# An OWL knowledge base for classifying and querying collections of physiological models: A prototype human physiome

Maxwell Lewis Neal[1,*], Daniel L. Cook[2,3] and John H. Gennari[3]
[1] Department of Bioengineering, University of Washington, Seattle, WA, USA
[2] Department of Physiology and Biophysics, University of Washington, Seattle, WA, USA
[3] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA

**ABSTRACT**

The human physiome is envisioned as the quantitative description of the whole of human physiology. Researchers actively working toward achieving this grand challenge have populated publicly available repositories of quantitative physiological models; however, no mechanism has been developed that integrates the information from these models into a single knowledge resource that can be said to represent a physiome. Here we present a mechanism for automatically generating a physiome knowledge base (SemPhysKB) from models collected from the CellML repository, the National Simulation Resource repository, and the BioModels database. Applying description logic axioms and automated classification, we demonstrate that the SemPhysKB can be queried to retrieve models, sub-models and mathematical formulations of interest and provides a set of logically defined physiological reference terms currently unavailable among biomedical ontologies.

## 1 INTRODUCTION

An organism's physiome is defined as the holistic, quantitative description of its physiological and pathophysiological processes (Bassingthwaighte, 2000; Hunter and Borg, 2003). Researchers currently working towards this grand challenge apply quantitative, computational models of normal and disease physiology to describe, mathematically, physiological phenomena across molecular, cellular, tissue and organ scales. If realized, the human physiome would provide a framework for understanding normal and pathological human physiology at a complex, systemic level. It would also provide modelers with a wide set of reusable, quantitative models. Rather than re-code existing models from scratch, or develop them anew, researchers could browse from a collection of curated models and repurpose appropriate models or model parts suited to their research.

Describing an organism's physiome is an obviously ambitious task that requires integrating quantitative physiological knowledge across biological scales. Currently, physiome researchers approach this challenge by creating manageable computational models of subsets of physiological phenomena and store them within publicly available repositories. Together, the models in these repositories could be said to represent a growing subset of multiple organisms' physiomes; however, there is no way to interrogate this broader collection of physiological models as a unified knowledge resource. In other words, while the models in these repositories may represent an integration of some aspects of a physiological system, it is difficult to determine the connections between individual models and to view the biological topology of an organism's physiome as a whole. As a more integrated, alternative approach, we envision the human physiome as a single, publicly available knowledge resource that contains all the quantitative and semantic information represented by all curated, quantitative models of human physiology in existence. In our vision it should be possible to query across all these curated models and interrogate the physiome as an integrated, continuous representation of physiology.

In this study we describe the creation of a prototype physiome, the Semantically-integrated Physiome Knowledge Base (SemPhysKB). We automatically constructed our prototype by collecting the quantitative and semantic information available in a set of 94 human physiology models obtained from the BioModels database (Le Novere, et al., 2006), the CellML model repository (Lloyd, et al., 2008) and the National Simulation Resource (NSR) model repository (http://www.physiome.org/Models). Our goal was to demonstrate the potential utility of the SemPhysKB by querying it for information useful to biosimulation modelers. We demonstrate that we can create and use the SemPhysKB to search for models and sub-models (model parts with distinct biological or computational meaning) of interest *across disparate repositories*, and that we can investigate the different mathematical formulations for particular biological concepts. These queries represent searches that a modeler might perform to discover existing models related to their research and to examine the models' underlying mathematical assumptions. We also demonstrate how the SemPhysKB can serve as a repository of physiological property terms that are formally defined using description logics in OWL. Many of these properties are fundamental to the domain of biomedicine, but have resisted formal representation in the corpus of existing biomedical ontologies.

## 2 METHODS

To create a knowledge resource that describes human physiology in quantitative terms, we must link mathematical representations of physiological phenomena to machine-readable semantic descriptions of those phenomena.

* To whom correspondence should be addressed: mneal@uw.edu

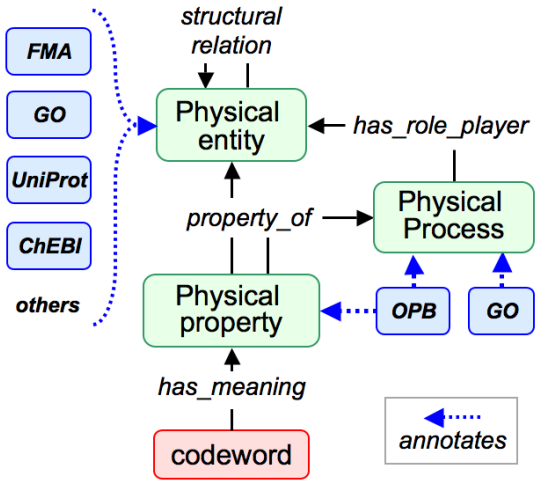


**Fig. 1.** The SemSim framework captures the biological semantics of parameters and variables (codewords) in biosimulation models using annotations against reference ontologies.

