Synergising ELIXIR resources for training in systems biology

Claire Rioualen[1], **Finn Bacall**[2], **Cristina Furlan**[3], **Matúš Kalaš**[4], **Brane Leskošek**[5], **Hervé Ménager**[1,6], **and Barbara Szomolay**[7]

1 Institut Français de Bioinformatique, CNRS UAR 3601, Évry, France **2** University of Manchester, United Kingdom **3** Laboratory of Systems and Synthetic Biology, Wageningen University and Research, Wageningen, Netherlands **4** ELIXIR Norway, and Department of Informatics, University of Bergen, Norway **5** University of Ljubljana, Faculty of Medicine, IBMI, Centre ELIXIR-SI, Slovenia **6** Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, 75015, Paris, France **7** Division of Infection and Immunity & Systems Immunity Research Institute, Cardiff University School of Medicine, United Kingdom

BioHackathon series:
BioHackathon Europe 2023
Barcelona, Spain, 2023
Project 32 - Synergising ELIXIR resources for training in systems biology

Submitted: 08 Apr 2024

Published by BioHackrXiv.org

Project abstract

Systems biology (SB) is a new ELIXIR community, that aims to utilise the ELIXIR ecosystem, such as the Training eSupport System (TeSS) and bio.tools, a registry of software tools and data resources for life sciences. One of the main initial objectives of the SB community is to create an SB-themed domain hosted by TeSS, encompassing SB-related ELIXIR services and events, in a fully automated way.

Most content in TeSS is sourced through automated aggregation ("scraping") of external sources containing resources marked up with semantic metadata, like Bioschemas. Currently, TeSS cannot recognize references to bio.tools identifiers from a Bioschemas-annotated resource, so the number of resources linked to bio.tools is relatively low.

In this project, we will focus on selected SB disciplines from the priority areas of the ELXIR SB community to integrate and cross-link related ELIXIR products - training events, training materials, computational and bioinformatics tools, databases and services from the bio.tools registry.

This will be achieved using suitable ontologies identified by the SB community and by careful curation of SB-related materials. We aim to extend this work to other ELIXIR products such as lists of trainers, related ELIXIR Innovation and Industry events and publications. This will serve as a pilot project leading to broader integration with other SB disciplines, and will be of interest to several other ELIXIR communities.

Introduction

Background

The systems biology community (Santos et al., 2022) is one of the most recently-created ELIXIR community, with the aim of answering several infrastructure needs identified by the community. Owing to its very nature, the field of systems biology relies not only on the development and use of modelling tools, but also on data storage solutions and community standards. For these reasons, the SB community set some of its main focuses on the interoperability of systems biology resources and the coordination of capacity building and training resources.

These objectives align within the scope of the ELIXIR Programme and ecosystem, and more specifically with the Training eSupport System (TeSS) (Beard et al., 2020) and the bio.tools registry (Ison et al., 2019). TeSS provides a platform for all sorts of life science-related training events and materials, mostly by aggregating data and metadata from identified content providers by means of HTML scraping, application programming interface (API) integration, and structured-data formats parsing (Schema.org) (Guha et al., 2016). Besides, TeSS is highly involved with Bioschemas (Gray et al., 2017) development, an initiative that aims at building upon Schema.org specifications, while providing better-tailored specification profiles for life sciences at large. The bio.tools platform is a community effort of curation of computational biology tools, answering the need for a consistent and up-to-date registry of existing tools and algorithms across all fields of life sciences. The bio.tools registry, also supported by ELIXIR, currently amounts to almost 30,000 tools, annotated by topics, operations, data formats, and many additional criteria, allowing users to navigate it rather straightforwardly.

The adequate annotation of both tools and training materials available in TeSS and bio.tools is granted and facilitated by using the EDAM ontology (Black et al., 2022). EDAM is precisely aimed at standardising terms and definitions for data analysis and management in the context of life sciences and beyond. Its goal is to define a controlled vocabulary to be used for several purposes such as the the semantic annotation and classification of relevant resources.

Problem

While the wealth of systems biology resources address interoperability and findability concerns, they often lack standard annotations and consistent metadata coverage. A second potential bottleneck is the absence of adequately fine-tuned EDAM concepts and terms, due to the uneven coverage of certain subdisciplines of biology, as well as the constant evolution of bioinformatics methods and topics deriving from advanced biotechnologies. In particular, the domain of systems biology in EDAM has not been populated extensively yet (beyond data formats and a few main concepts (Shaikh et al., 2022)), partly due to a lack of demand and appropriate expertise (*i.e.* experts interested in contributing).

