Briefings in Bioinformatics, 00(00), 2020, 1–12

doi: 10.1093/bib/bbaa046
Method Review

Setting the basis of best practices and standards for curation and annotation of logical models in biology—highlights of the [BC]2 2019 CoLoMoTo/SysMod Workshop

Anna Niarakis, Martin Kuiper, Marek Ostaszewski, Rahuman S. Malik Sheriff, Cristina Casals-Casas, Denis Thieffry, Tom C. Freeman, Paul Thomas, Vasundra Touré, Vincent Noël, Gautier Stoll, Julio Saez-Rodriguez, Aurélien Naldi, Eugenia Oshurko, Ioannis Xenarios, Sylvain Soliman, Claudine Chaouiya, Tomáš Helikar† and Laurence Calzone†

Corresponding author: Anna Niarakis, Univ. Évry, Université Paris-Saclay, Genopole, 91025 Évry, France. E-mail: anna.niaraki@univ-evry.fr
†These authors contributed equally to this work.

Abstract

The fast accumulation of biological data calls for their integration, analysis and exploitation through more systematic approaches. The generation of novel, relevant hypotheses from this enormous quantity of data remains challenging. Logical models have long been used to answer a variety of questions regarding the dynamical behaviours of regulatory networks. As the number of published logical models increases, there is a pressing need for systematic model annotation, referencing and curation in community-supported and standardised formats. This article summarises the key topics and future directions of a meeting entitled 'Annotation and curation of computational models in biology', organised as part of the 2019 [BC]2 conference. The purpose of the meeting was to develop and drive forward a plan towards the standardised annotation of logical models, review and connect various ongoing projects of experts from different communities involved in the modelling and annotation of molecular biological entities, interactions, pathways and models. This article defines a roadmap towards the annotation and curation of logical models, including milestones for best practices and minimum standard requirements.

Key words: biocuration; logical modelling; reproducibility; model reusability; annotation standards

Anna Niarakis is an Associate Professor at Univ Evry, University of Paris-Saclay. Her research focuses on the application of computational systems biology approaches in human diseases, including the construction of disease maps, tool development for model inference, network integration and dynamical modelling.
Martin Kuiper is a Professor in systems biology at the Department of Biology of the Norwegian University of Science and Technology. His group develops enabling technology for data analysis and knowledge management, among others for the construction of logical models for systems medicine.
Marek Ostaszewski is a Scientist and a Project Manager at the LCSB, working on knowledge management and visualisation in systems biomedicine, in particular in Parkinson's disease, including clinical research.
Rahuman S. Malik Sheriff is a Project Leader at the European Bioinformatics Institute (EMBL-EBI). He leads the development of BioModels infrastructure as well as curation and annotation of computational models in the standard formats. He is one of the members of the SBML Editorial Board. His interest includes building tools and resources for system biology modelling.
Submitted: 16 December 2019; **Received (in revised form):** 20 February 2020

Introduction

Reproducibility of research findings constitutes a key concern of the scientific community as multiple reports show that published results in various scientific domains cannot be replicated [1]. In the field of computational systems biology, where scientists combine prior knowledge based on experimental evidence and computational approaches, the reproducibility of results can be fostered through the use of consensual practices and standards, extensive annotation, code sharing, as well as by depositing of the resulting models in dedicated repositories. Logical (or logic) models (Boolean, multivalued, or other variants) have been widely used for studying and analysing in-depth regulatory mechanisms and biological processes for which kinetic data are scarce. Some repositories for this type of models exist already, including GINsim repository [2] and Cell Collective, a platform for building, analysing and visualising models [3, 4].

In the GINsim repository, one can find models built with the software GINsim and used for simulations in peer-reviewed articles. Models are stored in their zginml format, and a summary along with a link to the supporting scientific article is provided. In Cell Collective, models have been manually curated by reconstruction, re-simulation and analysis to ensure that their dynamics correspond to published results. Efforts are further made to include logical models in BioModels, a repository of mathematical models of biological and biomedical systems [5]. Annotation practices, accuracy and reproducibility checks made by the BioModels team will facilitate consistent quality control of these models.

To facilitate the exchanges of logical models and communication between tools, previous work by the CoLoMoTo consortium and Systems Biology Markup Language (SBML) teams was focused on the standardisation of model formats by developing a specific package of the SBML level 3 (SBML L3) [6], SBML-*qual* [7, 8].

However, specific minimum requirements for the annotation and level of curation of logical models remain to be defined. Even when results are reproducible, models often fail to be reusable because of the lack of explicit referencing to the sources used for their construction (organism, experimental setting and type of data, published manuscript sources, identifiers to relevant database entries, etc.).

To address the pressing need to propose and develop best practices and standards in the annotation and curation of logical models in biology, Anna Niarakis, Laurence Calzone and Tomáš Helikar (representatives of the CoLoMoTo [9] and SysMod [10] communities) organised a workshop in the context of the [BC]2 conference recently held in Basel [11], with the aim to bring together logical modellers and curators. The workshop, entitled 'Annotation and curation of computational models in biology' [12] is the most recent of a series of workshops organised by the logical modelling community over the past years, in the context of prominent international conferences such as ECCB 2014 (Strasbourg, France), ICSB 2015 (Singapore), ECMTB 2016 (Nottingham, UK), [BC]2 2017 (Basel, Switzerland) and ECCB 2018 (Athens, Greece).

The meeting was divided into four sessions, which highlighted the key challenges faced by the modelling community

