CERN IdeaSquare Journal of Experimental Innovation, 2025; 9(1): 42-48 ***ORIGINAL ARTICLE***
DOI: https://doi.org/10.23726/cij.2025.1597

The Impact of Research Data Infrastructures: The Case of the AlphaFold Database

Angelo Kenneth Romasanta[1], Jonathan Wareham[1], Laia Pujol Priego[1]

[1]Esade Business School, Ramon Llull University, Barcelona, Spain
Corresponding author: angelokenneth.romasanta@esade.edu

ABSTRACT

While the scientific output of research infrastructures is well documented, the broader effects of their secondary outputs, such as computational resources and datasets, remain poorly understood. To better understand the benefits of these public resources, this study explores the AlphaFold (AFDB) database, a collaboration between DeepMind and the European Molecular Biology Laboratory (EMBL) that democratizes access to protein structure data. Employing a quantitative case study strategy using bibliometric analysis, this study compares publications indexed in the Web of Science Core Collection citing the original AF paper (Jumper et al., 2021) (n=13,049) with those citing the AlphaFold database (Varadi et al., 2022) (n=659), covering publications up to August 2024. We examine the impact of the EMBL AlphaFold database on research themes, collaboration patterns, and scientific impact. Our exploratory analysis identifies several impacts: studies leveraging the AF database investigate application-focused themes and require collaboration between fewer institutions. This research highlights the wide-ranging impacts of research infrastructures, emphasizing the need for comprehensive impact assessments to inform future research policy and funding decisions.

Keywords: Research infrastructure; AlphaFold; bibliometrics; scientific impact; research evaluation.

Received: September 2024. Accepted: April 2025.

INTRODUCTION

Research Infrastructures (RIs) are essential pillars of modern science, facilitating the discovery of breakthroughs and innovations (Reed, et al., 2021; Scarrà & Piccaluga, 2022; Florio & Sirtori, 2016). Traditionally, their success has been measured by publications and patents; however, in an increasingly data-driven research landscape, this narrow view often overlooks the significant ripple effects from their secondary outputs including datasets, software, computational resources and tools (Wareham et al., 2022; D'ippolito & Rüling, 2019, Pujol Priego and Wareham, 2023). These investments represent significant public expenditure, exemplified by recent European funding calls allocating figures such as €221.5 million for new RI projects (European Research Executive Agency, 2023). As neglecting these secondary outputs can lead to an incomplete picture of an RI's value, it is then crucial to take them into account to holistically assess the return on investment in RIs and to guide future science policy.

The recent development of AlphaFold (AF), an artificial intelligence system for protein structure prediction, provides a unique opportunity to explore the impact of research infrastructure outputs (Jumper et al., 2021). Created by DeepMind, AlphaFold has been hailed as one of the most important breakthroughs in biology, but the computational resource requirements initially limited its widespread use. To address this constraint, a collaboration between DeepMind and the European Molecular Biology Laboratory (EMBL) led to the creation of the AlphaFold database (AFDB), a shared scientific data infrastructure for the millions of protein structures predicted by AlphaFold (Varadi et al., 2022).

This distinction between the algorithmic breakthrough (AF) and the readily accessible database (AFDB) provides a valuable quasi-experimental setting. While AF represents the fundamental breakthrough, AFDB is the secondary output by an RI designed to democratize access to the results of that development. The database adds significant value by making these predictions readily accessible to researchers without the need for extensive computational resources and expertise. Investigating the downstream scientific outcomes citing AFDB, in comparison to those citing the original AF paper, allows us to explore further the specific impact of such open data infrastructures.

Understanding the differential impact of the AFDB is critical for RI policy and funding. Does making complex data readily accessible truly foster wider use? Does it enable different research themes? Does it stimulate impactful, downstream research? Without such studies, the full value of these resources may not be adequately accounted for in policy decisions, potentially hindering wider investments in similar infrastructures. In summary, we seek to answer the following research question: *How*

does the AlphaFold database impact downstream scientific activity – specifically research themes, collaboration patterns and scientific impact – compared to the original algorithmic breakthrough? We conduct a quantitative bibliometric analysis using data from Web of Science. This research serves as an important foundation for the COMPUTE IMPACT project funded by ATTRACT, which aims to holistically evaluate the impact of computational tools and datasets generated by research infrastructures.

