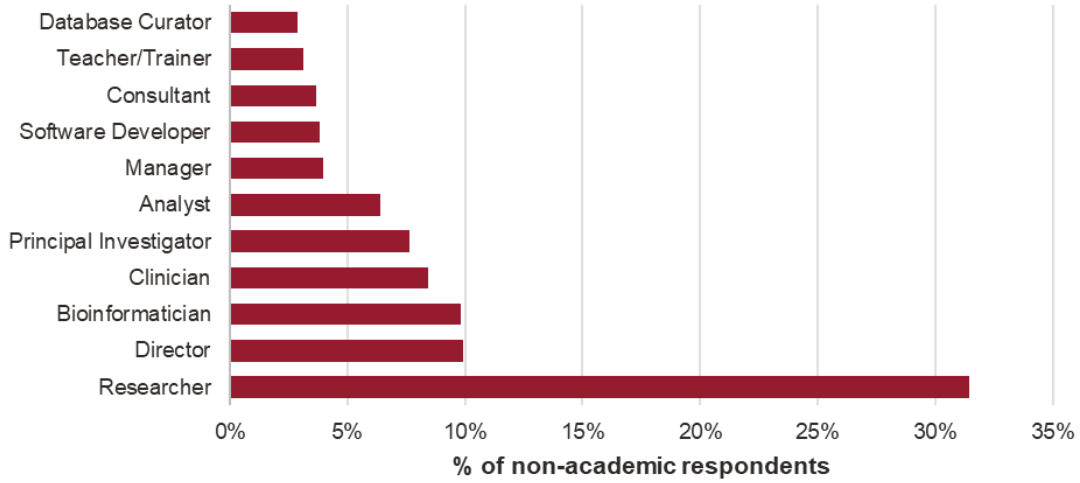


Source: *2025 EMBL-EBI User Survey*

Note: *Of those who reported that their main affiliation is the academic sector, question 3 of the survey asked respondents "What is your main role within this affiliation?" (N = 1,772)*

Figure 36 shows that, of the user survey respondents affiliated with the non-academic sector, researchers (31%), directors (10%) and bioinformaticians (10%) were the most commonly listed roles. Since the Beagrie (2021) survey, the distribution of survey respondents between job roles remains broadly similar, with the most notable differences being an increase in the share of non-academic respondents reporting that they are researchers (by around 1 percentage point) and a decrease in the share of bioinformaticians (by around 5 percentage points).

Figure 36 Main role within the non-academic sector



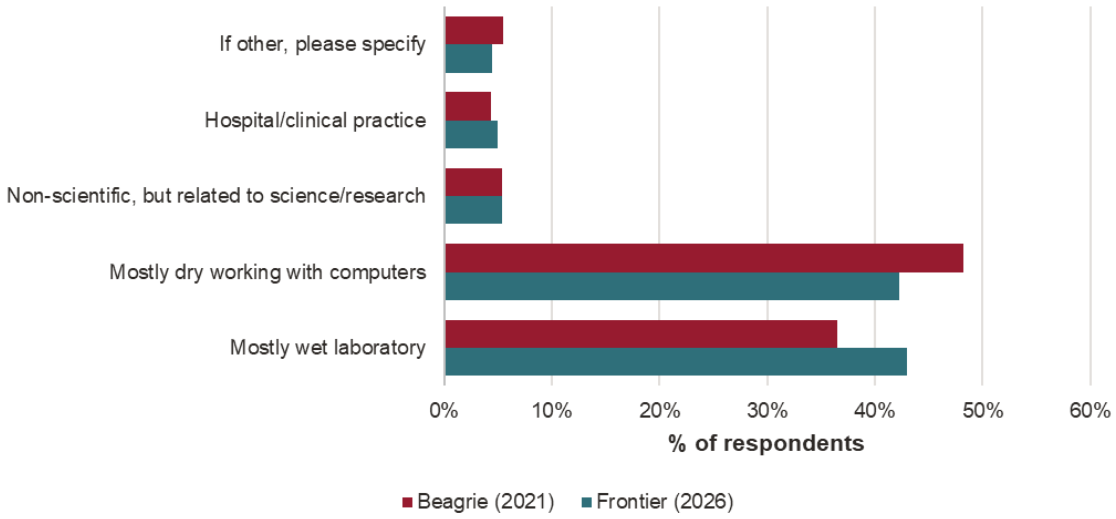
Source: *2025 EMBL-EBI User Survey*

Note: *Of those who reported that their main affiliation is the non-academic sector, question 4 of the survey asked respondents "What is your main role within this affiliation?" (N = 735)*

Q5: Which of the following most closely describes the nature of your work or study? (N = 2,443)

42% of respondents to the 2025 survey reported mostly dry working with computers, compared to 48% of respondents to the 2021 survey. In addition, 43% of respondents to the 2025 survey reported mostly wet laboratory working, compared to 37% of the respondents to the 2021 survey. Otherwise, the distribution of survey respondents between types of work remains broadly similar.

Figure 37 Nature of work or study



Source: *2025 EMBL-EBI User Survey*

Note: *Question 5 of the survey asked respondents "Which of the following most closely describes the nature of your work or study?" (N = 2,443)*

A.2 Questions about the resources that respondents use

Q6 to Q17 asked respondents which EMBL-EBI data resources they use and how frequently they access and/or download them.

When asking respondents about the frequency with which they accessed or downloaded each resource, the questionnaire offered the following options, which we converted to approximate annual frequencies using Beagrie's (2021) method below:

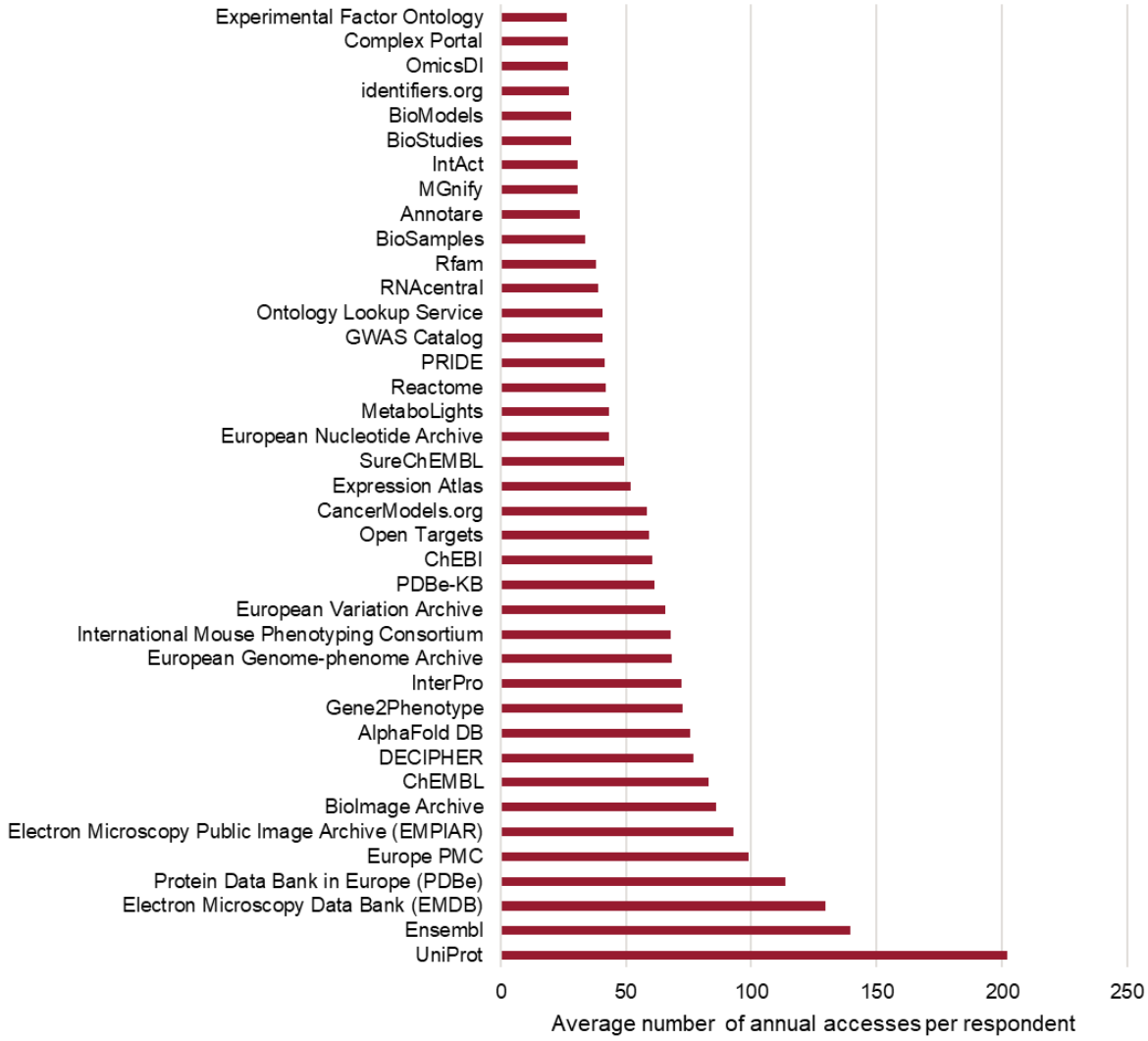
- > 5 times a day (*converted to 1,320 accesses per year*);
- 2-5 times a day (*converted to 770 accesses per year*);
- Daily (*converted to 220 accesses per year*);

- Weekly (*converted to 45 accesses per year*);
- Fortnight (*converted to 22 accesses per year*);
- Monthly (*converted to 12 accesses per year*);
- Quarterly (*converted to 4 accesses per year*);
- Once or twice a year (*converted to 1.5 accesses per year*); and
- Not used (*converted to 0 accesses per year*).

