Standard Formats in Computational Modeling

Matthias König

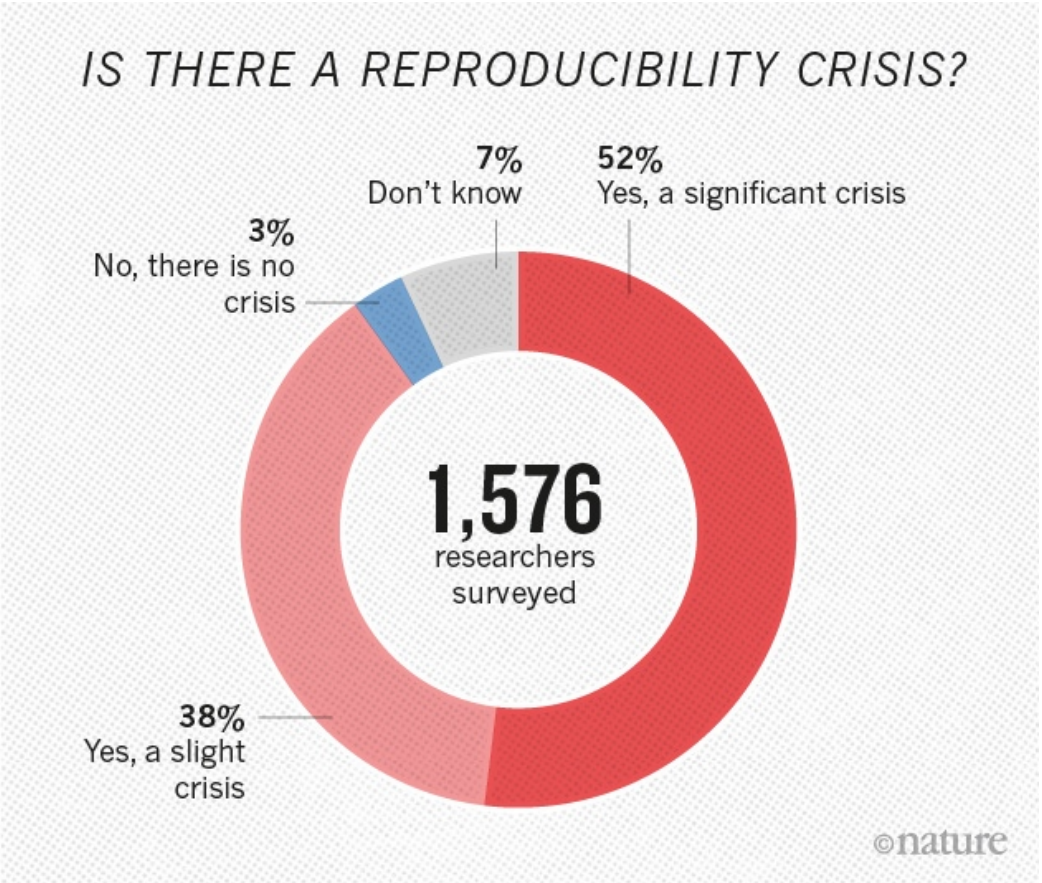
SCIENTIFIC MEDIA

PUBLIC MEDIA

The science 'reproducibility crisis' – and what can be done about it

March 15, 2017 by Ottoline Leyser, Danny Kingsley And Jim Grange, The Conversation

Reproducibility crisis



Baker, Nature May 2016, Vol 533, 453

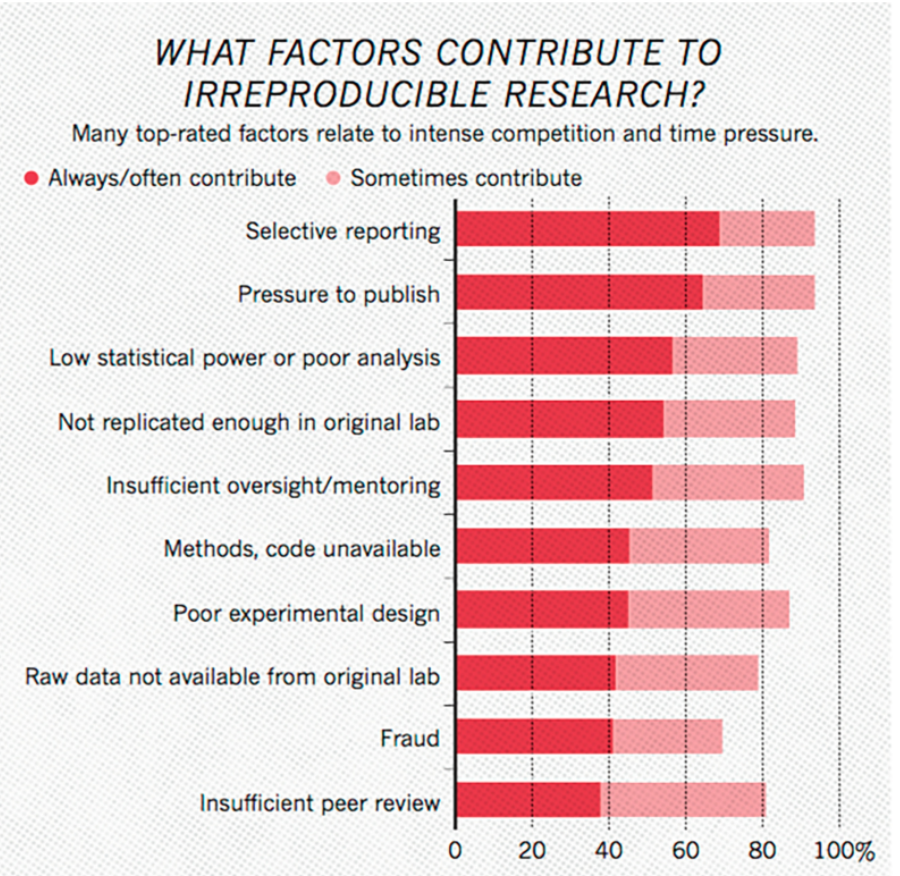
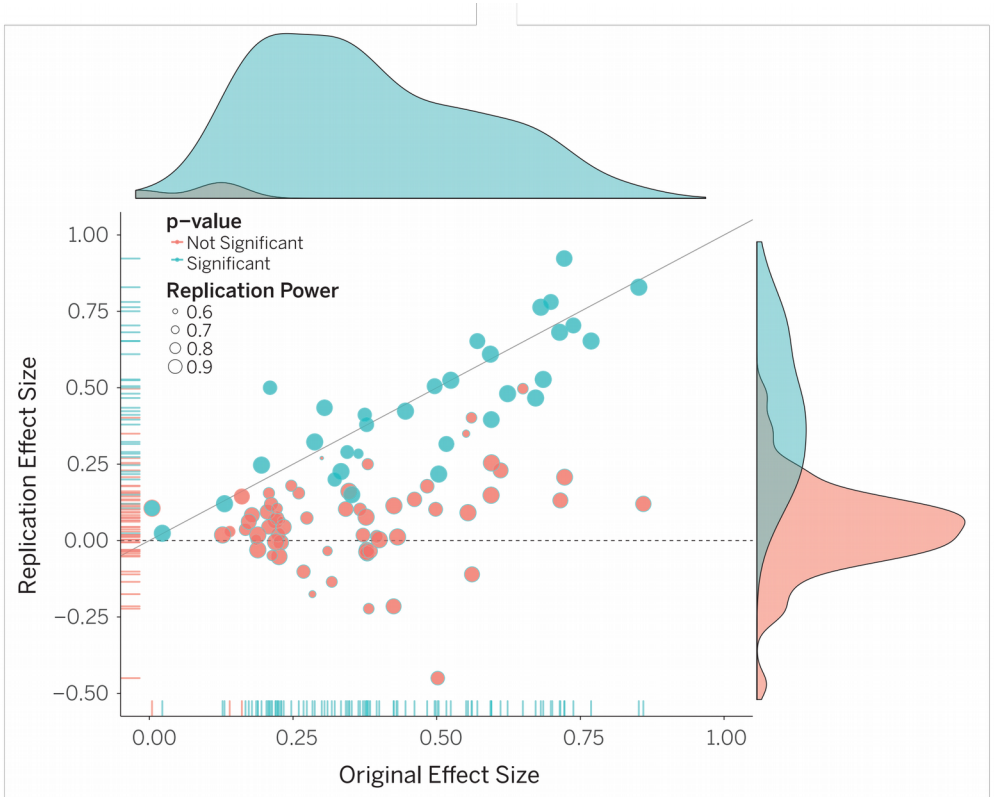


