https://doi.org/10.1038/s41587-024-02534-3

A platform for the biomedical application of large language models

Generative artificial intelligence (AI) has advanced considerably in recent years, particularly in the domain of language. However, despite its rapid commodification, its use in biomedical research is still in its infancy[1,2]. The two main avenues for using large language models (LLMs) are end-user-ready platforms, which are usually provided by large corporations, and custom solutions developed by individual researchers with programming knowledge. Both use cases have significant limitations. Commercial platforms do not meet the transparency standards required for reproducible research; none are open source, and only a few provide (superficial) scientific descriptions of their algorithms[3]. They are also subject to privacy concerns (reuse of user data) and to considerable commercial pressures. In addition, they are not fully customizable to accommodate a specific research domain or workflow.

Individual solutions, on the other hand, are not accessible to most biomedical researchers. They require many specialized skills in addition to the researcher's domain-specific knowledge, such as programming, data management, machine learning knowledge, technical expertise in deployment and frameworking, and management of software versions in a rapidly changing environment. This, in turn, prevents robust and reproducible results owing to the many technical challenges involved. As a result, applications of LLMs in biomedical research are still at the level of individual case studies[2,4], in contrast to the imaging domain, which boasts several open-source AI frameworks and approved medical devices[1].

To bridge the gap between complex custom solutions and closed-source commercial

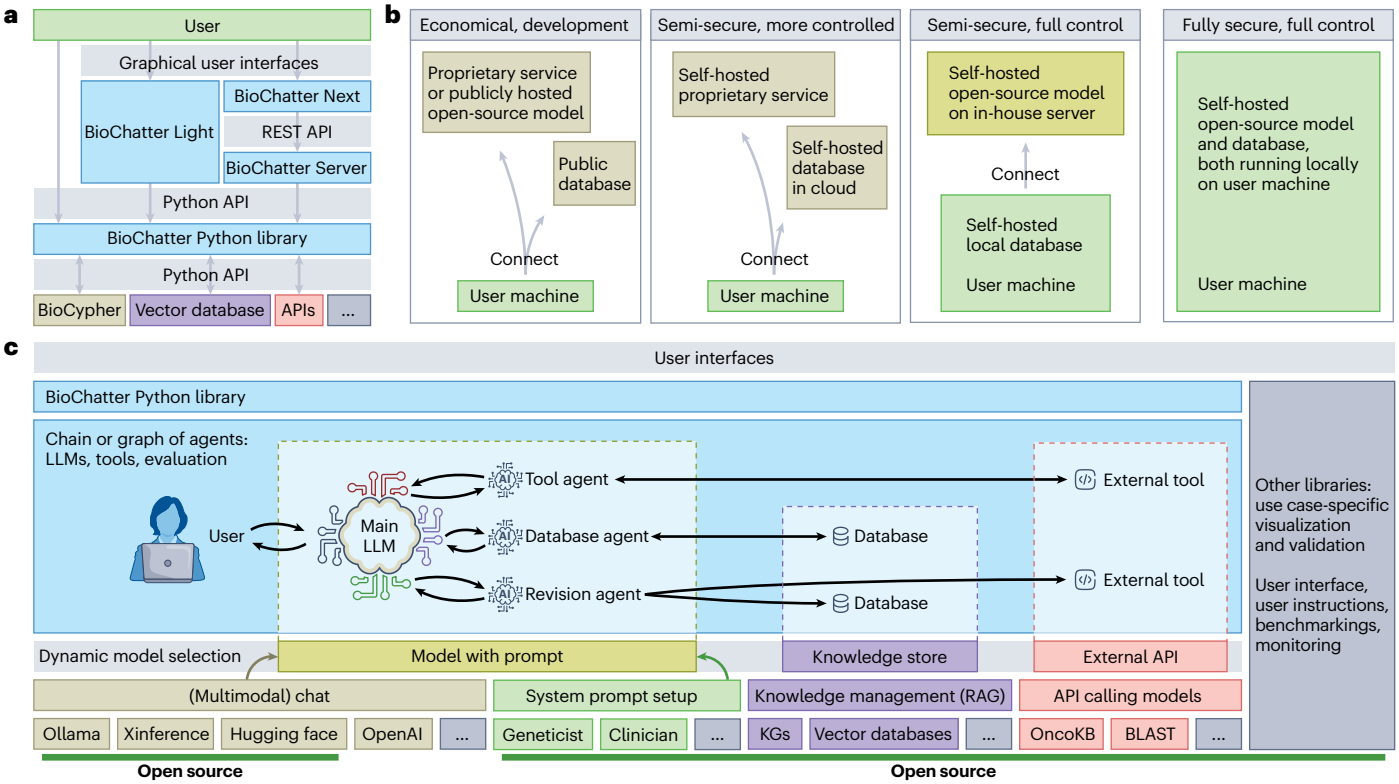

Fig. 1 | The modular BioChatter platform architecture. a, BioChatter provides a selection of diverse APIs for various use cases (Python, REST) and two graphical user interfaces (the Python-based "Light" for rapid prototyping and the more full-featured JavaScript app "Next"). **b**, BioChatter facilitates the creation of custom deployments on a spectrum of tradeoff between simplicity/economy (left) and security (right). **c**, BioChatter harmonizes the APIs of open-source LLM deployment tools and proprietary LLM providers (brown), knowledge management systems such as knowledge graphs and vector databases (purple), public APIs (red) of databases (such as OncoKB[14]), and software (such as BLAST[15]). In addition, the LLM can be specifically instructed according to the user's context via customizable system prompts (green). Each use case is then an individualized combination of these components, combined by either manual or semiautomated agentic workflows, and adapted to the user's needs, including use-case-specific validation for robustness. API, application programming interface; KG, knowledge graph; LLM, large language model; REST, representational state transfer.