THEORETICAL BACKGROUND

Impact assessment of RIs has traditionally been focused on primary scientific outputs such as publications and patents (Heidler, & Hallonsten, 2015; Mayernik et al., 2021). However, this focus fails to capture the full spectrum of their contributions (Florio & Sirtori, 2016; Autio et al., 2004), particularly the increasingly important role played by secondary outputs including curated datasets and computational tools (e.g. Beagrie & Houghton, 2021). In an era increasingly defined by big data and increasing calls for Open Science (Vicente-Saez & Martinez-Fuentes, 2018), a more comprehensive understanding of the diverse ways RI investments spread beyond their immediate scientific purview is paramount for scientific policy.

The Open Innovation in Science framework (Beck et al., 2021, 2022) offers a useful lens for understanding these broader impacts. OIS refers to the process of managing knowledge flows across boundaries throughout the entire scientific research process, from conceptualization to dissemination and adoption by industry (Beck et al., 2020). Within this framework, open RI outputs like AFDB can be conceptualized as a *"shared scientific infrastructure"*, facilitating knowledge sharing between developers of scientific breakthroughs (e.g. DeepMind) and downstream users across various scientific and industrial domains. By lowering barriers to access and use, it is expected that these infrastructures democratize participation in science and accelerate knowledge diffusion.

Drawing on the OIS framework, we can theorize various potential impacts of AFDB in lowering barriers to entry and democratizing access to scientific resources. However, the specific effects of such open data infrastructures are not predetermined and can be multifaceted.

First, by providing readily usable data, such research infrastructure might reduce the time, resources and expertise needed for certain research phases (Romasanta et al., 2022; Fabre et al., 2021). As theorized by OIS, such accessible infrastructures can facilitate knowledge flows into new application domains. This democratization of science could allow a wider range of topics to be explored more easily and shift research towards downstream applications. Alternatively, the primary effect might be simply enabling more research within established thematic areas, effectively amortizing the costs fundamental research across a broader base (Pujol Priego and Wareham 2024). In this scenario, the distribution of research themes might not differ significantly with AFDB primarily acting as an accelerator for existing research lines.

Second, open resources can enable researchers from resource-limited institutions to contribute meaningfully to scientific discourse (Pujol Priego et al., 2022). We can then theorize that increased accessibility and reduced resource requirements may allow smaller teams and resource-limited institutions to contribute much more readily to scientific progress. Similarly, AFDB could function as a boundary object – a shared resource that facilitates interaction between diverse groups. Providing a common data foundation could lower the barriers to collaboration, potentially leading to papers involving more contributing groups (Olson, Zimmerman & Bos, 2008).

Third, open datasets and tools may contribute to increased scientific impact by enabling more researchers with unique perspectives to build upon existing work (Gold, 2021). Such open infrastructures may shape the pace of scientific discovery by reducing duplication of effort and allowing researchers to focus on novel aspects of their studies. In turn, we could speculate how this could lead to these papers having higher citations (Colavizza et al., 2024, Piwowar and Vision, 2013) Alternatively, the very openness of the database could dilute the perceived contribution of any single application paper. Work building on the pre-computed AFDB data might be perceived by the scientific community as less novel or transformative, leading to lower citations.

This study uses bibliometric analysis to empirically investigate these competing possibilities. While the OIS framework provides a strong theoretical grounding, empirical evidence of the impact of research infrastructure datasets and tools remains limited Addressing this gap, this bibliometric study aims to provide early evidence of the impact of such RI secondary outputs on research themes, collaboration, and scientific impact, towards a more holistic view of RIs' impacts.

METHOD AND DATA

This study employs a quantitative case study strategy using bibliometric methods to explore the understudied impacts of Ris' secondary outputs in their real-world contexts. We selected the AFDB (Varadi et al., 2022) as the case due to its timeliness, relevance, and specific nature as a secondary RI output explicitly designed to democratize access to the original AF breakthrough (Jumper et al., 2021). This setup allows us to explore the distinct impacts of the database, particularly regarding

potential democratization effects, by comparing publications citing the database to those citing only the original algorithm.

