Democratising knowledge representation with BioCypher

Sebastian Lobentanzer[1,*], Patrick Aloy[2,3], Jan Baumbach[4], Balazs Bohar[5,6], Pornpimol Charoentong[8,9], Katharina Danhauser[10], Tunca Doğan[11,12], Johann Dreo[13,14], Ian Dunham[15,16], Adrià Fernandez-Torras[2], Benjamin M. Gyori[17], Michael Hartung[4], Charles Tapley Hoyt[17], Christoph Klein[10], Tamas Korcsmaros[5,18,19], Andreas Maier[4], Matthias Mann[20,21], David Ochoa[15,16], Elena Pareja-Lorente[2], Ferdinand Popp[22], Martin Preusse[23], Niklas Probul[4], Benno Schwikowski[13], Bünyamin Sen[11,12], Maximilian T. Strauss[20], Denes Turei[1], Erva Ulusoy[11,12], Dagmar Waltemath[24], Judith A. H. Wodke[24], Julio Saez-Rodriguez[1,*]

[1] Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany

[2] Institute for Research in Biomedicine (IRB Barcelona), the Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain

[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

[4] Institute for Computational Systems Biology, University of Hamburg, Germany

[5] Earlham Institute, Norwich, UK

[6] Biological Research Centre, Szeged, Hungary

[7]

[8] Centre for Quantitative Analysis of Molecular and Cellular Biosystems (Bioquant), Heidelberg University, Im Neuenheimer Feld 267, 69120, Heidelberg, Germany

[9] Department of Medical Oncology, National Centre for Tumour Diseases (NCT), Heidelberg University Hospital (UKHD), Im Neuenheimer Feld 460, 69120, Heidelberg, Germany

[10] Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, LMU Munich, Germany

[11] Biological Data Science Lab, Department of Computer Engineering, Hacettepe University, Ankara, Turkey

[12] Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Turkey

[13] Computational Systems Biomedicine Lab, Department of Computational Biology, Institut Pasteur, Université Paris Cité, Paris, France

[14] Bioinformatics and Biostatistics Hub, Institut Pasteur, Université Paris Cité, Paris, France

[15] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

[16] Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

[17] Laboratory of Systems Pharmacology, Harvard Medical School, Boston, USA

[18] Imperial College London, London, UK

[19] Quadram Institute Bioscience, Norwich, UK

[20] Proteomics Program, Novo Nordisk Foundation Centre for Protein Research, University of Copenhagen, Copenhagen, Denmark

[21] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

[22] Applied Tumour Immunity Clinical Cooperation Unit, National Centre for Tumour Diseases (NCT), German Cancer Research Centre (DKFZ), Im Neuenheimer Feld 460, 69120, Heidelberg, Germany

[23] German Centre for Diabetes Research (DZD), Neuherberg, Germany

[24] Medical Informatics Laboratory, University Medicine Greifswald, Germany

* Corresponding author: pub.saez@uni-heidelberg.de

* Co-corresponding: sebastian.lobentanzer@uni-heidelberg.de

All authors except first and last are listed alphabetically.

[AU: edits ok? Please adjust numbering of references after finalizing your edits]

To the editor

Biomedical data are amassed at an ever-increasing rate, and machine learning tools that use prior knowledge in combination with biomedical big data are gaining much traction [1,2]. Knowledge graphs (KGs) are rapidly becoming the dominant form of knowledge representation However, for many research groups, building their own biomedical KG is prohibitively expensive. This motivated us to build the BioCypher framework to support users in creating KGs (https://biocypher.org).

The ability to build a task-specific KG is important, since directly standardising the representation of biomedical knowledge is not appropriate for the diverse research tasks in the community. While human researchers can contextualise and abstract concepts easily, the same does not apply to algorithms. For example, drug discovery tasks (viewing genes as functional ancestors of protein targets) require a different KG structure and content compared to the implementation of a molecular tumour board (genes as clinical markers), which is different still from research into cell type-contextualised gene regulatory network inference (genes as targets of regulatory mechanisms). Even for similar tasks, the KG structure or subtle decisions about included resources lead to different results for many modern analytic methods [2]. In addition, decisions about how to represent knowledge at each primary resource pose problems in their integration, for instance via the use of different identifier namespaces, levels of granularity, or licences [4,5].

The current landscape of biomedical KGs is not easily navigated; neither the KGs themselves, nor the pipelines used to build them, consistently adhere to FAIR (Findable, Accessible, Interoperable, and Reusable) [6] and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) [7] principles. Understandably, the overhead required to implement these principles may not be justified when building a one-off task-specific KG for research. Thus, many KGs are built manually for specific applications, which leads to issues in their reuse and integration [4]. For downstream users, the resulting KGs are too distinct to easily compare or combine [5]. Maintaining KGs for the community is additional work; once maintenance stops, they quickly deteriorate, leading to reusability and reproducibility issues [4] (**Supplementary Note 1**).

