Formalizing Systems Biology Models with Biomedical Ontologies

work by:
Robert Hoehndorf, Michel Dumontier,
John H. Gennari, Sarala Wimalaratne,
Bernard de Bono, Dan Cook, George Gkoutos

Cambridge University, Carleton University,
EBI, University of Washington

Harmony @ NYC. April 21,2011

Systems Biology

We create and simulate models to :
- gain biological insight into the structure and function of biochemical networks
- reveal capabilities and predict phenotypes
- undertake metabolic engineering maximize desired product

To do this, we need
- to manage our data & knowledge in a coherent, scalable and machine understandable manner
- use efficient software to execute simulations

Designed by Donald E. Nicholson, Department of Biochemistry and Molecular Biology, The University of Leeds, England, and Sigma

Computational Knowledge Discovery

- Terminological resources increasingly being used to annotate biomolecular models
  - easier to explore or find models

- converting models into formal representations of knowledge
  - check the annotation consistency
  - infer knowledge explicit in terminological resources
  - discover biological implications inherent in the models.

SBML

XML-based representation of biochemical models, their components (compartments, species, reactions, events), descriptors (rules, constraints, functions, units)

Consider the following enzymatic reaction:

$$E + S \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} ES \xrightarrow{k_{\text{cat}}} E + P$$

SBML is an XML-based format

$$E + S \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} ES \xrightarrow{k_{\mathrm{cat}}} E + P$$

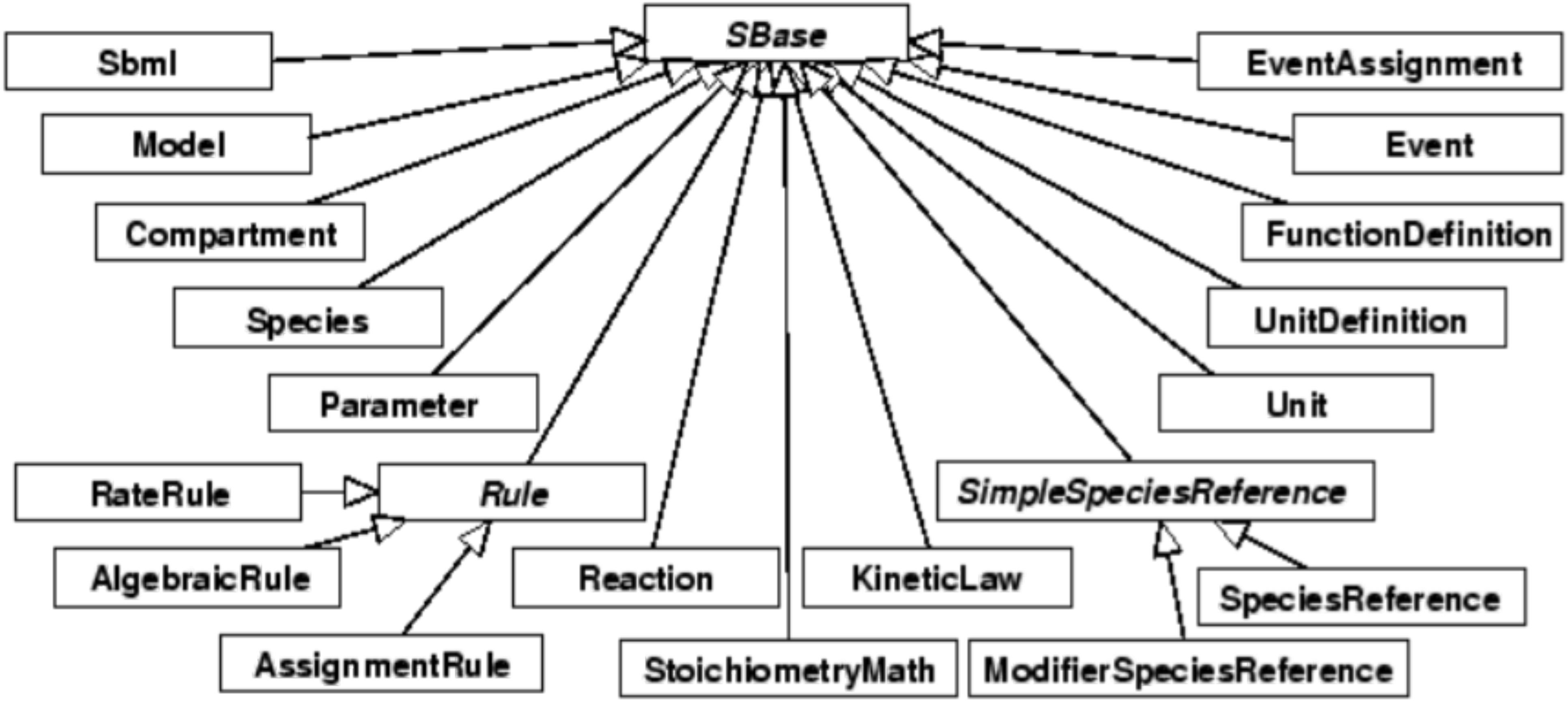
```
<?xml version="1.0" encoding="UTF-8"?>
<sbml level="2" version="3" xmlns="http://www.sbml.org/sbml/level2/version3">
  <model name="EnzymaticReaction">
    <listOfUnitDefinitions>
      <unitDefinition id="per_second">
        <listOfUnits>
          <unit kind="second" exponent="-1"/>
        </listOfUnits>
      </unitDefinition>
      <unitDefinition id="litre_per_mole_per_second">
        <listOfUnits>
          <unit kind="mole"   exponent="-1"/>
          <unit kind="litre"  exponent="1"/>
          <unit kind="second" exponent="-1"/>
        </listOfUnits>
      </unitDefinition>
    </listOfUnitDefinitions>
    <listOfCompartments>
      <compartment id="cytosol" size="1e-14"/>
    </listOfCompartments>
    <listOfSpecies>
      <species compartment="cytosol" id="ES" initialAmount="0"     name="ES"/>
      <species compartment="cytosol" id="P"  initialAmount="0"     name="P"/>
      <species compartment="cytosol" id="S"  initialAmount="1e-20" name="S"/>
      <species compartment="cytosol" id="E"  initialAmount="5e-21" name="E"/>
    </listOfSpecies>
```

```
<listOfReactions>
   <reaction id="veq">
      <listOfReactants>
         <speciesReference species="E"/>
         <speciesReference species="S"/>
      </listOfReactants>
      <listOfProducts>
         <speciesReference species="ES"/>
      </listOfProducts>
      <kineticLaw>
         <math xmlns="http://www.w3.org/1998/Math/MathML">
            <apply>
               <times/>
               <ci>cytosol</ci>
               <apply>
                  <minus/>
                  <apply>
                     <times/>
                     <ci>kon</ci>
                     <ci>E</ci>
                     <ci>S</ci>
                  </apply>
                  <apply>
                     <times/>
                     <ci>koff</ci>
                     <ci>ES</ci>
                  </apply>
               </apply>
            </apply>
         </math>
         <listOfParameters>
            <parameter id="kon"  value="1000000" units="litre_per_mole_per_second"/>
            <parameter id="koff" value="0.2"     units="per_second"/>
         </listOfParameters>
      </kineticLaw>
   </reaction>
```