FIGURE 5: Plot Representing What Researchers Believe are the Primary Causes of Reproducibility Failure in the Sciences.

Reproducibility efforts



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

- Replication studies of 100 experimental and correlation studies (psychology studies)
- **97%** original statistically significant results, **37%** of replications
- Replication **effects half the magnitude of original**

Open Science Collaboration, Science, August 2015, Vol 349, Issue 6251

Publications are advertisement

“An article about (computational) science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship.

The actual scholarship is the complete ... set of instructions and data which generated the figures.”

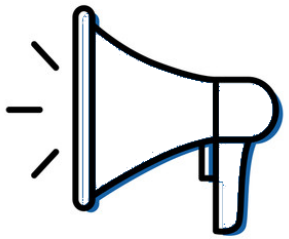
David Donoho, 1998

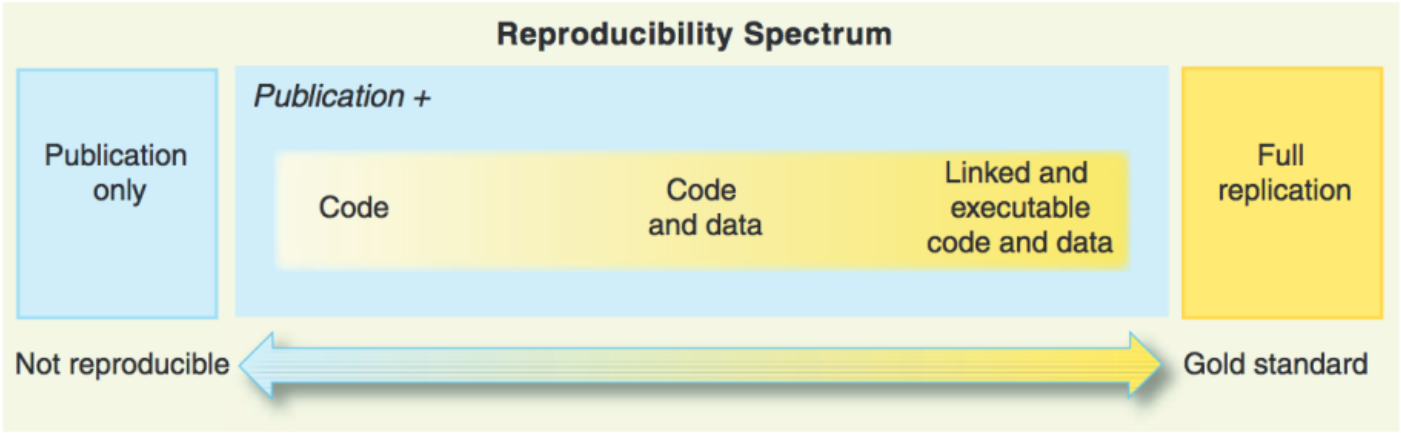
Publications are advertisement

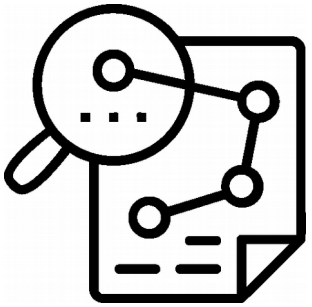
"An article about (computational) science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship.

The actual scholarship is the complete ... set of instructions and data which generated the figures."

David Donoho, 1998







Standard Formats

- Exchangability, reusability, interoperability
  - Encoding information in computer readable format
- Annotations to ontologies
  - Knowledge integration (biological, computational)
  - Documentation (what is my model component)
- Reproducibility
  - Identical results in multiple tools (roadrunner, COPASI, JWS)
- Model quality
  - Minimal Information for models and simulation (MIRIAM, MIASE)
  - Automatic validation (unit checking, model consistency)
- Large ecosystems of tools
  - Simulation, parameter fitting, model analysis, visualization, …