Based on these conversions, Figure 38 shows the average annual frequency with which respondents accessed or downloaded EMBL-EBI data resources. The figure shows that the most accessed resources were UniProt, Ensembl, EMDB and PDBe.

Based on the total number of annual accesses reported by each respondent across all listed resources, we calculated the total number of EMBL-EBI accesses per respondent per year. The mean number of times that survey respondents accessed or downloaded EMBL-EBI data resources was 937 times per year, and the median number of annual accesses or downloads was 199 times per year.

Figure 38 Average annual frequency of use of EMBL-EBI data resources



Source: *2025 EMBL-EBI User Survey*

Note: *Questions 7, 9, 11, 13, 15 and 17 of the survey asked respondents "In the last 12 months, which of these […] EMBL-EBI data resources did you use? And approximately how often did you access and/or download from each of these resources?"*

Q18: What would be the impact on your work or study if you could not access the EMBL-EBI data resources or tools you currently use? (N = 2,030)

The questionnaire offered the following options:

- No impact (7%);
- Slight negative impact (9%);
- Moderate negative impact (18%);
- Major negative impact (33%); and
- Severe negative impact (33%).

A comparison with the equivalent findings from the previous Beagrie (2016, 2021) studies is presented in Section 3.3.

The survey questionnaire also asked respondents to "please specify or provide examples." We have included examples of users' open text responses in Section 3.3.

Q19: Would you be able to take forward your work or study without the EMBL-EBI data resources or tools you currently use? (N = 2,015)

The questionnaire offered the following options:

- No – it would not be possible (27%);
- Yes – but with significant additional time and effort (44%);
- Yes – but with moderate additional time and effort (18%); and
- Yes – with similar time and effort (11%).

The survey questionnaire also asked respondents to "please specify or provide examples." We have included examples of users' open text responses in Section 3.3.

The previous Beagrie (2016, 2021) user surveys did not ask respondents this question.

A.3 Questions about obtaining equivalent data elsewhere or recreating similar datasets

Q20: Which EMBL-EBI data resource did you use most frequently in the last 12 months? (N = 1,917)

The top 20 EMBL-EBI data resources that users reported using most frequently are reported in Section 3.1.

Q21: Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. Approximately how long did it take you to find and obtain the EMBL-EBI data resource or tool you were looking for? (N = 1,172)

The survey questionnaire asked respondents to report the time taken to find or obtain EMBL-EBI data resources in terms of months, days, hours and minutes. Overall, responses tended to be subject to outliers and extreme values. Given that there is not an intuitive cut-off point for removing outliers, we exclude responses that were more than 2 standard deviations from the mean. There are 1,075 remaining responses following the removal of outliers, with a mean value of 16 hours and a median value of 5 minutes.

As described in Section 2.3.2, this survey question was modified in comparison to Beagrie (2016, 2021). Previously, the survey asked users how long it took them to find and obtain their most recently used EMBL-EBI resource. This was updated to ask about their most frequently

used resource, which is more likely to be representative of a typical access time. Despite the change in methodology, the median access time remains consistent with the Beagrie (2016, 2021).

Q22: Please explain how you have estimated your answer in the previous question. (N = 824)

The LLM analysis of the open text responses identified the following themes:

- **Experience of familiarity:** 62% of respondents based their time/value estimate on their direct, often extensive, experience with the resource. They mention familiarity, bookmarks, routine use, or knowing exactly where to look;
- **Uncertainty:** 12% of respondents express confusion, uncertainty, or inability to answer due to not understanding the question or lacking enough experience;
- **Guesswork or estimation:** 6% of respondents describe that their estimate is a guess, approximation, or not based on a specific recollection or method;
- **Uncategorised:** The remaining 21% of responses were unable to be categorised.

When calculating the average time spent finding and obtaining EMBL-EBI data resources (as estimated based on the responses to Question 21), on a conservative basis, we run a sensitivity analysis that excludes the responses where (i) respondents were uncertain or did not understand the question, and (ii) responses were based on guesswork or estimation. This sensitivity results in the mean access time reducing from 16 hours to 12 hours, with the median access time remaining unchanged at 5 minutes.

Q23: Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to find and obtain equivalent data or tools from another source? (N = 1,888)

The questionnaire offered the following options:

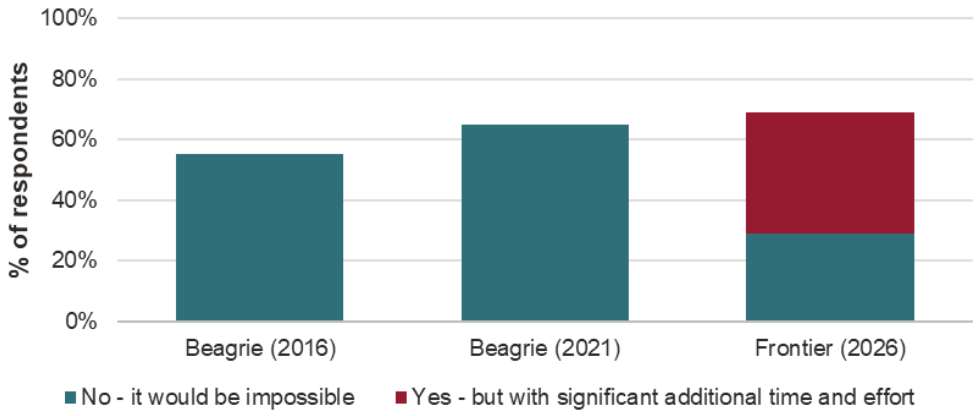
- No – it would not be possible (29%);
- Yes – but with significant additional time and effort (40%);
- Yes – but with moderate additional time and effort (18%); and
- Yes – with similar time and effort (13%).

While the 2025 user survey results are not directly comparable with the previous Beagrie (2016, 2021) studies,[66] the share of respondents who reported not being able to find or obtain

[66] They are not directly comparable because (i) the Beagrie surveys asked users about the EMBL-EBI resources they use in general rather than most frequently used, and (ii) the Beagrie surveys gave respondents two binary (i.e. yes / no) options, whereas the 2025 survey gave four options reflecting different levels of additional time and effort required to replicate the data. To identify trends over time, Figure 39 compares respondents answering that it would not be possible to obtain

equivalent data (without significant additional time and effort) has increased consistently over time from 55% in 2016 to 65% in 2021 and to 69% in 2025. This is shown by Figure 39.

Figure 39 The share of respondents who would not be able to find or obtain an equivalent data resource elsewhere; 2016-2025



Source: *EMBL-EBI User Survey; Beagrie (2016, 2021)*

Note: *Question 23 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to find and obtain equivalent data or tools from another source?" (N = 1,888).*

Q24: Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. Approximately how long would it take you to find and obtain an equivalent data resource or tool from another source? (N = 706)

The survey questionnaire asked respondents to report the time required to find or obtain equivalent data resources in terms of months, days, hours and minutes. Overall, responses tended to be subject to outliers and extreme values. Given that there is not an intuitive cut-off point for removing outliers, we exclude responses that were more than 2 standard deviations from the mean. We also remove all zeros, which are deemed an unrealistic access time. There are 686 remaining responses following the removal of outliers, with a mean value of 126 hours and a median value of 8 hours. This has increased compared to the previous findings from Beagrie (2021), where respondents suggested it would take them a mean of 64 hours to obtain equivalent data elsewhere, with a median of 1 hour.

This survey question was modified in comparison to Beagrie (2016, 2021). Previously, the survey asked users how long it took them to find and obtain their most recently used EMBL-EBI resource. This was updated to ask about their most frequently used resource, which is more likely to be representative of a typical access time.

elsewhere or re-create their EMBL-EBI data (in the Beagrie surveys) with 2025 survey respondents who reported that it would either (i) not be possible, or (ii) only be possible with significant time or effort.

Q25: Please explain how you have estimated your answer in the previous question. (N = 472)

The LLM analysis of the open text responses identified the following themes:

- **Process description or comparative analysis:** 49% of respondents estimated the length of time it would take them to find equivalent data resources based on a breakdown of the required process or a comparative analysis;
- **Hypothetical/Guesswork/Uncertain:** 20% of responses were based on hypothetical scenarios, guesswork, or expressed uncertainty or a lack of knowledge;
- **Based on Personal Experience:** 15% of respondents based their answers on personal experience; and
- **Uncategorised:** The remaining 16% of responses were unable to be categorised.

We run a sensitivity analysis that excludes the responses based on hypothetical scenarios, guesswork or uncertainty when calculating the average time spent finding and obtaining equivalent data resources (as estimated based on the responses to Question 24). This sensitivity results in the mean time reducing from 126 hours to 110 hours, whilst the median access time remains unchanged at 8 hours.