We collected bibliometric data from Web of Science Core Collection for all document types (articles, reviews, conference papers, etc.) citing both AF and AFDB indexed up to August 2024. The initial dataset comprised 15,700 articles citing AF and 3,310 citing AFDB. To isolate the distinct effects of each, we further filtered these datasets to create two core groups: the AF studies contained 13,049 articles citing Jumper et al. (2021) but not Varadi et al. (2022). The AFDB papers contained 659 articles citing Varadi et al. (2022) but not Jumper et al. (2021). A third group citing both papers was also identified but primarily used for supplementary context, not a direct comparison in the main analyses focused on distinct impacts.

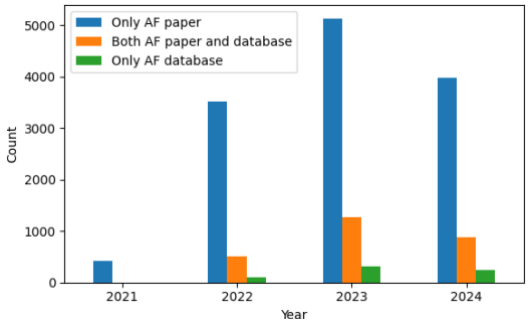


Fig. 1. Citing articles over time.

To explore the differences in research themes between the two sets of papers, we first looked at the top journal sources for both. Moreover, to visualize these topics, we also generated a cooccurrence map of the top 150 author keywords for each dataset using the software VosViewer (van Eck & Waltman, 2010). To dive deeper into their differences, we performed topic modelling with Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003). The text corpus for each article consisted of its combined title, abstract, author and Web of Science-assigned keywords. The text was then pre-processed by converting it to lowercase, removing standard English stopwords and tokenizing. We used the gensim library in Python to train the LDA model with 10 topics. We obtained the topic probability distribution for each document. We then used the non-parametric Mann-Whitney U test to determine if the distributions of topic probabilities differed significantly between the AF-only and Database-only groups for each of the 10 topics.

After comparing the distribution of research topics, we wanted to differentiate the collaboration patterns and scientific impact between these two sets. To mitigate potential confounding factors related to research fields and publication dates, we employed a matching procedure for better comparison. Given the substantial difference in the size of the AF-only (13,049) and AFDB-only (659) groups, this matching aimed to create comparable subsets. For each of the 659 articles citing only AFDB, we found a matching article from the AF-only group with the same publication year and journal. When an exact journal match was not possible, we expanded the criteria to match based on the same year and Web of Science Category. This process resulted in a matched dataset of 659 pairs of articles. While we acknowledge that this matching approach, based on year and journal/WoS category is a simplification and does not control for all potential confounding variables (e.g., funding, author seniority), it provides a pragmatic first step to explore potential differences.

Using the matched dataset of 659 pairs, we compared their collaboration patterns through co-authorship and scientific impact through citations. To get a sense of the top institutions for each set, using VosViewer, we generated the co-authorship map across institutions with at least 5 publications. To differentiate the extent of collaboration in each set, we compared the average number of co-authors and institutions between the AF database and AF-only groups using the non-parametric Mann-Whitney U test to assess statistical significance due to non-normal distributions typical in such data. Moreover, we compared the number of citations received by articles in both groups. Given the relatively short time since the publication of these papers, citation counts may not yet fully reflect their long-term impact. Additionally, in any set of papers, many papers typically will receive few or no citations. As such, we utilized a logarithmic scale to better visualize the distribution of citation counts. We applied the Mann-Whitney U test to compare the matched groups. We also performed a Fligner-Killeen test to check for differences in the variance of citations between the two groups.

Overall, this methodology provides an exploratory case study of AFDB's distinct downstream impacts, laying the groundwork for future, more rigorous, investigations.

RESULTS

Research themes

Comparing the top journals of papers citing AF and AFDB already reveals initial differences. The top journal sources for AF come from *Nature Communications* (676), *International Journal of Molecular Sciences* (350) and *PNAS* (243). In contrast, for AFDB, the top journals included *Nucleic Acids Research* (43), *International Journal of Molecular Sciences* (24) and *Nature Communications* (23). The top keywords for each set are shown in Figure 2. Visual examination of each network shows that the cluster on computational methods (green) for papers citing AF is larger compared to AFDB.

To achieve a more nuanced understanding of thematic differences, we applied topic modelling to reveal 10 distinct topics across the full set of AF-only

papers and AFDB-only papers. Table 1 shows the top keywords for each topic and indicates whether there were statistical differences (Mann-Whitney U test, $p < 0.05$) across the two groups.