Whereas available models can be reduced to a common mathematical syntax (despite differences in model code syntax), model integration across scales and domains requires a comprehensive semantics for expressing and computing on the biological meaning implicit in the mathematics. A purely mathematical representation of the human physiome would be un-interpretable without an implicit understanding of what each codeword or model subcomponent represented, biologically. To make these links explicit, we have developed the SemSim framework (Gennari, et al., 2008; Gennari, et al., 2011; Neal, et al., 2009), a logical architecture for describing biological models that formally links their mathematical statements to their biophysical meanings as represented in available ontologies.

## 2.1 The SemSim framework

The SemSim approach was motivated by a need to modularize biological models so they can be searched, decomposed and recomposed in biological, rather than mathematical, terms. In other words, SemSim was created to enable the semantic interoperability of a wide range of biosimulation models. Achieving this level of interoperability requires that all model parameters and variables (codewords) and subcomponents that represent distinct biological observables be precisely defined in a machine-readable manner. As illustrated in Figure 1, we annotate codewords against reference ontology terms to make these definitions explicit and machine-readable. However, precisely defining many of these codewords requires a level of semantic expressivity that is unavailable among existing biomedical ontologies. To address this limitation we have developed a grammar for constructing more descriptive, *composite* annotations (Gennari, et al., 2011) which link the physical property represented by the variable (e.g., chemical concentration) to the physical entity or physical process that it is a property of (e.g., glucose or glucokinase activity). For example, while a variable representing arterial blood pressure could be annotated using a singular ontology term from a resource such as SNOMED-CT (Ruch, et al., 2008) *aortic* blood pressure is, to our knowledge, not a concept available in biomedical ontologies. In the SemSim approach, we compose this term by linking existing ontology concepts from the Ontology of Physics for Biology (OPB) (Cook, et al., 2011) and the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) with formal relations from the OPB and the Relations Ontology (RO) (Smith, et al., 2005).

*OPB:Fluid pressure*
*<OPB:physicalPropertyOf>*
*FMA:Portion of blood*
*<RO:contained_in>*
*FMA:Lumen of aorta*

Annotating a SemSim model with such machine-readable annotations makes it possible to search for and retrieve model components using standard biological language (Gennari, et al., 2012), as opposed to relying on idiosyncratic naming schemes used by individual modelers. It also becomes possible to automatically extract model subcomponents based on their place within the biological organization of the model, and to automatically recognize biologically-consistent interfaces between merged models (Neal, 2010). This latter feature can make the construction of integrated models from multiple sources a more automated, less error-prone process (Beard, et al., 2012).

We have chosen to use OWL to implement SemSim models. While SemSim models could theoretically be encoded in less expressive, lighter-weight languages such as RDF-S or SKOS, we use OWL because one of our research goals has been to use automated reasoning for model retrieval, decomposition and composition tasks. The work presented here represents the first substantial application of automated classification and reasoning of SemSim models.

## 2.2 Constructing the SemPhysKB

We amassed 94 annotated SemSim models derived from human physiological simulation models to create the prototype physiome. Most of these were automatically converted from a set of curated Systems Biology Markup Language (SBML) (Hucka, et al., 2003) models from the BioModels database (n=83). We retrieved these models using the "Browse curated models using Taxonomy" feature on the BioModels website. We downloaded those models under the *Homo sapiens* heading, then used SemGen, our software tool for creating and composing SemSim models (Gennari, et al., 2011), to automatically convert them to the SemSim OWL format. The BioModels curators have annotated many of the chemical species, reactions and compartments repre-

sented in these models against reference ontologies and we preserve these annotations during SBML-to-SemSim conversion.