Q26: Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to re-create the same or similar dataset or tool yourself? (N = 1,799)

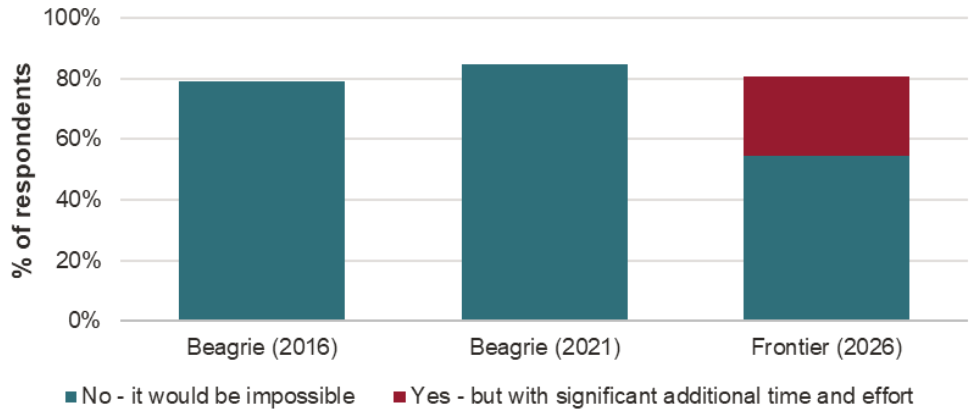
The questionnaire offered the following options:

- No – it would not be possible (54%);
- Yes – but with significant additional time and effort (26%);
- Yes – but with moderate additional time and effort (11%); and
- Yes – with similar time and effort (9%).

While the 2025 user survey results are not directly comparable with the previous Beagrie (2016, 2021) studies,[67] the share of respondents who reported not being able to recreate a similar dataset (without significant additional time and effort) has increased from 79% in 2016 to 81% in 2025, although has slightly decreased from 85% in 2021.

[67] They are not directly comparable because (i) the Beagrie surveys asked users about the EMBL-EBI resources they use in general rather than most frequently used, and (ii) the Beagrie surveys gave respondents two binary (i.e. yes / no) options, whereas the 2025 survey gave four options reflecting different levels of additional time and effort required to replicate the data. To identify trends over time, Figure 40 compares respondents answering that it would not be possible to obtain elsewhere or re-create their EMBL-EBI data (in the Beagrie surveys) with 2025 survey respondents who reported that it would either (i) not be possible, or (ii) only be possible with significant time or effort.

Figure 40 The share of respondents who would not be able to recreate a similar dataset; 2016-2025



Source: *EMBL-EBI User Survey; Beagrie (2016, 2021)*

Note: *Question 26 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to re-create the same or similar dataset or tool yourself?" (N = 1,799).*

Combining the responses to Q23 and Q26, of the respondents who answered both questions (N = 1,799), 67% reported that they would not be able to obtain an equivalent dataset elsewhere or recreate a similar dataset (without significant additional time and effort). As presented in Section 3.3, this value has increased from 45% in 2016 to 58% in 2021 and now 67% in 2025.

Q27: Approximately how much time would it take to collect and recreate the data or tool you used most frequently in the last 12 months? (N = 363)

The survey questionnaire asked respondents to report the time required to recreate similar datasets in terms of months, days, hours and minutes. Overall, responses tended to be subject to outliers and extreme values. Given that there is not an intuitive cut-off point for removing outliers, we exclude responses that were more than 2 standard deviations from the mean. We also remove all zeros, which are deemed an unrealistic access time. There are 337 remaining responses following the removal of outliers, with a mean value of 951 hours and a median value of 150 hours. The mean value reported has decreased compared to the previous findings from Beagrie (2021), where respondents suggested it would take them a mean of 1,675 hours to recreate their EMBL-EBI data resources. However, the median reported value remains consistent with the Beagrie (2021) findings.

This survey question was modified in comparison to Beagrie (2016, 2021). Previously, the survey asked users how long it would take them to recreate their most recently used EMBL-EBI resource. This was updated to ask about their most frequently used resource, which is more likely to be representative of the typical time spent accessing EMBL-EBI data resources.

Q28: Please explain how you have estimated your answer in the previous question. (N = 224)

The LLM analysis of the open text responses identified the following themes:

- **Subjective / general impression:** 40% of respondents estimated the length of time it would take them to recreate a similar dataset based on general, qualitative impressions of the effort required, using phrases such as "a lot of effort," "a long time," or "tremendous effort," without specific details or numbers;
- **Concrete / experience-based estimation:** 33% of responses were based on specific and often quantitative estimates based on their own prior attempts, time tracking, or a clear breakdown of the tasks required to recreate the resource/tool;
- **Uncertain / unable to estimate:** 19% of respondents explicitly stated they did not know, could not estimate, or did not understand the question; and
- **Uncategorised:** The remaining 8% of responses were unable to be categorised.

We run a sensitivity analysis that excludes the responses where respondents expressed that they could not estimate or understand the question when calculating the average time spent recreating similar data resources (as estimated based on the responses to Question 27). This sensitivity results in the mean time reducing from 951 hours to 937 hours, whilst the median time remains unchanged at 150 hours.

A.4 Questions about the time spent conducting research with EMBL-EBI data resources

Q29: Do you conduct research as part of your role? Here we define research to include academic and clinical research, commercial R&D, and similar public or private sector activities where life science data and knowledge creation play a significant role. (N = 1,770)

The survey questionnaire provided respondents with two binary (i.e., yes / no) options. 90% of respondents reported that they do conduct research as part of their role, and 10% reported that they do not.

Q30: In the last 12 months, on average, how many hours did you spend conducting research each day? (N = 1,382)

The survey questionnaire asked respondents to report the time they spend conducting research each day in terms of hours. We removed responses that reported spending more than 13 hours per day conducting research, in line with Beagrie (2016, 2021). There were 1,357 remaining responses following the removal of outliers, with a mean value of 6.2 hours and a median value of 6 hours.

We revised the framing of the survey question in comparison to the previous Beagrie (2016, 2021) questionnaire. The previous questionnaire asked respondents to estimate the average number of hours per working week spent doing research, whereas we asked respondents to estimate the number of hours per day. By asking respondents to estimate the value with reference to a shorter time period, we sought to improve the accuracy and precision of responses.

Q31: Of the hours per day you spent conducting research, approximately what percentage of your research time was spent… (N = 1,030)

The questionnaire offered the following options:

- Working with EMBL-EBI data and tools (e.g., finding and analysing data) (%);
- Working with non-EMBL-EBI data and tools (e.g., collecting, finding, and analysing data) (%); and
- Performing other tasks (e.g., reading and writing articles, management) (%).

On average, respondents reported that they spent 17.5% of their research time working with EMBL-EBI data and tools; 38.1% of their research time working with non-EMBL-EBI data and tools; and 45.1% of their research time performing other tasks.

Combining the responses to Q30 and Q31, we estimate that respondents spend, on average, 1.1 hours per day working with EMBL-EBI data and tools.

Q32: What do you think might be typical for others in the same role and research field as you? (N = 795)

The questionnaire offered the following options:

- Working with EMBL-EBI data and tools (e.g., finding and analysing data) (%);
- Working with non-EMBL-EBI data and tools (e.g., collecting, finding, and analysing data) (%); and
- Performing other tasks (e.g. reading and writing articles, management) (%).

On average, respondents reported that they expect others to spend 19.2% of their research time working with EMBL-EBI data and tools; 36.9% of their research time working with non-EMBL-EBI data and tools; and 44.4% of their research time performing other tasks. This is similar to how they spend their own time.

A.5 Questions about the efficiency savings as a result of having access to EMBL-EBI data resources

Q33: Have EMBL-EBI data resources or tools saved you time and effort which you have then used on other research tasks? (N = 1,382)

The questionnaire offered the following options:

- No – it has not saved me time and effort (3%);
- Yes – a small amount of time and effort saved (12%);
- Yes – a moderate amount of time and effort saved (21%); and
- Yes – a significant amount of time and effort saved (64%).

Q34: How much time each day do you save by having access to EMBL-EBI data resources and tools (compared to if you didn't have access)? (N = 803)

The survey questionnaire asked respondents to report the time they save each day on an in terms of hours and minutes. We removed responses that implied a negative time saving or a time saving of more than 12 hours per day, as these responses were deemed as unrealistic. We also removed cases where respondents reported an identical time period in terms of the number or hours and minutes (e.g., 4 hours and 240 minutes), as these responses implied a misunderstanding of the response framework and risked introducing bias to the results. There were 749 remaining responses following the removal of outliers, with a mean value of 2.2 hours per day and a median value of 2 hours per day.

As described in Annex B.3, this survey question was phrased differently compared to Beagrie (2016, 2021). Previously, efficiency estimates were asked as a percentage of a respondent's working day.[68] As reported by Beagrie (2021), the framing of this question created ambiguity for respondents, as it did not explicitly state a time period from which users should estimate the share of their time that is saved. To remove this ambiguity, we asked respondents to directly provide an hourly estimate of the time saved by having access to EMBL-EBI data resources.