Cristina Casals-Casas is a biocurator in the Swiss-Prot group at SIB. Her main responsibilities include the literature-based expert curation of mammalian proteins for UniProtKB and GO knowledge bases. She is also involved in research projects that require expert biocuration.
Denis Thieffry is currently a Professor of systems biology at the Ecole Normale Supérieure (ENS), Paris, France, and a part-time Invited Research Professor at the Cancer Science Institute of the National University of Singapore.
Tom C. Freeman is a Professor holding the Chair of Systems Immunology at the University of Edinburgh. He has a broad range of interests ranging from immune cell differentiation, transcriptomics, data visualisation to pathway modelling and simulation.
Paul Thomas is a Professor of preventive medicine and biological sciences at the University of Southern California. His research is in the area of computational analysis of genomic data, with an emphasis on gene function and evolution. In addition to founding and continuing development on the PANTHER project, Dr Thomas is a Director of the Gene Ontology Consortium.
Vasundra Touré is a PhD Student at the Norwegian University of Science and Technology. She is interested in the knowledge management and data extraction of molecular causal statements.
Vincent Noël is a post-doctoral fellow currently working in Institut Curie. His research interests concern modelling of biological systems, from the construction and the analysis of models to the optimisation of their simulations.
Gautier Stoll is a research engineer at INSERM U1138, in the group of G. Kroemer, a biology research group focused on metabolism, cancer & immunity. His research activity is focused on biological data analysis and modelling of heterogenous multicellular population.
Julio Saez-Rodriguez is a Professor of medical bioinformatics and data analysis at the Faculty of Medicine of Heidelberg University, the Director of the Institute for Computational Biomedicine, a Group Leader of the EMBL-Heidelberg University Molecular Medicine Partnership Unit and a Co-director of the DREAM challenges. His research focuses on computational methods to understand and treat the deregulation of cellular networks in disease.
Aurélien Naldi is a post-doctoral fellow currently working in the Ecole Normale Supérieure (ENS). His interests cover the integration of advanced computational methods into software tools for the analysis of complex biological models.
Eugenia Oshurko is a PhD Student at ENS Lyon. Her research interests include graph rewriting, graph-based databases, and knowledge representation and their application to the curation of biological knowledge relevant to building executable rule-based models.
Ioannis Xenarios is an Associate Professor of computational biology at the CHUV/UNIL/LICR. His research aims to model the tumour microenvironment of T cells infiltrates and monocytes. He develops both curated and data driven-based model to characterise and provide experimentally testable hypothesis using drugs and immunotherapies.
Sylvain Soliman is a researcher at Inria, Saclay. His research interests focus around computational biology and theoretical computer science. He has been one of the main developers and maintainers of the BIOCHAM modelling and analysis platform for more than 10 years.
Claudine Chaouiya is an Associate Professor at I2M, Aix-Marseille University. Her research focuses on developing appropriate methods for the analysis of large logical models of regulatory and signalling networks. Methodological advancements and tool development are motivated and further validated through modelling projects revolving around cell fate decisions, in single and multi-cellular contexts.
Tomáš Helikar is an Associate Professor in the Department of Biochemistry at the University of Nebraska-Lincoln. His research focuses on the use of integrative multi-approach modelling pipelines for dynamical analysis of biological networks. More precisely, his studies focus on understanding how aberrant changes in biological networks result in disease so that we could strategically develop more effective therapies.
Laurence Calzone is a researcher in Institut Curie. Her research activities revolve around building knowledge maps, constructing logical models and developing tools for improving and optimising the qualitative models and their predictions.

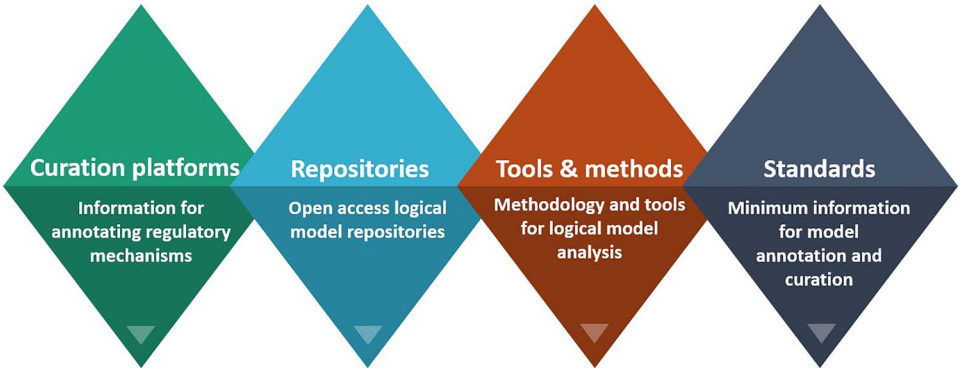

Figure 1. Four main thematic axes of the presentations and the round table discussion of the meeting. Biocuration platforms, available model repositories, tool development and integrative methodologies were the main subjects of the meeting. All presentations highlighted the need for standards in model annotation and curation.

(Figure 1), starting with curation platforms and model repositories. In particular, the need for establishing annotation criteria, quality control checks and the use of a common repository were extensively discussed. The following session was focused on recent methodological advancements to analyse logical models to ensure interoperability and reusability. Lastly, the afternoon sessions were focused on integrative approaches and tools. In Table 1, we have summarised briefly the topics discussed in each session. The presentations were followed by an extensive discussion between all speakers and participants on three key topics.

(i) **Reproducibility**, i.e. the ability to replicate scientific results using the same model.
(ii) **Reusability**, i.e. the ability to reuse an existing model using transparent biocuration processes, extensive annotations and references that increase model liability.
(iii) **Interoperability**, i.e. the ability to analyse the same model with multiple tools due to the use of standard formats.

Model curation and annotation, and available repositories

The first session was dedicated to annotation and curation approaches, together with relevant repositories, including the presentation of curation approaches and tools for the development of Boolean models for colon cancer and molecular causal interaction statements, the introduction to the complementary platforms BioKB [13] and MINERVA [14], followed by that of the BioModels repository. An example of an atherosclerotic plaque formation model demonstrated the necessity of proper annotation for optimal model-based predictions. The first session highlighted the necessity to annotate prior knowledge networks (PKNs) and logical models accurately for reusability and to enrich them with knowledge from heterogeneous resources to avoid potential ambiguities (e.g. UniProtKB [15], SIGNOR [16], HGNC [17], Gene Ontology (GO) [18] and REACTOME [19]).

Martin Kuiper (DrugLogics team, NTNU) presented work on a set of colon cancer logical models, named CASCADE (CAncer Signaling CAusality DatabasE [20]), and the development of a novel curation interface, named Visual Syntax Method (VSM [21]), which enables the curation of biological network information that includes causal molecular relationships. The VSM interface was tested extensively to annotate the full collection of experimentally analysed DNA binding transcription factors for human, mouse and rat [22] and is now being implemented in 'a curation platform for causal interaction statements' [23]. Causal interaction statements are basic representations of regulatory interactions between two biological entities that can be efficiently extracted from the literature, provided that proper annotation tools and curation guidelines are provided.

Marek Ostaszewski (Luxembourg Centre for Systems Biomedicine) presented 'BioKB and MINERVA: a workflow for curation and fast prototyping of annotated knowledge repositories' [13, 14]. To construct graphical models of molecular mechanisms, one needs to: (i) extract entities, interactions and relevant annotations from the literature; (ii) build a consistent graphical representation; and (iii) review and parameterise the model. BioKB [13, 24] is a platform initially designed for exploring text mining data, which currently allows combining human-provided and machine-identified elements and their interactions into 'facts'—human-curated relationships, annotated with sentences, literature and recognised identifiers. As BioKB is not a diagram editor, the biocurator can focus only on the accuracy of the extracted facts. This model's visualisation step, however, can be complemented with the MINERVA Platform, which allows API-driven [25] conversion of a layout-less model into an editable diagram (SBGN-ML [26]), that which can be further processed with various systems biology editors (e.g. CellDesigner [27], Newt [28], etc.). This enables realization of the final step of the model curation workflow: a curated diagram can be exported to the chosen systems biology format, refined and parameterised. Moreover, such API-based conversion makes it convenient to include in bigger bioinformatic workflows.