$$\dot{S}_1 = J_0 - v_1$$

$$\dot{S}_2 = v_1 - v_2$$

$$\dot{S}_3 = 2v_2 - v_3 - v_8$$

$$\dot{S}_4 = v_3 - v_4$$

$$\dot{S}_5 = v_4 - v_5$$

$$\dot{S}_6 = v_5 - v_6 - J$$

$$\dot{S}_6^{ex} = \varphi J - v_9$$

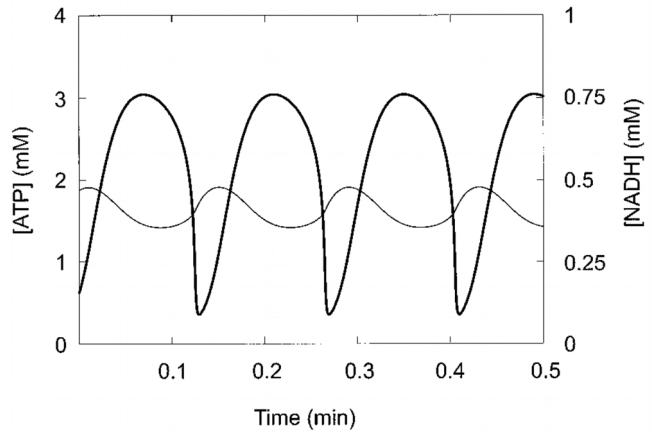
$$\dot{A}_3 = -2v_1 + v_3 + v_4 - v_7$$

$$\dot{N}_2 = v_3 - v_6 - v_8$$

TABLE 1 Parameter Values

Parameter	Value
J_0	50.0 mM · min^{-1}
k_1	550.0 mM^{-1} · min^{-1}
K_i	1.0 mM
k_2	9.8 min^{-1}
k_{GAPDH+}	323.8 mM^{-1} · min^{-1}
k_{GAPDH-}	57823.1 mM^{-1} · min^{-1}
k_{PGK+}	76411.1 mM^{-1} · min^{-1}
k_{PGK-}	23.7 mM^{-1} · min^{-1}
k_4	80.0 mM^{-1} · min^{-1}
k_5	9.7 min^{-1}
k_6	2000.0 mM^{-1} · min^{-1}
k_7	28.0 min^{-1}
k_8	85.7 mM^{-1} · min^{-1}
κ	375.0 min^{-1}
φ	0.1
A	4.0 mM
N	1.0 mM
n	4

$$v_3 = \frac{k_{GAPDH+}k_{PGK+}S_3N_1(A-A_3) - k_{GAPDH-}k_{PGK-}S_4A_3N_2}{k_{GAPDH-}N_2 + k_{PGK+}(A - A_3)}$$



combine the computational modeling in biology network

- initiative to coordinate the development of the various community standards and formats for computational models
- COMBINE meeting & HARMONY hackathon
- Core standards
  http://co.mbine.org/Standards#Core_COMBINE_standards
  - **SBML** Systems Biology Markup Language
  - **SED-ML** Simulation Experiment Description Language (SED-ML)
  - **SBGN** Systems Biology Graphical Notation
  - **CellML**
  - **SBOL** Synthetic Biology Open Language Data
  - **NeuroML**

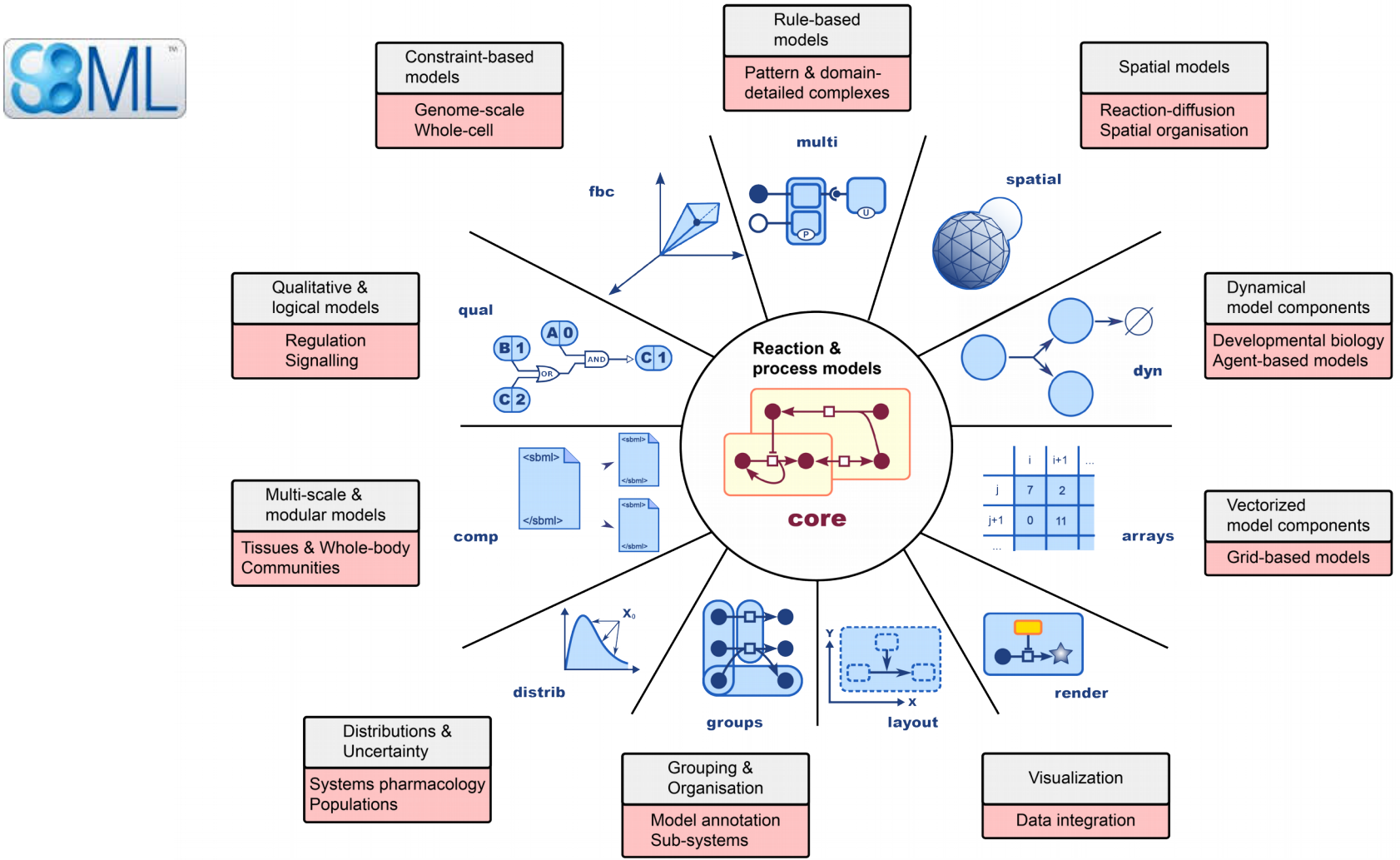
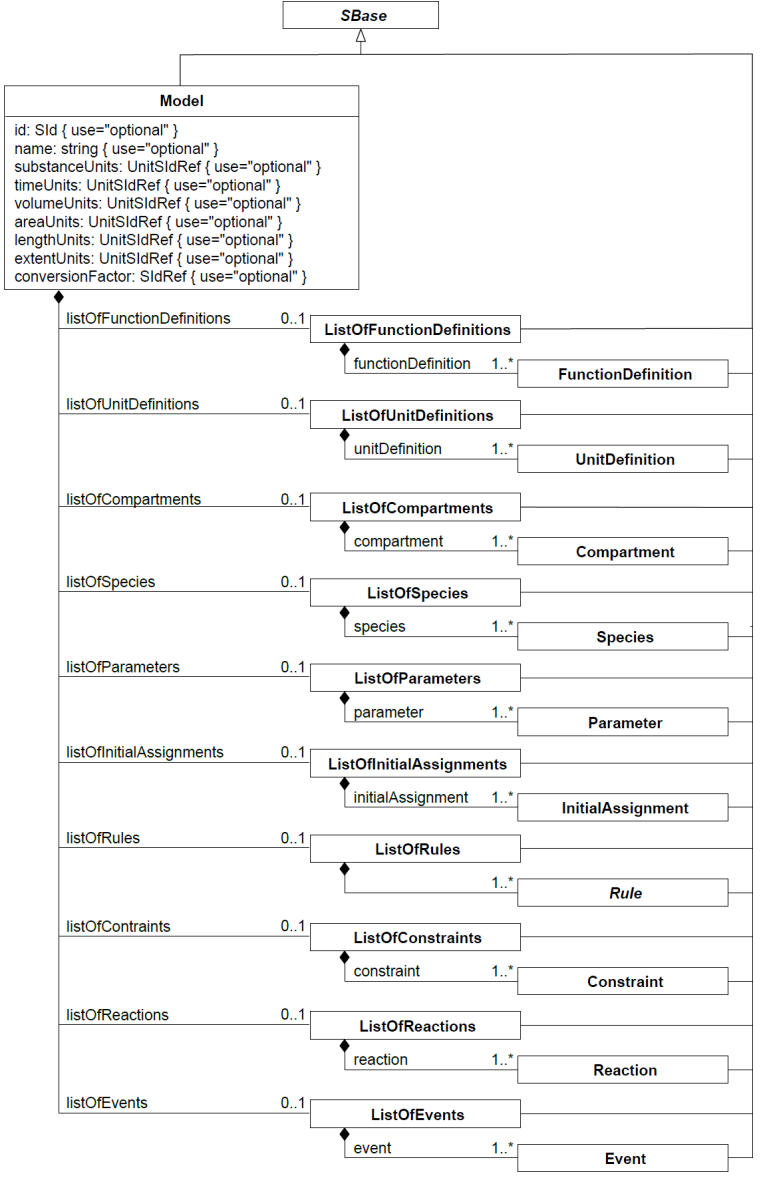