Q35: Please explain how you have estimated your answer in the previous question. (N = 398)

The LLM analysis of the open text responses identified the following themes:

- **Qualitative / experimental estimation:** 38% of respondents estimated the time savings based on a qualitative or experimental estimation, describing how EMBL-EBI resources

[68] For example, see question 32 of the user-survey at Appendix 1 of Beagrie (2021).

integrate into their work and how much more difficult or time-consuming tasks would be without them, but without explicit calculations.

- **Unable / unwilling to estimate:** 27% of respondents expressed that they were unwilling or unable to estimate a response. Respondents stated that they could not provide an estimate, found the question impossible, irrelevant, or refused to answer, often citing the essential nature of the resources or the difficulty of imagining an alternative.
- **Quantitative / comparative estimation:** 14% of responses were based on a quantitative or comparative estimation, using explicit calculations, comparisons, or time-tracking to estimate time saved, often referencing specific workflows, alternative methods, or time logs.
- **Guesswork / intuitive estimation:** 11% of responses were based on guesswork or intuitive estimation, often due to the hypothetical nature of the question or lack of direct comparison.
- **Uncategorised:** The remaining 11% of responses were unable to be categorised.

When calculating the average time saving (as estimated based on the responses to Question 34), we run a sensitivity analysis that excludes the cases where respondents expressed that (i) they were unwilling or unable to estimate a response, and (ii) their responses were based on guesswork. This sensitivity results in the mean time saving reducing slightly from 2.2 hours per day to 2.1 hours, whilst the median time remains unchanged at 2 hours per day.

A.6 Questions about secondary usage of EMBL-EBI data resources

Q36: Which of the following are part of your role (if any), and what is the impact of EMBL-EBI data resources or tools on those curated data resources? (N = 1,190)

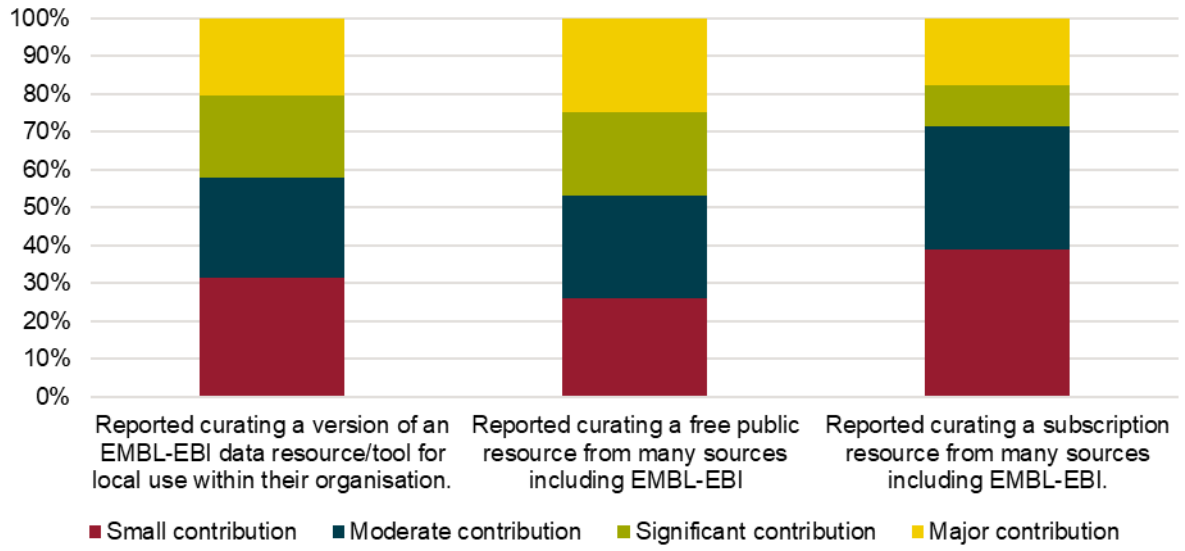
The questionnaire asked respondents to state whether the following activities were a part of their role:

- I curate a version of an EMBL-EBI data resource/tool for local use within my organisation;
- I curate a free public data resource from many resources including EMBL-EBI data resources; and
- I curate a subscription resource (user pays) from many sources including EMBL-EBI data resources.

As shown in Section 3.1, 12% of respondents reported curating a subscription resource based on EMBL-EBI data, 28% curated a free public data resource based on EMBL-EBI data, and 26% curated a version for local use within their organisation based on EMBL-EBI data. Section 3.1 also shows these trends over time based on the previous Beagrie (2016, 2021) findings.

Figure 41 also shows the reported contribution of EMBL-EBI data resources to those curated secondary resources.

Figure 41 Impact of EMBL-EBI data resources on curation of secondary resources



Source: *2025 EMBL-EBI User Survey; Beagrie (2016, 2021)*

Note: *Question 36 of the survey asked respondents "Which of the following are part of your role (if any), and what is the impact of EMBL-EBI data resources or tools on those curated data resources?" (N = 1,190)*

Q37: In the past 12 months, approximately how many people accessed the resource you are curating using EMBL-EBI data resources or tools? (N = 1,139)

The detailed results for this question are summarised in Section 3.1.

A.7 Questions about the value of EMBL-EBI data resources and tools

Q38: Imagine a scenario where access to all EMBL-EBI data resources and tools was no longer free. What is the maximum amount you would be willing to pay each year to retain your individual access to all EMBL-EBI data resources and tools? Please enter a whole number without punctuation. Your currency will be derived from the country you are currently working or studying in (which you provided in question 1). Note that this is a hypothetical scenario, and EMBL-EBI does not intend to charge for access to its services. (N = 798)

The questionnaire asked respondents to estimate their WTP in their local currency. Responses are converted into GBP using global average exchange rates for 2024.[69] To remove unrealistic or extreme values, we removed:

[69] We convert monetary values reported throughout the survey responses into GBP using data from the World Bank. Currencies are converted based on the latest full year of data (i.e. 2024). For exchange rates that do not have available data for 2024, we use the next full year of data.

- **All negative values:** one response reported a negative WTP;
- **Responses that reported a zero WTP on principle / a policy basis:** as described below, the LLM analysis of the open text section identified cases where the respondent expressed that scientific data should be freely accessible. Of these responses, we remove the cases where the reported WTP was zero. We apply this approach to maintain consistency with the previous Beagrie (2016, 2021) method; and
- **Observations more than 3 standard deviations away from the mean:** In the absence of an intuitive cutoff point, we employed a statistical approach that removes observations that are more than 3 standard deviations from the mean.

There are 761 remaining responses following the removal of outliers, resulting in a mean WTP of £3,352 per year and a median value of £89. The mean WTP has increased from £2,757 in 2021 (or £3,335 in 2025 real terms), whilst the median has decreased from £200 (or £241 in 2025 real terms).

Q39 How have you estimated the value you provided in the previous question? (N = 571)

The LLM analysis of the open text responses identified the following themes:

- **Personal / institutional budget constraints:** 35% of respondents here base their answer on what they or their institution could realistically afford, often referencing limited funding, grant restrictions, or personal income;
- **Comparison to other services / subscriptions:** 21% of responses set their value by benchmarking against the cost of other software, databases, or subscription services (e.g., ChatGPT, GraphPad, Netflix, commercial databases);
- **The value or benefit derived from EMBL-EBI:** 9% of respondents estimate value based on the utility, time saved, or critical importance of EMBL-EBI resources to their work, sometimes quantifying in terms of salary or productivity;
- **Principle / Policy (should be free / open access):** 8% of answers are driven by the belief that scientific data should be freely accessible, often referencing public funding, open science, or ethical arguments;
- **Guesswork / no clear basis:** 17% of respondents admit to guessing, using intuition, or not having a clear rationale for their answer;
- **Uncategorised:** the remaining 10% of responses are unable to be categorised.

We run a sensitivity analysis that excludes the responses based on guesswork or no clear basis when calculating the average WTP (as estimated based on the responses to Question 38). This sensitivity results in the mean annual WTP increasing from £3,352 to £3,484, whilst the median value increases from £89 to £100.

Q40: Imagine a scenario where you can either continue using EMBL-EBI data resources and tools or give up your access in exchange for a payment. What is the minimum amount you would have to be paid to give up your individual access to all EMBL-EBI data resources and tools for a year? Please enter a whole number without punctuation. Your currency will be derived from the country you are currently working or studying in (which you provided in question 1). Note that this is a hypothetical scenario, and EMBL-EBI does not intend to charge for access to its services. (N = 639)

The questionnaire asked respondents to estimate their willingness to accept (WTA) in their local currency. As with Q38, responses were converted into GBP using global average exchange rates for 2024, and we removed the following extreme or unrealistic values:

- All negative values;
- Responses that reported a zero WTA on principle / a policy basis; and
- Observations more than 3 standard deviations away from the mean.

There are 605 remaining responses following the removal of outliers, resulting in a mean WTA of £1m per year and a median value of £436. These values are notably higher than the average WTP given that survey respondents are no longer required to consider their budget constraints when reporting their WTA. On a conservative basis, we rely on the average WTP rather than the average WTA when estimating the overall value of EMBL-EBI data resources in Section 4.1.