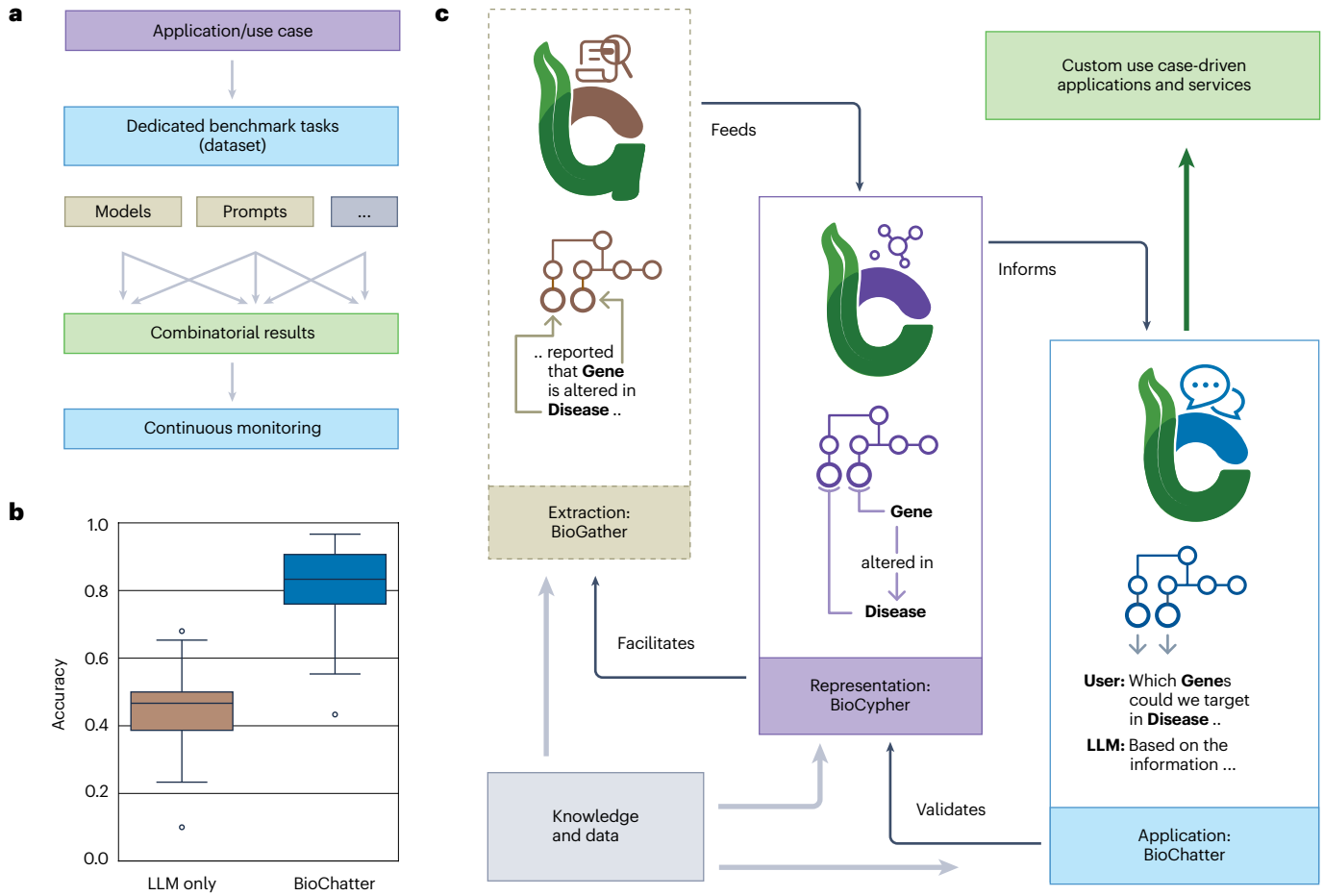

Fig. 2 | Benchmarking, monitoring, and outlook. a, The workflow of introducing use case-specific tests into the BioChatter benchmarking framework facilitates continuous monitoring. Dedicated benchmarks are run across a combination of models and other parameters. **b**, Comparison of two benchmark tasks for knowledge graph query generation show that BioChatter's prompt engine achieves much higher accuracy than the naive approach (measured as number of correct query components among all tested). The BioChatter variant involves a multistep procedure of constructing the query, while the "LLM only" variant receives the complete schema definition of a BioCypher knowledge graph (which BioChatter also uses as a basis for the prompt engine). The general instructions for both variants are otherwise the same (Supplementary Note: Benchmarking). The test includes all models, sizes and quantization levels ($N = 150$), and the performance is measured as the mean accuracy for the two tasks (0.486 ± 0.12 vs 0.844 ± 0.11, unpaired t-test $P < 0.001$, $t = 18.65$). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers. **c**, The BioCypher ecosystem will cover the process of knowledge management from extraction through representation to application. We will develop BioGather (denoted by dashed lines as ongoing work) to integrate natively with the BioCypher and BioChatter frameworks to allow flexible extraction of information from diverse resources using a unified API. This will achieve two pairs of bidirectional synergies: knowledge graphs to extraction pipelines and knowledge graphs to LLMs, respectively.

platforms, we present BioChatter (https://biochatter.org), an open-source Python framework designed to develop custom biomedical research software in line with open science principles[5]. BioChatter's modular architecture offers broad applicability across various biomedical research contexts (Fig. 1a). Its flexible composition supports a wide range of applications, from rapid prototyping to fully encapsulated deployment (Fig. 1b).