Objective

The overall vision of this project is to better synergise the ELIXIR ecosystem by working towards the integration of finer-grained SB concepts and their use for the annotation of software and training materials relevant to the field, and thus improve their findability and navigability.

Strategy

In developing this project and evaluating its feasibility, we undertook a step-by-step approach involving several key tasks needed to design the foreseen product. We initiated the process by identifying specific use cases, spanning various systems biology courses, relevant topics, tools, and search terms. This user-centred design approach provided insights into the requirements needed to refine our method and align it with our objectives.

Simultaneously, we implemented a data model utilising resources like EDAM, TeSS, and bio.tools. This allowed us to assess the availability and adequacy of metadata for representing and connecting different elements within the domain, such as events, materials, and software.

Furthermore, we focused on identifying ontologies needed to properly describe SB resources, particularly refining the EDAM ontology and other specialised ontologies like the Systems Biology Ontology (SBO; http://biomodels.net/SBO/) and the The Mathematical Modelling Ontology (MAMO; http://identifiers.org/MAMO/). This ensured the provision of an accurate and consistent representation of the available information, thus strengthening the effectiveness of our approach.

Additionally, we conducted a thorough analysis to identify any existing gaps in platforms like TeSS and bio.tools. This analysis served to guide improvements to these platforms, enhancing their usability and relevance for users in the field of systems biology.

BioHackathon results

1. Definition of use cases

We first aim to define use cases for this project, which will specify which features can be added or improved. In order to achieve this, we first selected a set of existing systems biology resources, both as training courses (Table 1) and software tools (Table 2). Additionally, a list of keywords commonly used by systems biology experts to query similar resources is listed (Table 3).

Table 1. List of systems biology target events or courses.

Type	Name	URL
Course	Systems biology: from large datasets to biological insight	https://www.ebi.ac.uk/training/events/systems-biology-large-datasets-biological-insight-2/
Course	Integrative analysis of multi-omics data	https://www.embl.org/about/info/course-and-conference-office/events/mmd24-01/

Table 2. List of systems biology software tools.

Software link
https://biofam.github.io/MOFA2/
https://bio.tools/cytoscape
https://bio.tools/cellnoptr
https://apps.cytoscape.org/apps/cytocopter
https://bio.tools/cosmos-omics
https://bio.tools/seurat
https://github.com/MLO-lab/MuVI
https://bio.tools/mixomics

Table 3. List of keywords used in order to search for target courses.

Keyword
machine learning
network analysis
data management
logic modelling
data integration
single cell omics
deep learning
open science
data heterogeneity
dimensionality reduction
multi-omics

The different resources, to facilitate their access, should be properly searchable (using e.g. the keywords mentioned above), and linked in the ELIXIR resources (e.g. Training eSupport System (TeSS) (Beard et al., 2020), the bio.tools registry (Ison et al., 2019), EDAM Browser (Brancotte et al., 2018)). Figure 2 illustrates a proposition for these use cases, backed by the data model described in the next section.

2. Data model

Most of the ELIXIR resources that serve the contents we aim to connect provide interoperable metadata using Schema.org, Bioschemas, and EDAM. We surveyed here how these metadata standards can enable the representation of the information necessary to describe different resources (*e.g.* software, training materials) so that they can be searched and connected.

To enhance the findability and navigation of SB resources, the first step was to connect TeSS and bio.tools entries using schema.org markup annotations and establish a semantic model for interoperability (Figure 1).