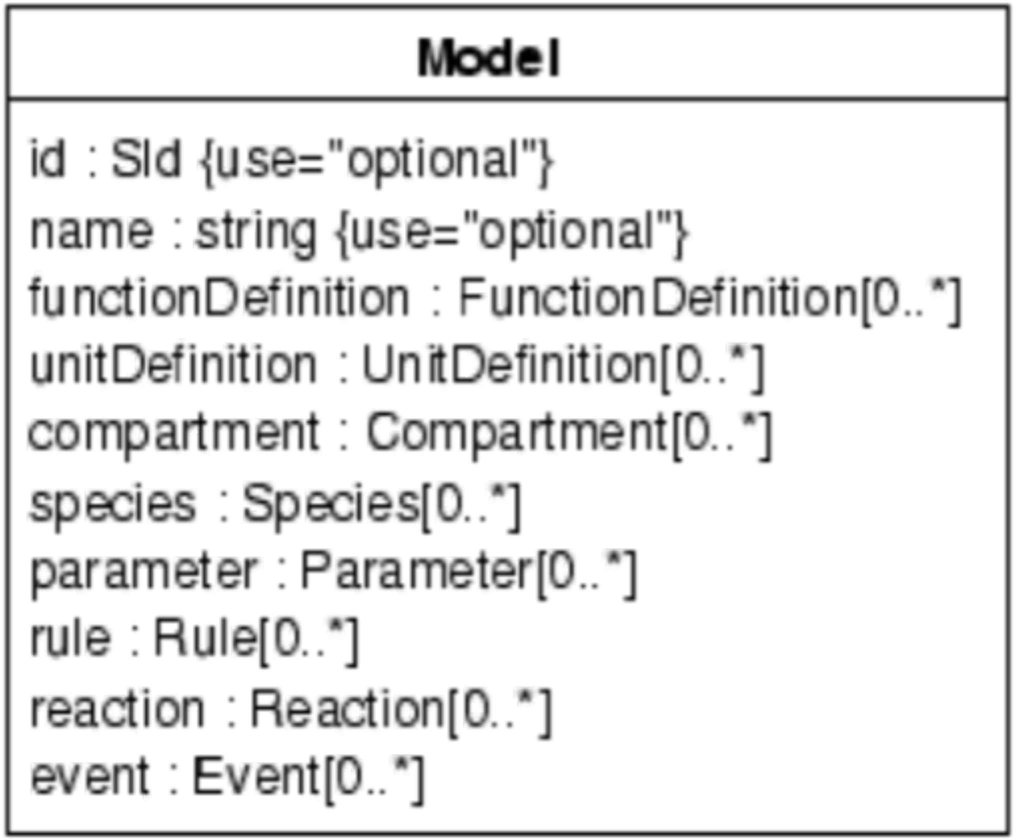
$$E + S \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} ES \xrightarrow{k_{\mathrm{cat}}} E + P$$

```
      <reaction id="vcat" reversible="false">
        <listOfReactants>
          <speciesReference species="ES"/>
        </listOfReactants>
        <listOfProducts>
          <speciesReference species="E"/>
          <speciesReference species="P"/>
        </listOfProducts>
        <kineticLaw>
          <math xmlns="http://www.w3.org/1998/Math/MathML">
            <apply>
              <times/>
              <ci>cytosol</ci>
              <ci>kcat</ci>
              <ci>ES</ci>
            </apply>
          </math>
          <listOfParameters>
            <parameter id="kcat" value="0.1" units="per_second"/>
          </listOfParameters>
        </kineticLaw>
      </reaction>
    </listOfReactions>
  </model>
</sbml>
```

SBML conceptualization



SBML specifies what models can have as attributes

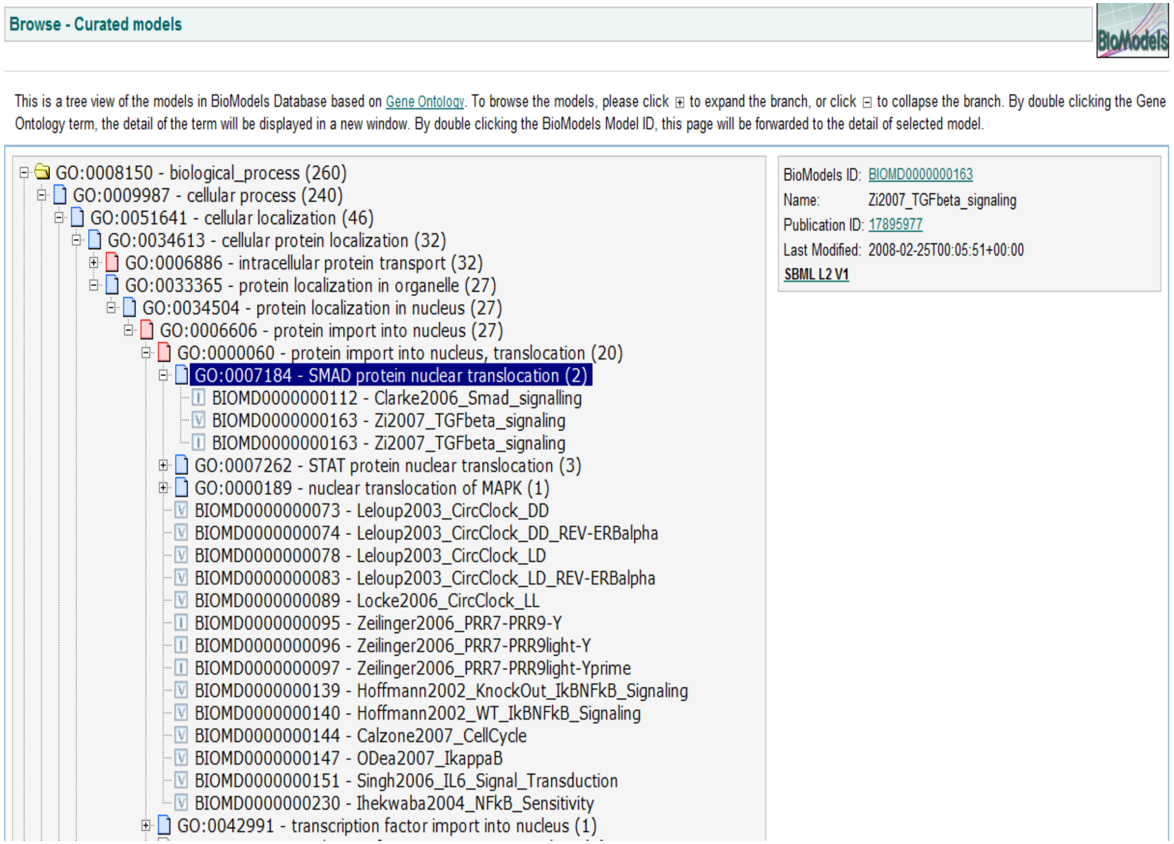


Model

id : SId {use="optional"}
name : string {use="optional"}
functionDefinition : FunctionDefinition[0..*]
unitDefinition : UnitDefinition[0..*]
compartment : Compartment[0..*]
species : Species[0..*]
parameter : Parameter[0..*]
rule : Rule[0..*]
reaction : Reaction[0..*]
event : Event[0..*]

Biomodels are semantically annotated SBML models

- EBI managed resource
- 600 + models available as SBML
- 269 models are curated with GO process, function and component terms, and has links to protein databases.
- Possible to browse by GO terms:



Browse - Curated models

This is a tree view of the models in BioModels Database based on Gene Ontology. To browse the models, please click ⊞ to expand the branch, or click ⊟ to collapse the branch. By double clicking the Gene Ontology term, the detail of the term will be displayed in a new window. By double clicking the BioModels Model ID, this page will be forwarded to the detail of selected model.