Q41: How have you estimated the value you provided in the previous question? (N = 440)

The LLM analysis of the open text responses identified the following themes:

- **Rational / calculated estimate:** 28% of respondents made a conscious effort to estimate the value based on time, salary, cost of alternatives, or other rational calculations;
- **Value-based / principled refusal:** 18% of respondents refused to assign a monetary value, often citing principles of open science, irreplaceable value, or ethical objections;
- **Arbitrary / guesswork**: 19% of respondents did not use a specific methodology, instead providing a number based on intuition, randomness, or lack of understanding;
- **Confusion:** 18% of respondents expressed confusion, irrelevance, or inability to answer, often due to misunderstanding the scenario or lack of applicability;
- **Uncategorised:** the remaining 18% of responses were unable to be categorised.

Q42: What impact, if any, has access to EMBL-EBI data resources and tools had on your ability to… (N = 1,072)

The questionnaire asked this question with respect to the following practices:

- Train and/or evaluate machine learning or other AI models;
- Comply with funder/publisher open science requirements;
- Collaborate on research across disciplines;
- Provide or receive credit/citation for data outputs; and
- Validate the reproducibility of data outputs.

Section 3.3 provides an overview of responses in terms of the share of respondents that reported that EMBL-EBI data resources contribute to these practices. Table 2 provides a further breakdown of the share of respondents that reported that EMBL-EBI data resources made no contribution, a small contribution, a moderate contribution, or a significant contribution to these areas. The remainder of respondents reported that the question was "not applicable" to them, signalling that they do not conduct the practice at all.

Table 2 Impact of EMBL-EBI data resources on different areas

	Not applicable	No contribution	Small contribution	Moderate contribution	Significant contribution	Major contribution
Collaborate on research across disciplines	17%	5%	11%	22%	25%	20%
Validate the reproducibility of data outputs	21%	7%	12%	18%	22%	20%
Comply with funder/publisher open science requirements	30%	7%	10%	14%	17%	23%
Provide or receive credit/citation for data outputs	32%	8%	15%	15%	15%	14%
Train and/or evaluate machine-learning or other AI models	51%	6%	9%	11%	11%	11%

Source: *2025 EMBL-EBI User Survey*

Note: *Question 42 of the survey asked respondents "What impact, if any, has access to EMBL-EBI data resources and tools had on your ability to…" (N = 1,072)*

Annex B – Methodological approach for the impact modelling

The impact modelling estimates four headline value indicators quantifying the overall value of EMBL-EBI data resources, including:

- Willingness to pay for EMBL-EBI data resources;
- Self-reported efficiency impact achieved as a result of having access to EMBL-EBI data resources;
- The value of time users spend interacting with EMBL-EBI data resources; and
- The return on R&D enabled by access to EMBL-EBI data resources.

As described in Section 2.3.2, the process for estimating these value indicators follows four steps:

- Step 1: Estimate the user base;
- Step 2: Analyse the user survey;
- Step 3: Scale up the user survey results to the wider user population; and
- Step 4: Calculate the BCRs.

This annex sets out the detailed methodological approach for these four steps.

B.1 Estimated user base

We estimate the total size of the EMBL-EBI user base based on EMBL-EBI's annual web-log data. In line with Beagrie's (2016, 2021) methodology, this involves converting unique IP addresses accessing EMBL-EBI data resources to users, with a few adjustments. To estimate the user base for the latest calendar year (2024), we take a four step approach:

- **Step 1: Establish the ratio of users to unique IP addresses:** Fomitchev (2010) analysed website traffic logs and found that counts based on unique IP addresses or cookies systematically overstate the true number of individual visitors by a ratio of 1:80.[70] The degree of overestimation increases linearly with the duration of data collection, arising from factors such as users changing IP addresses, employing multiple devices, and accessing from different network locations. We take this 1:80 ratio as our starting point for estimating the user base, as in Beagrie (2016, 2021).
- **Step 2: Account for other methods of accessing EMBL-EBI data resources:** accesses via other methods, such as the use of FTP or Aspera, are not captured by the internal EMBL-EBI IP data. Based on consultation with the IT team at EMBL-EBI, they indicated that the share of users using these alternative access methods accounts for around 5% of all users. To capture these additional users, we uplift the ratio between

[70] Fomitchev (2010), "How Google Analytics and conventional cookie tracking techniques overestimate unique visitors."

unique users and the unique IP addresses by 5%. This is higher than the 2.5% figure in Beagrie (2021).

- **Step 3: Account for cases where multiple users sit behind a single IP address:** in some instances, there may be multiple users accessing EMBL-EBI from a single IP address if, for example, they use a common institutional or university access point. The user survey received 2,563 responses from 2,450 unique IP addresses, (i.e., 1.05 respondents per unique IP), suggesting that the ratio between unique users and the unique IP addresses should be uplifted by 5%. This is slightly lower than the 6% figure in Beagrie (2021).
- **Step 4: Account for the increase in hybrid working since the pandemic:** the COVID-19 pandemic had a systematic impact on the rate of hybrid working. The Fomitchev (2010) study pre-dates the pandemic, so does not account for some individuals now using multiple devices in multiple locations. This has the effect of artificially increasing the number of IP addresses. In the absence of post-pandemic IP to user ratio data to replace Fomitchev (2010), we apply a final adjustment to account for the increase in hybrid working since the pandemic. Aggregated EU data from Eurostat suggests that the share of the population working on a hybrid basis in 2024 was 9 percentage points higher than the five year average prior to the pandemic (i.e., between 2015 and 2019).[71] Assuming that aggregate EU-wide trends are representative of the EMBL-EBI user population, we scale down the ratio of unique IP addresses to unique users by 9 percentage points.[72] Future analyses should seek to apply an updated ratio of users to unique IP addresses based on post-2020 data to account for the influence of the pandemic.[73]

Internal EMBL-EBI data suggests that 41m unique IP addresses accessed EMBL-EBI's online portal in 2024. Based on the above steps, this results in a user base estimate for 2024 of around 520,000 unique users. This marks a 4% to 15% increase in the user base since the previous Beagrie (2021) study, which reported a user base of between 450,000 and 500,000.

[71] See data from Eurostat on Employed persons working from home by professional status - % of total employment. We use the annual share of employed persons who "sometimes" or "usually" work from home as a proxy for the annual share of employed persons who work on a hybrid basis.

[72] This approach using aggregated EU data reflects the best available evidence on trends in hybrid working both before and after the pandemic. Industry-level time series data measuring hybrid working trends both before and after the pandemic were not available.

[73] We tested multiple approaches to account for the impact of the pandemic on the ratio of unique IP addresses to unique users. An alternative approach involved building upon Beagrie's (2021) approach, which attributed half of the 2019-20 growth rate in IP addresses to the working from home phenomenon (as described at page 21 of Beagrie (2021)). Based on a review of various industry-level UK and US data sources, we find that between 20% and 60% (midpoint: 40%) of people in industries related to life science research worked on a hybrid basis in 2024. Assuming that 100% of these people worked from multiple locations (i.e. on a hybrid basis) during 2020 as workers switched from office or lab based work to working remotely, an alternative approach involves scaling back Beagrie's pandemic adjustment by 60% (i.e. 100% minus 40%). This approach also results in a final user base estimate of around 520k, providing confidence in our estimate.

The 520,000 estimate is conservative. More recent evidence from Mishra et al. (2020) indicates that the yearly ratio used in Step 1 above could be as low as 1:40.[74] Using this ratio would double the user base estimate, thereby doubling the value indicators estimated throughout the report. In addition, our estimate of EMBL-EBI's direct user base does not capture the usage of EMBL-EBI data resources through secondary resources. If the value of EMBL-EBI to its secondary users were also considered, the indicators presented throughout this section would also increase. The user base estimate used in this report should therefore be considered a lower bound.

Based on consultation with the EMBL-EBI team, we understand that users increasingly rely on modern access routes that are not captured by EMBL-EBI's internal web-log data, making it increasingly difficult to accurately measure the user base. For example, cloud providers (behind which multiple users may sit) appear as a single IP in EMBL-EBI's usage data. Whilst we provide an indicative adjustment for this in our estimates, future work should more robustly assess the number of users accessing data via these new routes.

B.2 Willingness to pay

The willingness to pay (WTP) indicator is based on the user survey. It asked the amount survey respondents would be willing to pay each year to retain access to all EMBL-EBI data and tools, to which respondents reported an average value of £3,352.[75,76] It provides an indication of the direct benefits users derive from having access to EMBL-EBI data resources, and their monetary valuation of that.

The aggregate willingness to pay is calculated by scaling up the mean user-level willingness to pay across the wider EMBL-EBI user population, using the formula below.

$$Aggregate\ willingness\ to\ pay = \\ mean\ willingness\ to\ pay * estimated\ user\ base$$

As in the previous Beagrie studies, survey respondents were also asked the minimum amount that they would be willing to accept each year to give up access to EMBL-EBI data resources and tools.[77] This question removes the budgetary constraints associated with the willingness

[74] Mishra et al. (2020), "Don't count me out: On the relevance of IP addresses in the tracking ecosystem." The authors suggest that users retain a unique IP address for, on average, 9.3 days. To derive the annual ratio of users to unique IP addresses, we divide 365 days by 9.3 days per unique IP address (i.e. approximately 1:40).