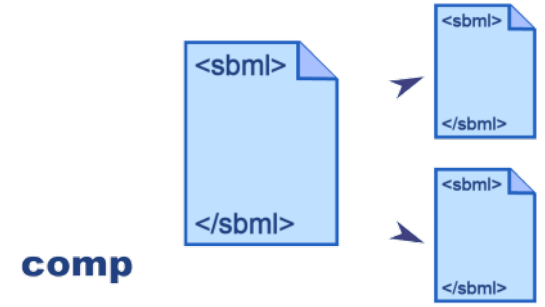
Figure 1: SBML Level 3 consists of a core (center) and specialized SBML Level 3 Packages *(in blue) providing new syntactical constructs and cover new modeling approaches. The Packages support new types of modeling (in gray) needed for large and complex models such as used in various domains and fields of biology (in red).*

- **De facto standard** for encoding computational models
- **Libraries**: libsbml (C++, python, R, JavaScript, ... & JSBML (Java)
- **TestSuite** and **Validators**
- Components
  - UnitDefinitions
  - FunctionDefinitions
  - Compartments
  - Species
  - Parameters
  - InitialAssignments
  - Rules
  - Constraints
  - Reactions
  - Events



SBML comp

- Hierarchical model composition
- Coupling of models



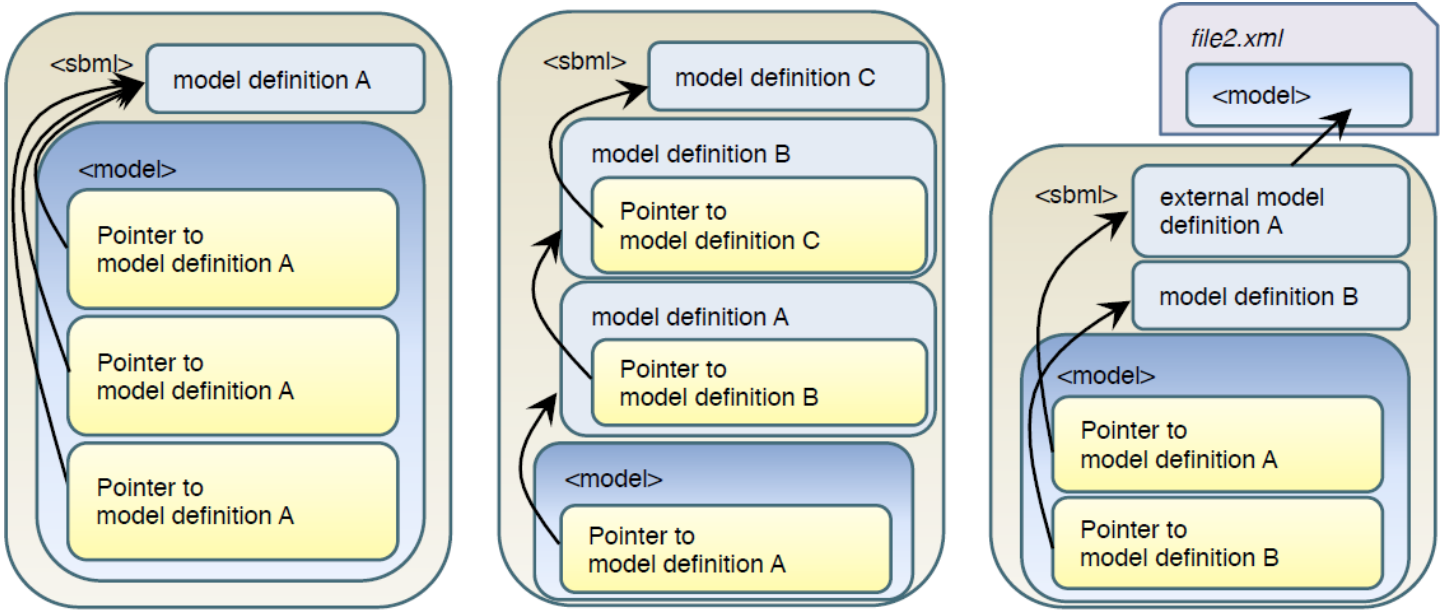


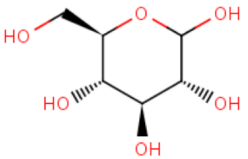
Figure 1: *Three different examples of model composition scenarios. From left to right: (1) a model composed of multiple instances of a single, internally-defined submodel definition; (2) a model composed of a submodel that is itself composed of submodels; and (3) a model composed of submodels, one of which is defined in an external file.*

Annotations

- Ontology
  - definition of controlled vocabulary with clear meaning and relationships
  - allows to precisely describe objects
- Annotation
  - process of attaching ontology terms to objects
  - important for mapping data onto models
  - important for automatic methods (model merging, reuse of components
- RDF triples
  - (subject, verb, object)
- Examples
  - CHEBI (chemical entities)
  - UniProt (proteins)
  - Ontology Lookup Service