- GO:0008150 - biological_process (260)
  - GO:0009987 - cellular process (240)
    - GO:0051641 - cellular localization (46)
      - GO:0034613 - cellular protein localization (32)
        - GO:0006886 - intracellular protein transport (32)
        - GO:0033365 - protein localization in organelle (27)
          - GO:0034504 - protein localization in nucleus (27)
            - GO:0006606 - protein import into nucleus (27)
              - GO:0000060 - protein import into nucleus, translocation (20)
                - GO:0007184 - SMAD protein nuclear translocation (2)
                  - BIOMD0000000112 - Clarke2006_Smad_signalling
                  - BIOMD0000000163 - Zi2007_TGFbeta_signaling
                  - BIOMD0000000163 - Zi2007_TGFbeta_signaling
                - GO:0007262 - STAT protein nuclear translocation (3)
                - GO:0000189 - nuclear translocation of MAPK (1)
                - BIOMD0000000073 - Leloup2003_CircClock_DD
                - BIOMD0000000074 - Leloup2003_CircClock_DD_REV-ERBalpha
                - BIOMD0000000078 - Leloup2003_CircClock_LD
                - BIOMD0000000083 - Leloup2003_CircClock_LD_REV-ERBalpha
                - BIOMD0000000089 - Locke2006_CircClock_LL
                - BIOMD0000000095 - Zeilinger2006_PRR7-PRR9-Y
                - BIOMD0000000096 - Zeilinger2006_PRR7-PRR9light-Y
                - BIOMD0000000097 - Zeilinger2006_PRR7-PRR9light-Yprime
                - BIOMD0000000139 - Hoffmann2002_KnockOut_IkBNFkB_Signaling
                - BIOMD0000000140 - Hoffmann2002_WT_IkBNFkB_Signaling
                - BIOMD0000000144 - Calzone2007_CellCycle
                - BIOMD0000000147 - ODea2007_IkappaB
                - BIOMD0000000151 - Singh2006_IL6_Signal_Transduction
                - BIOMD0000000230 - Ihekwaba2004_NFkB_Sensitivity
              - GO:0042991 - transcription factor import into nucleus (1)

BioModels ID: BIOMD0000000163
Name: Zi2007_TGFbeta_signaling
Publication ID: 17895977
Last Modified: 2008-02-25T00:05:51+00:00
SBML L2 V1

http://www.ebi.ac.uk/biomodels-main/

Biomodels are semantically annotated SBML models

```
<?xml version="1.0" encoding="UTF-8"?>
<sbml xmlns="http://www.sbml.org/sbml/level2" metaid="metaid_0000001" level="2" version="1">
 <model metaid="metaid_0000002" id="Proctor2005_Hsp90" name="Hsp90model_basis510">
  <annotation>
   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:bqbiol="http://biomodels.net/biology-qualifiers/"
      xmlns:bqmodel="http://biomodels.net/model-qualifiers/">
     <bqbiol:isVersionOf>
       <rdf:Bag>
        <rdf:li rdf:resource="urn:miriam:obo.go:GO%3A0051085"/>
        <rdf:li rdf:resource="urn:miriam:obo.go:GO%3A0007569"/>
        <rdf:li rdf:resource="urn:miriam:obo.go:GO%3A0009408"/>
       </rdf:Bag>
      </bqbiol:isVersionOf>
     </rdf:RDF>
   </annotation>
  </model>
</sbml>
```

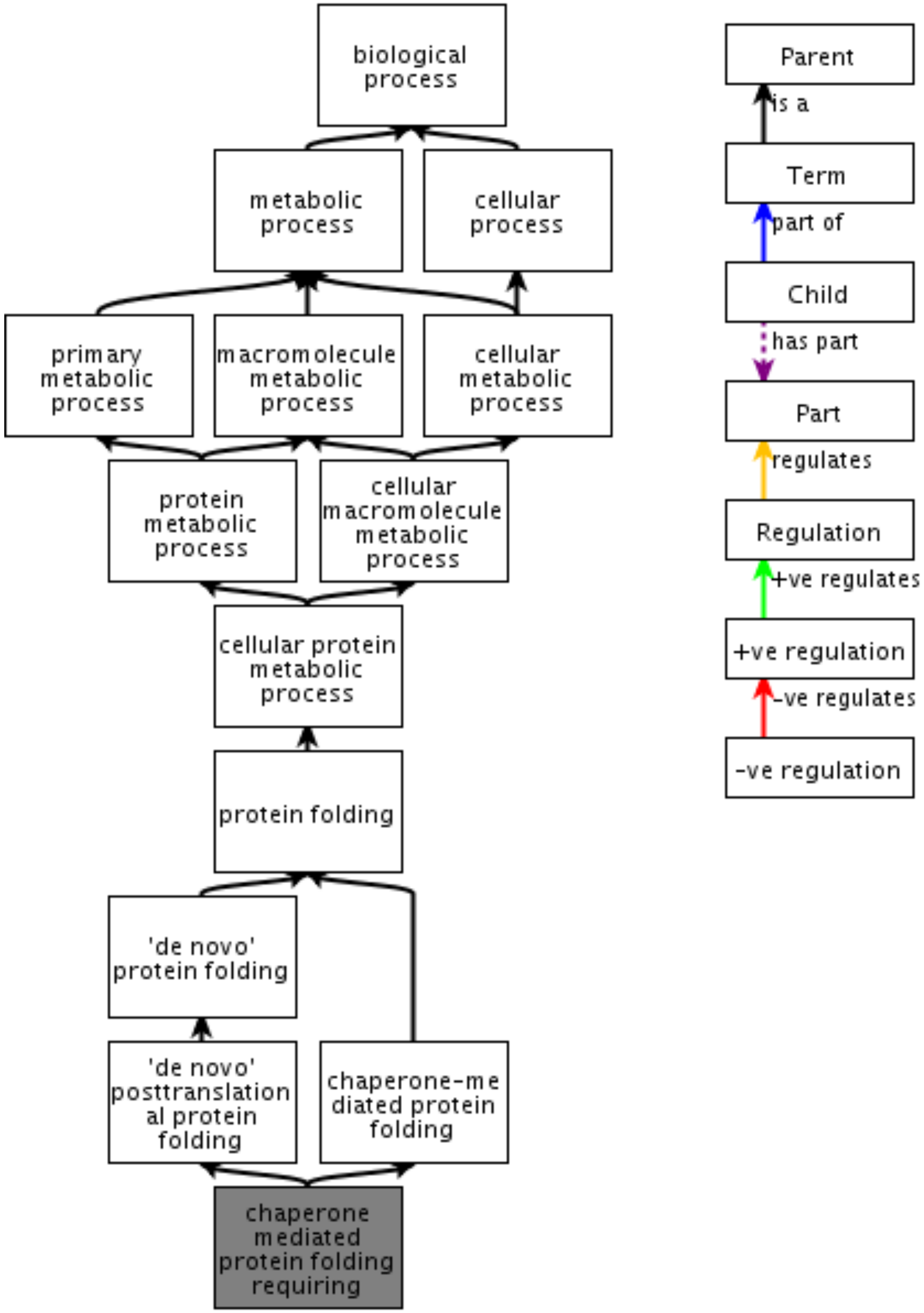
GO:0051085
chaperone mediated protein folding requiring cofactor