We collected 11 models from either the CellML model repository or the NSR repository that were used in previous SemSim-related projects (Beard, et al., 2012; Gennari, et al., 2008; Gennari, et al., 2011; Neal, 2010; Neal, et al., 2009). These include simulations of cardiac electrophysiology, cardiovascular dynamics, respiratory mechanics, blood gas handling, blood-tissue gas exchange, hemostatic blood product dynamics, the baroreflex, the chemoreflex, and hormone dynamics of renal regulation of blood pressure. As shown in the workflow of Figure 2, in step 1 we used SemGen to convert these models into the SemSim OWL format; however, because models from these repositories are not generally annotated against reference ontologies, we annotated them according to their original reference publications and direct communication with model authors. Combined with the 83 BioModels models, the 94 SemSim models in our collection represent physiological processes across physical and temporal scales and across several research domains.

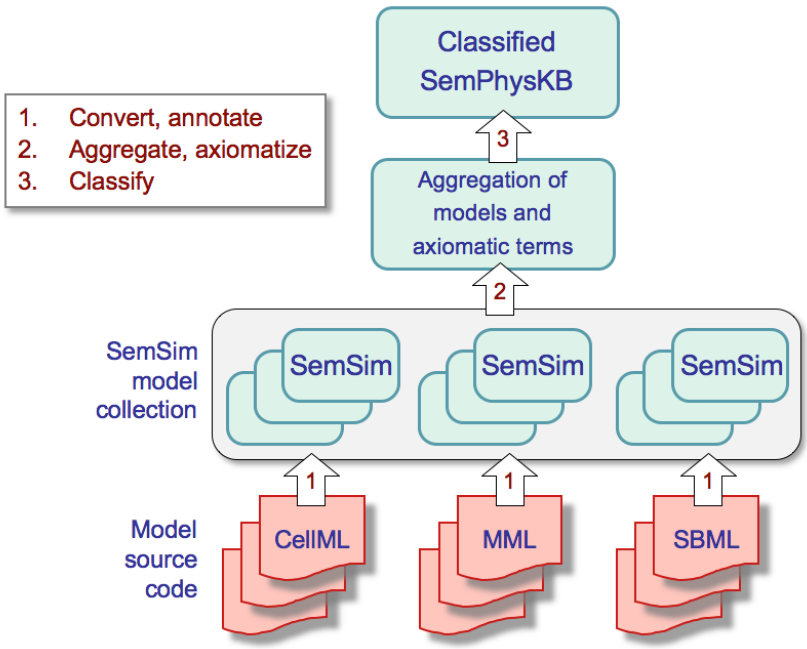


**Fig. 2.** Workflow for creating the SemPhysKB. Models coded in CellML, the Mathematical Modeling Language (MML) used in the NSR repository, and SBML are converted into the SemSim format and annotated. Axioms from individual models are aggregated into a single OWL file, along with axiomatic compilations of their composite annotations. The file is then classified for querying.

We created a Java program that uses the OWL API (http://owlapi.sourceforge.net/) and a prototype SemSim API to aggregate all our SemSim models into a new, single ontology. For each model processed, we created a new OWL individual that represents that particular model and made it a child of *SemSim_model,* a new class created for the SemPhysKB. To associate these instances of specific models with the logical axioms that describe their SemSim implementation, we link the model individuals to the *SemSim:Data_structure* individuals encoded in the model. This linking is needed to query the SemPhysKB for models and sub-models of interest, as described in section 2.3.

For each composite annotation in our model set, we created a new, singular class within the SemPhysKB that is logically defined according to the composite annotation. For example, the composite annotation for aortic blood pressure presented in section 2.1 was asserted to be a new subclass of *OPB:Fluid pressure* as defined according to an OWL "equivalent classes" axiom:

```
OPB:Fluid pressure and
(OPB:physicalPropertyOf some FMA:Portion of
blood and (RO:contained_in some FMA:Lumen
of aorta))
```

To make these axiomatic terms more user-friendly, we provide human-readable names for them using the free-text definitions of the model codewords that represent these properties. (For example, "Aortic blood pressure" is one human-readable name associated with the logically defined class above.) We generated these singular classes for each composite annotation so these biological concepts can be reused for annotating simulation models and data sources. Rather than require annotators to create composite annotations anew, the SemPhysKB can provide their single-term equivalents. Figure 2 diagrams the overall workflow we used to construct the knowledge base.