BioCypher has been built with continuous consideration of the FAIR and TRUST principles, yielding benefits to the entire community in multiple respects:

1) **Modularity:** To rationalise efforts across the community, we propose a modular architecture that maximises reuse of data and code in three ways: input, ontology, and output (**Figure 1A**). Input adapters allow delegating maintenance work to one central place for each resource, ontology adapters give access to the wealth of structured information curated by the ontology community, and output adapters allow benchmarking and selection of database management systems. Together, these mechanisms enable a workflow that reduces the time and effort to develop and deploy custom KGs.
2) **Harmonisation:** By using ontologies as expertly crafted repositories of conceptual hierarchies, we facilitate harmonisation from a biological perspective. We aid with the

technical aspects of using and manipulating ontologies, for instance by flexibly extending or hybridising complementary ontologies.

3) **Reproducibility:** By sharing the mapping of KG contents to ontologies, we facilitate reproduction of the structure of the corresponding database without access to the primary data, which may be prohibited by licence or privacy issues. We also enable extraction of subgraphs, effectively converting storage-oriented to task-specific KGs, which due to their reduced sizes are easier to share alongside analyses.
4) **Reusability and accessibility:** Finally, the sustainability of research software is strongly related to adoption in – and contributions from – the community. BioCypher is developed as a TRUSTworthy open-source software, applying methods of continuous integration and deployment, and including a diverse community of researchers and developers from the beginning. This facilitates workflows that are tested end-to-end, including the integrity of the scientific data. We operate under the permissive MIT licence and provide community members with guidelines for their contributions and a code of conduct (https://github.com/biocypher).

Different measures further increase the accessibility and FAIRness of our framework. For example, we provide a template repository for a BioCypher pipeline with adapters, including a *Docker Compose* setup. To enable learning by example, we curate existing pipelines, as well as all adapters they use, in our GitHub organisation. Using the GitHub API and a BioCypher pipeline, we build a "meta-graph" for the simple browsing and analysis of BioCypher workflows (https://meta.biocypher.org). To inform the contents of this meta-graph, we have reactivated and now maintain the Biomedical Resource Ontology (BRO [8]), which helps to categorise pipelines and adapters into research areas, data types, and purposes (**Supplementary Note 2**).

BioCypher is implemented as a Python library that provides a low-code access point to data processing and ontology manipulation, emphasising the reuse of existing resources to the highest extent possible. By our design principles and the automation of data management tasks, we aim to free up developer time and guide decision making on how to represent knowledge, bridging the gap between the field of biomedical ontology and the broad application of databases in research.

By abstracting the KG build process as a combination of modular *input* adapters, we save developer time in the maintenance of integrative resources built from overlapping primary sources (**Figure 1B**), for instance OmniPath [9], Bioteque [2], CROssBAR DB [10], and the Clinical Knowledge Graph [11].

By mapping the contents of those resources onto a common *ontological* space, we gain interoperability between the different biomedical domains (**Figure 1C**). BioCypher helps with the mapping procedure by providing examples and an interface, as well as numerous user-friendliness measures. By using the industry standard Web Ontology Language (OWL) format, we provide access to the majority of available ontologies. Separating the ontology framework from the modelled data enables the implementation of reasoning applications at the ontology level, for instance the ad-hoc harmonisation of disease ontologies.

By providing access to a range of modular *output* adapters, we facilitate the project-specific benchmarking and selection of suitable database management systems. For instance, a Neo4j adapter provides rapid access to extensive databases for maintenance of knowledge and enables queries from analysis (Jupyter) notebooks. Switching to alternative graph or relational databases (e.g., ArangoDB or PostgreSQL) allows for task-specific performance optimisation. A CSV-writer and Python-native adapters (e.g., Pandas, sparse matrix, or NetworkX formats) yield knowledge representations that can directly be used programmatically by a wide range of machine learning frameworks. Due to BioCypher's modular nature, additional output adapters can quickly be added.

Application programming interfaces (APIs) built on top of the BioCypher KGs enable complex and versatile queries and simplify the interaction of users with the knowledge. For example, web widgets and apps (such as drug discovery and repositioning with https://crossbar.kansil.org and analysis workflows with https://drugst.one) allow researchers to browse and customise the database, and to plug it into standard pipelines. Additionally, a structured, semantically enriched knowledge representation facilitates connection to and improves performance of modern natural language processing applications such as GPT [12] [13]. The use of common standards enables sharing of tools across projects and communities or in cloud-based services that preserve sensitive patient data (**Supplementary Note 3**).