To eliminate redundancies from the frequent reimplementation of basic workflows, we offer a permissively licensed software package that integrates and maintains basic open-source components (Supplementary Methods). This not only eases the software development burden but also unifies common LLM-driven workflows in biomedicine, making them available through a single, cohesive application programming interface (API, Fig. 1c). We harmonize the distinct APIs of LLM deployment tools and proprietary LLM providers, allowing the user to switch between different models and providers without changing their code.

We integrate with existing open-source infrastructure such as BioCypher[6] and other databases, allowing injection of domain knowledge to perform retrieval-augmented generation (Supplementary Note: Retrieval-Augmented Generation). We also facilitate the integration of live services (including web-based APIs) via the LLMs' ability to parameterize API queries, made robust by custom implementations of each API service in BioChatter. Finally, our customizable platform allows users to align the LLM to their context via system prompts, as well as advanced workflows using agent-based systems (Supplementary Methods). Simplifying these customizations is a major benefit of BioChatter, often achievable by changing a

simple configuration file (for examples, see Supplementary Note: Customization).

We provide abstract classes for the implementation of multi-agent systems based on reflection[7]. These systems can be configured to perform simple reflective tasks, such as fact checking, or more complex tasks, such as iteratively improving knowledge graph queries on the basis of their results (Supplementary Methods). Prospectively, these agent systems have the potential to scale to considerable levels of complexity. However, this will require thorough evaluation and balancing between performance and computing cost[8].