## 2.3 Querying the SemPhysKB

We loaded the raw SemPhysKB OWL file into Protégé version 4.1.0 for classification and querying. We used the Pellet reasoner (Sirin, et al., 2007) to classify the knowledge base and the DL Query tab for test queries. We created three example queries that represent searches a modeler might perform during the model construction process. The first retrieves all models and sub-models that represent some blood pressure. The second retrieves all models and sub-models that represent some concentration of calcium. These two queries were designed to test model retrieval across physical scales, from the macroscopic (blood dynamics) to the microscopic (calcium kinetics), and across model repositories. The third query retrieves the different mathematical

formulations for left ventricular pressure. This query exposes whether model variables are computed in a way that harmonizes with prevailing modeling assumptions.

## 3 RESULTS

Our prototype human physiome is 59 MB, takes approximately 13 minutes to classify with Pellet, and returns query results in approximately four seconds on a 2.8 GHz Intel Core 2 Duo MacBook Pro. During the initial classification of the knowledge base the reasoner returned an error indicating that individual members of the Gene Ontology class "protein binding" (GO:0005515) could not be classified consistently. We determined that the inconsistency occurred because two models, BIOMD#186 and BIOMD#187, annotated a chemical species against this GO term. In SemSim models, reference ontology terms (e.g., GO:0005515) are sub-classed under either of the disjoint *SemSim:Physical entity*, *SemSim:Physical process* or *SemSim:Physical property* classes. The inconsistency arose because the GO term was sub-classed as a physical entity for BIOMD#186 and BIOMD#187 but was sub-classed as a physical *process* for a third model, BIOMD#88, which uses "protein binding," appropriately, as an annotation for a reaction. Because *SemSim:Physical entity* and *SemSim:Physical process* are disjoint, the reasoner could not subclass the GO term and its OWL individuals under both. This was an unexpected result demonstrating that the SemPhysKB can act as a tool for checking the consistency of biosimulation model annotations. Accordingly, the BioModels curation team has confirmed and corrected the inconsistencies.

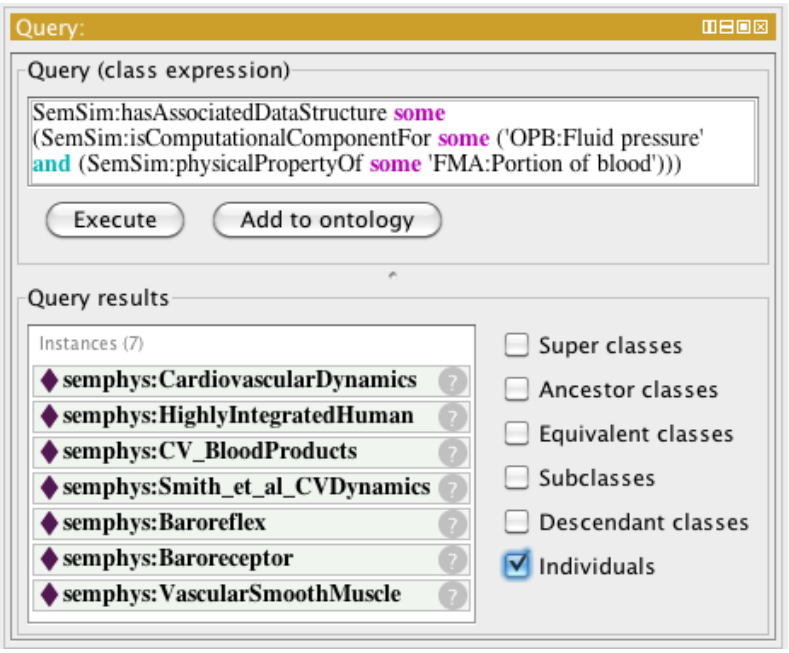


**Fig. 3.** Query formulation and results for retrieving models and sub-models in the SemPhysKB that simulate blood pressure.

### 3.1 Query results

Figure 3 shows the results of the query for models and sub-models that represent some blood pressure. The query returned seven models, all of which are from the group of 11 SemSim models that were collected from either the CellML or NSR repositories. Five of these models focus on hemodynamics of the cardiovascular system and two simulate baroreceptor regulation of blood pressure. None of these models currently have sub-models specified and so the query returned no sub-models.

Figure 4 shows the results of the query for models and sub-models that simulate the concentration of calcium. This query returned 33 instances, some from each of the BioModels, CellML, and NSR repositories.