There have been numerous attempts at standardising KGs and making biomedical data stores more interoperable. We can identify three general types of approaches, in increasing order of abstraction: centrally maintained databases, explicit standard formats (modelling languages), and KG frameworks. With BioCypher, we aim to improve user-friendliness on all three levels of abstraction; for an in-depth discussion, see **Supplementary Note 4.** Despite many efforts, there is no widely accepted solution. Very often, resources take the "path of least resistance" in adopting their own, arbitrary formats of representation. To our knowledge, no framework provides easy access to state-of-the-art KGs to the average biomedical researcher, a gap that BioCypher aims to fill. We demonstrate some key advantages of BioCypher by case studies in **Supplementary Note 5**.

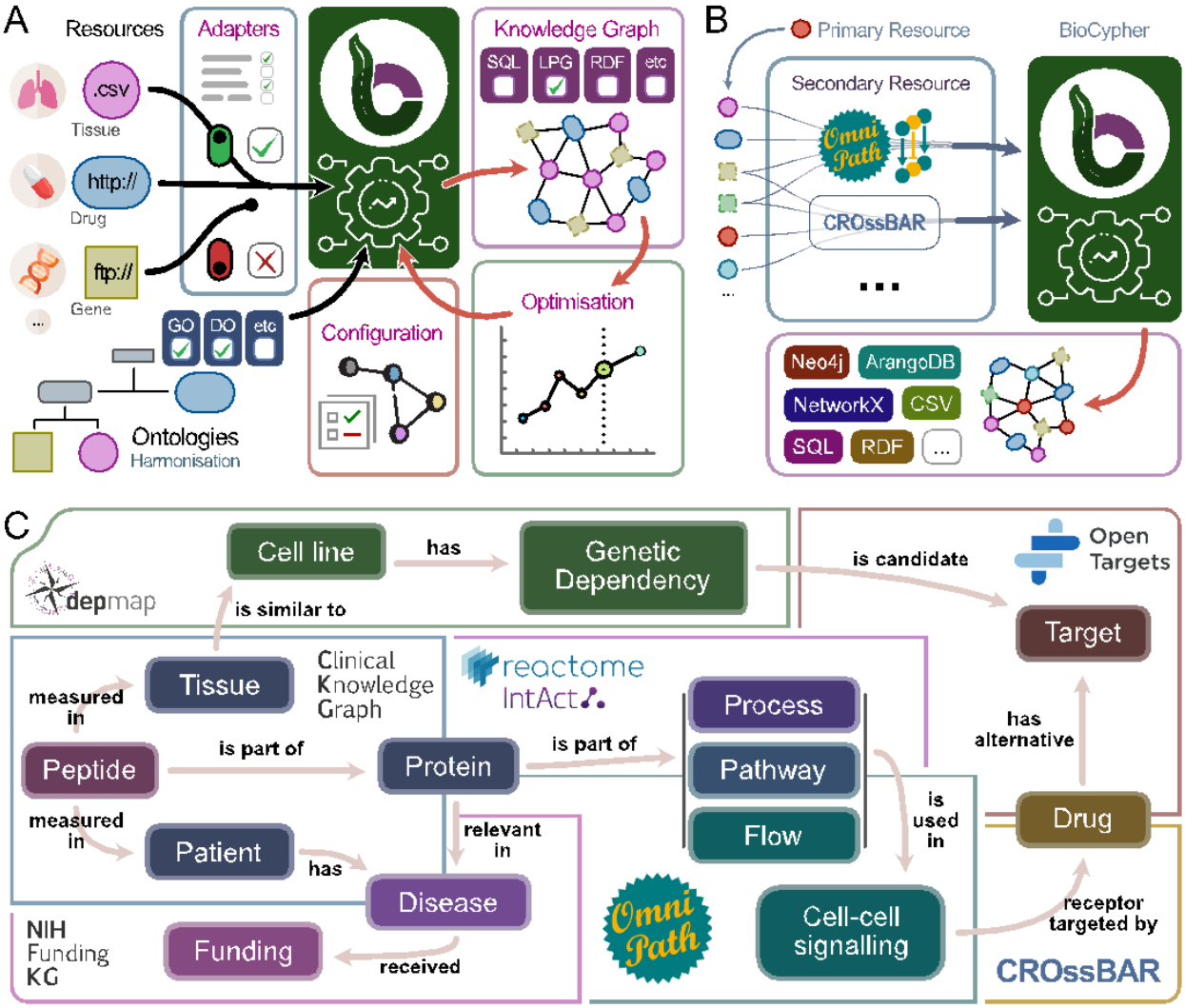

Figure 1: The BioCypher framework. *A) Threefold modularity: Resources (left) and ontologies (bottom left) are combined to yield a knowledge graph (right). The mapping of entities to ontology concepts is realised by shareable configuration, which can be iteratively optimised. B) Initially, we transform commonly used, curated "secondary" resources into configurable, task-specific knowledge graphs in various output formats. Incrementally, these secondary adapters will be replaced by primary resource adapters (see* ***Figure S1****). Coloured panels in A and B indicate parts of the BioCypher ecosystem. C) Agreeing on a common representational framework allows harmonisation of task-specific data sources to answer complex queries across biomedical domains. For instance, starting at mass spectrometry measurements of a patient's tumour (left), one could go through clinical annotations to genetic dependencies from the Dependency Map project to identify potential drug targets, or through pathway / process annotations in Reactome and IntAct, identify relevant ligand-receptor pairs using OmniPath, and use CROssBAR to perform drug discovery or repurposing for these receptors. Panels correspond to resources; although we work on most of the displayed resources, the figure is used for illustrative purposes and does not depict an existing pipeline.*

We believe that creating a more interoperable biomedical research community is as much a social effort as it is a scientific software problem. To facilitate adoption of any approach, the

process must be made as simple as possible, and it must yield tangible rewards, such as significant savings in developer time. We will provide hands-on training for all interested researchers, and we invite all database and tool developers to join our collective effort.

Editor's note: This article has been peer-reviewed