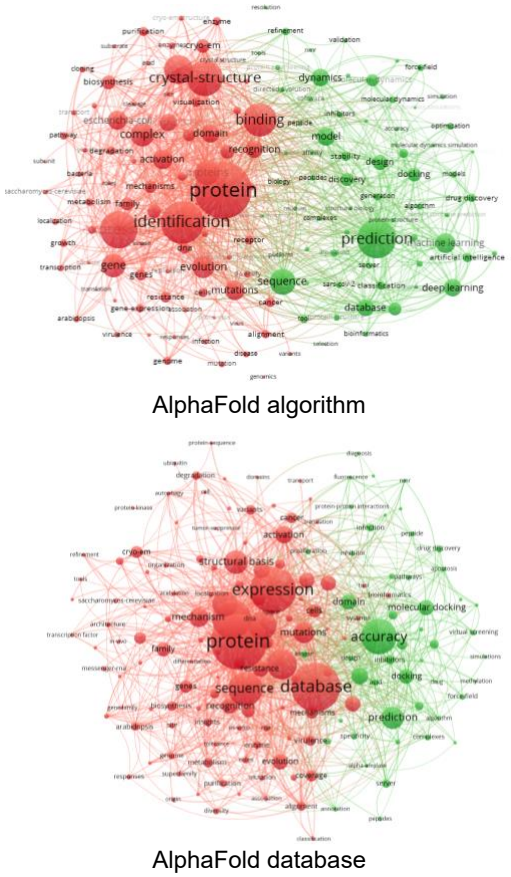


Fig. 2. Top keywords from papers citing the AlphaFold algorithm and AlphaFold database paper.

Table 1 shows that five topics showed significant differences in their prevalence between the two groups. AFDB papers were significantly more focused on Drug Discovery, Disease Mechanisms, and Macromolecular Complexes, suggesting a strong orientation toward applying the readily available structural data to specific biological problems. Conversely, AF papers had significantly higher probabilities associated with Protein Prediction and Machine Learning, reflecting a greater focus on the algorithmic, methodological, and computational aspects of protein structure prediction.

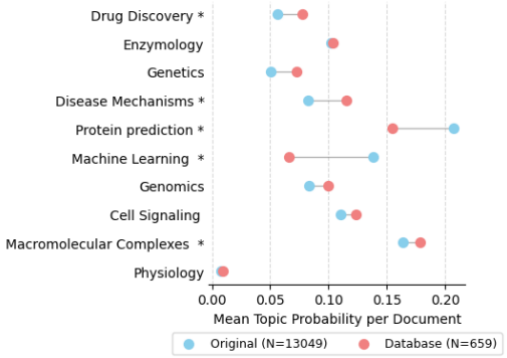


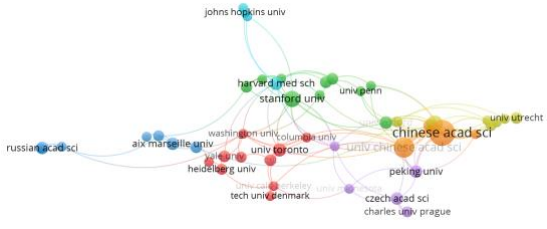
Fig. 3. Differences in topic distributions.

Table 1. Keyword analysis.

Topic	Top keywords	Difference
Drug Discovery	Drug, molecular, binding, peptide, docking, target, inhibitors, discovery, receptor	AFDB higher
Enzymology	Enzyme, activity, acid, substrate, biosynthesis, catalytic, engineering	
Genetics	Variants, mutations, gene, variant, disease, genetic, associated, patients, missense	
Disease Mechanisms	Cancer, cell, cells, virus, cov, human, immune, disease, expression, sars, vaccine	AFDB higher
Protein prediction	Protein, structure, proteins, prediction, based, structures, structural, sequence, model	AFDB lower
Machine Learning	Learning, data, machine, deep, models, design, neural, methods, artificial	AFDB lower
Genomics	Gene, genes, genome, species, evolution, plant, expression, resistance, host	
Cell Signaling	Membrane, protein, proteins, cell, bacterial, binding, domain, transport, system	
Macromolecular Complexes	Protein, dna, complex, domain, binding, proteins, rna, structure, structural, cryo	AFDB higher
Physiology	Sperm, male, venom, sex, insulin, odorant, fertilization, egg, olfactory	

Collaboration Networks

Figure 4 shows the collaboration among the organizations which published at least 5 papers in our matched set.