To address the challenges of reproducibility in LLM workflows, we developed a continuous benchmarking system that allows the community to monitor the performance of all included models on specific tasks. Whenever we add a feature, such as knowledge graph query generation, we introduce a battery of tests to validate its functionality based on community-driven use cases (Fig. 2a). The benchmarking framework runs these tests across all models and relevant parameters and reports the results to the community through our website (Supplementary Note: Benchmarking). We rely on open-source software to rapidly incorporate new models as they are released, effectively keeping pace with rapid developments in the field by distributing technical tasks across the community (Supplementary Methods).

In the case of knowledge graph connectivity, we see a large increase of performance across all LLMs due to the native interaction of BioChatter with BioCypher knowledge graphs (Fig. 2b). Through the detailed description of knowledge graph components in the BioCypher schema configuration, we can effectively guide the LLM in using the knowledge graph (Supplementary Methods). In future work, we plan to extend this approach to extracting information from text and images, and we have begun developing a new framework, BioGather, to support this effort (Fig. 2c). A key advantage of this integrated approach is its ability to guide the model in extracting information from unstructured sources, aligning with the knowledge graph's schema for each use case. This promotes data harmonization and leads to multiple synergies: LLMs make knowledge more accessible and show superior extraction performance owing to their context-awareness while knowledge graphs make LLMs more reliable and facilitate the harmonization of extracted information.

Demands have been made for regulation, an international forum, and benchmarks of LLMs applied to biomedical research[9,10], but practical solutions rooted in the scientific community are still lacking. In contrast to biomedical image analysis, which boasts numerous open frameworks to facilitate access to AI methods, language-based research tasks are at the stage of exploration, case studies, perspectives, and recommendations for manual application[1,2,4,11]. A prevailing perspective on the medical application of LLMs is that their development and evaluation will be shaped and driven by regulators and closed-source companies[11], which is likely to exclude actors outside the Global North. We argue that, instead, the open science community should lead the development and evaluation of these new and rapidly evolving technologies in a fully transparent, open-source manner, to allow an approach inclusive of all groups of stakeholders and the global community[3]. We believe this to be essential for approaching the emerging challenges in the field, such as ensuring safety in sensitive applications. Rather than relying solely on benchmarking suites that highlight model abilities, a framework based on Popperian falsification – designed to push the limits of these models – will be paramount[8].

We prioritize pragmatic execution over attempting to create a one-size-fits-all solution. By offering a flexible, modular platform tailored to developers of custom solutions in the biomedical field, we aim to reduce development and maintenance burdens while enhancing the robustness of resulting applications. BioChatter is not designed to compete with existing infrastructure or consumer products. Instead, it leverages open-source infrastructure to efficiently address the specific demands of biomedical research, distinguishing itself from proprietary consumer-oriented products through its commitment to openness and transparency.

Our ultimate aim is to harmonize APIs not only for LLMs but across the entire scientific knowledge management ecosystem. This includes everything from extraction of information from text and images, through knowledge representation, to the application of this knowledge in decision-making, data analysis, hypothesis generation and scientific communication. We focus on facilitating research tasks that are manual and repetitive, freeing up more time for creative thinking and complex reasoning[4]. To promote early collaboration, we follow a completely open-source development model. We have initiated projects to tackle challenges in research software support, knowledge management, publishing and large-scale drug discovery through the BioChatter consortium (Supplementary Note: The BioChatter Consortium).

In the future, generative AI models will be contrastively trained to synthesize information from multiple relevant modalities, including text, images, and molecular measurements such as genomics and transcriptomics[2,12]. This approach is expected to enhance their ability to aid in reasoning[8,13]. While some of our benchmarks already demonstrate the excellent capabilities of current-generation LLMs in extracting multimodal information from text and images (Supplementary Note: Benchmarking), research and applications in this area are still in their early stages. We are committed to updating BioChatter and its ecosystem to support new developments in the field. We encourage the community to engage with these advancements by requesting features, contributing code and sharing their research and applications.