GO:0007569
cell aging

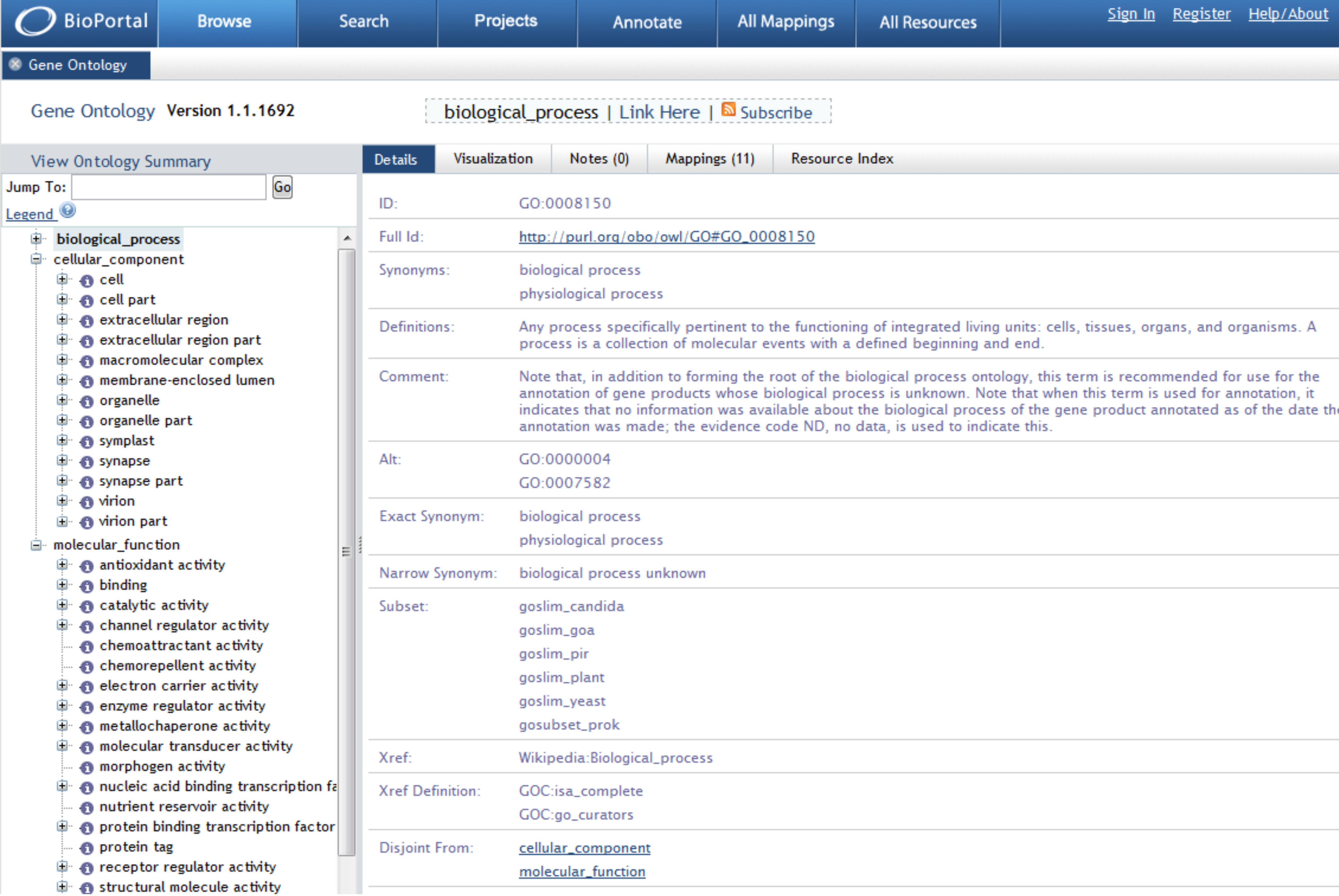
GO:0009408
response to heat

Gene Ontology

- nearly 30,000 terms
- *covers*
  - biological processes
  - molecular functions
  - cellular components
- terms organized around "is a" hierarchy
- terms further described with 'has part'/'part of'; 'regulates' and '+ regulates', '- regulates'



QuickGO - http://www.ebi.ac.uk/QuickGO



- biological_process
- cellular_component
  - cell
  - cell part
  - extracellular region
  - extracellular region part
  - macromolecular complex
  - membrane-enclosed lumen
  - organelle
  - organelle part
  - symplast
  - synapse
  - synapse part
  - virion
  - virion part
- molecular_function
  - antioxidant activity
  - binding
  - catalytic activity
  - channel regulator activity
  - chemoattractant activity
  - chemorepellent activity
  - electron carrier activity
  - enzyme regulator activity
  - metallochaperone activity
  - molecular transducer activity
  - morphogen activity
  - nucleic acid binding transcription fa
  - nutrient reservoir activity
  - protein binding transcription factor
  - protein tag
  - receptor regulator activity
  - structural molecule activity

ID:	GO:0008150
Full Id:	http://purl.org/obo/owl/GO#GO_0008150
Synonyms:	biological process physiological process
Definitions:	Any process specifically pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. A process is a collection of molecular events with a defined beginning and end.
Comment:	Note that, in addition to forming the root of the biological process ontology, this term is recommended for use for the annotation of gene products whose biological process is unknown. Note that when this term is used for annotation, it indicates that no information was available about the biological process of the gene product annotated as of the date th annotation was made; the evidence code ND, no data, is used to indicate this.
Alt:	GO:0000004 GO:0007582
Exact Synonym:	biological process physiological process
Narrow Synonym:	biological process unknown
Subset:	goslim_candida goslim_goa goslim_pir goslim_plant goslim_yeast gosubset_prok
Xref:	Wikipedia:Biological_process
Xref Definition:	GOC:isa_complete GOC:go_curators
Disjoint From:	cellular_component molecular_function

Species may be annotated with UniProt, KEGG, ChEBI terms

```
    <species metaid="metaid_0000038" id="ROS" name="ROS" compartment="
compartment" initialAmount="100" hasOnlySubstanceUnits="true">
      <annotation>
        <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
           xmlns:bqbiol="http://biomodels.net/biology-qualifiers/"
           xmlns:bqmodel="http://biomodels.net/model-qualifiers/">
          <rdf:Description rdf:about="#metaid_0000038">
            <bqbiol:isVersionOf>
              <rdf:Bag>
                <rdf:li rdf:resource="urn:miriam:obo.chebi:CHEBI%3A26523"/>
              </rdf:Bag>
            </bqbiol:isVersionOf>
          </rdf:Description>
        </rdf:RDF>
      </annotation>
    </species>
```