CHEBI:4167 - D-glucopyranose

Main | ChEBI Ontology | Automatic Xrefs | Reactions | Pathways | Models



ChEBI Name	**D-glucopyranose**
ChEBI ID	**CHEBI:4167**
ChEBI ASCII Name	D-glucopyranose
Definition	A glucopyranose having D-configuration.
Stars	⭐⭐⭐ This entity has been manually annotated by the ChEBI Team.
Supplier Information	eMolecules:711823, eMolecules:29536451, MolPort-021-782-999
Download	Molfile XML SDF

- Find compounds which contain this structure
- Find compounds which resemble this structure
- Take structure to the Advanced Search

more structures >>

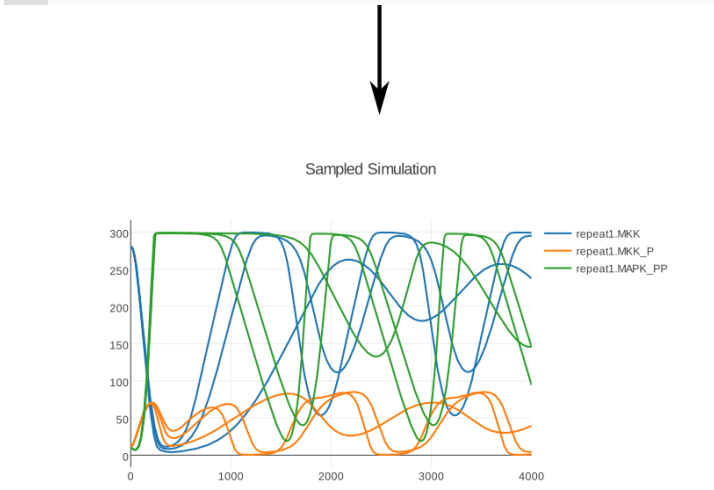
Wikipedia | License

Glucose (also called dextrose) is a simple sugar with the molecular formula C6H12O6. Glucose is the most abundant monosaccharide, a subcategory of carbohydrates. Glucose is mainly made by plants and most algae during photosynthesis from water and carbon dioxide, using energy from sunlight. There it is used to make cellulose in cell walls, which is the most abundant carbohydrate. In energy metabolism, glucose is the most important source of energy in all organisms. Glucose for metabolism is partially stored as a polymer, in plants mainly as starch and amylopectin and in animals as glycogen. Glucose circulates in the blood of animals as blood sugar. The naturally occurring form of glucose is D-glucose, while L-glucose is produced synthetically in comparably small amounts and is of lesser importance. Glucose, as intravenous sugar solution, is on the World Health Organization's List of Essential Medicines, the most important medications needed in a basic health system. The name glucose derives through the French from the Greek γλυκός, which means "sweet," in reference to must, the sweet, first press of grapes in the making of wine. The suffix "-ose" is a chemical classifier, denoting a sugar.

Read full article at Wikipedia

Formula	C6H12O6
Net Charge	0
Average Mass	180.15588
Monoisotopic Mass	180.063
InChI	InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1
InChIKey	WQZGKKKJIJFFOK-GASJEMHNSA-N
SMILES	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O

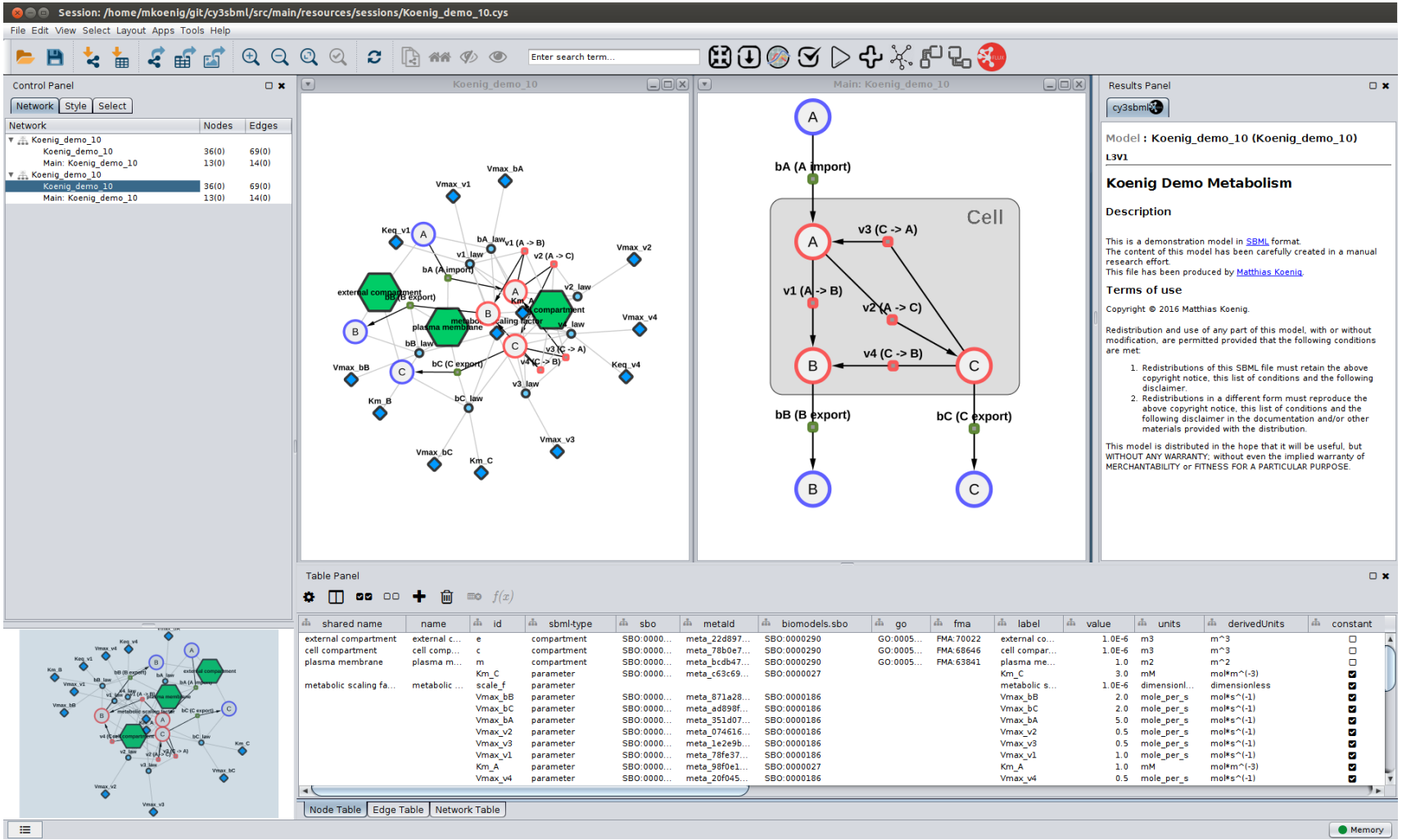
[3]
```
// -- Begin Antimony block converted from MAPKcascade.xml
// Created by libAntimony v2.9.3
model *MAPKcascade()
  ...
  // Reactions:
  J0: MKKK => MKKK_P; J0_V1*MKKK/((1 + (MAPK_PP/J0_Ki)^J0_n)*(J0_K1 + MKKK));
  J1: MKKK_P => MKKK; J1_V2*MKKK_P/(J1_KK2 + MKKK_P);
  J2: MKK => MKK_P; J2_k3*MKKK_P*MKK/(J2_KK3 + MKK);
  J3: MKK_P => MKK_PP; J3_k4*MKKK_P*MKK_P/(J3_KK4 + MKK_P);
  J4: MKK_PP => MKK_P; J4_V5*MKK_PP/(J4_KK5 + MKK_PP);
  J5: MKK_P => MKK; J5_V6*MKK_P/(J5_KK6 + MKK_P);
  J6: MAPK => MAPK_P; J6_k7*MKK_PP*MAPK/(J6_KK7 + MAPK);
  J7: MAPK_P => MAPK_PP; J7_k8*MKK_PP*MAPK_P/(J7_KK8 + MAPK_P);
  J8: MAPK_PP => MAPK_P; J8_V9*MAPK_PP/(J8_KK9 + MAPK_PP);
  J9: MAPK_P => MAPK; J9_V10*MAPK_P/(J9_KK10 + MAPK_P);
  ...
end
// -- End Antimony block

// -- Begin PhraSEDML block converted from main.xml
// Created by libphrasedml v1.0.7
// Models
model1 = model "MAPKcascade"

// Simulations
sim1 = simulate uniform(0, 4000, 1000)

// Tasks
task1 = run sim1 on model1

// Repeated Tasks
repeat1 = repeat task1 for model1.J1_KK2 in [1, 10, 40], reset=true

// Outputs
plot "Sampled Simulation" repeat1.time vs repeat1.MKK, repeat1.MKK_P, repeat1.MAPK_PP
// -- End PhraSEDML block
```