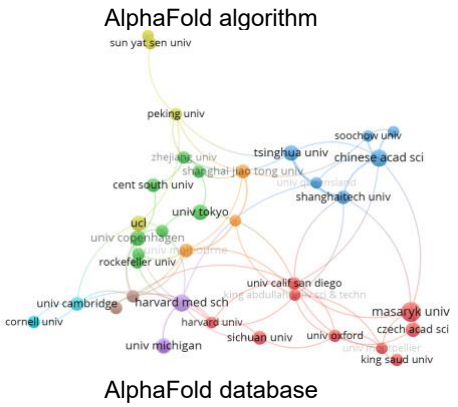


Fig. 4. Co-authorship networks for papers citing the AlphaFold algorithm and the AlphaFold database.

To examine the impact of AFDB on collaborations, we compared the average number of authors and organizations between papers citing AF and AFDB. The results reveal slight differences in team composition. Articles citing only the AF database had a slightly lower average number of authors (7.7) compared to the matched AF paper-only group (8.0), however, this difference was not statistically significant (Mann-Whitney U test, p = 0.666). In contrast, the average number of organizations per paper was significantly lower for the database-only group (4.9 vs. 5.6 for paper-only; Mann-Whitney U test, p = 0.032). This statistically significant difference suggests that research utilizing the pre-computed structures in the database tends to involve collaborations spanning fewer institutions. This finding might tentatively support the idea that the database reduces the need for multi-institutional consortia often required to pool diverse expertise or resources to run complex models like AF.

Table 2. Collaborations.

	Only AF database	Both	Only AF paper
Average authors per paper	7.7	8.1	8.0
Average organizations per paper	4.9	5.6	5.6
Top citing organizations	Masaryk University Brno: 45	Univ. Calif. System: 56	Chinese Acad. of Sci.: 79
	CNRS: 44	Harvard University: 46	CNRS: 62
	Harvard Univ.: 41	Chinese Acad. Sci.: 43	Univ. Calif. System: 58
	Univ. Texas System: 38	Univ. of Toronto: 37	Harvard Univ.: 38
	Univ. Calif. System: 30	CNRS: 31	NIH USA: 30

Further analysis of the unmatched datasets reveals interesting differences in the most prolific institutions. Masaryk University Brno led the number of publications for AFDB, while the typical research powerhouses such as the University of California System, Chinese Academy of Sciences, CNRS and Harvard University topped the articles referencing the original AF paper. This difference hints that the infrastructure might be particularly enabling for institutions beyond the top global research producers who might have more resources to run the original model.

Scientific Impact

We compared the citation counts of the AF and AFDB groups. As shown in Fig. 5, the median citations for the three datasets is 1. While the mean citation count was higher for the database-only group (7.0) compared to the paper-only group (4.5), this difference was not statistically significant (Mann-Whitney U test, p = 0.697). While the means are similar, it could still be the case that they have different variations especially since it seemed like there were more outliers in AFDB with unusually high larger number of citations. Thus, we ran a subsequent analysis to compare the variance in their distributions. However, a test for heterogeneity of variances (Fligner-Killeen test statistic = 0.32, p-value = 0.57) indicated no significant difference in the spread of citation counts between the two groups. These results suggest that, within this relatively short timeframe for citations to accumulate up to August 2024, studies leveraging AFDB did not lead to a measurable difference in citation impact compared to building directly on the original AF paper.

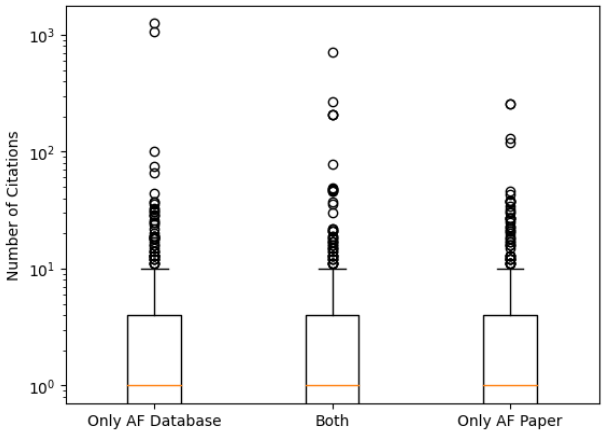


Fig. 5. Distribution of citations.