Following the effort towards the transparency of the different steps leading to model construction and the reusability of these models, Rahuman S. Malik-Sheriff [European Bioinformatics Institute (EMBL-EBI)] discussed how 'Curation and annotation of models in BioModels repository promote reproducibility and reusability'. Established in 2005, BioModels provides a platform to support sharing, easy accessibility and reproducibility of mathematical models of biological processes [5, 29]. Models submitted to BioModels are verified and curated by expert in-house curators. In 2011, an effort was made to extend SBML to the logical formalism and SBML-qual was defined [7, 8], allowing the inclusion of logical models in the database. Following Minimum Information Requested In the Annotation of Models (MIRIAM) guidelines, curated models are encoded in the standard SBML format and semantically enriched with controlled vocabularies [30]. Model entities are linked to several data resources (e.g. UniProt [15], Ensembl [31], the NCBI Taxonomy Database [32]) as well as ontologies, such as GO [18], ChEBI [33], Mathematical Modelling Ontology [34], Systems Biology

Table 1. Summary of different topics and presentations

Workshop sessions and chairs	Presentations and speakers
Model curation and annotation and available repositories **Chairs: Anna Niarakis and Denis Thieffry**	• **Martin Kuiper:** Towards a curation platform for causal interaction statements. • **Marek Ostaszewski:** BioKB and MINERVA: a workflow for curation and quick prototyping of annotated knowledge repositories • **Rahuman S. Malik Sheriff:** Curation and annotation of models in BioModels repository promotes reproducibility and reusability • **Cristina Casals:** SysVasc PKN: an example of biocuration for Boolean modelling
Community standards development and interoperability/reusability **Chairs: Marek Ostaszewski and Laurence Calzone**	• **Denis Thieffry:** Computational verification of large logical models: application to the prediction of T cell response to checkpoint inhibitors • **Tom Freeman:** A graphical and computational model of the renal mammalian circadian clock • **Paul Thomas:** Gene Ontology Causal Activity Modelling • **Anna Niarakis:** Automated inference of annotated Boolean models from molecular interaction maps using CaSQ
Tools (I) **Chair: Julio Saez Rodriguez**	• **Tomáš Helikar:** Cell Collective modelling platform • **Gaultier Stoll and Vincent Noel:** MaBoSS ecosystem • **Vasundra Touré:** The Minimum Information about a Molecular Interaction Causal Statement (MI2CAST): a guideline for the annotation of molecular causal interactions
Tools (II) **Chair: Tomáš Helikar**	• **Julio Saez Rodriguez**: CellNOpt • **Aurélien Naldi:** The CoLoMoTo Interactive Notebook: accessible and reproducible computational analyses for qualitative biological networks • **Eugenia Oshurko:** KAMIStudio

Ontology (SBO) [35] and Brenda Tissue Ontology [36]. Such annotations allow the unambiguous identification of model components and processes. BioModels currently hosts over 900 curated models, becoming the world's largest repository of curated models. BioModels team will soon start to systematically curate logical models. To date, however, only 17 logical models, 3 curated and 14 non-curated are included in the BioModels' collection.

Cristina Casals-Casas from (Swiss-Prot) presented 'SysVasc prior knowledge network (PKN): an example of biocuration for dynamical modelling'. As a case study, Cristina Casals-Casas and collaborators have built a PKN to allow dynamical modelling of atherosclerotic plaque formation [37]. The expert curation strategy was centred on regulatory interactions between biological entities (gene products, chemical compounds and processes) interacting with each other in a complex manner, and exhibiting conditional dependencies between co-regulators. Biological entities were defined using strictly controlled vocabulary terms, retrieved from UniProtKB, HGNC, ChEBI or GO, among others. The resulting PKN includes 729 components linked by 3406 interactions, of which 1841 are complex regulations encoded with logical operators, while 1565 are simply activatory or inhibitory interactions. For each component, they demonstrated how the description of complex signalling functions and their integration are essential to correctly predict health and disease states. Their work highlighted the essential role of expert curation to correctly identify and encode complex regulatory interactions from experimental literature. Failure to encode these relationships correctly can alter significantly the behaviour of the model and the derived predictions. Dynamical models should be fine-tuned by contextualisation to the specific biological system under study, and for this, proper annotation and expert curation are essential.

Community standards development and interoperability/reusability of existing models

The second session of the meeting was dedicated to the interoperability and reusability of models and provided examples using three different model applications. Novel dynamical analysis methods and a framework for GO annotations for supporting model building were also presented. All these approaches take advantage of existing databases to assist modellers and automatise error-prone and cumbersome tasks, currently still often performed manually, in order to optimise iterative modelling.

Denis Thieffry (Ecole Normale Supérieure, Paris) presented novel methods for the 'Computational verification of large logical models', with an 'application to the prediction of T cell response to checkpoint inhibitors'. A first approach enables the formalisation and automatic verification of validation criteria for whole models or defined subparts, thereby greatly facilitating model development and correction. A second approach consists in computing the impact of specific environmental or genetic perturbations on model dynamics by propagating their impact on model logical rules. These methods were applied to the analysis of the impact of T lymphocyte checkpoint inhibitors, and their use was integrated and illustrated in the CoLoMoTo Interactive Notebook [38] (presented by Aurelien Naldi in the afternoon session) to foster transparent and reproducible analyses.

Tom Freeman (Roslin Institute) presented a 'graphical and computational model of the renal mammalian circadian clock'. A comprehensive graphical model of the circadian pathway was constructed using the modified Edinburgh Pathway Notation scheme [39] and used to analyse the diurnal pattern of gene expression in the mouse kidney [40] using a stochastic Petri net-based approach [41]. The model encapsulates the interactions between 69 molecular species and contains 2013

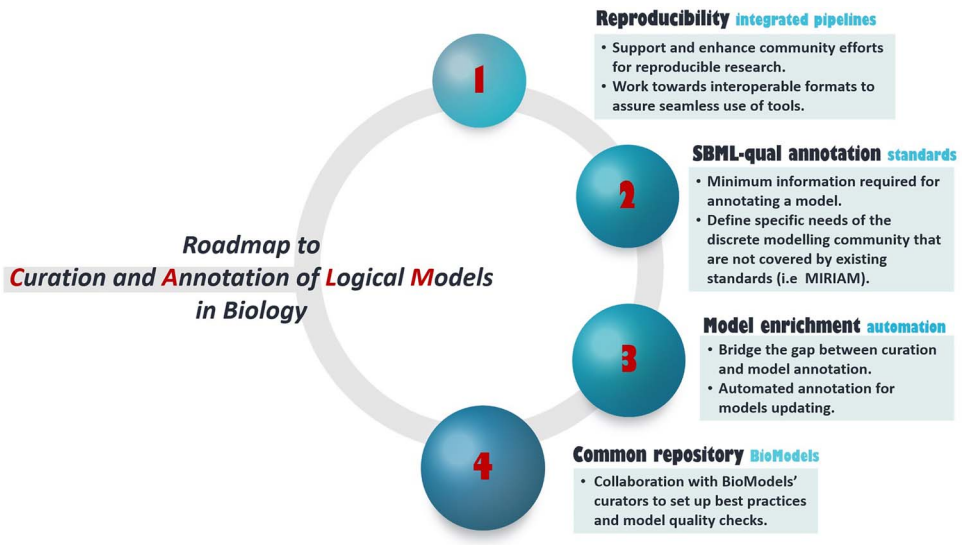

Figure 2. Roadmap to CALMs in Biology. Four milestones were identified as key steps in the roadmap to best practices for logical models annotation and curation: integrated pipelines for reproducible research, standards for SBML qual annotations, automation of model enrichment, and the use of a common repository.

