Published online 7 December 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D9–D17***
https://doi.org/10.1093/nar/gkac1098

EMBL's European Bioinformatics Institute (EMBL-EBI) in 2022

Matthew Thakur [1,*], Alex Bateman [1], Cath Brooksbank[1], Mallory Freeberg [1], Melissa Harrison[1], Matthew Hartley[1], Thomas Keane [1], Gerard Kleywegt[1], Andrew Leach[1], Mariia Levchenko [1], Sarah Morgan[1], Ellen M. McDonagh [1,2], Sandra Orchard [1], Irene Papatheodorou [1], Sameer Velankar [1], Juan Antonio Vizcaino [1], Rick Witham[1], Barbara Zdrazil[1] and Johanna McEntyre[1,*]

[1]Data Services Teams, EMBL's European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, UK and [2]OpenTargets, EMBL's European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, UK

Received October 13, 2022; Revised October 21, 2022; Editorial Decision October 24, 2022; Accepted October 31, 2022

ABSTRACT

The European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) is one of the world's leading sources of public biomolecular data. Based at the Wellcome Genome Campus in Hinxton, UK, EMBL-EBI is one of six sites of the European Molecular Biology Laboratory (EMBL), Europe's only intergovernmental life sciences organisation. This overview summarises the status of services that EMBL-EBI data resources provide to scientific communities globally. The scale, openness, rich metadata and extensive curation of EMBL-EBI added-value databases makes them particularly well-suited as training sets for deep learning, machine learning and artificial intelligence applications, a selection of which are described here. The data resources at EMBL-EBI can catalyse such developments because they offer sustainable, high-quality data, collected in some cases over decades and made openly availability to any researcher, globally. Our aim is for EMBL-EBI data resources to keep providing the foundations for tools and research insights that transform fields across the life sciences.

INTRODUCTION

The European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) is one of the world's leading sources of public biomolecular data. Based at the Wellcome Genome Campus in Hinxton, UK, EMBL-EBI is one of six sites of the European Molecular Biology Laboratory (EMBL), Europe's only intergovernmental life sciences organisation, whose world-class research infrastructure and services support cutting-edge science globally.

EMBL-EBI enables life science research and its translation to medicine, agriculture, industry and society by:

- freely providing data and bioinformatics services to the scientific community in ways that promote scientific progress.
- contributing to the advancement of biology through investigator-driven research.
- providing bioinformatics training to scientists at all levels.
- disseminating cutting-edge technologies to industry and applications of science.
- supporting, as an ELIXIR Node, the coordination of biomolecular data provision in Europe.

EMBL-EBI contributes to EMBL's 2022–2026 'Molecules to Ecosystems' programme, which aims to establish the molecular basis of life in context, to gain new knowledge that is relevant to understanding life on Earth, and to provide translational potential to support advances in human and planetary health

This overview focuses on services that EMBL-EBI data resources provide to scientific communities globally, describing related training and industry applications where relevant. As many other EMBL-EBI data resources have dedicated articles elsewhere in this special issue, this overview focuses primarily on major changes to data resources not described elsewhere.

EMBL-EBI data resources comprise: deposition databases, which archive experimental data; added-value databases, which provide annotation, curation, reanalysis and integration of deposited data; and open source software tools, that enable reuse of these resources. Deposition

*To whom correspondence should be addressed. Email: mthakur@ebi.ac.uk
Correspondence may also be addressed to Johanna McEntyre. Email: mcentyre@ebi.ac.uk

databases, added-value databases and tools are described and accessed via the EMBL-EBI services web portal. All EMBL-EBI data resources and many software systems can be downloaded and installed locally, and are made available on an open and free basis for reuse. Many services offer further bulk and machine-readable access including via API, FTP, Aspera and Globus services.

Co-housing these data resources at one institute results in close integration between resources, demonstrated by the high degree of between-resource data flow, and the availability of integrated tools like pan-resource EBI-Search (1). Data resources also benefit from institutional support with technical and infrastructure management. EMBL-EBI resources serve as foundations for hundreds of external resources and tools (with many recent developments described below). Europe's flagship bioscience data coordination programme ELIXIR identifies the Core Data Resources and Deposition Databases of most fundamental importance to the wider life-science community and the long-term preservation of biological data. Many EMBL-EBI resources have achieved this designation.

The scale, openness, rich metadata and extensive curation of EMBL-EBI added-value databases makes them particularly well-suited as training sets for deep learning, machine learning and artificial intelligence applications (abbreviated here under the umbrella term AI applications). A recent major AI application is the DeepMind AlphaFold system for predicting previously unknown 3D structures, which was trained on openly available experimentally verified protein structure data from the Protein Databank (2), jointly delivered by EMBL-EBI and other partners (the wwPDB consortium), as well as protein sequences and annotation from Uniprot (3) and metagenomics data from MGnify (4). As of September 2022, the outputs of AlphaFold, hosted by EMBL-EBI as AlphaFold-DB (5), included 214 684 311 predicted structures, with 48 complete proteomes available for bulk download. Over 500 000 researchers in 190 countries used AlphaFoldDB in its first year of operation. The data is already enabling researchers to progress a number of previously intractable research questions (6). Further examples of how EMBL-EBI resources are enabling AI applications in proteomics, drug discovery, imaging and other areas are included below.

The impact of EMBL-EBI data resources

EMBL-EBI tracks the use of data resources through metrics including the number of web requests and unique IP addresses visiting service websites, the volume of data deposited, and the number of open citations EMBL-EBI data resources receive in scientific publications. While each metric has limitations and cannot provide an exact quantification of use, considered together they give an indication of the scale and trend in usage.

Demand from researchers for EMBL-EBI data resources increased considerably in 2020, particularly during the second quarter (April–June 2020) which coincided with the beginning of the global COVID-19 pandemic. Usage continued to grow throughout the rest of 2020 and into 2021, as many researchers transitioned back from remote to in-person or hybrid working (Figure 1). Demand has remained high in the first two quarters of 2022 with an average of 3.1 billion web requests and 5.2 million unique IPs per month in the second quarter of the year. This is ∼100% higher than user demand for the equivalent period in 2018. EMBL-EBI data resources have a global reach, with every UN member state country represented in our user base in 2021, and the data available for the current year to date suggests similar global user demand in 2022.