Data availability

All data used in this study are available in the repository at https://github.com/biocypher/biochatter. In addition, the repository is DOI-indexed at Zenodo/OpenAIRE (https://doi.org/10.5281/zenodo.10777945).

Code availability

All code used in this study is available in the repository at https://github.com/biocypher/biochatter. In addition, the repository is DOI-indexed at Zenodo/OpenAIRE (https://doi.org/10.5281/zenodo.10777945).

Sebastian Lobentanzer[1,2] ✉, **Shaohong Feng**[3], **Noah Bruderer**[4], **Andreas Maier**[5], **The BioChatter Consortium***, **Cankun Wang**[3], **Jan Baumbach**[5,6], **Jorge Abreu-Vicente**[7], **Nils Krehl**[1], **Qin Ma**[3], **Thomas Lemberger**[7] & **Julio Saez-Rodriguez**[1,8] ✉

[1]Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany. [2]Open Targets, European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK. [3]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA. [4]Michael Sars Centre, University of Bergen, Bergen, Norway. [5]Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany. [6]Computational Biomedicine Lab, Department of Mathematics and Computer

Science, University of Southern Denmark, Odense, Denmark. [7]EMBO, Heidelberg, Germany. [8]European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK. *A list of authors and their affiliations appears at the end of the paper.
✉e-mail: sebastian.lobentanzer@gmail.com; saezlab@ebi.ac.uk

Published online: 22 January 2025

References

1. Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F. & Kather, J. N. *Nat. Rev. Cancer* **24**, 427–441 (2024).
2. Simon, E., Swanson, K. & Zou, J. *Nat. Methods* **21**, 1422–1429 (2024).
3. Liesenfeld, A. & Dingemanse, M. Rethinking open source generative AI: open-washing and the EU AI Act. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, https://doi.org/10.1145/3630106.3659005 (ACM, 2024).
4. Pividori, M. *Nature* https://doi.org/10.1038/d41586-024-02630-z (2024).
5. UNESCO. UNESCO Recommendation on Open Science. *UNESCO* https://doi.org/10.54677/mnmh8546 (2021).
6. Lobentanzer, S. et al. *Nat. Biotechnol.* **41**, 1056–1059 (2023).
7. Shinn, N. et al. Preprint at https://doi.org/10.48550/arxiv.2303.11366 (2023).
8. Nezhurina, M., Cipolina-Kun, L., Cherti, M. & Jitsev, J. Preprint at https://doi.org/10.48550/arxiv.2406.02061 (2024).
9. van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. L. *Nature* **614**, 224–226 (2023).
10. Bockting, C. L., van Dis, E. A. M., van Rooij, R., Zuidema, W. & Bollen, J. *Nature* **622**, 693–696 (2023).
11. Lee, P., Goldberg, C. & Kohane, I. *The AI Revolution in Medicine: GPT-4 and Beyond* (Pearson, 2023).
12. Schaefer, M. et al. Joint embedding of transcriptomes and text enables interactive single-cell RNA-seq data exploration via natural language. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations* https://openreview.net/forum?id=yWiZaE4k3K (2024).
13. Lobentanzer, S., Rodriguez-Mier, P., Bauer, S. & Saez-Rodriguez, J. *Mol. Syst. Biol.* **20**, 848–858 (2024).
14. Chakravarty, D. et al. *JCO Precis. Oncol.* https://doi.org/10.1200/po.17.00011 (2017).
15. Camacho, C. et al. *BMC Bioinformatics* **10**, 421 (2009).

Acknowledgements

We thank H. Schumacher, D. Dimitrov and P. Badia i Mompel for feedback on the original draft of the manuscript and the software. This work was supported by funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 965193 for DECIDER (JSR), awards U54AG075931 and R01DK138504 (QM) from the National Institutes of Health, and the Pelotonia Institute for Immuno-Oncology. This manuscript was written using Manubot (https://github.com/manubot) and partially revised using LLMs. The entire manuscript was double-checked for correctness, and the responsibility for the final content lies with the authors only. This project is funded by the European Union under grant agreement 101057619. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. This work was also partly supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract 22.00115.