components and 2100 interactions. All pathway components are labelled using standard nomenclature (HGNC gene id), and any modifications to those components are explicitly stated in their labels. Moreover, proteins, genes and biochemicals are hyperlinked to online resources, e.g. NCBI gene, ChEMBL, and interactions between components (process nodes) are annotated with publications providing supporting evidence. In this respect, models can also serve as descriptive diagrams of known events that can be easily evaluated and reused.

Reinforcing this idea, Paul Thomas introduced 'gene ontology activity modeling'. GO annotations are the most comprehensive structured representation of gene function and are widely used in the interpretation of genome-wide experimental data. However, an individual GO annotation associates a single gene product with a single GO term, which limits the expressiveness of annotations and their application in the computational analysis of experimental data. To address this limitation, Thomas *et al.* have developed a novel framework, GO Causal Activity Modelling (GO-CAM), for linking multiple GO annotations into an integrated model of a biological system. GO-CAM supports modelling at multiple levels, from individual gene products to complex regulatory and metabolic pathways, and can be applied in network analysis and systems biology modelling or converted into standard GO annotations for traditional GO-based analyses. Paul Thomas further presented the Noctua Modelling Tool used by GO Consortium curators to create GO-CAM models, from existing GO annotations or from scratch.

Finally, Anna Niarakis (Univ Evry, University of Paris-Saclay) closed the session by introducing the 'automated inference of annotated Boolean models from molecular interaction maps using CaSQ'. She proposed a methodology to convert complex molecular maps into computable logical models. Molecular interaction maps have emerged as a useful way of representing biological mechanisms, based on information mining and human curation [40]. Nevertheless, their static nature does not allow for *in silico* simulations. With Sylvain Soliman (INRIA, Paris-Saclay), they have developed CaSQ [42], a tool that infers preliminary Boolean rules based on the topology and semantics of the molecular interaction maps, transforming these maps into executable Boolean models. They used a state-of-the-art molecular interaction map for Rheumatoid Arthritis [43, 44] as a case study, but the tool can handle various maps differing in size and complexity, and supports the SBGN standard. CaSQ-inferred models are encoded in SBML *qual*, while references, annotations and layout are retained, thereby facilitating interoperability and model reusability.

Tools and modelling platforms for dynamical analysis of logical models

The afternoon sessions highlighted the efforts of the community to develop methodologies and software that address the issues of interoperability, reproducibility and reusability of modelling efforts. The level of annotation and the amount of curation are highly dependent on the modeller and on the capabilities of the existing tools to support this type of information in both human- and machine-readable formats.

Tomáš Helikar (University of Nebraska-Lincoln) introduced 'Cell Collective - Enabling accessible and collaborative construction and analysis of comprehensive and annotated models'. Cell Collective is a computational modelling platform for the collaborative construction, simulation and analyses of large-scale dynamic (logical) models of biological and biochemical processes [3, 4, 45]. It contains nearly 100 public, peer-reviewed logical models of various biological and biochemical processes. To ease the reuse and expansion of existing models, every component and interaction is annotated to track the biological data used to build the model. Models in Cell Collective can be created either *de novo* or imported using SBML-*qual*. Models are accessible in Cell Collective, where they can be simulated and further developed or can be downloaded in SBML *qual* format, including via its public API [46].

Gaultier Stoll (Centre de Recherche des Cordeliers, INSERM) and Vincent Noël (Institut Curie) presented 'MaBoSS (Markovian Boolean Stochastic Simulator) ecosystem'. MaBoSS is a tool for simulating logical models with continuous-time Markov processes [47]. Stochastic simulations allow the computation of the probabilities of each state of the model over time. Over the years, MaBoSS was extended [48] and various tools were developed, including UPMaBoSS, enabling the study of the dynamical

Table 2. Suggestion of minimum qualifiers for the annotation of logical models. The newly proposed hasState qualifier could be added to account for a node's state (qualitative levels)

Model annotation levels	Minimum qualifiers	Examples of knowledge sources stored in RDFs
Model	**Model qualifiers: bqmodel** **is, identity** This qualifier might be used to link an encoded model to a database of models. **isDescribedBy, description** This relation might be used to link a model to the literature that describes it. **hasTaxon, taxon** This qualifier might be used to indicate taxonomy/organism (i.e. human, plant, animal). **isVersionOf, version** This qualifier can be used to link a model to the gene ontology terms regarding the biological function described. **hasProperty, property** This relation could be used to indicate mathematical formalism. **isDerivedFrom, origin** This relation may be used to express a refinement or adaptation in usage for a previously described model	PMID, BioModels ID, doi, CC ID, GINsim ID, GO
Qualitative species	**Biology qualifiers: bqbiol** **is, identity** This relation might be used to link a biological entity to its exact counterpart in a database. **isDescribedBy, description** This relation should be used to link a species to the literature that describes the role of that species or its presence in the system of interest. **hasVersion, version** This relation may be used to represent an isoform or modified form of a biological entity. **hasState, state** This relation could be used to describe the state of a biological entity.	GO, UniProt, HGCN, PMID
Causal interactions/-transitions	**Biology qualifiers: bqbiol** **hasProperty, property** This relation might be used when a biological entity exhibits a certain enzymatic activity or exerts a specific function. **isDescribedBy, description** This relation should be used, for instance, to link a reaction to the literature that describes it.	KEGG, REACTOME, PMID

behaviour of cell populations (including its size), and PhysiBoSS, based on an agent-based formalism where each agent is a logical model run with MaBoSS. A model of cell fate decision was used to showcase different ways of running the tools: through the command line, through the CoLoMoTo Jupyter interactive notebook, showing the interoperability of the tool and using the python library 'pymaboss' [49].

Vasundra Touré (DrugLogics group, NTNU) presented 'The Minimum Information about a Molecular Interaction Causal Statement (MI2CAST): a set of guidelines for the annotation of molecular causal interactions' [23]. The NTNU group proposes MI2CAST as a standard for representing causal statements that can serve as a checklist that can be followed in curation processes for capturing the essential contextual information about a causal relationship, to ensure clarity, uniformity and reusability of the data across resources. MI2CAST has been developed in collaboration with the International Molecular Exchange (IMEx) consortium [50] and Human Proteome Organization–Proteomics Standards Initiative (HUPO-PSI) [51]. The NTNU group has also implemented the MI2CAST guidelines and annotation terms in a prototype curation tool based on the VSM foundation [21], named causalBuilder [52].