The rate of data deposition by volume into EMBL-EBI's archival resources continues to accelerate, with over 25 PB of data deposited in 2021, bringing the cumulative total storage up to approximately 75 Petabytes (Figure 2). The two largest archival resources are European Nucleotide Archive (ENA) (7) and European Genome-phenome Archive (EGA) (8), between them accounting for over 90% of total data deposited to date. Notably rapid data growth in recent years has been in imaging data resources - BioImage Archive (BIA) (9); and the electron microscopy imaging resources Electron Microscopy Public Image Archive (EMPIAR) (10) and Electron Microscopy Databank (EMDB) (11)

In 2021, an independent study estimated the economic value and impact of EMBL-EBI data resources. The study found that researchers spent 140 million hours/year using EMBL-EBI data resources, with a value equivalent to £5.5 billion.

MAJOR CHANGES IN THE EMBL-EBI DATA RESOURCE PORTFOLIO

Federated EGA network officially launched

Until recently, most of the individual-level human omics data made discoverable on the EMBL-EBI European Genome-phenome Archive (EGA) (8) were generated by research consortia, not in healthcare settings. Many countries now have personalised medicine programmes that are generating data from national or regional initiatives, resulting in a shift from research-driven to healthcare-driven genomic data. Data generated in a healthcare context can be subject to different governance and national data protection legislation than research data, and these access controls risk blocking reuse for research. If reuse for research were not possible, the potential value and impact of emerging healthcare genomic data would be significantly reduced. Federated EGA uses a distributed network of international repositories to ensure genomic data accelerates research by enabling transnational discovery of and access to human data, while also respecting jurisdictional data protection regulations, thus enabling scale and more powerful research insights.

One of the first applications of Federated EGA is to provide transnational data discovery and access infrastructure for the European 1+ Million Genomes and Genome Data Infrastructure projects. The subsequent Beyond One Million Genomes EC coordination and support actions project will demonstrate how Federated EGA enables rare disease federated discovery and access.

The Federated EGA network was officially launched in 2022 with the signing of the first legal agreements with inaugural nodes in Sweden, Norway, Germany, Finland, and Spain. Dozens of additional nodes across Europe and the

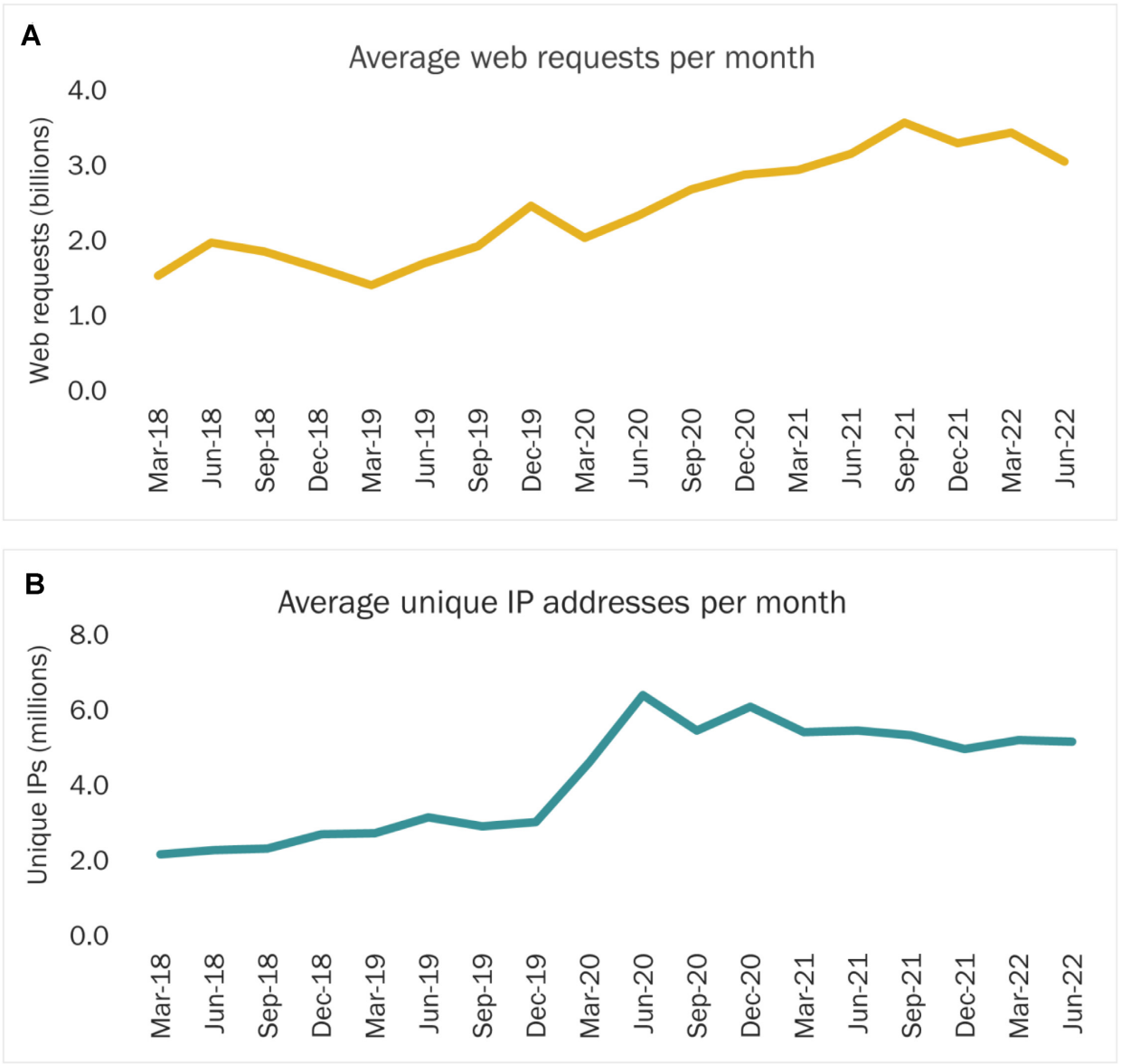

Figure 1. Web requests (yellow, **A**) and Unique IP visits (blue, **B**) to EMBL-EBI data resources, 2018–2022.

world are working towards joining the Federated EGA network, with a shared vision of establishing a truly global resource for sensitive human data discovery and sharing.