DISCUSSION AND CONCLUSIONS

This study provides an exploratory assessment of the impact of AFDB, a significant secondary output from the research infrastructure EMBL, in comparison to the original AF breakthrough. Our findings, interpreted through the lens of the Open Innovation in Science (OIS)

framework (Beck et al., 2022), provide preliminary empirical evidence that accessible data infrastructures like AFDB can shape scientific trajectories differently than the foundational innovations they derive from.

First, the statistically significant divergence in research themes provides support for the OIS principle that accessible infrastructures facilitate knowledge flows into new domains. We found that papers citing AFDB showed a stronger orientation towards downstream application areas like drug discovery and disease mechanisms. In contrast, the original AF paper maintained a stronger focus on machine learning and protein prediction. This aligns with the OIS view of accessible resources enabling knowledge to flow into new application domains and potentially accelerating the translation of foundational breakthroughs into downstream, practical use cases.

Second, our findings on collaboration patterns offer partial alignment with the OIS concept of democratization through lowered barriers to entry. The statistically significant lower average number of participating organizations in papers citing only AFDB provides tentative support for the notion that such resources can enable contributions from a potentially wider or different institutional base, possibly reducing the need for large consortia with specialized computational resources or expertise. While the difference in author count was not significant, the reduction in institutional span suggests the database might lower coordination needs or dependency on specific institutional capabilities required to run the original model. The differing profiles of top publishing institutions between the groups further hint at this potential broadening of access.

Third, the lack of a significant difference in citation impact between the matched groups presents a nuanced finding. While OIS suggests open resources might accelerate discovery and impact, this may not immediately translate into higher citation counts. This suggests several possibilities. First, citation impact takes longer to accrue, especially for application-focused work and thus, would need longer-term monitoring. Second, the perceived novelty of these studies using pre-computed data might be reduced. Third, the focus on citations might not be granular enough to fully characterize differences in scientific impact.

Thus, while many of our findings look promising, it is crucial to acknowledge the limitations of this study. First, as a primarily exploratory study, the observed differences may be influenced by confounding factors not accounted for in our simple matching procedure; future work could employ methods like propensity score matching (PSM). Second, the relatively short time frame may not fully capture its long-term impact on the research landscape. To build upon these initial findings, longitudinal analyses tracking these cohorts over several more years could provide insights into how the impact of the AF database evolves over time. Third, bibliometric data such as citation counts only capture a small dimension of the scientific enterprise. They do not fully encompass knowledge translation, societal impact, or changes in research practices. Qualitative research, such as interviews with researchers using the database, could offer deeper insights into how it influences research practices and collaborations.

Despite these limitations, this study offers preliminary theoretical and practical contributions. First, this study provides initial insights on the ripple impact of large-scale, open-access resources (like AFDB) that reach beyond the initial scientific breakthrough. Other researchers could build upon our findings on OIS in the case of scientific infrastructures to better understand how such outputs impact downstream science. Second, for research infrastructure funders, our results underscore the significant downstream value generated by secondary RI outputs like open-access databases. Our findings suggest that AFDB is not just facilitating more research, but potentially different kinds of research with broader institutional involvement. Evaluations of RI impact should develop methodologies to capture the value of activities such as data sharing, tool provision, knowledge democratization, and patents. Third, experimental researchers now have access to a broad range of public research databanks with numerous opportunities to enrich and accelerate their workflows (Pujol Priego & Wareham 2024).

In conclusion, while acknowledging the exploratory nature of our analysis, this research provides initial evidence suggesting that the EMBL AlphaFold Database has distinct impacts on the scientific landscape compared to the original AlphaFold algorithm. While more sophisticated methods and longer timeframes are needed for future research, these initial findings indicate the need for a more holistic assessment of research infrastructure impacts. Correctly recognizing these broader impacts is essential for enabling the increasingly data-driven and open scientific landscape.

ACKNOWLEDGMENTS

This project was funded through the socioeconomic grant call of ATTRACT Phase 2, funded by European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004462. We would like to thank our colleagues from EMBL and DeepMind for their constant engagement in our research.