Author contributions

Authors between consortium and last author are ordered alphabetically by first name. S.L. conceptualized and developed the platform, coordinated the consortium, and wrote the manuscript. S.F. implemented BioChatter functionality and developed both frontend and backend components for the BioChatter Next server. N.B. developed the API calling module with S.F. and S.L. A.M. implemented the local deployment functionality. The BioChatter consortium members contributed to the development of the platform and provided feedback on the manuscript. C.W. architected the BioChatter Next server infrastructure. J.B. provided guidance and supervision as well as hardware resources for local LLM use and contributed to performance benchmarking. J.A.-V. developed text extraction benchmarking procedures. N.K. implemented benchmarking procedures. Q.M. oversaw the development and deployment of the BioChatter Next server environment. T.L. oversaw the text extraction work and acquired funding. J.S.-R. supervised the project, revised the manuscript, and acquired funding. All authors read and approved the final manuscript.

Competing interests

J.S.-R. reports funding from GSK, Pfizer and Sanofi and fees or honoraria from Travere Therapeutics, Stadapharm, Pfizer, Grunenthal, Owkin, Moderna and Astex Pharmaceuticals.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-024-02534-3.

The BioChatter Consortium

Adrián G. Díaz[9,10], Amy Strange[11], Andreas Maier[5], Anis Ismail[12], Anton Kulaga[13], Aurelien Dugourd[1], Barbara Zdrazil[14], Bastien Chassagnol[15], Cankun Wang[3], Cyril Pommier[16,17], Daniele Lucarelli[18,19], Ellen M. McDonagh[2], Emma Verkinderen[20], Fernando M. Delgado-Chaves[5], Georg Fuellen[21], Hannah Sonntag[7], Jan Baumbach[5,6], Jorge Abreu-Vicente[7], Jonatan Menger[22], Julio Saez-Rodriguez[1,8], Lionel Christiaen[4], Ludwig Geistlinger[23], Luna Zacharias Zetsche[24], Marlis Engelke[24], Megan McNutt[3], Melissa Harrison[8], Melissa Hizli[25], Nikolai Usanov[26], Nils Krehl[1], Noah Bruderer[4], Patrick Baracho[24], Qin Ma[3], Sebastian Beier[27], Sebastian Lobentanzer[1,2], Shaohong Feng[3], Stefan Boeing[28], Taru A. Muranen[29], Thomas Lemberger[7], Trang T. Le[30], Valeriia Dragan[1], Xiao-Ran Zhou[27], Yasmin Nielsen-Tehranchian[25] & Yuyao Song[8]

[9]Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium. [10]Structural Biology Brussels, Vrije Universiteit Brussels, Brussels, Belgium. [11]The Francis Crick Institute, London, UK. [12]Laboratory of Multi-Omic Integrative Bioinformatics, Center for Human Genetics, Faculty of Medicine, Katholieke Universiteit Leuven, Leuven, Belgium. [13]Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, Rostock, Germany. [14]Chemical Biology Services, European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK. [15]Laboratory of Probability, Statistics, and Modelling, Sorbonne University, Paris, France. [16]Université Paris-Saclay, INRAE, BioinfOmics, Plant Bioinformatics Facility, Versailles, France. [17]Université Paris-Saclay, INRAE, URGI, Versailles, France. [18]Institute of Translational Cancer Research and Experimental Cancer Therapy, Technical University of Munich, Munich, Germany. [19]Department of Computational Health, Institute of Computational Biology, Helmholtz, Munich, Germany. [20]Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles-Vrije Universiteit Brussel, Brussels, Belgium. [21]Institute for Biostatistics and Informatics in Medicine and Ageing Research (IBIMA), Rostock University Medical Center, Rostock, Germany. [22]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany. [23]Core for Computational Biomedicine, Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [24]Heilbronn University of Applied Sciences, Heilbronn, Germany. [25]GECKO Institute, Heilbronn University of Applied Sciences, Heilbronn, Germany. [26]HEAlthy Life Extension Society (HEALES), Brussels, Belgium. [27]Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Bioeconomy Science Center (BioSC), CEPLAS, Forschungszentrum Jülich, Jülich, Germany. [28]Bioinformatics and Biostatistics Science Technology Platform and Software Engineering and AI Science Technology Platform, The Francis Crick Institute, London, UK. [29]Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [30]Bristol Myers Squibb, Cambridge, MA, USA.