Pfam merging into InterPro

EMBL-EBI hosts two major protein family resources, Pfam (12) and InterPro (13). The Pfam database of protein families was formerly hosted at the Sanger Institute until 2012, when it was migrated to EMBL-EBI. Although similar in scope, there are important differences between the two resources. InterPro provides a comprehensive view across most of the world's protein family resources by aggregating data from 13 other resources, including Pfam (13). Although it brings protein family data together it does not generate the signatures for identifying a particular protein family, such as a profile-hidden Markov model. The signatures are provided by the 13 member databases. Pfam provides EMBL-EBI the ability to create new family signatures as well as update existing ones and thus provides an important complementary functionality to InterPro. To make the production and dissemination of these two resources as efficient and scalable as possible, the functionality of the Pfam website was merged into Interpro. The Pfam website was decommissioned in January 2023, but all of its data and functionality continue to be provided via InterPro.

ArrayExpress migrating into BioStudies

The BioStudies Database (14) is a resource for encapsulating all the data associated with a biological study, which may exist across a number of different data resources. One of the goals of BioStudies is to manage data generated in experiments that can be characterized as 'multi-omics'. In-

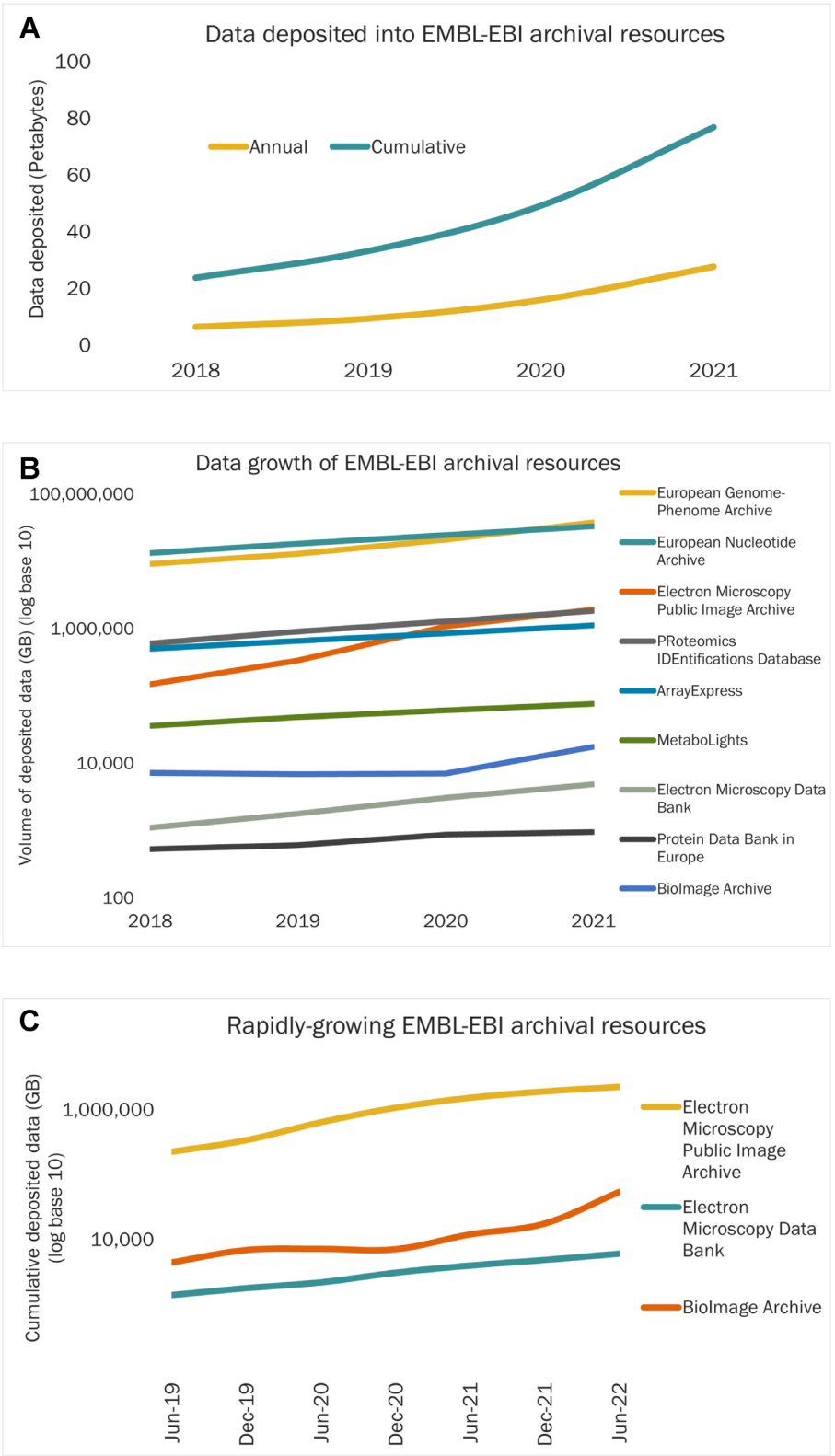

Figure 2. (**A**) Annual (yellow) and cumulative (blue) data deposition into EMBL-EBI archival data resources. (**B**) Annual deposition into nine archival resources. Note the logarithmic scale. and rapid rate of growth for the Imaging and cryo-electron microscopy resources Bioimage Archive, EMPIAR and EMDB. (**C**) Annual quarterly data growth for the Bioimage Archive, EMPIAR and EMDB imaging and cryo-electron microscopy data resources. Note the logarithmic scale.

creasingly, many experiments that used to belong to the domains of transcriptomics or functional genomics are now multi-modal, resulting in decreased depositions into the array-specific ArrayExpress data resource (15). Since 2020, to streamline the data submission processes and data representation at EMBL-EBI, data served from ArrayExpress has been migrated to BioStudies, under the 'ArrayExpress collection'. Following positive user feedback on the new pipelines and processes, the ArrayExpress interface was decommissioned in September 2022 and all new functional genomics submissions are now processed and loaded in BioStudies, before flowing on into other data resources including ENA [7].