REFERENCES

Autio, E., Hameri, A. P., & Vuola, O. (2004). A framework of industrial knowledge spillovers in big-science centers. *Research Policy*, 33(1), 107-126.

Beagrie, N., & Houghton, J. (2021). The value and impact of EMBL-EBI managed data resources. European

Bioinformatics Institute (EMBL-EBI). https://www.embl.org/documents/document/embl-ebi-impact-report-2021

Beck, S., Bergenholtz, C., Bogers, M., Brasseur, T. M., Conradsen, M. L., Di Marco, D., ... & Xu, S. M. (2022). The Open Innovation in Science research field: a collaborative conceptualisation approach. *Industry and Innovation*, *29*(2), 136-185.

Beck, S., Bercovitz, J., Bergenholtz, C., Brasseur, T., D'Este, P., Dorn, A., ... & Zyontz, S. (2021). Experimenting with Open Innovation in Science (OIS) practices: A novel approach to co-developing research proposals. *CERN IdeaSquare Journal of Experimental Innovation*, *5*(2), 28-49.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Colavizza, G., Cadwallader, L., LaFlamme, M., Dozot, G., Lecorney, S., Rappo, D., & Hrynaszkiewicz, I. (2024). An analysis of the effects of sharing research data, code, and preprints on citations. *Plos one*, *19*(10), e0311493.

D'ippolito, B., & Rüling, C. C. (2019). Research collaboration in Large Scale Research Infrastructures: Collaboration types and policy implications. *Research Policy*, *48*(5), 1282-1296.

European Research Executive Agency. (2023, December 5). Over €220 million to fund new Research Infrastructures projects. https://rea.ec.europa.eu/news/over-eu220-million-fund-new-research-infrastructures-projects-2023-12-05_en

Fabre, R., Egret, D., Schöpfel, J., & Azeroual, O. (2021). Evaluating the scientific impact of research infrastructures: The role of current research information systems. *Quantitative Science Studies*, *2*(1), 42-64.

Florio, M., & Sirtori, E. (2016). Social benefits and costs of large scale research infrastructures. *Technological Forecasting and Social Change*, *112*, 65-78.

Heidler, R., & Hallonsten, O. (2015). Qualifying the performance evaluation of Big Science beyond productivity, impact and costs. *Scientometrics, 104*, 295-312.

Gold, E. R. (2021). The fall of the innovation empire and its possible rise through open science. *Research Policy*, *50*(5), 104226.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Mayernik, M. S., Hart, D. L., Maull, K. E., & Weber, N. M. (2017). Assessing and tracing the outcomes and impact of research infrastructures. *Journal of the Association for Information Science and Technology*, *68*(6), 1341-1359.

Olson, G. M., Zimmerman, A., & Bos, N. (2008). *Scientific collaboration on the Internet*. The MIT Press.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175.

Pujol Priego, L., Wareham, J., & Romasanta, A. K. S. (2022). The puzzle of sharing scientific data. *Industry and Innovation*, 29(2), 219-250.

Pujol Priego, L., & Wareham, J. (2023). From bits to atoms: Open source hardware at CERN. *MIS Quarterly*, *47*(2), 639-668.

Pujol Priego, L. & Wareham, J. (2024). Data Commoning in the Life Sciences. *MIS Quarterly*, 48(2) 491-520.

Reed, M. S., Ferré, M., Martin-Ortega, J., Blanche, R., Lawford-Rolfe, R., Dallimer, M., & Holden, J. (2021). Evaluating impact from research: A methodological framework. *Research Policy*, *50*(4), 104147.

Romasanta, A., Ahmadova, G., & Wareham, J. (2022). From potential to realized impacts: the bridging role of digital infrastructures in fair data. *European Conference on Information Systems.*

Scarrà, D., & Piccaluga, A. (2022). The impact of technology transfer and knowledge spillover from Big Science: a literature review. *Technovation*, *116*, 102165.

Van Eck, N., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439-D444.

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of business research*, *88*, 428-436.

Wareham, J., Priego, L. P., Romasanta, A. K., Mathiassen, T. W., Nordberg, M., & Tello, P. G. (2022). Systematizing serendipity for big science infrastructures: The ATTRACT project. *Technovation*, *116*, 102374.