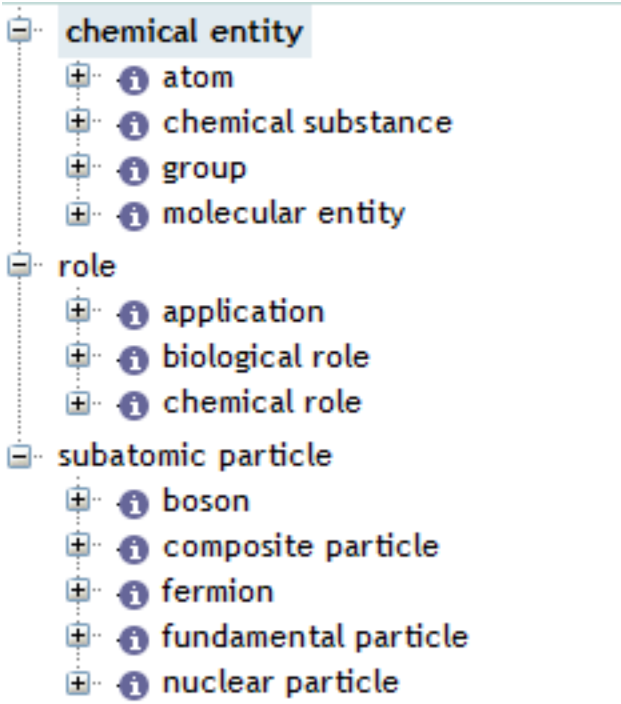
CHEBI:26523 reactive oxygen species

Chemical Entities of Biological Interest (ChEBI)

recently refactored to be in line with formal (reasoning capable) ontology

scope includes chemical entities (atoms, substances, groups, molecules), roles and subatomic particles

large numbers of curated molecules



- chemical entity
  - atom
  - chemical substance
  - group
  - molecular entity
- role
  - application
  - biological role
  - chemical role
- subatomic particle
  - boson
  - composite particle
  - fermion
  - fundamental particle
  - nuclear particle

Approach

The idea is to create sophisticated OWL ontologies from biomolecular models represented using the Systems Biology Markup Language (SBML).

Features

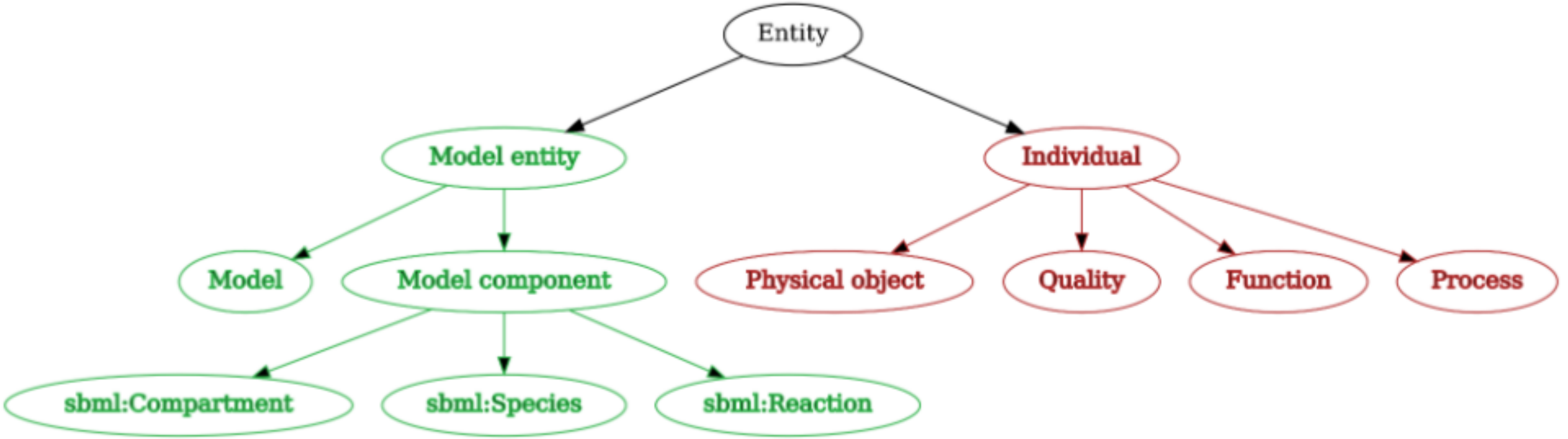
- ontological commitment: terms in a vocabulary become formally defined classes and relations in an ontology
- upper level ontology to distinguish and constrain model entities the spatio-temporal entities they represent
- basic relations are used to describe (constrain) entities in terms of the attributes and relationships they hold

code: http://code.google.com/p/sbmlharvester/

Conceptualization (SBML)

- 2 kinds of entities:
    - in silico: model components
    - in vivo: the entities represented by a model

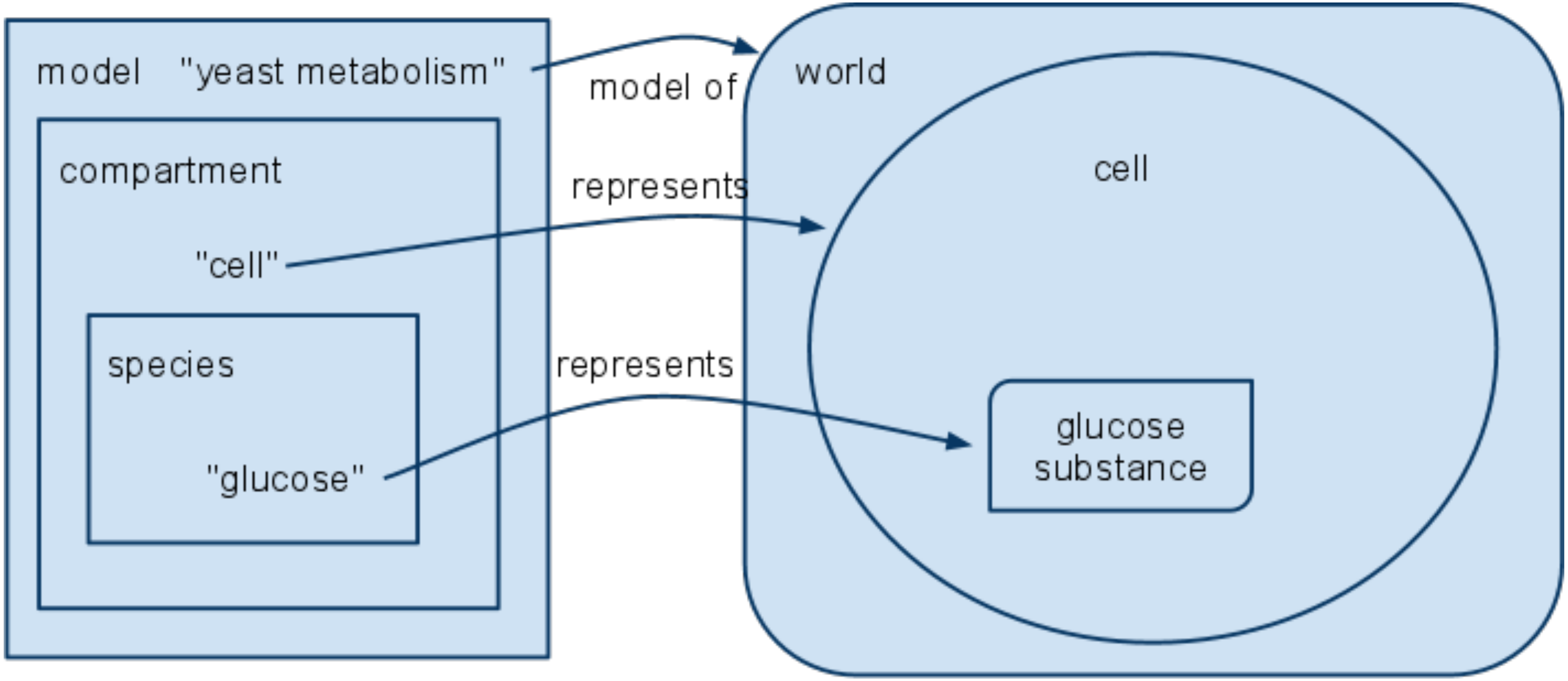
An Upper Level Ontology distinguishes models from the entities they represent