NEW FEATURES AND AI APPLICATIONS OF EXISTING DATA RESOURCES

UniProt annotates 50 million previously uncharacterised proteins using machine-learning

The UniProt Knowledgebase of protein sequence and function (16) combines both automated and expert-curated annotations of protein function. Expert biocurators link UniProtKB/Swiss-Prot entries to a summary of experimentally verified, or computationally predicted, functional information about each protein. Information is added to non-reviewed entries in the UniProtKB/TrEMBL system using annotation transfer from reviewed entries by automated systems (17).

A UniProt-organised challenge in 2022 asked competitors from the machine learning community to develop software tools and algorithms which predict metal binding sites in proteins with accuracy and at scale. The best of these tools will be incorporated into future production pipelines.

UniProt data is made available in formats suitable for researchers to develop their own tools and resources. As of release 22_05, sequence embeddings for all of UniProtKB/Swiss-Prot are released on the UniProt ftp site. Longer-term plans are to make records more readily machine-readable, for example by increasing usage of ontologies, to support utility as positive training sets in AI applications.

Mass spectrometry-based proteomics datasets drive AI applications

Proteomics Identification Database (PRIDE) is the world-leading database for mass spectrometry (MS)-based proteomics datasets (18) and is one of the founding members of the International ProteomeXchange Consortium of proteomics resources (19). On average, ∼500 datasets were submitted to PRIDE every month in 2022. The unprecedented availability of proteomics datasets in the public domain is driving multiple applications reusing this data. AI applications have been applied to improve every step in the proteomics analytical workflow—for a recent review see (20). These approaches are enabling new biological findings, for instance in the context of protein phosphorylation (21), identification of antimicrobial peptides (22) and prediction of antigen presentation of HLA molecules (23). Multi-omics approaches involving proteomics data are an area where further applications of public datasets will generate novel tools. All PRIDE datasets that have been used for AI applications, including training and evaluating models, are tagged using the term 'Machine Learning'. This shows the enormous value of public datasets with high quality annotation to enable novel 'big data' approaches in proteomics (24).

Applications of small molecule bioactivity data for in silico drug discovery

The ChEMBL database (25) is a large-scale open resource of small molecule bioactivity data which was first launched in 2009. It mainly hosts curated data extracted from the medicinal chemistry literature as well as deposited datasets, and has grown significantly in size and complexity since its first release. The current release of ChEMBL (version 31, prepared in July 2022) hosts around 20 million bioactivity data points for 2.3 million compounds corresponding to 1.5 million assays, 15 000 targets, and 85 000 documents.

Prior to the launch of ChEMBL, only large private organisations were able to access diverse and high-quality (proprietary or commercial) bioactivity data sets for a wide range of biological targets at scale. Data from ChEMBL has proven indispensable for the development (26), validation and benchmarking (27–29) of a wide range of AI and other *in silico* applications, including those described below.

Given its vast size and coverage of medicinal chemistry space, ChEMBL is frequently leveraged for chemical space analysis, either driven by a focus on drugs (30), a chemotype-centric view (31,32), or a specific area of drug discovery research (33–35). ChEMBL also facilitates large-scale comparison of species differences in bioactivity data (36), and assay and bioactivity endpoint comparisons (37,38). Insights from such analyses influence how predictive models are built and applied, and guide experimental design when searching for new chemical matter.

ChEMBL facilitates the development of in silico target prediction algorithms (39–41) and molecular de novo design (42,43). Bioactivity data from ChEMBL, in conjunction with other data types such as pathway and disease information, is a foundational part of knowledge graph-based discovery tools, with applications such as phenotypic assay target deconvolution (44,45).

AI-ready imaging datasets and interoperability standards

The BioImage Archive (9) is EMBL-EBI's deposition database for life sciences imaging data associated with publications, as well as reference imaging datasets. AI applications are revolutionising the process of analysing and gaining insight from biological images. However, such techniques often generate 'black box' models. Understanding how these models function, what biases they may contain and what type of data they can be safely applied to, is very difficult without access to original training data.

To support reproducibility of AI applications in image analysis, the BioImage Archive supports deposition of both images and ground truth annotations used in training datasets. The archive already makes available over 30 imaging datasets with these AI-suitable annotations, which allows method developers to use existing data to accelerate development. Work is underway to enhance support

for these 'AI-ready' datasets, through dedicated deposition pipelines, and developer friendly presentations. Bioimage Archive is playing an active role in the development of the community standards for interoperable segmentation, image categorisation and other annotation required to enable widespread imaging data sharing for AI applications (46).

The Electron Microscopy Public Image Archive EMPIAR (10), is a public resource for raw images underpinning 3D cryo-EM maps and tomograms (the latter archived in the Electron Microscopy Databank, EMDB (11)). EMPIAR also accommodates 3D datasets obtained with volume EM techniques, and soft and hard X-ray tomography. All data archived in EMPIAR can be re-used freely without any conditions or restrictions via a 'CC0' license model, making it an easily accessed source of data for AI applications in image analysis. EMPIAR released two datasets in 2022 specifically developed to support machine learning – CEM-MitoLab, a dataset of ∼22K cellular EM 2D images with label maps of ∼135K mitochondrial instances, and CEM1.5M: a cellular EM dataset containing ∼1.5M unlabeled 2D image patches curated for deep learning.