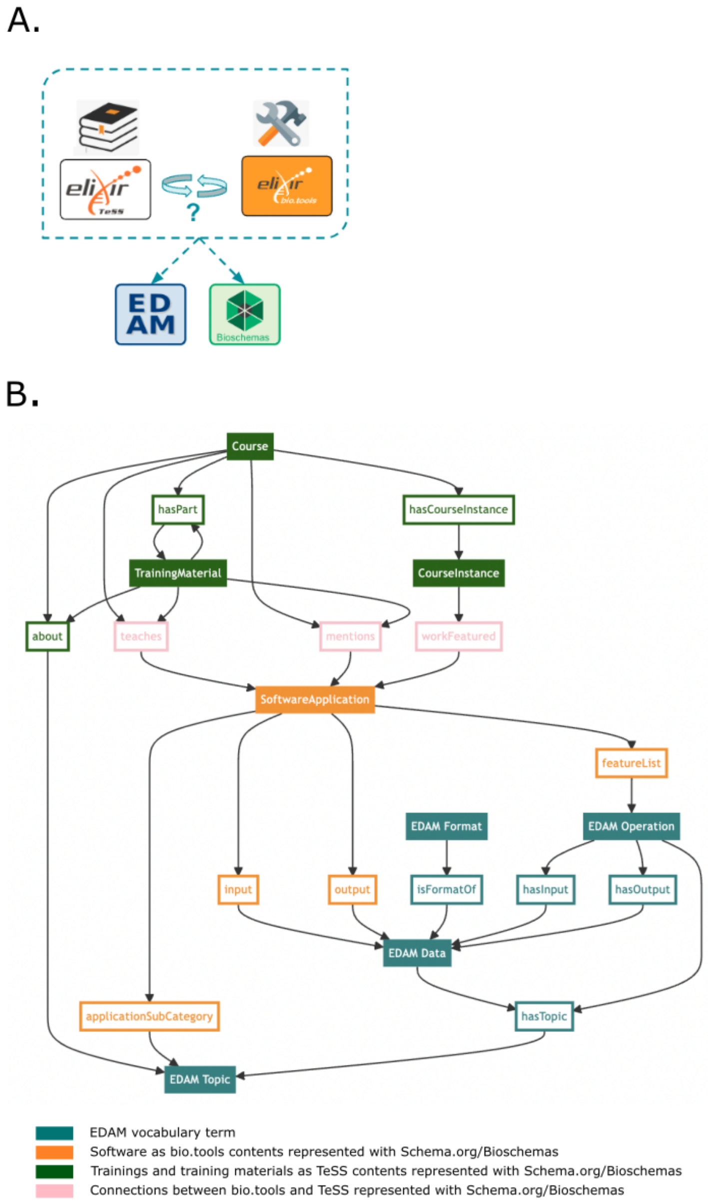


Figure 1: A. Connecting events and tools from TeSS and bio.tools via Bioschemas markup and the EDAM ontology implementation is feasible, but currently not straightforward. B. This data model represents the data objects and links linking training and software tools as well as EDAM with Bioschemas links. It summarises information that can be linked in an interoperable way between bio.tools and TeSS, using EDAM and Bioschemas/Schema.org.

Although the connection is functionally feasible, in practice it relies on the proper use of markup annotations, as recommended by the FAIR guiding principles. Specifically, it demands ongoing efforts to utilise, develop, and sustain a systems biology vocabulary within the EDAM ontology (developer-dependent), while also necessitating efforts to maximise the utilisation of markup annotations offered by Bioschemas (content maker-dependent).

3. Knowledge gap: immediate improvements

Following our curation of existing ontologies and keywords list, we were able to make immediate improvements to our main issue: the accessibility of training resources for systems biology. It involved 3 components: the EDAM ontology, the bio.tools registry, and finally the TeSS platform.

EDAM

First, we identified concepts from our keywords list that were already available in EDAM for the annotation of resources, and when relevant, added or edited relevant attributes. Then we identified terms to be added to the ontology, whether they are topics, operations, or data types, and their parents in the ontology, concise definition, URL, common synonyms, *etc.* (Table 4).

Table 4. Selection of terms that may be used for a user to search for SB courses or training materials in TeSS. Some of those terms are already available in EDAM and can be further annotated with relevant synonyms, some terms are available in specialised ontologies (SBO, MAMO), and finally some terms are yet to be added to and defined in EDAM. More detail can be found in the supplementary table S1.

Label	Ontology term	Sub-ontology
machine learning	http://edamontology.org/topic_3474	Topic
deep learning	https://bioportal.bioontology.org/ontologies/EDAM-BIOIMAGING/?p=classes&conceptid=http%3A%2F%2Fedamontology.org%2Ftopic_____Deep_learning	Topic
dimensionality reduction	http://edamontology.org/operation_3935	Operation
logic modelling	http://identifiers.org/mamo/MAMO_0000030	Data
single-cell omics	http://edamontology.org/topic_4028 (dev. version)	Topic

bio.tools

Tools that were already available in the bio.tools registry were edited to include more accurate annotations of their related EDAM topics, operations, as well as data types when needed. Plus, a few tools were added to the registry (Table 5).