Julio Saez-Rodriguez (Heidelberg University) focused on 'Integrating knowledge and experimental data to build context-specific logic models'. The general pipeline involves obtaining existing prior knowledge on pathways from available public resources using OmniPath [52], building a logic model from this prior knowledge and training it to data with tools such as CellNOpt (for targeted readouts [53]), PHONEMeS (for untargeted mass spectrometry [54]) and CARNIVAL (for gene expression [54]). Regarding annotations, OmniPath provides information about localisation, function, disease relationships, proteins and complexes based on 36 resources. Collectively, Omnipath provides 2 200 000 annotation entries for 20 000

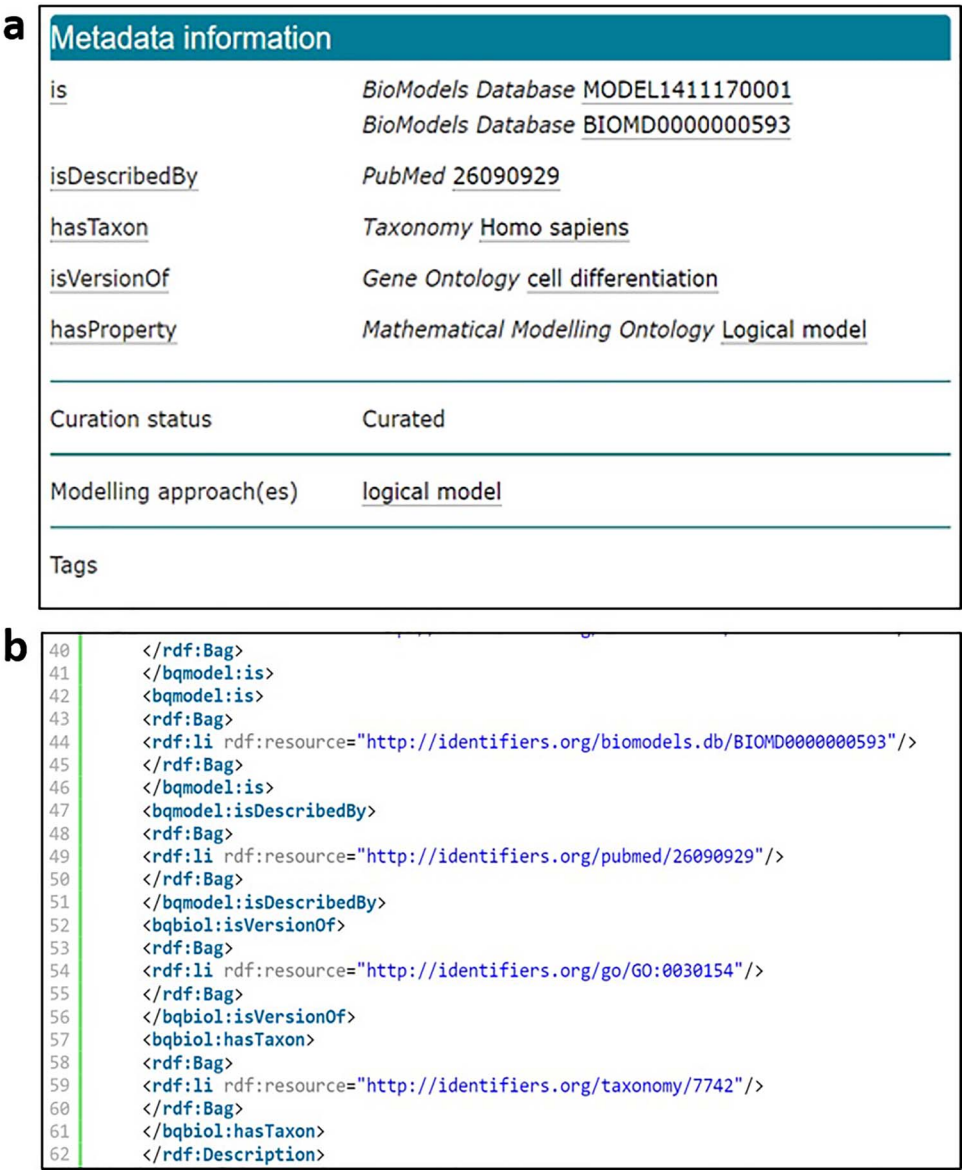

Figure 3. A logical model in Biomodels database. Metadata information for the curated logical model in BioModels database (**a**) and the corresponding block code (**b**).

human proteins and 16 500 complexes and is available via a Python module, an R package, as a web service, or from Cytoscape [55, 56].

Aurelien Naldi (Ecole Normale Supérieure, Paris) presented 'The CoLoMoTo Interactive Notebook', which provides a unified environment to edit, execute, share and reproduce analyses of Boolean and multi-valued models of biological networks. This framework combines the power of different software tools to ensure reproducibility and to reduce their learning curve. The CoLoMoTo Interactive Notebook currently eases access to GINsim, BioLQM [57], Pint [58], MaBoSS, and Cell Collective. More tools will be included in the future. Computational workflows can be edited through a web interface based on the Jupyter notebook, enabling the inclusion of textual annotations, along with the explicit code to execute, as well as the visualisation of the results. The framework is distributed as a Docker image with the tools ready to use without any installation step besides Docker, which can run on Linux, macOS and Microsoft Windows systems.

Lastly, Eugenia Oshurko (Ecole Normale Supérieure, Lyon) presented 'KAMIStudio: an environment for biocuration of cellular signalling knowledge' [59] suitable for rule-based modelling languages, such as Kappa [60] and BioNetGet [61]. The KAMIStudio environment is based on the KAMI biocuration framework that aims to decouple knowledge curation from model building [62]. The KAMIStudio environment can be used for semi-automatic curation of large corpora of cellular signalling knowledge and for automatic generation of dynamical models.

Round table discussion

The general discussion highlighted four important aspects, namely (i) the need to provide annotated models that would include textual annotations, bibliographic references and crosslinks to knowledge resources through the use of common identifiers, (ii) the importance of creating interfaces for automatic integration of annotations by leveraging the wealth of curated interactions in dedicated databases, (iii) the utility

of agreeing on best practices, use of standards and on the minimum information required to ensure model reproducibility and reusability, and lastly (iv) the use of common repositories for logical models that would foster interactions and facilitate exchanges between scientists interested in reusing models. The need to encourage researchers to submit systematically novel publications that include logical models to one of the model repositories was also discussed, as this would increase visibility, ease reproducibility and promote reusability of logical models.

Roadmap to best practices for the curation and annotation of logical models (CALMs) in biology

Based on these discussions, four interdependent milestones were identified for the roadmap to CALMs in biology (Figure 2).