Pre-print corpus for text-mining and AI applications

Europe Pubmed Central (Europe PMC) (47) provides open access to a worldwide collection of life science preprints and peer-reviewed journal articles. Following the COVID-19 full text preprints initiative, since April 2022 Europe PMC makes the full text of preprints supported by its 37 funders available for search, reading, and reuse, both on the Europe PMC website and programmatically in standard JATS XML format. As of September 2022 there were over 450 000 preprints indexed in Europe PMC from 24 preprint servers and nearly 32 000 of these are available as full text. Of the full text preprints 98% have an open access licence and are available via bulk download for text analytics and machine learning applications. To further possibilities for large-scale meta analyses, preprints in Europe PMC are linked to underlying research data, open peer review materials, citations, grants and other useful resources. The preprint corpus will increase discoverability, ensure the continued access to findings presented in preprints and enable new analytical possibilities including AI applications.

Improving preprint transparency and tracking

The ability to improve and correct the manuscript through new versions is an important part of preprints' appeal. However, changes to preprints can be difficult to track, especially across many different preprint servers and journals. Researchers working with preprints need to know which version should be cited, how a preprint differs from its published version, and whether conclusions presented in a preprint are valid after a preprint has been withdrawn or removed. To address these issues Europe PMC now offers a way to check for preprints updates. The **Article Status Monitor** is a Europe PMC tool that allows users to check if a preprint has been withdrawn, removed, published in a journal, or updated with a new version. Updates can be retrieved using a simple website tool, email alert or programmatically via the status-update-search module of the Articles API.

Big data for target-disease association and disease-causing genes now available in the Cloud

The Open Targets consortium is a pre-competitive partnership between EMBL-EBI, the Wellcome Sanger Institute, and pharmaceutical company partners GSK, Sanofi, BMS—with Pfizer joining in 2022. The consortium generates data and builds informatics tools to enhance the identification and prioritisation of targets that will ultimately lead to more effective and safer drugs. Open Targets produce two open source informatics resources: the Open Targets Platform (48), which provides a knowledgebase and tools for target-disease association evidence and prioritisation; and Open Targets Genetics (49), developed to address the challenge of identifying disease-causing genes (and thus potential drug targets) from Genome-Wide Association Studies. These are increasingly being adopted as reference databases in their own right, but in addition provide structured data to enable other data integration and AI applications.

In May 2022, the resources were made available in the cloud via Google BigQuery and AWS Open Data. This integration and accessibility enables the data to be used in AI applications, such as machine learning to identify novel target-disease associations (50), building knowledge graphs for different biological insights (51–54) and for benchmarking new computational methods for drug target prioritisation (55,56).

As the code base is open source, separate instances of the Platform can be created and adapted to user requirements—an example is the recent release of the NIH Childhood Cancer Data Initiative Molecular Targets Platform, which integrates tumor gene expression and somatic alteration data (Figure 3).

Open data standards for proteomics

EMBL-EBI continues to lead many activities of the Proteomics Standards Initiative (PSI), the organisation in charge of developing open data standards in proteomics (57). Among these activities, during 2022, in collaboration with the Consortium for Top-Down Proteomics, the PSI released the ProForma 2.0 notation (58), providing a standard way to represent peptidoforms and proteoforms (combinations of protein sequences plus protein modifications).

ProForma can be used in conjunction with Universal Spectrum Identifiers (59), a PSI standard released in 2021 that provides a unique identifier for mass spectra in ProteomeXchange repositories (including PRIDE).

TRAINING

Recent years have provided many challenges in terms of training development and delivery, but 2022 saw the reintroduction of in-person training at EMBL-EBI as well as the retention of an extensive virtual programme.

EMBL-EBI's training programme focuses on empowering scientists to get the most out of openly accessible data resources and services, and to develop key bioinformatics analysis skills. This goal has been supported even further in 2022 through the addition of training in key principles for data management and open data in all EMBL-EBI live

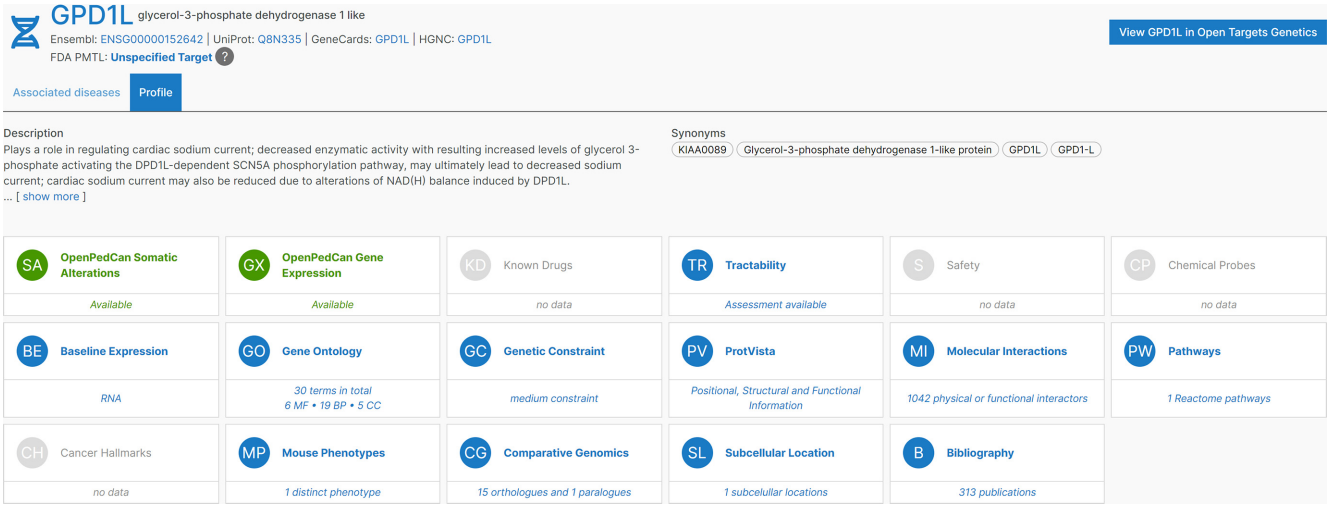

Figure 3. Example results from the Molecular Targets Platform, a US National Cancer Institute-supported instance of the Open Targets Platform with a focus on preclinical paediatric oncology data.

courses, as emphasised by EMBL's updated open science policy.