[75] Question 38 of the survey asked respondents "Imagine a scenario where access to all EMBL-EBI data resources and tools was no longer free. What is the maximum amount you would be willing to pay each year to retain your individual access to all EMBL-EBI data resources and tools? [...] Note that this is a hypothetical scenario, and EMBL-EBI does not intend to charge for access to its services."

[76] Survey respondents were asked to report monetary values in their local currency. All values in this report have been converted to Great British Pounds (GBP), using global average exchange rates from the World Bank as of 2024.

[77] Question 40 of the survey asked respondents "Imagine a scenario where you can either continue using EMBL-EBI data resources and tools or give up your access in exchange for a payment. What is the minimum amount you would have to

to pay question. For this reason, the reported values were particularly high.[78] In the interest of providing conservative value estimates throughout the impact modelling, we rely on users' stated willingness to pay rather than their willingness to accept.

B.3 Efficiency impact

The efficiency impact indicator assesses the total value of time saved by users from having access to EMBL-EBI data resources. This measure of productivity is based on the responses to the user survey, where respondents were asked how much time they save each day as a result of having access to EMBL-EBI data resources and tools.[79] The average reported figure was 2.2 hours per day, or 11 hours per working week.

The survey question underpinning this indicator was phrased differently to the Beagrie reports. Previously, efficiency estimates were asked as a percentage of a respondent's working day.[80] As reported by Beagrie (2021), the framing of this question created ambiguity for respondents, as it did not explicitly state a time period from which users should estimate the share of their time that is saved. Beagrie therefore reported a range of possible values for the overall efficiency saving to overcome this. To remove this ambiguity, we asked respondents to directly provide an hourly estimate of the time saved by having access to EMBL-EBI data resources. A range is therefore no longer required for this indicator.

To calculate this indicator, average user-level time savings (in hours per day) were converted to a pound value by applying an estimated average hourly value of user time. This per-user value was then scaled up to the wider EMBL-EBI user population, using the formula below.

$$Efficiency\ impact = (mean\ daily\ hours\ saved\ by\ having\ access\ to\ EMBL-EBI) * hourly\ value\ of\ time * working\ days\ per\ year * estimated\ user\ base$$

The hourly value of EMBL-EBI users' time is based on the estimate from the previous Beagrie (2021) study, uplifted for inflation.[81] This gives a figure of £46 per hour for 2025. While more recent data was sought to update the original Beagrie analysis, such data was not available.[82]

be paid to give up your individual access to all EMBL-EBI data resources and tools for a year? […] Note that this is a hypothetical scenario, and EMBL-EBI does not intend to charge for access to its services."

[78] When survey respondents were asked about their willingness to accept to give up access to all EMBL-EBI data resources and tools for a year, 14% of responses exceeded £100k, 7% of responses exceeded £1m and 3% of responses exceeded £10m.

[79] Question 34 of the survey asked respondents "How much time each day do you save by having access to EMBL-EBI data resources and tools (compared to if you didn't have access)?"

[80] For example, see question 32 of the user-survey at Appendix 1 of Beagrie (2021).

[81] We uplift values to account for inflation using the UK GDP deflator index.

[82] The original hourly value of £38 was calculated by mapping individual respondents to different salary bands (based on their reported job type), using a range of UK data sources. For more detail, see page 22 of Beagrie (2021).

However, as the demographic composition of the most recent user survey is similar to the Beagrie (2021) user survey (as described in Section 3), the average value of users' time is unlikely to have changed much besides inflation (which we adjust for).

B.4 Value of time using EMBL-EBI data resources

This indicator estimates the value users derive from EMBL-EBI data resources based on their time spent using it. This is because users must, as a minimum, value the resource equal to their time spent using it, otherwise it would not be rational for them to use the resource (as they would spend their time elsewhere).

The value of time spent using data resources is an imperfect proxy of value. On the one hand, increases in time spent could reflect users deriving more value from EMBL-EBI data resources (as measured by the willingness to pay indicator). On the other, decreases in time spent could reflect productivity improvements from EMBL-EBI data resources (as measured by the efficiency indicator). The latter would lead to a decrease in the value indicator, despite this being driven by an improvement in the service. Given the potential ambiguity, the reported figures should only be assessed with the context of the other value indicators presented throughout this report.

In line with the previous Beagrie (2016, 2021) studies, we provide multiple estimates for the value of time spent using EMBL-EBI data resources based on the user survey:

- The value of access time captures the median time respondents spent accessing EMBL-EBI data resources each year; and
- The value of use time measures the mean time they spent working with EMBL-EBI data each year.[83]

The value of access time is calculated by multiplying the median time taken for users to access EMBL-EBI data resources[84] (5 minutes) by the median number of accesses per user per year (199).[85] As per the formula below, this user-level estimate is scaled up across the wider user population, and transformed into a monetary value by applying the same value of time assumption from the efficiency impact indicator (£46 per hour).

[83] The median time taken to access EMBL-EBI resources is used instead of the mean value to reduce the influence of cognitive estimation bias on the results. The median number of annual accesses is used instead of the mean to minimise the influence of outliers and to maintain consistency with the previous Beagrie (2016, 2021) approach.

[84] Question 21 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. Approximately how long did it take you to find and obtain the EMBL-EBI data resource or tool you were looking for?"

[85] Questions 7, 9, 11, 13, 15 and 17 of the survey asked respondents "In the last 12 months, which of these […] EMBL-EBI data resources did you use? And approximately how often did you access and/or download from each of these resources?"

$$Value\ of\ access\ time\ =$$
$$(median\ hours\ taken\ to\ access\ most\ frequently\ used\ resources$$
$$*\ median\ number\ of\ annual\ accesses)\ *$$
$$hourly\ value\ of\ time\ *\ estimated\ user\ base$$

The survey question to capture the median time taken to access EMBL-EBI data resources was modified in comparison to Beagrie (2016, 2021). Previously, the survey asked users how long it took them to find and obtain their most recently used EMBL-EBI data resource. This was updated to ask about their most frequently used, which is more likely to be representative of a typical access time. Despite the change in methodology, the median access time is the same as in Beagrie (2016, 2021) at 5 minutes per access.

A similar approach was taken to calculating the value of use time. Survey respondents were asked to estimate how many hours, on average, they spend conducting research per day (6.2), and what percentage of this time was spent working with EMBL-EBI data and tools (18%).[86] As per the formula below, this was scaled up across the wider user population, and converted into a monetary value by applying the same value of time assumption (£46 per hour).

$$Value\ of\ use\ time\ =$$
$$\begin{pmatrix} mean\ hours\ spent\ with\ research\ per\ day\ * \\ share\ of\ research\ time\ spent\ with\ EMBL-EBI\ data \end{pmatrix} *$$
$$hourly\ value\ of\ time\ *\ working\ days\ per\ year\ *\ estimated\ user\ base$$

B.5 Return on R&D enabled by EMBL-EBI data resources

This indicator captures the wider societal impacts from researchers having access to EMBL-EBI data resources. It measures the social returns on the research that is directly enabled by EMBL-EBI, including:

- **The research time spent using EMBL-EBI data resources that otherwise could not have occurred.** This is calculated by taking the value of time indicator above, multiplying this by the share of that research time that could not have occurred if users did not have access to EMBL-EBI, and applying a rate of return to R&D for this portion of time. The share is based on the proportion of survey respondents who reported that they would not have been able to (i) find or obtain an equivalent version of their most frequently used

[86] Question 30 of the survey asked respondents "In the last 12 months, on average, how many hours did you spend conducting research each day?" Question 31 of the survey asked respondents "Of the hours per day you spent conducting research, approximately what percentage of your research time was spent working with EMBL-EBI data and tools (e.g. finding and analysing data)?"

EMBL-EBI data resource or tool from another source,[87] or (ii) re-create the same or a similar version of their most frequently used EMBL-EBI data resource or tool.[88]

- **The time saved by users from having access to EMBL-EBI data resources.** This is equal to the time that could be spent on other productive work activities as a result of EMBL-EBI users working more productively. This is equivalent to applying a rate of return to R&D to the efficiency impact indicator described above.

The calculation approach is outlined in the formula below. This approach to capturing the returns on enabled R&D builds on the previous Beagrie (2016, 2021) method as it also incorporates the impact of research efficiency.[89] To aid comparability with the previous studies, we reconstruct equivalent historical values for Beagrie (2016, 2021) to allow a comparison over time.

$$\begin{gathered} Return\ on\ enabled\ R\&D = \\ Rate\ of\ return\ on\ R\&D * \\ (value\ of\ research\ time\ that\ otherwise\ could\ not\ have\ occurred\ + efficiency\ impact), \end{gathered}$$

$$\begin{gathered} where: \\ value\ of\ research\ time\ that\ otherwise\ could\ not\ have\ occurred = \\ (mean\ daily\ hours\ spent\ with\ EMBL-EBI\ data * \\ \%\ of\ usage\ that\ could\ not\ have\ been\ obtained\ elsewhere\ or\ recreated) * \\ hourly\ value\ of\ time * working\ days\ per\ year * estimated\ user\ base \end{gathered}$$

The rate of return to R&D parameter above captures two things: 1) the private return (e.g., the profits derived by a firm as a result of their research) as well as 2) wider spillover effects (e.g., spillover effects on the output of firms that were not involved in the R&D yet still benefit from the knowledge and data created. Beagrie (2016, 2021) reviewed the empirical literature estimating the value of this parameter, and concluded that the average return was around 40% (albeit with a high degree of variance). Having reviewed more recent literature, we consider the 40% figure to still be appropriate, and therefore apply this to our estimates.[90,91] To appropriately capture the value of all future returns on enabled R&D, we also report the NPV of the annualised returns. This ensures that we comprehensively capture the returns on

[87] Question 23 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to find and obtain equivalent data or tools from another source?"