Modeling Tools

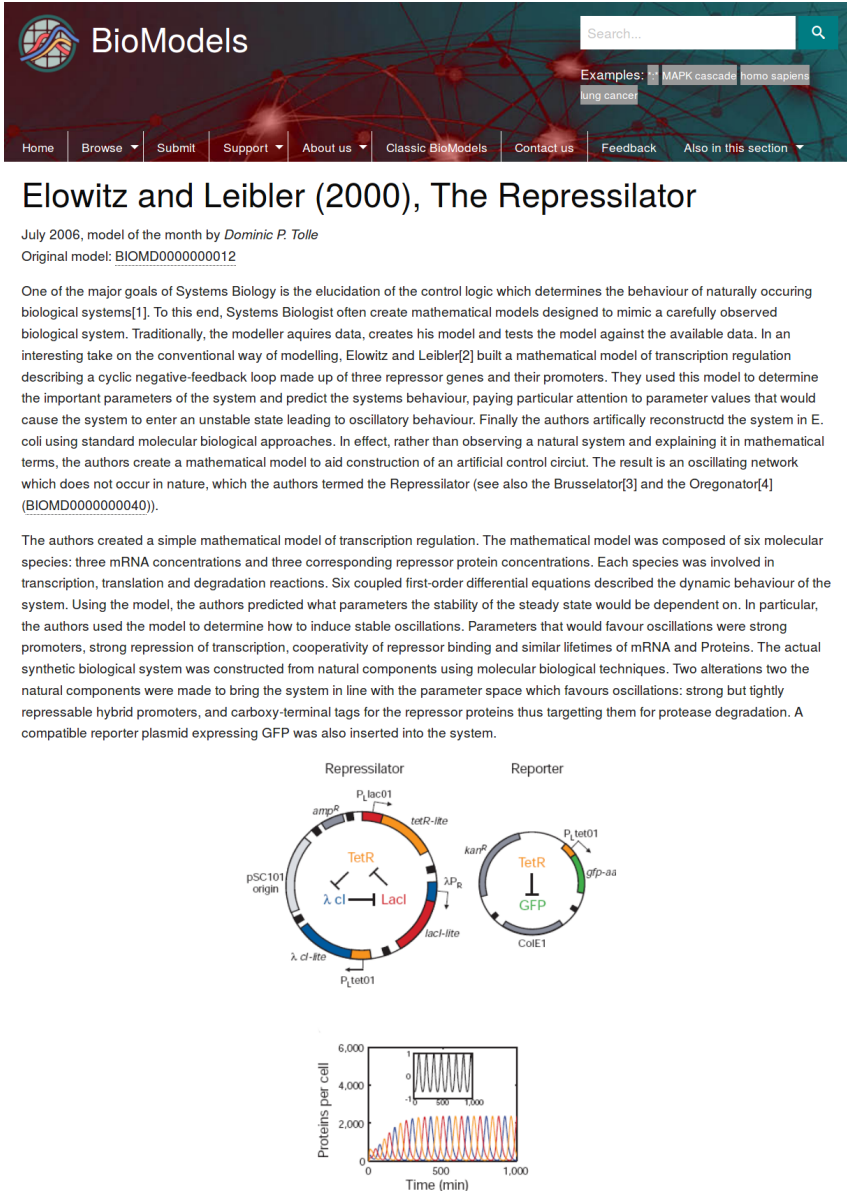
- **libRoadRunner**: High performance SBML simulator
- **tellurium**: Python based modeling environment (library & notebook)
- **COPASI**: GUI based tool for working with SBML models
- **JWS:** web based tool for simulations