This enables us to encourage all scientists we train to deposit their data in open resources, ensuring they know how the deposition process works, how to get started and where to find the support required at all points. We are also working with countries where deposition rates have traditionally been low, e.g. LMICs, to determine barriers to deposition and how these can be overcome. Roughly 500 scientists participate in in-person courses per year, with the majority of these reporting that they go on to pass on their learning to others. Web-based on-demand training sees around 500 000 unique IP users per year, of which 80% rate webinars as excellent or very good.

2022 has seen the culmination of a three-year project to completely redesign the training website. Testing indicated a great improvement in user experience for those seeking training courses and content for their own use, or in the training of others. A key development in 2022 was ensuring those undertaking self-directed learning through EMBL-EBI on-demand materials can easily track their progress, keep a record of courses they have completed, and plan their future studies. New Personal Account functionality enables EMBL-EBI trainees to tag favourite courses and manually log their progress through a course, whilst also recording their quiz results and maintaining a record of completion.

New on-demand formats, such as curated collections and learning pathways provide trainees with a more structured approach to learning for a particular topic. Created from a mixture of on-demand tutorials and webinars, alongside excerpts of video and practical exercises from live courses, this guided learning is further reinforced through the EMBL-EBI webinar programme, with the addition of expert panel Q&A sessions.

A final piece of work has been to further improve the FAIRness of EMBL-EBI live training materials, by creating an openly accessible training material set for each course. Materials have always been available via FTP post course, but their reuse was limited as the context of the sessions was often lost. EMBL-EBI new course material sets provide a complete overview of each course and allow for easier reuse by both trainers and trainees.

Finally, EMBL-EBI have set up a trainer-specific space to further build capacity for bioinformatics trainers and educators and support the teaching of EMBL-EBI resources by external trainers using expert written materials.

CONCLUSION

As the world and the scientific community recover from the ongoing COVID-19 pandemic, there is ample opportunity to reflect on the importance of open science and open data. Open data resources such as those at the EMBL-EBI need to continuously evolve and engage with their user communities to meet changing scientific needs. Many of the developments described above reflect the emerging need to prepare data resources for use in AI applications, which are already starting to transform many scientific fields. Building for AI is reflected in the collection and curation of reference datasets, and the development of community-driven data standards and guidelines that support the reuse of data beyond the bounds of the experiment that generated them.

The transformative potential of AI applications has been amply demonstrated with DeepMind's development of AlphaFold, which predicted protein structure for almost all 200M protein sequences in UniProt, and lead to myriad scientific uses of this new data across many different fields. The data resources at EMBL-EBI can catalyse such developments because they offer sustainable, open availability of high-quality data resources, collected in some cases over decades. Our aim is for EMBL-EBI data resources to keep providing the foundations for tools and research insights that transform fields across the life sciences.

DATA AVAILABILITY

All of the data resources described above are freely available to access and reuse at https://www.ebi.ac.uk/services.

ACKNOWLEDGEMENTS

The authors and staff of EMBL-EBI are indebted to the hundreds of thousands of scientists who submit data and annotation to these shared data resources. This article's listed authors are direct contributors to the text, but all developments to our services are the work of the much broader services teams stewarding the data resources hosted by EMBL-EBI, whose efforts we gratefully acknowledge here.