Table 5. Selection of tools commonly used for SB analyses. Some tools are referenced in bio.tools but not necessarily thoroughly annotated, while others are yet to be added. More detail can be found in the supplementary table S2.

Name	bio.tools ID	Topic	Operation
Cytoscape	https://bio.tools/cytoscape	Systems biology	Network analysis
Seurat	https://bio.tools/seurat	Transcriptomics	Data integration
MOFA2	-	Multi-omics	

TeSS

Courses that are in the TeSS catalogue cannot currently be edited by users to add markup annotations. New entries can however be annotated with EDAM topics and operations using an integrated drop-down menu (Table 6).

Table 6. Selection of events or courses related to systems biology. More detail can be found in the supplementary table S3.

Type	Name	Topics	Operations	Tools
Course	Systems biology: from large datasets to biological insight	Systems biology, Multi-omics	Data integration	Cytoscape
Course	Integrative analysis of multi-omics data	Multi-omics	Dimensionality reduction	MOFA2, Seurat

The addition of these terms and annotations across the EDAM-bio.tools-TeSS ecosystem allows us to connect tools and courses through our semantic model, thus facilitating users' search for appropriate courses and training material in the field of systems biology. As shown in Figure 2, the connection of events and tools.

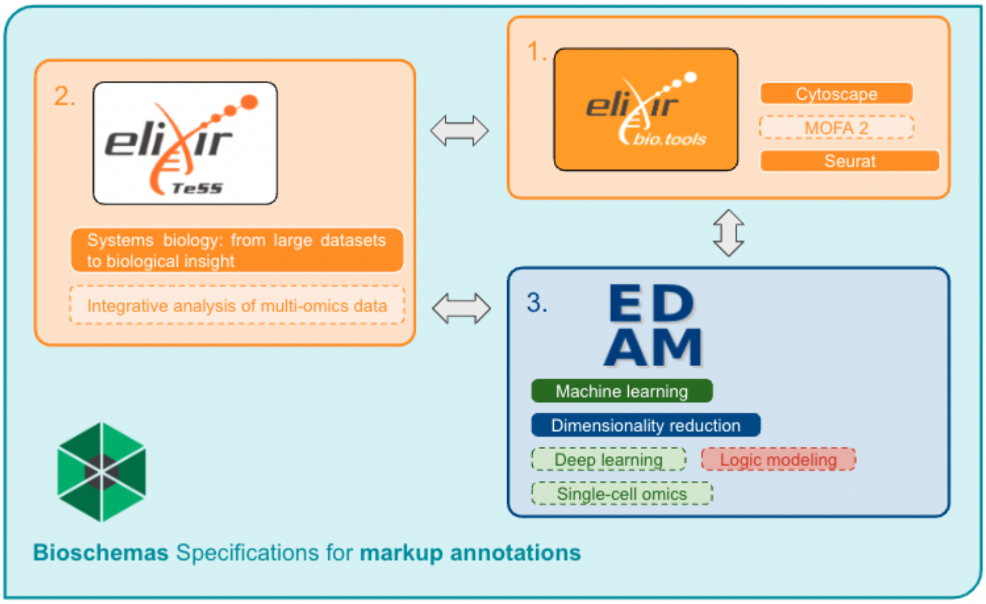


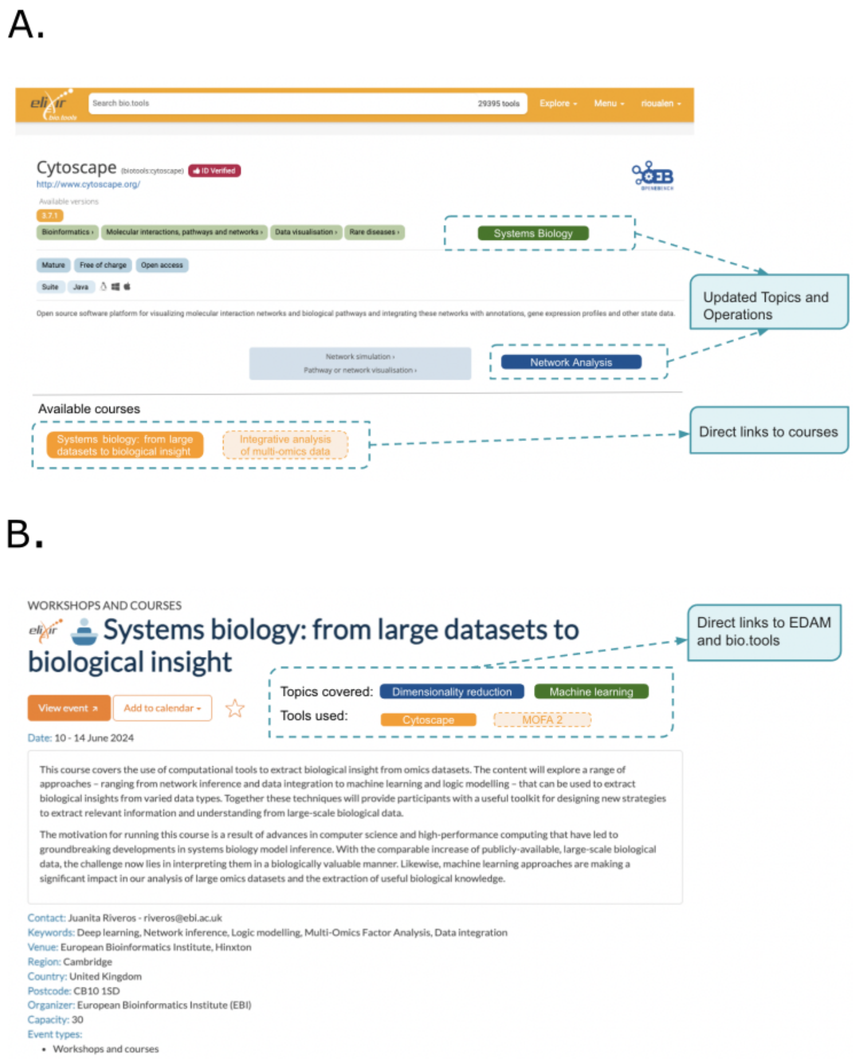
Figure 2: The update of existing terms and tools in the ecosystem (dark-coloured tags) and the addition of new terms and tools (pale-coloured tags) allow the connection of events and tools through Bioschemas markup annotations, specifically, by using (1) the "ComputationalTool" profile (Group Tools), (2) the "Course", "CourseInstance" and "TrainingMaterial" profiles (Group Training) and (3) the EDAM ontology (controlled vocabulary).

In a wider fashion, the adoption of detailed markup annotation as recommended guidelines will contribute to a mutual enrichment among ELIXIR platforms as well as their content providers, through bidirectional metadata scraping.

Discussion

The general goal of the project presented here was to synchronise systems biology resources with the ELIXIR ecosystem. The work conducted during the 2023 Biohackathon Europe event focused on a pilot study, to assess more precisely the feasibility of our goal.