Analysis and Visualization



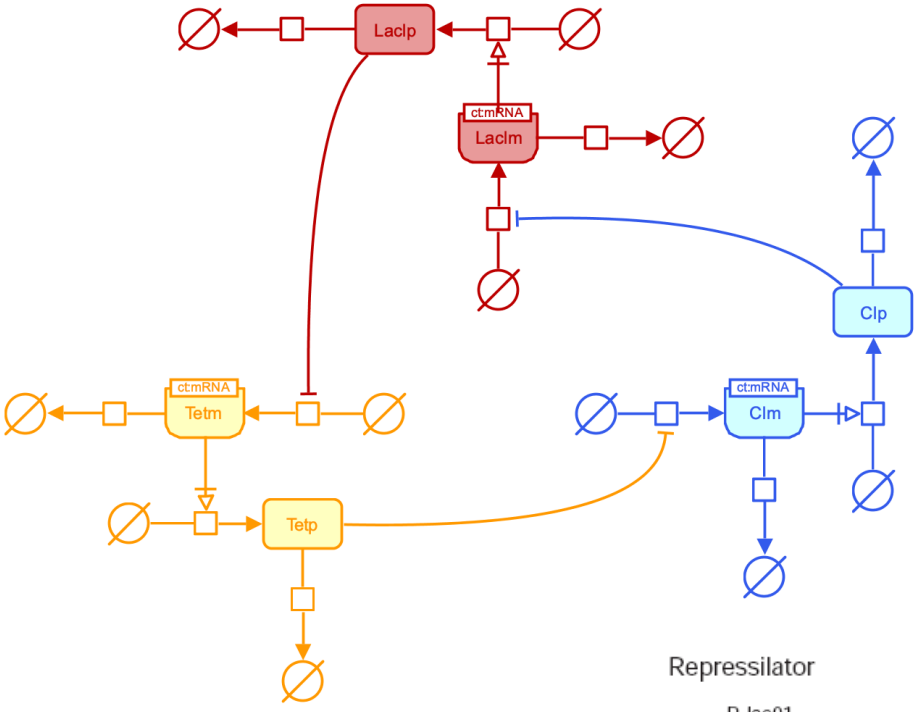
Model databases

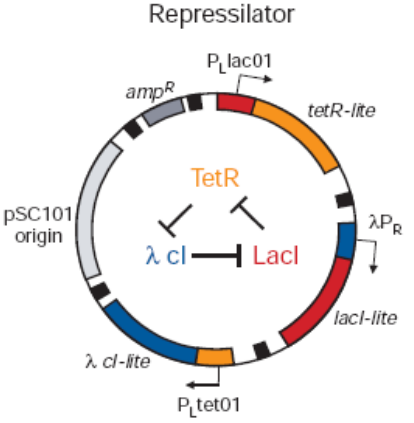
- Biomodels
  - large collection of freely available models (SBML and others)
  - curated & uncurated
  - https://biomodels.org
- JWS
  - similar database, allows for online simulations
  - https://jjj.bio.vu.nl/models/experiments/elowitz2000_fig1c/simulate



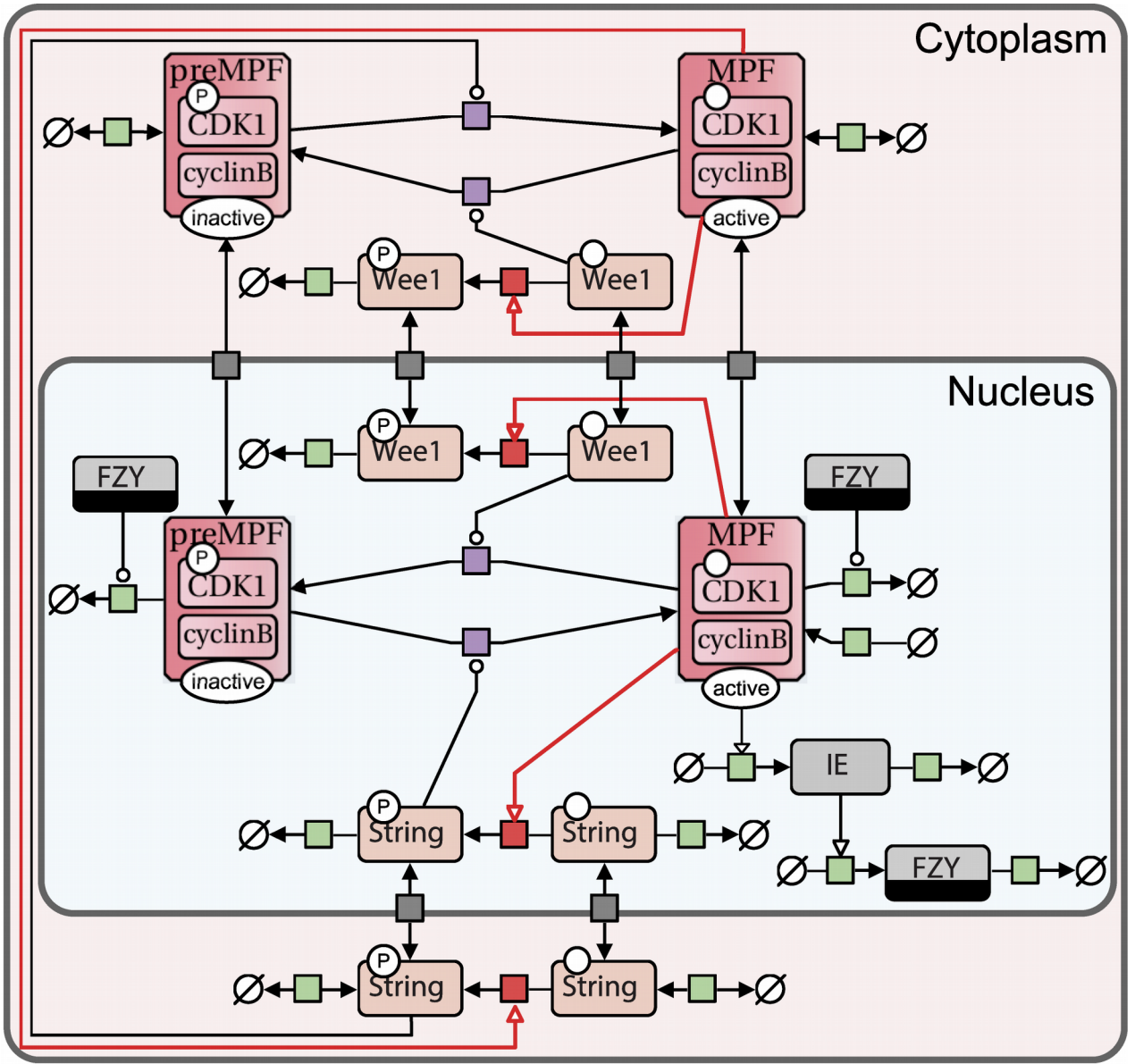
SBGN

- high quality, standard graphical languages for representing biological processes and interactions
  - PD: process description
  - AF: activity flow
  - ER: entity relationship
- http://sbgn.github.io/sbgn/about