Ontological commitment

- Assumption 1: Every model represents a material entity (Model SubClassOf: represents some MaterialEntity)
- a Model annotated with class *C* represents a C that is a subclass of *MaterialEntity*

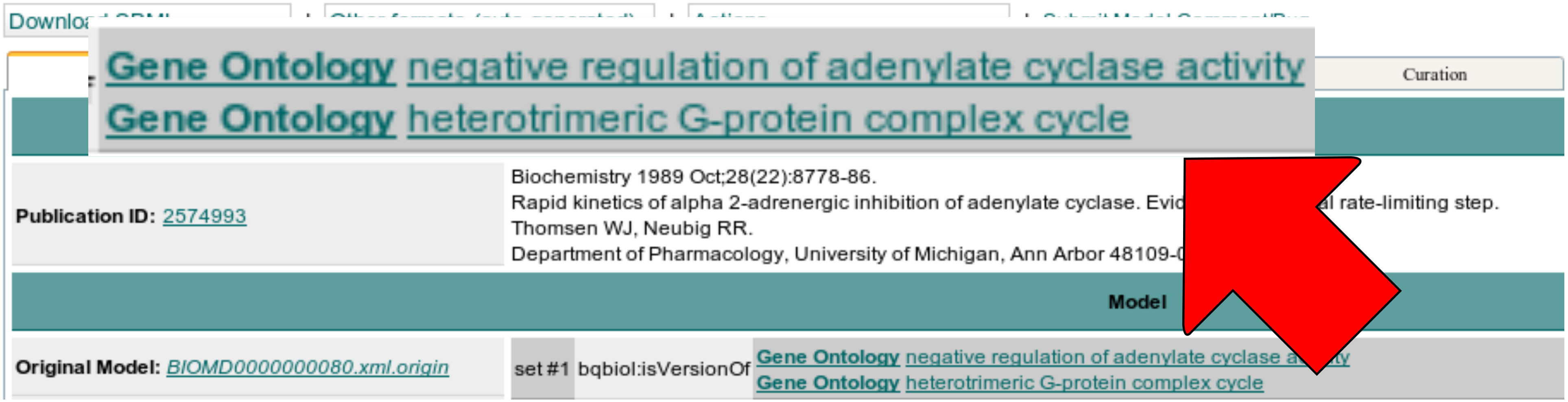
Model entities (models and model components) are distinguished from the entities they represent



every element E of the SBML language represents a class Rep(E) and we assert that E subClassOf: represents some Rep(E)

Ontological commitment

In addition to annotation to physical entities, we note that annotations include functions and processes



Downloa

Gene Ontology negative regulation of adenylate cyclase activity
Gene Ontology heterotrimeric G-protein complex cycle

Curation

Publication ID: 2574993

Biochemistry 1989 Oct;28(22):8778-86.
Rapid kinetics of alpha 2-adrenergic inhibition of adenylate cyclase. Evid … al rate-limiting step.
Thomsen WJ, Neubig RR.
Department of Pharmacology, University of Michigan, Ann Arbor 48109-0

Model

Original Model: *BIOMD0000000080.xml.origin*

set #1 bqbiol:isVersionOf
Gene Ontology negative regulation of adenylate cyclase a … ty
Gene Ontology heterotrimeric G-protein complex cycle

Ontological commitment

- Assumption 1: Every model represents a material entity (Model SubClassOf: represents some MaterialEntity)
- a Model annotated with class *C* represents a C that is a subclass of *MaterialEntity* or a Thing that has-function some C or a Thing that has a function that is realized by only C's