In addition to the various specific editions and additions listed above, a number of evolutions or new features could potentially facilitate the maintenance and curation of bioinformatics resources listed and described with systems and references such as bio.tools, TeSS, or EDAM. In particular, a concrete and rather easy-to-implement feature would to enable a direct navigation from tools referenced in the bio.tools registry to courses or training materials that feature them (Figure 3A), and conversely, from courses or events referenced in TeSS to bio.tools entries relevant to those (Figure 3B).



Open source software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data.

Date: 10 - 14 June 2024

This course covers the use of computational tools to extract biological insight from omics datasets. The content will explore a range of approaches – ranging from network inference and data integration to machine learning and logic modelling – that can be used to extract biological insights from varied data types. Together these techniques will provide participants with a useful toolkit for designing new strategies to extract relevant information and understanding from large-scale biological data.

The motivation for running this course is a result of advances in computer science and high-performance computing that have led to groundbreaking developments in systems biology model inference. With the comparable increase of publicly-available, large-scale biological data, the challenge now lies in interpreting them in a biologically valuable manner. Likewise, machine learning approaches are making a significant impact in our analysis of large omics datasets and the extraction of useful biological knowledge.

Contact: Juanita Riveros - riveros@ebi.ac.uk
Keywords: Deep learning, Network inference, Logic modelling, Multi-Omics Factor Analysis, Data integration
Venue: European Bioinformatics Institute, Hinxton
Region: Cambridge
Country: United Kingdom
Postcode: CB10 1SD
Organizer: European Bioinformatics Institute (EBI)
Capacity: 30
Event types:

- Workshops and courses

Figure 3: A. From bio.tools to TeSS: enable browsing on found tools in bio.tools and provide links that will search with the same terms on TeSS. B. From TeSS to bio.tools: enable browsing on found materials in TeSS and provide links that will search with the same terms on bio.tools.