[88] Question 26 of the survey asked respondents "Think back to the EMBL-EBI data resource or tool you used most frequently in the last 12 months. If you no longer had access to this EMBL-EBI data resource or tool, would you be able to re-create the same or similar dataset or tool yourself?"

[89] See pages 27-29 of Beagrie (2021) for an overview of their approach to capturing the returns on enabled R&D.

[90] See Frontier Economics (2023), "Rate of Return to Investment in R&D." The authors suggest that, on a conservative basis, one might expect an average social rate of return on R&D of around 40%.

[91] See Frontier Economics (2024), "Returns to Public R&D." The authors suggest that the average rate of return on public R&D in terms of increasing private sector productivity growth is up to 40%.

R&D enabled by EMBL-EBI data resources for as long as the knowledge or data created by the R&D delivers benefits to society.

Calculating the NPV requires estimating how the value of the returns on R&D evolve over time. This is subject to uncertainty and depends on the share of the R&D facilitated by EMBL-EBI data resources that takes place in the public sector versus the private sector.[92] To reflect this uncertainty, we provide a range:

- **Upper bound NPV:** we use the same parameters as in the previous Beagrie (2016, 2021) studies. Based on a review of the literature, Beagrie assume that (i) the data / knowledge created by the R&D will remain useful for an average of 30 years, (ii) the returns will ramp up uniformly over the course of the first nine years; and (iii) the data / knowledge created will depreciate at an annual rate of 5%. These parameters are consistent with EMBL-EBI data supporting research conducted largely by the public sector.[93] A discount rate of 3.5% is applied to measure the future returns in 2025 real terms.[94]
- **Lower bound NPV:** this aims to capture a scenario where the knowledge created by the R&D depreciates more quickly and delivers benefits over a shorter time horizon. Based on a review of the literature, we assume that (i) the data / knowledge created by the R&D will remain useful for an average of 20 years, (ii) the returns will ramp up uniformly over the course of the first nine years; and (iii) the data / knowledge created will depreciate at a rate of 15% over time. These assumptions are more consistent with EMBL-EBI data supporting research conducted largely by the private sector.[95] A discount rate of 3.5% is applied to measure the future returns in 2025 real terms.

As shown by the demography of the user survey respondents presented in Section 3, EMBL-EBI users are split across the public and private sectors, with some users working for corporations and some users working for hospitals or in the academic sector. This mix of the EMBL-EBI researcher population suggests that the NPV of the returns on enabled R&D should lie somewhere between our lower bound and upper bound values.

---

[92] Our review of the literature suggests that, in comparison to private sector research, public sector research might deliver benefits for a longer time period and depreciate more slowly if it is aimed at building general purpose knowledge that can be re-used and built upon. For example, Frontier Economics (2024) report that "public knowledge is often thought of being built on incrementally rather than being rendered obsolete by new innovations (as may happen in the case of firm-specific knowledge and R&D)".

[93] For example, a recent study for the UK Department for Science, Innovation & Technology (2025) suggests that a depreciation rate of 5% may be suitable for rates of return to public R&D. The study also suggests that an appraisal period of as long as 60 years is suitable when considering the benefits of public R&D. On a conservative basis, we use a 30 year appraisal period in line with the previous Beagrie (2016, 2021) approach.

[94] A discount rate of 3.5% is compliant with HM Treasury's Green Book guidance.

[95] For example, a recent study for the UK Department for Science, Innovation & Technology (2025) suggests that a depreciation rate of 15% may be suitable for rates of return to private R&D. Frontier Economics (2023) also suggest that "the majority of studies assume a 15% depreciation rate for returns to private knowledge". By increasing the depreciation rate for our lower bound estimate, we already account for the fact that private R&D generates benefits over a shorter time period compared to public R&D. On a conservative basis, we also reduce the service life to 20 years when estimating the lower bound NPV.

B.6 Benefit-cost ratios

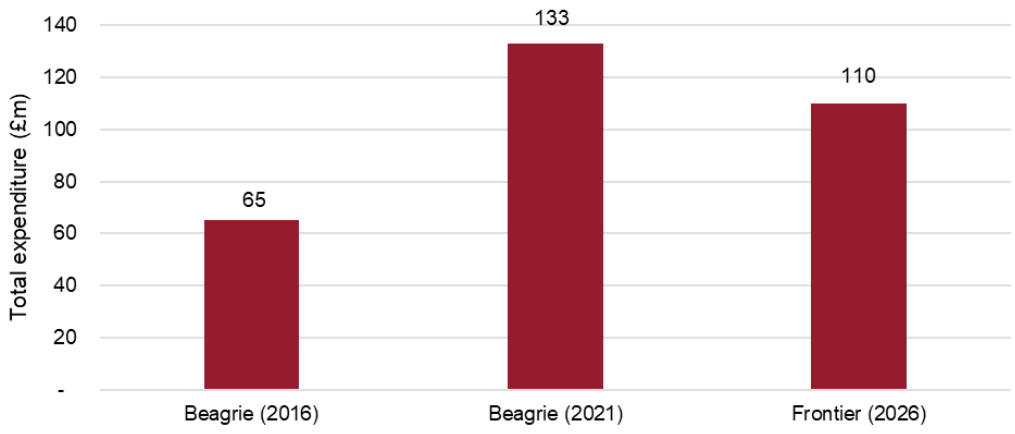
In line with HM Treasury's Green Book guidance, we calculate BCRs for each of the four headline indicators above. This provides a robust basis for establishing the value for money of investments in EMBL-EBI data resources.

To calculate BCRs, the total estimated benefits for each indicator (in £s) in a given year are divided by the total cost of maintaining and making available the EMBL-EBI data resources in that year. Cost data for the calendar year 2024 (£75m) was provided by EMBL-EBI, and includes all expenditures relating to the running of EMBL-EBI's data services, including shares of annualised capital expenditure, central services and other overheads. We convert these figures into 2025 real terms using the UK GDP deflator, in line with Green Book guidance. In line with Beagrie (2016, 2021), the cost associated with the original creation and deposition of data (as part of the research process) that is then hosted by EMBL-EBI is however treated as 'sunk', and therefore excluded.

In-kind contributions from collaborators and partners are not however captured in the EMBL-EBI cost data. To account for the full cost of providing these services, an indicative 30 percent uplift is applied to the EMBL-EBI cost data to account for these contributions. This is based on internal consultations with EMBL-EBI's finance department, as per the Beagrie (2016, 2021) reports.

Following this approach, we estimate total annual expenditure on EMBL-EBI data resources to be around £110 million in 2025 real terms. Figure 42 shows how this cost compares to the historical figures from the previous Beagrie (2016, 2021) analyses in 2025 real terms. Whilst the total expenditure has nearly doubled since the 2016 study (for which costs are reported using the average annual expenditure between 2012-14), real expenditure has decreased by over £20 million since the 2021 study (for which costs are reported for 2020).

Figure 42 Total costs of running EMBL-EBI data services



Source: Frontier analysis of EMBL-EBI cost data; Beagrie (2021); Beagrie (2016).
Note: All values are shown in 2025 real terms based on an adjustment using the latest UK GDP deflators.

Annex C – LLM analysis methodology

A large language model (LLM) is an advanced machine learning tool that analyses text to identify patterns, meanings and relationships between words. The model is trained on very large volumes of written material and learns to recognise common themes, sentiments and associations. It can then be used to analyse new text and to group, label and summarise it according to its content.

We use LLM methods to uncover themes and insights from all open text responses to the user survey, allowing us to extract insights that would be difficult or impractical to obtain using more conventional approaches. The LLM methods employed in this report involve applying a 4-stage, multi-agent approach to open-text responses:

- **Clustering agent:** Suggests clusters with which we can categorise the open text responses. These categories help give structure and identify insights from the unstructured responses.
- **Classification agent:** Suggests an initial categorisation for each response based on both the open-ended question and related questions. For example, when analysing the responses to the open-ended question "how did you estimate your time saved from using EMBL-EBI tools", the agent also considers the corresponding time saving estimates.
- **Judging agent:** Reviews the respondents' answers and the classification agent's categorisation, returning a final categorisation, rationale, confidence level, and highlighting any errors in the initial categorisation.
- **Human evaluation:** Finally, the final categorisations are reviewed by a human. Priority is given to responses with low confidence and errors.

The result is a categorisation for all open text survey responses. It provides a structured and transparent way to analyse a large volume of open-text data, improving the robustness of our interpretation and enabling clearer identification of user perspectives.