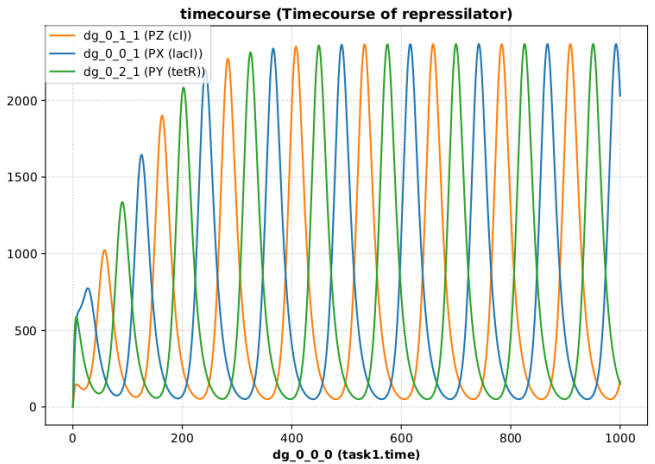


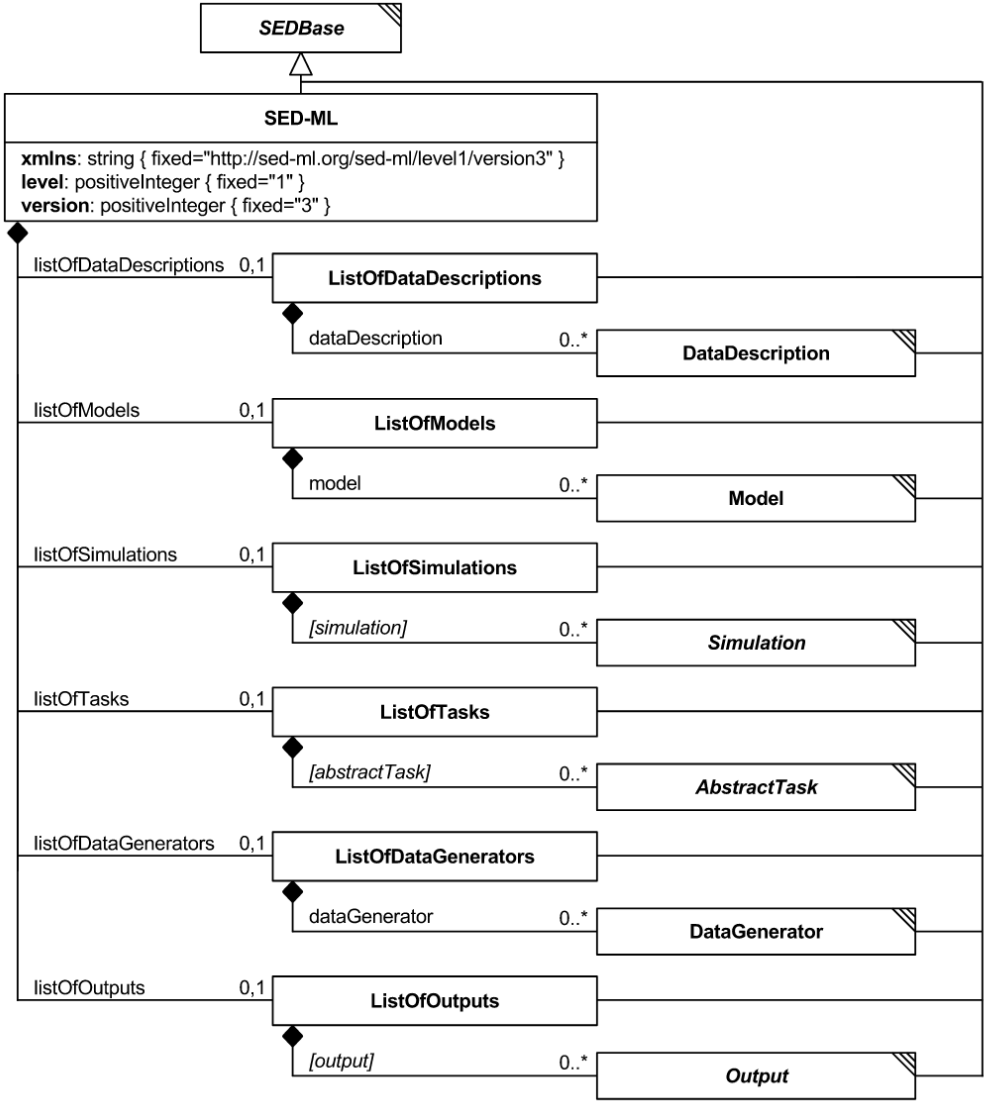
- Map of drosophila cell cycle



SED-ML

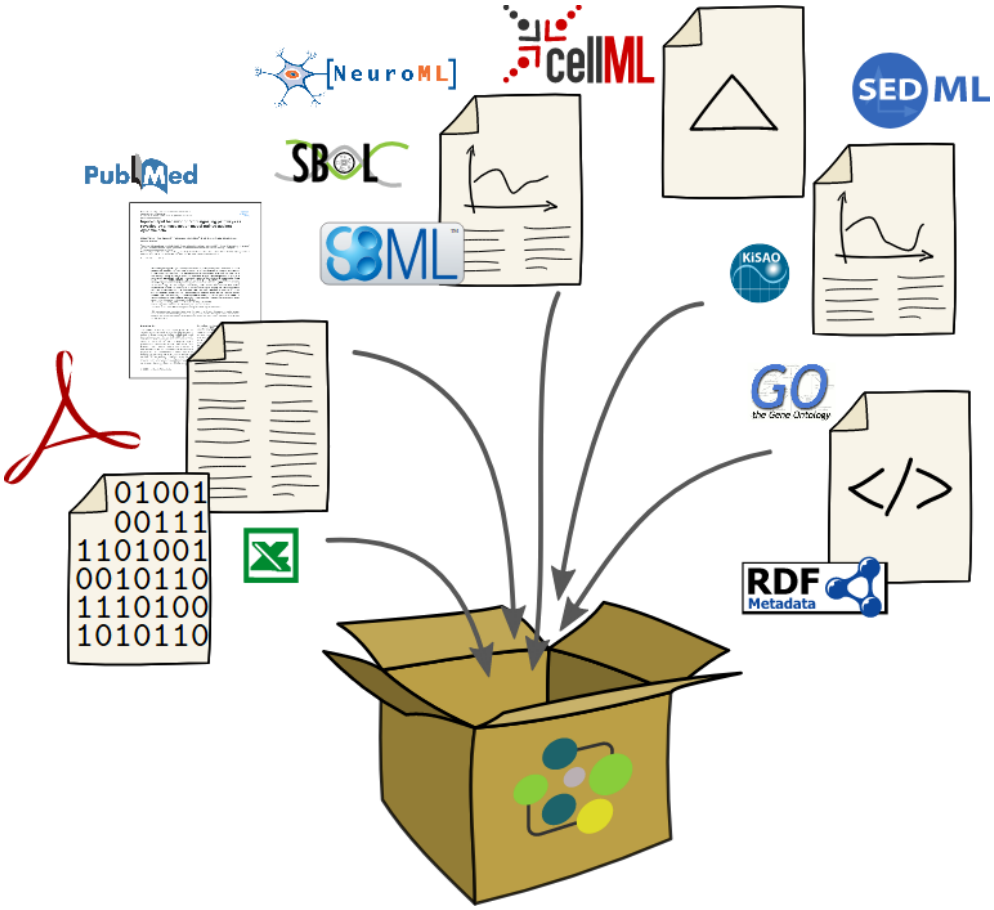
- **S**imulation **E**xperiment **D**escription **M**arkup **L**anguage (SED-ML)
  https://sed-ml.github.io
- SED-ML is an XML-based format for encoding simulation setups, to ensure exchangeability and reproducibility of simulation experiments.





COMBINE archive

- A COMBINE archive is a single file bundling the various documents necessary for a modeling and simulation project.
- The archive is encoded using the **Open Modeling EXchange format (OMEX).**



Executable simulation model

https://matthiaskoenig.github.io/exsimo/