Long-term goals include the extension the automated framework to other SB-related ELIXIR domains and services by adapting ontologies and exploring ontology mappings (e.g. between EDAM and SBO or MAMO), the annotation of SB-related products by a set of controlled and relational vocabularies, by using selected SB disciplines and related TeSS and bio.tools products (training events, training materials, computational, bioinformatics tools, databases, services), the integration of TeSS and bio.tools by extending TeSS' Bioschemas parser, and finally, the exploration of a possible extension to other ELIXIR resources.

Supplementary data

The supplementary tables are provided with the preprint's github repository (https://github.com/rioualen/BioHackEU23_preprint/).

Acknowledgements

This work was performed during the ELIXIR BioHackathon Europe 2023 organised by ELIXIR in November 2023. This work was supported by ELIXIR, the research infrastructure for life science data. CR is part of the *Institut Français de Bioinformatique* (IFB, UAR 3601), funded by the *Programme d'Investissements d'Avenir* subsidised by the *Agence Nationale de la Recherche*, number ANR-11-INBS-0013.

References

Beard, N., Bacall, F., Nenadic, A., Thurston, M., Goble, C. A., Sansone, S.-A., & Attwood, T. K. (2020). TeSS: A platform for discovering life-science training opportunities. *Bioinformatics*, *36*(10), 3290–3291. https://doi.org/10.1093/bioinformatics/btaa047 **[cito:citesAsAuthority]**

Black, M., Lamothe, L., Hager Eldakroury, Kierkegaard, M., Ankita Priya, Machinda, A., Khanduja, U. S., Drashti Patoliya, Rashika Rathi, Tawah Peggy Che Nico, Umutesi, G., Blankenburg, C., Op, A., Chieke, P., Omodolapo Babatunde, Laurie, S., Neumann, S., Schwämmle, V., Kuzmin, I., … Kalaš, M. (2022). EDAM: The bioscientific data analysis ontology (update 2021)[version 1; not peer reviewed]. *F1000Research*, *11*(ISCB Comm J), 1. https://doi.org/10.7490/f1000research.1118900.1 **[cito:citesAsAuthority]**

Brancotte, B., Blanchet, C., & Ménager, H. (2018). A reusable tree-based web-visualization to browse EDAM ontology, and contribute to it. *Journal of Open Source Software*, *3*(27), 698. **[cito:citesAsAuthority]**

Gray, A. J., Goble, C. A., & Jimenez, R. (2017). Bioschemas: From Potato Salad to Protein Annotation. *ISWC (Posters, Demos & Industry Tracks)*. **[cito:citesAsAuthority]**

Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.org: Evolution of structured data on the web. *Communications of the ACM*, *59*(2), 44–51. https://doi.org/10.1145/2844544 **[cito:usesMethodIn]**

Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Ménager, H., Kalaš, M., Schwämmle, V., Grüning, B., Beard, N., Lopez, R., Duvaud, S., Stockinger, H., Persson, B., Vařeková, R. S., Raček, T., Vondrášek, J., Peterson, H., Salumets, A., Jonassen, I., … Brunak, S. (2019). The bio.tools registry of software tools and data resources for the life sciences. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1772-6 **[cito:citesAsAuthority]**

Santos, V. M. dos, Anton, M., Szomolay, B., Ostaszewski, M., Arts, I., Benfeitas, R., Angel, V. D. D., Ferk, P., Fey, D., Goble, C., Golebiewski, M., Gruden, K., Heil, K. F., Hermjakob, H., Kahlem, P., Klapa, M. I., Koehorst, J., Kolodkin, A., Kutmon, M., … Hancock, J. M. (2022). *Systems Biology in ELIXIR: Modelling in the spotlight* (No. 11:1265). F1000Research. https://doi.org/10.12688/f1000research.126734.1 **[cito:citesAsRelated]**

Shaikh, B., Smith, L. P., Vasilescu, D., Marupilla, G., Wilson, M., Agmon, E., Agnew, H., Andrews, S. S., Anwar, A., Beber, M. E., Bergmann, F. T., Brooks, D., Brusch, L., Calzone, L., Choi, K., Cooper, J., Detloff, J., Drawert, B., Dumontier, M., … Karr, J. R. (2022). BioSimulators: a central registry of simulation engines and services for recommending specific tools. *Nucleic Acids Research*, *50*(W1), W108–W114. https://doi.org/10.1093/nar/gkac331 **[cito:citesAsRelated]**