The LLM analysis provided an overview of the main themes in respondents' comments and an indication of where uncertainty or bias may have influenced responses. For example, the analysis identified:

- comments showing that some users' willingness-to-pay estimates were shaped by personal or institutional budget limits, highlighting that some responses should be interpreted as a lower bound estimate;
- references to guesswork, uncertainty, or difficulty in estimation when reporting values, highlighting areas where users' responses should be interpreted with caution; and
- positive sentiments about the value and impact of EMBL-EBI resources, which supported our broader findings on the benefits of access to open data.

Further details of the insights from the LLM analysis are included throughout the report as well as in Annex A .

Annex D – Open data case study: quantitative methods

To analyse the distributional effect of the AlphaFold Database, we use two divergence metrics: the Hellinger Distance (HD) and the Number of Effective Categories (N_{eff}). This annex outlines the equations for each of these metrics, what they measure, and our sampling approach.

D.1 Hellinger Distance

Hellinger Distance (HD) is a statistical measure of dissimilarity between two probability distributions, which ranges from 0 (identical distributions) to 1 (completely disjoint). The equation for HD is:

$$HD(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$$

Given two probability distributions, *P* and *Q*, where in our analysis the distribution is the proportion of papers represented by each unique research topic in a dataset of research papers:

1. Take the square root of the probability for each topic in both distributions.
2. Find the difference between the square roots for each topic i, $\sqrt{p_i} - \sqrt{q_i}$
3. Square each of those differences to make them positive.
4. Add up all those squared differences across all topics. This gives you a measure of how far apart the two distributions are overall.
5. Take the square root of that sum.
6. Finally, multiply by $\frac{1}{\sqrt{2}}$ to normalise the distance so it always falls between 0 and 1.

We use the Hellinger Distance to measure difference in the research topics between two sets of papers, because:

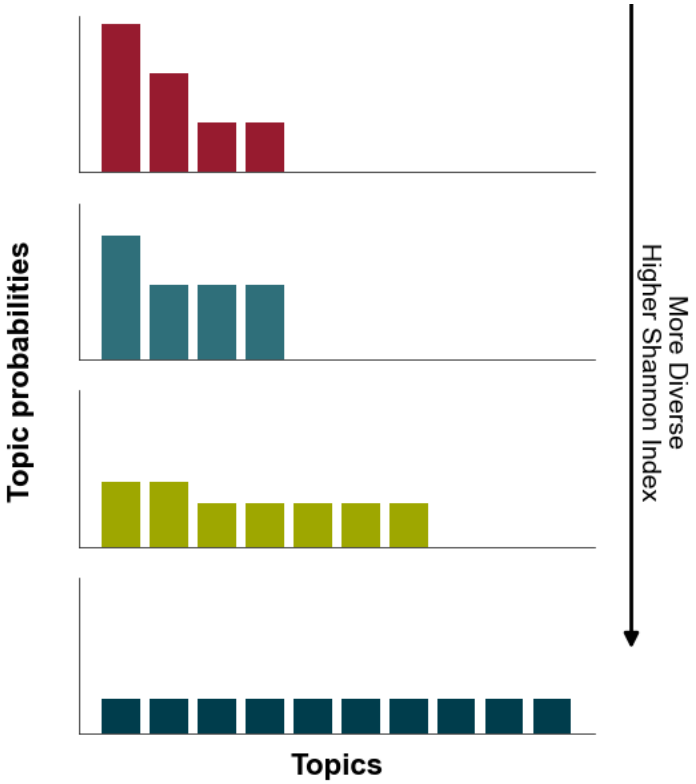
- It is a robust analytical method for comparing differences across all topics researched using different tools.
- Handles long-tailed distributions effectively (like the many rare topics from OpenAlex).
- Provides a single, interpretable value of similarity that can be used to compare differences across different pairs of resources.

D.2 Shannon Index / Number of Effective Categories

The Shannon Index (SI) is a measure of diversity in a distribution, where higher values imply more evenly spread categories, and lower values imply that the distribution is dominated by

fewer larger categories.[96] Figure 43 illustrates how the Shannon Index changes with the shape of the probability distribution when we use research topics as our variable of interest.

Figure 43 Shannon Index diagram



Source: Frontier Economics

The equation for the Shannon Index is:

$$SI(P) = -\sum_{i=1}^{k} p_i \ln(p_i)$$

Given a probability distribution, *P*, where in our analysis the distribution is the proportion of papers represented by each unique research topic or published location in a dataset of research papers:

1. Take each probability p_i in the distribution – that's the probability of each category i.
2. Compute the logarithm of that probability $\ln(p_i)$. This tells you how "surprising" or "informative" that category would be if it occurred.
3. Multiply $p_i \times \ln(p_i)$. This weights the information content of each outcome by how likely it is to happen.

[96] In this section, "categories" can be taken to represent research topics and published locations, depending on the variable of interest.

4. Sum all those $p_i \ln(p_i)$ values across every category. This gives you the expected (average) information content.
5. Negate the total (multiply by −1) to make the result positive.

In our analysis, we use the Number of Effective Categories (N_{eff}), which makes the Shannon Index more interpretable, and can be interpreted as the effective number of equally common categories that would generate the same diversity. It is calculated by exponentiating the SI:

$$N_{\text{eff}} = e^{SI}$$

We use N_{eff} to measure the diversity of categories, because:

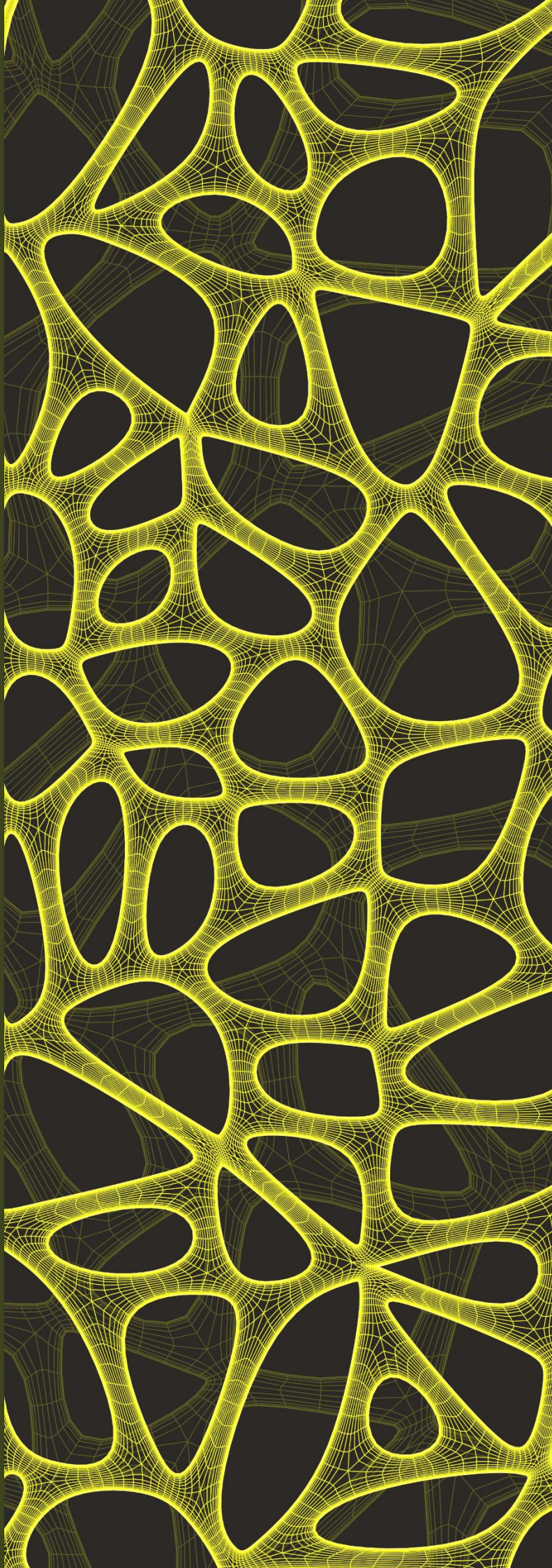
- It captures how evenly papers are distributed across categories.
- It complements the Hellinger Distance as a single, comparable measure of diversity.
- It is well-suited for long-tailed topic distributions with many rare categories (like topics and published locations from OpenAlex).

D.3 Sampling approach

A key challenge in applying these metrics to our bibliometric dataset is the imbalance in the number of papers for each tool: the AlphaFold Database (4,925 papers) and AlphaFold 2 algorithm are mentioned by far fewer papers than PDB (148,692 papers). To ensure a fair comparison between the tools, we down-sample the set of papers for each tool to the number of papers in the sample with the smallest size (i.e., the subset of papers mentioning only the AlphaFold Database). In other words, we take a representative sample of papers so there are an equivalent number of papers used to calculate the HD and N_{eff} for each tool.

To assess whether any observed differences across the dimensions are statistically significant, we bootstrap our calculations over 1,000 runs and calculate 95% confidence intervals for both absolute values of and differences in Hellinger Distances and the Number of Effective Categories.[97]

[97] Bootstrapping is a statistical resampling method where new datasets are repeatedly generated by sampling with replacement from the original data. We use it to estimate confidence intervals around each metric without relying on strong assumptions about the underlying distribution.



WWW.FRONTIER-ECONOMICS.COM