Model annotations are converted into ontology axioms

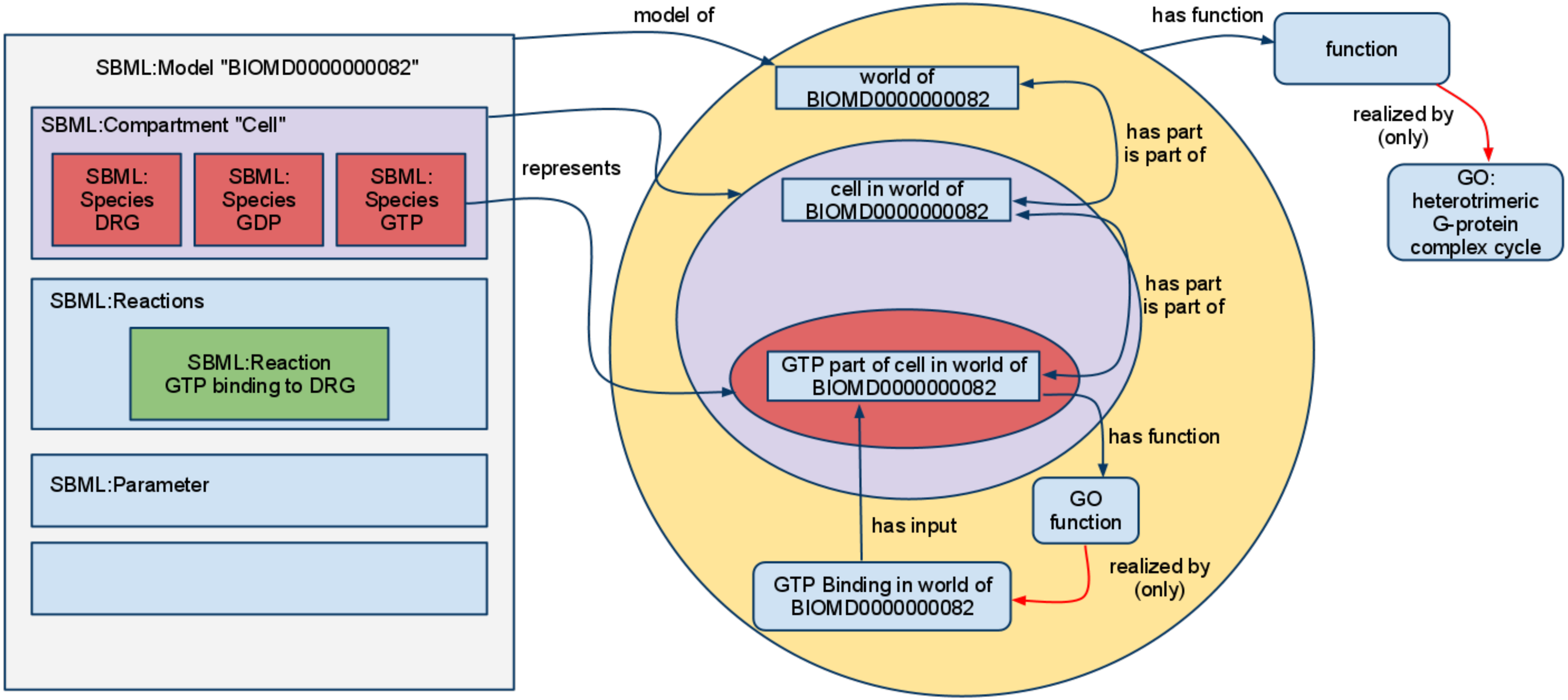
model annotation: **ontology axioms:**

'is' | 'isVersionOf' | 'isVariantOf' *subClassOf*

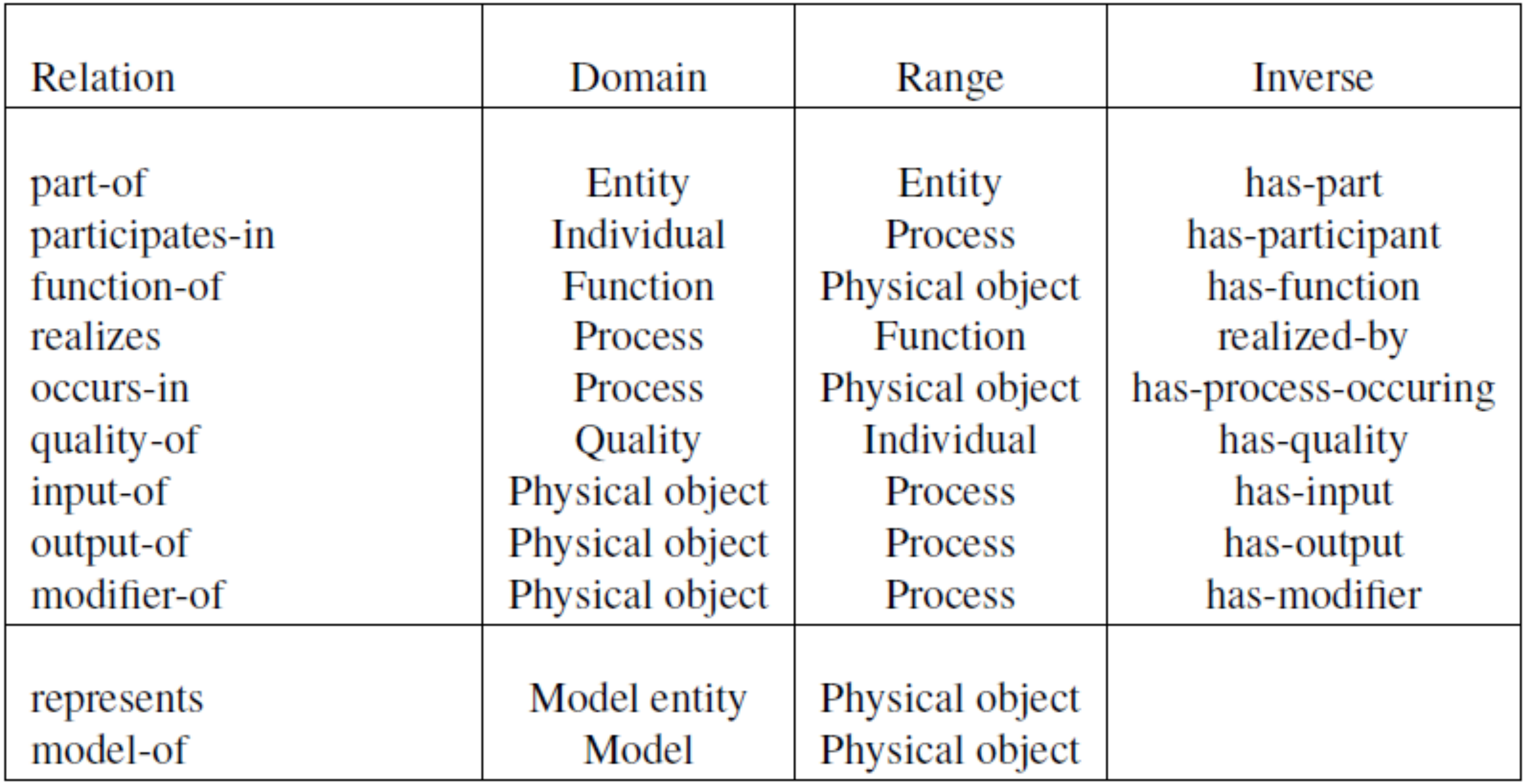
* physical entity --------------------> has-part some physical-entity

* molecular function ---------------> has-part some (has-function some function)

* biological process ---------------> has-part some (
has-function some (realized-by only
process))


SBML:Model "BIOMD0000000082"
SBML:Compartment "Cell"
SBML: Species DRG
SBML: Species GDP
SBML: Species GTP
SBML:Reactions
SBML:Reaction GTP binding to DRG
SBML:Parameter
model of
represents
world of BIOMD0000000082
has part
is part of
cell in world of BIOMD0000000082
has part
is part of
GTP part of cell in world of BIOMD0000000082
has function
GO function
realized by (only)
has input
GTP Binding in world of BIOMD0000000082
has function
function
realized by (only)
GO: heterotrimeric G-protein complex cycle

3. Relations impose additional constraints, such that inconsistencies arise when incorrectly used



Relation	Domain	Range	Inverse
part-of	Entity	Entity	has-part
participates-in	Individual	Process	has-participant
function-of	Function	Physical object	has-function
realizes	Process	Function	realized-by
occurs-in	Process	Physical object	has-process-occuring
quality-of	Quality	Individual	has-quality
input-of	Physical object	Process	has-input
output-of	Physical object	Process	has-output
modifier-of	Physical object	Process	has-modifier
represents	Model entity	Physical object	
model-of	Model	Physical object	

Ontological commitment

Assertion:
M SubClassOf: represents some C or represents some (has-function some C) or represents some (has-function some (realized-by only C))

C SubClassOf: MaterialEntity
Then:

- represents some C is satisfiable
- represents some (has-function some C) and represents some (has-function some (realized-by only C)) are unsatisfiable

Ontological commitment

Assertion:
M SubClassOf: represents some C or represents some (has-function some C) or represents some (has-function some (realized-by only C))

C SubClassOf: Function
Then:

- represents some (has-function some C) is satisfiable
- represents some C and represents some (has-function some (realized-by only C)) are unsatisfiable