(i) The first milestone concerns the reproducibility of the analyses of discrete models. The use of common, standardised formats (e.g. SBML packages *qual*, *layout*, *render*, etc.) would greatly facilitate the interoperability between different tools and the development of integrative pipelines. For example, the CoLoMoTo notebook could be expanded to include more tools, offering a flexible way of performing dynamical analyses in a fully transparent and reproducible manner. To achieve this goal, the logical modelling community aims to work closely with the communities developing standards, such as SBML, the Simulation Experiment Description Markup Language (SED-ML) and Computational Modelling in Biology Network (COMBINE) to contribute to community efforts and make sure that the standards developed are in line with the specificities of the logical formalism.

(ii) The second milestone concerns the minimum information for annotating a model, and also new mechanisms to encode such information in SBML-*qual*. The information should be stored in human- and machine-readable form, for example, by using Resource Description Framework (RDF) tags [63]. SBML format also provides the possibility to associate SBO terms outside of RDFs; however, unified storage of all model annotations in RDF could provide a simple, yet an efficient standard way of annotating logical

Use case of a curated and annotated logic model (BIOMD0000000593) in BioModels.

- Cross-references to well-established ontologies like GO, UNIPROT, SBO etc. are added to NODES (SPECIES) and CAUSAL INTERACTIONS (TRANSITIONS) using RDF.
- Use of COMBINE qualifiers http://co.mbine.org/standards/qualifiers where possible.

Annotations are added at two levels:

1. **Model level annotation**

```
<bqmodel:isDerivedFrom>
    <rdf:Bag>
        <rdf:li rdf:resource="http://identifiers.org/pubmed/12871957"/>
        <rdf:li rdf:resource="http://identifiers.org/pubmed/16314431"/>
    </rdf:Bag>
</bqmodel:isDerivedFrom>
```

2. **Model component level annotation:**
   a. **NODES (SPECIES)**

```
<qual:qualitativeSpecies metaid="species11" qual:compartment="default"
qual:constant="false" qual:id="IL21" qual:maxLevel="1" qual:name="IL21">
    <annotation>
        <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
            <rdf:Description rdf:about="#species11">
                <bqbiol:is>
                    <rdf:Bag>
                        <rdf:li rdf:resource="http://identifiers.org/uniprot/Q9HBE4" />
                    </rdf:Bag>
                </bqbiol:is>
            </rdf:Description>
        </rdf:RDF>
    </annotation>
</qual:qualitativeSpecies>
```

Box 1. An example of annotating a logical model using RDFs. BioModels propose a two-level annotation, model and model component. Model components are in turn annotated in two levels: nodes (species) and causal interactions (transitions). A colour code is used to highlight the different code blocks that refer to the each level of annotation. Code blocks are excerpts from a syntactically valid SBML *qual* file.

2. Model component level annotation:

b. CAUSAL INTERACTIONS (TRANSITIONS)

```
<qual:transition metaid="rxn01" qual:id="tr_TBET" qual:name="Interactions targeting TBET">
    <qual:listOfInputs>
        <qual:input metaid="rxn01_input01" qual:qualitativeSpecies="IL4"
        qual:transitionEffect="none"/>
    </qual:listOfInputs>
    <qual:listOfOutputs>
        <qual:output metaid="rxn01_output01" qual:qualitativeSpecies="TBET"
        qual:transitionEffect="assignmentLevel"/>
    </qual:listOfOutputs>
    <qual:listOfFunctionTerms>
        <qual:functionTerm metaid="rxn01_function01" qual:resultLevel="1">
            <math xmlns="http://www.w3.org/1998/Math/MathML">
                <apply>
                    <not/>
                    <apply>
                        <or
                            />
                        <apply>
                            <eq/>
                            <ci> IL4 </ci>
                            <cn type="integer"> 1 </cn>
                        </apply>
                    </apply>
                </apply>
            </math>

            <annotation>
                <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                xmlns:dc="http://purl.org/dc/elements/1.1/"
                xmlns:dcterms="http://purl.org/dc/terms/"
                xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
                    <rdf:Description rdf:about="#_rxnannotation1">
                        <bqbiol:hasProperty>
                            <rdf:Bag>
                                <rdf:li
                                rdf:resource="http://identifiers.org/SBO/SBO:0000169"/>
                            </rdf:Bag>
                        </bqbiol:hasProperty>
                    </rdf:Description>
                </rdf:RDF>
            </annotation>
        </qual:functionTerm>
        <qual:defaultTerm metaid="_defaultlevel" qual:resultLevel="0">
        </qual:defaultTerm>
    </qual:listOfFunctionTerms>
</qual:transition>
```

Box 1. Continue

models. Supported by larger computational modelling communities (e.g. COMBINE), RDF is considered as the *de facto* standard for encoding annotations [64]. The community should discuss and agree on the best way of integrating annotations in SBML-*qual*, notably which tags and which SBML elements to use, while also leveraging the experience of the SBML community and BioModels curation practices. Notably, the SBML specification documents [7] already propose a systematic way of annotation that can be adapted to logical models. Additionally, the logical modelling community should define specific needs that are not covered yet by existing standards (i.e. MIRIAM identifiers and BioModels.net qualifiers [65]) and propose feasible solutions. The minimum information for annotation could be proposed as a prerequisite for publishing a logical model in peer-reviewed journals. Table 2 lists suggested minimum qualifiers that could be used in order to annotate a model, in line with MIRIAM and BioModels suggestions. Furthermore, to aid model developers and curators, new tools need to be developed to to enrich of models with as many relevant and useful annotations as possible. The metadata information for one of the three curated logical models currently available in BioModels and the corresponding code block of the XML file is exemplified in Figure 3. While the logical modelling community has made progress towards identifying the important aspects of annotations, much work remains to be done. For example, the community is currently discussing the appropriate 'depth' of annotations for each logical function. For example, does each variable and operator between variables in a logical function need to

be annotated (such as currently available in Cell Collective)? While this level of annotations can add to the workload of the modeller/curator, one might argue that providing citable experimental evidence with of such aspects for the regulatory mechanism of each component will only increase the transparency of the model. The *qual*: transition component in the SBML model could be proposed for the annotation of causal interaction; however, this choice (already employed by some tools, i.e. CaSQ, Cell Collective) raises issues concerning the cases where a more precise annotation would be needed.

(iii) The third milestone refers to the collaboration between modellers and curators to bridge the gap between storing information and reusing information. Automated procedures for model annotation and enrichment could further help to maintain models up to date. Keeping track of literature information used to derive logical formulae can further foster model accuracy and enhance reusability. To make steps forward, the logical modelling community aims to work closely with biocurators and knowledge platform developers to identify best practices. An obvious way would be to agree on the use of common and well-established identifiers, such as UniProt, GO, HGCN and SBO, which would allow unambiguous identification of a model component with simultaneous access to the knowledge resource through crosslinks. This direct linking of model annotations to curated knowledge sources via standard identifiers could help significantly in establishing quality control checks regarding annotation and biological content.