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme (grant agreement No 965193 [DECIDER] and 116030 [TransQST]), the German Federal Ministry of Education and Research (BMBF, Computational Life Sciences grant No 031L0181B and MSCoreSys research initiative research core SMART-CARE 031L0212A), the Defense Advanced Research Projects Agency (DARPA) Young Faculty Award [W911NF-20-1-0255], and the Medical Informatics Initiative Germany, MIRACUM consortium, FKZ: 01ZZ2019.

We thank Henning Hermjakob, Benjamin Haibe-Kains, Pablo Rodriguez-Mier, Daniel Dimitrov, and Olga Ivanova for feedback on the manuscript, and Ben Hitz and Pedro Assis for feedback on their use of BioCypher.

Author Contributions

The project was conceived by SL and JSR. The software was developed by SL with input from DT. The manuscript was drafted by SL, edited by JSR, and jointly revised by all co-authors. All co-authors as members of the BioCypher Consortium contributed to the case studies in development and writing and gave feedback for software development, which was coordinated and integrated by SL.

Conflict of interest

JSR reports funding from GSK, Pfizer and Sanofi and fees from Travere Therapeutics and Astex Pharmaceuticals.

Bibliography

1. Li, M. M., Huang, K. & Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nat. Biomed. Eng.* **6**, 1353–1369 (2022).
2. Fernández-Torras, A., Duran-Frigola, M., Bertoni, M., Locatelli, M. & Aloy, P. Integrating and

formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque. *Nat. Commun.* **13**, 5304 (2022).

3. Tiddi, I. & Schlobach, S. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **302**, 103627 (2022).
4. Bonner, S. *et al.* A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Brief. Bioinformatics* **23**, (2022).
5. Callahan, T. J., Tripodi, I. J., Pielke-Lombardo, H. & Hunter, L. E. Knowledge-Based Biomedical Data Science. *Annu. Rev. Biomed. Data Sci.* **3**, 23–41 (2020).
6. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
7. Lin, D. *et al.* The TRUST Principles for digital repositories. *Sci. Data* **7**, 144 (2020).
8. Tenenbaum, J. D. *et al.* The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. *J. Biomed. Inform.* **44**, 137–145 (2011).
9. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
10. Doğan, T. *et al.* CROssBAR: comprehensive resource of biomedical relations with knowledge graph representations. *Nucleic Acids Res.* **49**, e96 (2021).
11. Santos, A. *et al.* A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol.* **40**, 692–702 (2022).
12. Andrus, B. R., Nasiri, Y., Cui, S., Cullen, B. & Fulda, N. Enhanced Story Comprehension for Large Language Models through Dynamic Document-Based Knowledge Graphs. *AAAI* **36**, 10436–10444 (2022).
13. Lobentanzer, S. & Saez-Rodriguez, J. A Platform for the Biomedical Application of Large Language Models. *arXiv* (2023) doi:10.48550/arxiv.2305.06488.