Ontological commitment

Assertion:
M SubClassOf: represents some C or represents some (has-function some C) or represents some (has-function some (realized-by only C))

C SubClassOf: Process
Then:

- represents some (has-function some (realized-by only C)) is satisfiable
- represents some C and represents some (has-function some C) are unsatisfiable

SBML2OWL: Implementation

- Combine libSBML and OWLAPI
- Use libSBML to access model structure
  - extract MIRIAM annotations in RDF
  - use Jena RDF API to parse RDF annotations in SBML models
  - use OWLAPI to perform conversion of SBML structure
  - combine with top-level ontology

SBML2OWL: Implementation

Application to BioModels repository yields:

- OWL ontology with more than 800,000 axioms
- includes all referenced ontologies
  - GO
  - ChEBI
  - Celltype
  - FMA
  - PATO

Model verification

After reasoning, we found 27 models to be inconsistent

reasons

1. our representation - functions sometimes found in the place of physical entities (e.g. entities that secrete insulin). better to constrain with appropriate relations
2. SBML abused - species used as a measure of time
3. constraints in the ontologies themselves mean that the annotation is simply not possible

Finding inconsistencies with axiomatically enhanced ontologies

recent work treats function as process and axioms state that an ATPase activity (GO:0004002) is a Catalytic activity that has Water and ATP as input, ADP and phosphate as output and is a part of an ATP catabolic process.

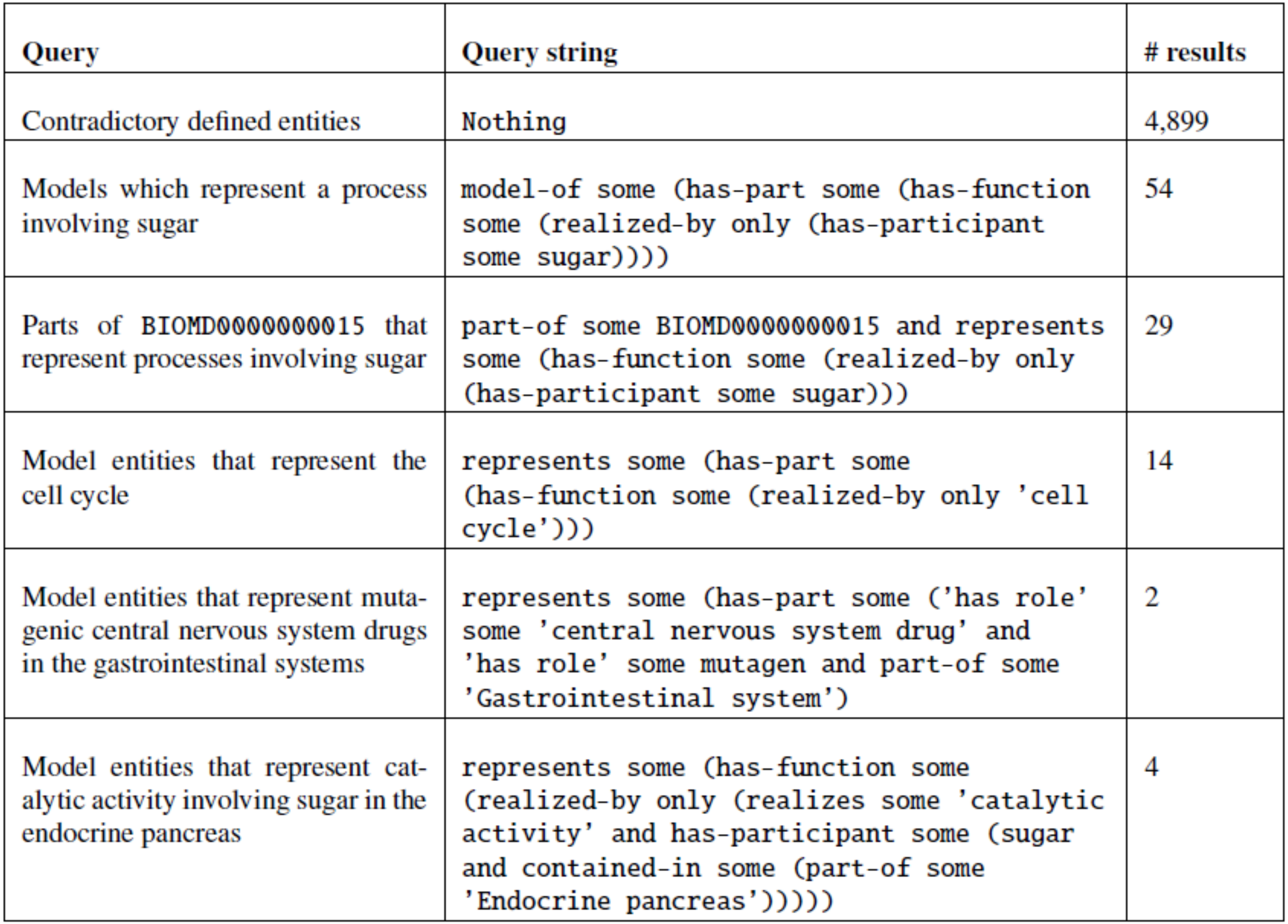
Tho this, we add:

- GO: ATP + Water the only inputs (universal quantification)
- ChEBI: Water, ATP, alpha-D-glucose 6-phosphate are all different (disjointness)

BIOMD0000000176 and BIOMD0000000177 models of anaerobic glycolysis in yeast.

- “ATP” input to “ATPase” reaction, which is annotated with ATPase activity. The species “ATP”, however, is mis-annotated with Alpha-D-glucose 6-phosphate (CHEBI: 17665), not with ATP.

Answering questions



Query	Query string	# results
Contradictory defined entities	`Nothing`	4,899
Models which represent a process involving sugar	`model-of some (has-part some (has-function some (realized-by only (has-participant some sugar))))`	54
Parts of `BIOMD0000000015` that represent processes involving sugar	`part-of some BIOMD0000000015 and represents some (has-function some (realized-by only (has-participant some sugar)))`	29
Model entities that represent the cell cycle	`represents some (has-part some (has-function some (realized-by only ’cell cycle’)))`	14
Model entities that represent mutagenic central nervous system drugs in the gastrointestinal systems	`represents some (has-part some (’has role’ some ’central nervous system drug’ and ’has role’ some mutagen and part-of some ’Gastrointestinal system’)`	2
Model entities that represent catalytic activity involving sugar in the endocrine pancreas	`represents some (has-function some (realized-by only (realizes some ’catalytic activity’ and has-participant some (sugar and contained-in some (part-of some ’Endocrine pancreas’)))))`	4

Outcomes

The SBML-derived ontologies can be

i) checked for their consistency, thereby uncovering erroneous curations

ii) infer attributes and relations of the substances, compartments and reactions beyond what was originally described in the models

iii) answer sophisticated questions across a model knowledge base

Current work

- Add more of SBML annotations (aka qualifiers) to the ontology
  - Specify the role of species in the processes that they participate in by extracting the roles from the semantically annotated kinetic expressions
- Simultaneously query knowledge and simulation results
  - currently time course
- Increase performance
  - Fit the transformation into one of the more computationally efficient OWL2 profiles

Questions?