(iv) The last milestone concerns fully leveraging available model repositories. Several logical model repositories exist, including Cell Collective, GINsim and PyBoolNet [66]. The Cell Collective provides models in several formats, including SBML *qual*. The GINsim model repository provides models in the GINML format, which can be converted to SBML *qual* (and other formats) using GINsim and BioLQM. Simultaneously, BioModels is one of the largest repositories of mathematical, SBML-encoded models. However, it has been traditionally focused on models described with other mathematical frameworks and lacks processes to curate logical models. Indeed, the logical modelling community has started to work closely with BioModels team to set up best practices and model quality checks that will be applicable to logical models. The aim is to create a dedicated collection of logical models within BioModels, which would provide an additional resource with curated logical models. In Box 1, we show a curated logical model stored in BioModels (BIOMD0000000593) annotated as a sample case.

The logical modelling community should also decide if the suggestions of the COMBINE community, as stated in Neal *et al.* [64], regarding the storage of annotations in a separate file could be adopted. While this would allow for more flexibility in terms of knowledge resources' choices for model annotation, i.e. one model file with several annotation files with different sources, it would add the extra burden of file synchronisation. However, dissociating model from model annotation could be in line with the approaches and methodologies presented in the first session of the meeting regarding the separation of the biocuration from the model layout and refinement. An additional point to consider is the simulation settings and their specifications through an established standard such as SED-ML [67, 68], which will likely require some adaptation to suit logical model simulations. In this respect, the COMBINE Archive format could offer a possible solution, as it provides a standardised way to bundle different files together [64].

Outcomes and outlook

The [BC]2 workshop on annotation and curation of logical models in biology brought together people from different communities, such as biocurators, modellers, methodology and software developers. The round table discussion clarified common objectives together with milestones on the roadmap to best practices. Presentations and discussions highlighted efforts and resources that can be used for enhancing reproducibility and model contextualisation. The authors have started to form working groups and will continue to foster communication and exchanges, first among the logical modelling community, but also by reaching out to other communities with similar interests, to attain these collective goals.

The complete list of abstracts can be found in the Supplementary Data *Abstract_Booklet*.

Key Points

- The identified milestones will help the community of logical modelling to coordinate efforts for reproducible research.
- Standards for minimum curation will help unify formats and annotations, in an effort to provide models of better accuracy and quality.
- Transparency in curation and standardised annotations will enhance model reusability.
- Format harmonisation will facilitate interoperability and integration of existing tools in seamless pipelines.
- Collaboration between modellers and curators will foster model enrichment and updating, taking advantage of the wealth of information stored in databases and knowledge bases.
- The use of a common repository will reinforce quality protocols and checks for models, which could even be used prior to publication.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

The authors would like to thank the SysMod community for its support, the [BC]2 organisers for supporting and hosting the meeting/workshop in Basel, and Krishna Kumar Tiwari, Scientific Database Curator in BioModels, for his help in producing the curated example depicted in Box 1.

Funding

TH is supported by the NIH 1R35GM119770-04 grant. PDT is supported by the NIH/NHGRI HG002273 grant. RSMS is supported by the BBSRC MultiMod BB/N019482/1 grant.

References

1. Allison DB, Shiffrin RM, Stodden V. Reproducibility of research: issues and proposed remedies. *Proc Natl Acad Sci* 2018;**115**:2561–2.
2. Naldi A, Hernandez C, Abou-Jaoudé W, *et al*. Logical modeling and analysis of cellular regulatory networks with GINsim 3.0. *Front Physiol* 2018;**9**:646.
3. Helikar T, Kowal B, McClenathan S, *et al*. The cell collective: toward an open and collaborative approach to systems biology. *BMC Syst Biol* 2012;**6**:96.
4. Helikar T, Kowal B, Rogers JA. A cell simulator platform: the cell collective. *Clin Pharmacol Ther* 2013;**93**:393–5.
5. Malik-Sheriff RS, Glont M, Nguyen TVN, *et al*. BioModels—15 years of sharing computational models in life science. *Nucleic Acids Res* 2020;**48**(D1):D407–D15. doi: 10.1093/nar/gkz1055.
6. Hucka M, Finney A, Sauro HM, *et al*. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;**19**:524–31.
7. Chaouiya C, Keating SM, Berenguier D, *et al*. SBML level 3 package: qualitative models, version 1, release 1. *J Integr Bioinform* 2015;**12**:2.
8. Chaouiya C, Bérenguier D, Keating SM, *et al*. SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Syst Biol* 2013;**7**:135.
9. Naldi A, Monteiro PT, Müssel C, *et al*. Cooperative development of logical modelling standards and tools with CoLoMoTo. *Bioinformatics* 2015;**31**(7):1154–9.
10. SysMod. *SysMod Community of Special Interest*. http://sysmod.info
11. BASEL LIFE 2019, 9–12 September 2019, Congress Center Basel, Switzerland. https://www.basellife.org/2019.html
12. BaselLife 2019 *Workshops and tutorials*. https://www.baselllife.org/2019/basel-life-structure/bc2/programme/workshops-and-tutorials.html#c68317
13. BioKB. https://biokb.lcsb.uni.lu/
14. Gawron P, Ostaszewski M, Satagopam V, *et al*. MINERVA-a platform for visualization and curation of molecular interaction networks. *NPJ Syst Biol Appl* 2016;**2**:16020.
15. Bateman A, Martin MJ, O'Donovan C, *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158–69.
16. Perfetto L, Briganti L, Calderone A, *et al*. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res* 2016;**44**:D548–54.
17. Braschi B, Denny P, Gray K, *et al*. Genenames.Org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res* 2019;**47**:D786–92.
18. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;**47**:D330–8.
19. Fabregat A, Jupe S, Matthews L, *et al*. The Reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55.
20. The Druglogics Initiative. https://www.druglogics.eu/
21. vsmjs. https://github.com/vsmjs. GitHub 2019
22. SciCura. http://demo.scicura.org/
23. MI2CAST. https://github.com/MI2CAST
24. Biryukov M, Groues V, Satagopam V. *BioKB - Text Mining and Semantic Technologies for the Biomedical Content Discovery*. University of Luxembourg, 2019. https://pdfs.semanticscholar.org/a286/dabfaa6356ac0634a9ace69e0d6ac289dfb4.pdf
25. Hoksza D, Gawron P, Ostaszewski M, *et al*. MINERVA API and plugins: opening molecular network analysis and visualization to the community. *Bioinformatics* 2019;**35**:4496–8.
26. Hoksza D, Gawron P, Ostaszewski M, *et al*. Closing the gap between formats for storing layout information in systems biology. *Brief Bioinform* 2020;**pii**:bbaa030. https://doi.org/10.1093/bib/bbz067.
27. Kitano H, Funahashi A, Matsuoka Y, *et al*. Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol* 2005;**23**:961–6.
28. Newt Editor. http://newteditor.org/index.html
29. Glont M, Nguyen TVN, Graesslin M, *et al*. BioModels: expanding horizons to include more modelling approaches and formats. *Nucleic Acids Res* 2018;**46**:D1248–53.
30. Le Novère N, Finney A, Hucka M, *et al*. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* 2005;**23**:1509–15.
31. Ensembl 2020. *Nucleic Acids Res* 2020;**48**(D1):D682–D88. https://doi.org/10.1093/nar/gkz966.
32. Federhen S. The NCBI taxonomy database. *Nucleic Acids Res* 2012;**40**:D136–43.
33. Degtyarenko K, de Matos P, Ennis M, *et al*. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;**36**:D344–50.
34. Mathematical Modelling Ontology. https://www.ebi.ac.uk/ols/ontologies/mamo
35. Systems Biology Ontology. http://www.ebi.ac.uk/sbo/main/
36. BRENDA tissue/enzyme source. https://www.ebi.ac.uk/ols/ontologies/bto
37. Bekkar A, Estreicher A, Niknejad A, *et al*. Expert curation for building network-based dynamical models: a case study on atherosclerotic plaque formation. *Database (Oxford)* 2018;2018;**2018**.
38. Naldi A, Hernandez C, Levy N, *et al*. The CoLoMoTo interactive notebook: accessible and reproducible computational analyses for qualitative biological networks. *Front Physiol* 2018;**9**:680. Published 2018 Jun 19. doi: 10.3389/fphys.2018.00680.
39. Freeman TC, Raza S, Theocharidis A, *et al*. The mEPN scheme: an intuitive and flexible graphical system for rendering biological pathways. *BMC Syst Biol* 2010;**4**:65.
40. Ivy Jessica R., Shih Barbara, Hogenesch John B., Mullins John J., Freeman Tom C. A detailed graphical and computational model of the mammalian renal circadian clock. *bioRxiv* 2019;795906. doi: https://doi.org/10.1101/795906.
41. Livigni A, O'Hara L, Polak ME, *et al*. Petri net-based graphical and computational modelling of biological systems. *bioRxiv* 2016;047043.
42. Aghamiri SS, Singh V, Naldi A, *et al*. Automated inference of Boolean models from molecular interaction maps using CaSQ. https://gitlab.inria.fr/soliman/casq
43. Singh V, Ostaszewski M, Kalliolias GD, *et al*. Computational systems biology approach for the study of rheumatoid arthritis: from a molecular map to a dynamical model. *Genomics Comput Biol* 2018;**4**:1.
44. Singh V, Kalliolias GD, Ostaszewski M, *et al*. RA-map: building a state-of-the-art interactive knowledge base for rheumatoid arthritis. *Database* 2020;**2020**:1–18. doi: 10.1093/database/baaa017.
45. Helikar T, Cutucache CE, Dahlquist LM, *et al*. Integrating interactive computational modeling in biology curricula. *PLoS Comput Biol* 2015;**11**:e1004131.