FUNDING

EMBL-EBI is indebted to its funders, including the EMBL member states; European Commission; Wellcome; UK Research and Innovation; US National Institutes of Health; our Industry Programme and many others. Continued growth in the data services EMBL-EBI can offer users was made possible by dedicated UK government funding for the EMBL-EBI Data Infrastructure programme, currently via the Strategic Priorities Fund. Funding for open access charge: EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Madeira,F., Pearce,M., Tivey,A.R.N., Basutkar,P., Lee,J., Edbali,O., Madhusoodanan,N., Kolesnikov,A. and Lopez,R. (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.*, **50**, W276–W279.
2. wwPDB consortium (2019) Protein data bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
3. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
4. Mitchell,A.L., Almeida,A., Beracochea,M., Boland,M., Burgin,J., Cochrane,G., Crusoe,M.R., Kale,V., Potter,S.C., Richardson,L.J. *et al.* (2020) MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**, D570–D578.
5. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A. *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
6. Mosalaganti,S., Obarska-Kosinska,A., Siggel,M., Taniguchi,R., Turoňová,B., Zimmerli,C.E., Buczak,K., Schmidt,F.H., Margiotta,E., Mackmull,M.-T. *et al.* (2022) AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science*, **376**, eabm9506.
7. Cummins,C., Ahamed,A., Aslam,R., Burgin,J., Devraj,R., Edbali,O., Gupta,D., Harrison,P.W., Haseeb,M., Holt,S. *et al.* (2022) The european nucleotide archive in 2021. *Nucleic Acids Res.*, **50**, D106–D110.
8. Freeberg,M.A., Fromont,L.A., D'Altri,T., Romero,A.F., Ciges,J.I., Jene,A., Kerry,G., Moldes,M., Ariosa,R., Bahena,S. *et al.* (2022) The european Genome-phenome archive in 2021. *Nucleic Acids Res.*, **50**, D980–D987.
9. Ellenberg,J., Swedlow,J.R., Barlow,M., Cook,C.E., Sarkans,U., Patwardhan,A., Brazma,A. and Birney,E. (2018) A call for public archives for biological image data. *Nat. Methods*, **15**, 849–854.
10. Iudin,A., Korir,P.K., Somasundharam,S., Weyand,S., Cattavitello,C., Fonseca,N., Salih,O., Kleywegt,G.J. and Patwardhan,A. (2022) EMPIAR: The Electron Microscopy Public Image Archive. *Nucleic Acids Res.*, https://doi.org/10.1093/nar/gkac1062.
11. Lawson,C.L., Patwardhan,A., Baker,M.L., Hryc,C., Garcia,E.S., Hudson,B.P., Lagerstedt,I., Ludtke,S.J., Pintilie,G., Sala,R. *et al.* (2016) EMDataBank unified data resource for 3DEM. *Nucleic Acids Res.*, **44**, D396–D403.
12. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
13. Blum,M., Chang,H.-Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S. *et al.* (2021) The interpro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
14. Sarkans,U., Gostev,M., Athar,A., Behrangi,E., Melnichuk,O., Ali,A., Minguet,J., Rada,J.C., Snow,C., Tikhonov,A. *et al.* (2018) The biostudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.*, **46**, D1266–D1270.
15. Sarkans,U., Füllgrabe,A., Ali,A., Athar,A., Behrangi,E., Diaz,N., Fexova,S., George,N., Iqbal,H., Kurri,S. *et al.* (2021) From arrayexpress to biostudies. *Nucleic Acids Res.*, **49**, D1502–D1506.
16. The UniProt Consortium (2022) UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, https://doi.org/10.1093/nar/gkac1052.
17. MacDougall,A., Volynkin,V., Saidi,R., Poggioli,D., Zellner,H., Hatton-Ellis,E., Joshi,V., O'Donovan,C., Orchard,S., Auchincloss,A.H. *et al.* (2020) UniRule: a unified rule resource for automatic annotation in the uniprot knowledgebase. *Bioinformatics*, **36**, 4643–4648.
18. Perez-Riverol,Y., Bai,J., Bandla,C., García-Seisdedos,D., Hewapathirana,S., Kamatchinathan,S., Kundu,D.J., Prakash,A., Frericks-Zipper,A., Eisenacher,M. *et al.* (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.*, **50**, D543–D552.
19. Deutsch,E.W., Bandeira,N., Perez-Riverol,Y., Sharma,V., Carver,J.J., Mendoza,L., Kundu,D.J., Wang,S., Bandla,C., Kamatchinathan,S. *et al.* (2022) The ProteomeXchange Consortium at 10 years: 2023 update. *Nucleic Acids Res.*, https://doi.org/10.1093/nar/gkac1040.
20. Mann,M., Kumar,C., Zeng,W.-F. and Strauss,M.T. (2021) Artificial intelligence for proteomics and biomarker discovery. *Cell Syst.*, **12**, 759–770.
21. Ochoa,D., Jarnuczak,A.F., Viéitez,C., Gehre,M., Soucheray,M., Mateus,A., Kleefeldt,A.A., Hill,A., Garcia-Alonso,L., Stein,F. *et al.* (2020) The functional landscape of the human phosphoproteome. *Nat. Biotechnol.*, **38**, 365–373.
22. Ma,Y., Guo,Z., Xia,B., Zhang,Y., Liu,X., Yu,Y., Tang,N., Tong,X., Wang,M., Ye,X. *et al.* (2022) Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.*, **40**, 921–931.
23. Chen,B., Khodadoust,M.S., Olsson,N., Wagar,L.E., Fast,E., Liu,C.L., Muftuoglu,Y., Sworder,B.J., Diehn,M., Levy,R. *et al.* (2019) Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.*, **37**, 1332–1343.
24. Dai,C., Füllgrabe,A., Pfeuffer,J., Solovyeva,E.M., Deng,J., Moreno,P., Kamatchinathan,S., Kundu,D.J., George,N., Fexova,S. *et al.* (2021) A proteomics sample metadata representation for multiomics integration and big data analysis. *Nat. Commun.*, **12**, 5854.
25. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Félix,E., Magariños,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
26. Walter,M., Allen,L.N., de la Vega de León,A., Webb,S.J. and Gillet,V.J. (2022) Analysis of the benefits of imputation models over traditional QSAR models for toxicity prediction. *J. Cheminform.*, **14**, 32.
27. Lenselink,E.B., ten Dijke,N., Bongers,B., Papadatos,G., van Vlijmen,H.W.T., Kowalczyk,W., IJzerman,A.P. and van Westen,G.J.P. (2017) Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.*, **9**, 45.
28. Mayr,A., Klambauer,G., Unterthiner,T., Steijaert,M., Wegner,J.K., Ceulemans,H., Clevert,D.-A. and Hochreiter,S. (2018) Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.*, **9**, 5441–5451.
29. Brown,N., Fiscato,M., Segler,M.H.S. and Vaucher,A.C. (2019) GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.*, **59**, 1096–1108.
30. Leeson,P.D., Bento,A.P., Gaulton,A., Hersey,A., Manners,E.J., Radoux,C.J. and Leach,A.R. (2021) Target-Based evaluation of 'Drug-Like' properties and ligand efficiencies. *J. Med. Chem.*, **64**, 7210–7230.