46. Kowal BM, Schreier TR, Dauer JT, *et al*. Programmatic access to logical models in the Cell Collective modeling environment via a REST API. *Biosystems* 2016;**139**: 12–6.
47. Stoll G, Caron B, Viara E, *et al*. MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics* 2017;**33**: 2226–8.
48. Letort G, Montagud A, Stoll G, *et al*. PhysiBoSS: a multi-scale agent-based modelling framework integrating physical dimension and cell signalling. *Bioinformatics* 2019; **35**:1188–96.
49. pyMaBoSS — pyMaBoSS 0.3.2 documentation. https://pymaboss.readthedocs.io/en/latest/
50. Orchard S, Kerrien S, Abbani S, *et al*. Protein interaction data curation: the international molecular exchange (IMEx) consortium. *Nat Methods* 2012;**9**:345–50.
51. Taylor CF, Hermjakob H, Julian RK, *et al*. The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). *Omics J Integr Biol* 2006;**10**:145–51.
52. causalBuilder. https://mi2cast.github.io/causalBuilder/.
53. Terfve CDA, Wilkes EH, Casado P, *et al*. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat Commun* 2015;**6**:1–11.
54. Liu A, Trairatphisan P, Gjerga E, *et al*. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *Npj Syst Biol Appl* 2019;**5**:1–10.
55. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* 2016;**13**:966–7.
56. Ceccarelli F, Turei D, Gabor A, *et al*. Bringing data from curated pathway resources to Cytoscape with OmniPath. *Bioinformatics* 2019;btz968.
57. Naldi A. BioLQM: a java toolkit for the manipulation and conversion of logical qualitative models of biological networks. *Front Physiol* 2018;**9**:1605. Published 2018 Nov 19. doi: 10.3389/fphys.2018.01605.
58. Paulevé L. *Pint: A Static Analyzer for Transient Dynamics of Qualitative Networks with IPython Interface*. 2017. https://hal.archives-ouvertes.fr/hal-01589248
59. Harmer R, Oshurko E. KAMIStudio: An Environment for Biocuration of Cellular Signalling Knowledge. Computational Methods in Systems Biology. CMSB 2019. In: *Lecture Notes in Computer Science* **11773**: pp 322–28. Springer, Cham.
60. Boutillier P, Maasha M, Li X, *et al*. The kappa platform for rule-based modeling. *Bioinformatics* 2018;**34**:i583–92.
61. Faeder JR, Blinov ML, Hlavacek WS. Rule-based modeling of biochemical systems with BioNetGen. *Methods Mol Biol* 2009;**500**:113–67. doi: 10.1007/978-1-59745-525-1_5.
62. Kappa-Dev/KAMI. https://github.com/Kappa-Dev/KAMI
63. Novère NL, Finney A. A simple scheme for annotating SBML with references to controlled vocabularies and database entries. 2005;**13**. https://scholar.google.com/scholar?q=Le+Nov%C3%A8re+N.+Finney+A.++(2005+)++A+simple+scheme+for+annotating+SBML+with+references+to+controlled+vocabularies+and+database+entries+. ++http://www.ebi.ac.uk/compneur-srv/sbml/proposals
64. Neal ML, König M, Nickerson D, *et al*. Harmonizing semantic annotations for computational models in biology. *Brief Bioinform* 2019;**20**:540–50.
65. BioModels.net Qualifiers | COMBINE. http://co.mbine.org/specifications/qualifiers
66. Klarner H, Streck A, Siebert H. PyBoolNet: a python package for the generation, analysis and visualization of boolean networks. *Bioinforma Oxf Engl* 2017;**33**:770–2.
67. Bergmann FT, Cooper J, König M, *et al*. Simulation experiment description markup language (SED-ML) level 1 version 3 (L1V3). *J Integr Bioinform* 2018;**15**(1):20170086. Published 2018 Mar 19. doi: 10.1515/jib-2017-0086.
68. Waltemath D, Adams R, Bergmann FT, *et al*. Reproducible computational biology experiments with SED-ML - the simulation experiment description markup language. *BMC Syst Biol* 2011;**5**:198.