31. Zdrazil,B. and Guha,R. (2018) The rise and fall of a scaffold: a trend analysis of scaffolds in the medicinal chemistry literature. *J. Med. Chem.*, **61**, 4688–4703.
32. Jasial,S., Hu,Y. and Bajorath,J. (2016) Assessing the growth of bioactive compounds and scaffolds over time: implications for lead discovery and scaffold hopping. *J. Chem. Inf. Model.*, **56**, 300–307.
33. Horvath,D., Orlov,A., Osolodkin,D.I., Ishmukhametov,A.A., Marcou,G. and Varnek,A. (2020) A chemographic audit of anti-Coronavirus Structure-activity information from public databases (ChEMBL). *Mol Inform*, **39**, e2000080.
34. Klimenko,K., Marcou,G., Horvath,D. and Varnek,A. (2016) Chemical space mapping and structure–activity analysis of the ChEMBL antiviral compound set. *J. Chem. Inf. Model.*, **56**, 1438–1454.
35. Orlov,A.A., Zherebker,A., Eletskaya,A.A., Chernikov,V.S., Kozlovskaya,L.I., Zhernov,Y.V., Kostyukevich,Y., Palyulin,V.A., Nikolaev,E.N., Osolodkin,D.I. *et al.* (2019) Examination of molecular space and feasible structures of bioactive components of humic substances by FTICR MS data mining in ChEMBL database. *Sci. Rep.*, **9**, 12066.
36. Mervin,L.H., Bulusu,K.C., Kalash,L., Afzal,A.M., Svensson,F., Firth,M.A., Barrett,I., Engkvist,O. and Bender,A. (2018) Orthologue chemical space and its influence on target prediction. *Bioinformatics*, **34**, 72–79.
37. Zdrazil,B., Pinto,M., Vasanthanathan,P., Williams,A.J., Balderud,L.Z., Engkvist,O., Chichester,C., Hersey,A., Overington,J.P. and Ecker,G.F. (2012) Annotating human P-Glycoprotein bioassay data. *Mol. Inf.*, **31**, 599–609.
38. Kalliokoski,T., Kramer,C., Vulpetti,A. and Gedeck,P. (2013) Comparability of mixed IC_{50} data - a statistical analysis. *PLoS One*, **8**, e61007.
39. Bosc,N., Atkinson,F., Felix,E., Gaulton,A., Hersey,A. and Leach,A.R. (2019) Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *J. Cheminform.*, **11**, 4.
40. Awale,M. and Reymond,J.-L. (2017) The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. *J. Cheminform.*, **9**, 11.
41. Koutsoukas,A., Lowe,R., Kalantarmotamedi,Y., Mussa,H.Y., Klaffke,W., Mitchell,J.B.O., Glen,R.C. and Bender,A. (2013) In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass naïve bayes and parzen-rosenblatt window. *J. Chem. Inf. Model.*, **53**, 1957–1966.
42. Olivecrona,M., Blaschke,T., Engkvist,O. and Chen,H. (2017) Molecular de-novo design through deep reinforcement learning. *J. Cheminform.*, **9**, 48.
43. Kerstjens,A. and De Winter,H. (2022) LEADD: lamarckian evolutionary algorithm for de novo drug design. *J. Cheminform.*, **14**, 3.
44. Zahoránszky-Kőhalmi,G., Sheils,T. and Oprea,T.I. (2020) SmartGraph: a network pharmacology investigation platform. *J. Cheminform.*, **12**, 5.
45. Dafniet,B., Cerisier,N., Boezio,B., Clary,A., Ducrot,P., Dorval,T., Gohier,A., Brown,D., Audouze,K. and Taboureau,O. (2021) Development of a chemogenomics library for phenotypic screening. *J. Cheminform.*, **13**, 91.
46. Sarkans,U., Chiu,W., Collinson,L., Darrow,M.C., Ellenberg,J., Grunwald,D., Hériché,J.-K., Iudin,A., Martins,G.G., Meehan,T. *et al.* (2021) REMBI: recommended metadata for biological Images—enabling reuse of microscopy data in biology. *Nat. Methods*, **18**, 1418–1422.
47. Ferguson,C., Araújo,D., Faulk,L., Gou,Y., Hamelers,A., Huang,Z., Ide-Smith,M., Levchenko,M., Marinos,N., Nambiar,R. *et al.* (2021) Europe PMC in 2020. *Nucleic Acids Res.*, **49**, D1507–D1514.
48. Ochoa,D., Hercules,A., Carmona,M., Suveges,D., Gonzalez-Uriarte,A., Malangone,C., Miranda,A., Fumis,L., Carvalho-Silva,D., Spitzer,M. *et al.* (2021) Open targets platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res.*, **49**, D1302–D1310.
49. Ghoussaini,M., Mountjoy,E., Carmona,M., Peat,G., Schmidt,E.M., Hercules,A., Fumis,L., Miranda,A., Carvalho-Silva,D., Buniello,A. *et al.* (2021) Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.*, **49**, D1311–D1320.
50. Han,Y., Klinger,K., Rajpal,D.K., Zhu,C. and Teeple,E. (2022) Empowering the discovery of novel target-disease associations via machine learning approaches in the open targets platform. *BMC Bioinf.*, **23**, 232.
51. Gogleva,A., Polychronopoulos,D., Pfeifer,M., Poroshin,V., Ughetto,M., Martin,M.J., Thorpe,H., Bornot,A., Smith,P.D., Sidders,B. *et al.* (2022) Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nat. Commun.*, **13**, 1667.
52. Ye,C., Swiers,R., Bonner,S. and Barrett,I. (2022) A knowledge graph-enhanced tensor factorisation model for discovering drug targets. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, https://doi.org/10.1109/TCBB.2022.3197320.
53. Geleta,D., Nikolov,A., Edwards,G., Gogleva,A., Jackson,R., Jansson,E., Lamov,A., Nilsson,S., Pettersson,M., Poroshin,V. *et al.* (2021) Biological insights knowledge graph: an integrated knowledge graph to support drug development. bioRxiv doi: https://doi.org/10.1101/2021.10.28.466262, 01 November 2021, preprint: not peer reviewed.
54. Fernández-Torras,A., Duran-Frigola,M., Bertoni,M., Locatelli,M. and Aloy,P. (2022) Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque. *Nat. Commun.*, **13**, 5304.
55. Failli,M., Paananen,J. and Fortino,V. (2019) Prioritizing target-disease associations with novel safety and efficacy scoring methods. *Sci. Rep.*, **9**, 9852.
56. Paliwal,S., de Giorgio,A., Neil,D., Michel,J.-B. and Lacoste,A.M. (2020) Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs. *Sci. Rep.*, **10**, 18250.
57. Deutsch,E.W., Orchard,S., Binz,P.-A., Bittremieux,W., Eisenacher,M., Hermjakob,H., Kawano,S., Lam,H., Mayer,G., Menschaert,G. *et al.* (2017) Proteomics standards initiative: fifteen years of progress and future work. *J. Proteome Res.*, **16**, 4288–4298.
58. LeDuc,R.D., Deutsch,E.W., Binz,P.-A., Fellers,R.T., Cesnik,A.J., Klein,J.A., Van Den Bossche,T., Gabriels,R., Yalavarthi,A., Perez-Riverol,Y. *et al.* (2022) Proteomics standards initiative's proforma 2.0: unifying the encoding of proteoforms and peptidoforms. *J. Proteome Res.*, **21**, 1189–1195.
59. Deutsch,E.W., Perez-Riverol,Y., Carver,J., Kawano,S., Mendoza,L., Van Den Bossche,T., Gabriels,R., Binz,P.-A., Pullman,B., Sun,Z. *et al.* (2021) Universal spectrum identifier for mass spectra. *Nat. Methods*, **18**, 768–770.