Figure 5 shows the results of the query for various mathematical formulations that calculate left ventricular pressure. The query returned four different members of the *SemSim:Computation* class representing four different instances of equations for left ventricular pressure, two of which are shown; one instance depends on pericardial pressure and the other does not.

### 3.2 Singular forms of composite annotations

The SemPhysKB defines 429 classes that represent physiological properties compiled from composite reference ontology terms. Examples include the aortic blood pressure class described in section 2.2, concentration of cytosolic potassium in the cardiac myocyte, molar flow rate of glucokinase activity, and the partial pressure of $CO_2$ in respiratory air.

## 4 DISCUSSION

In this study we have described an OWL knowledge base that integrates quantitative physiological information across modeling languages, physical scales, and research domains. Previously, we assisted in the development of the SBML Harvester, a tool for integrating knowledge from a large-scale biosimulation model repository into a single ontology (Hoehndorf, et al., 2011). Our subsequent efforts with the SemPhysKB presented here are the first attempt at integrating such information across model repositories and across physical scales and domains. Although our demonstration focuses exclusively on models of human physiology, our methods can be used to create physiome knowledge bases for non-human organisms.

Whereas typical reference knowledge bases and ontologies rely on mining expert knowledge as articulated in textbooks, journal articles and other lexical sources, the SemPhysKB mines the physiological hypotheses articulated mathematically in biosimulation models - a fundamentally

different expression of formal physiological knowledge. Two key benefits accrue from this approach. First, mapping model contents to ontology terms exposes the models' biophysical meanings. Second, mapping computational dependencies amongst variables as logical dependencies amongst physical properties exposes differences in how modelers conceptualize the interplay between biological entities and processes.

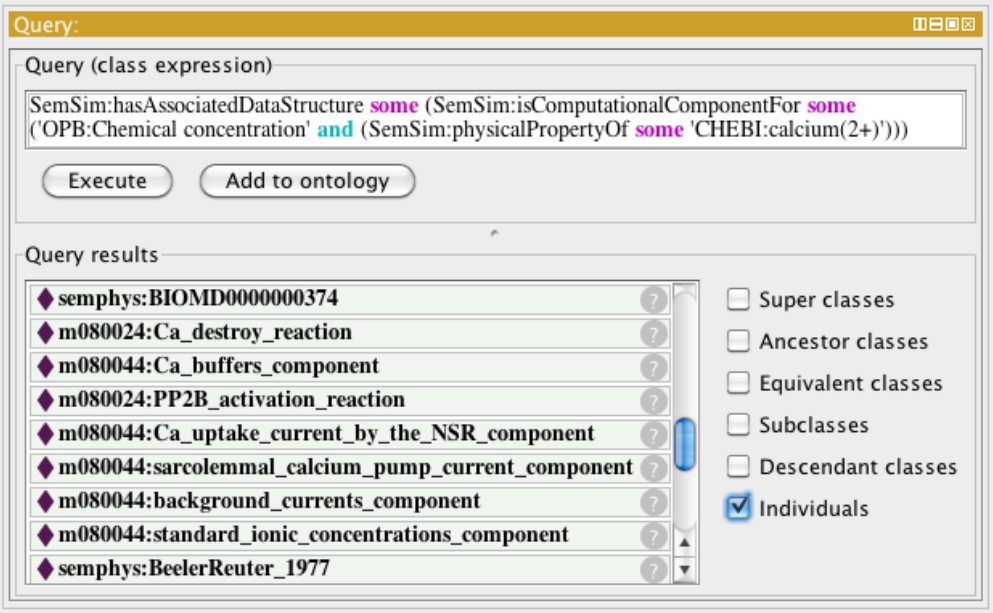


**Fig. 4.** Query formulation and results for retrieving models and sub-models that simulate calcium concentration. The "semphys" prefix indicates a model; the other prefixes indicate sub-models.

The SemPhysKB contains all the information necessary to extract fully functional versions of the source models. Thus, the SemPhysKB can act as a repository of reusable, physiological models and modeling parts, each with a unique namespace for its computational and biological elements. In future work we plan to test the programmatic extraction of full models and sub-models from the SemPhysKB.

As a prototype human physiome, the SemPhysKB contains quantitative descriptions of human physiology coupled to logical axioms that describe the computational dependencies amongst models. Thus, the SemPhysKB architecture integrates the physiological content of curated models across repositories so that investigators can also discover and explore the *qualitative* links between physiological phenomena. Rather than a simple flat list of models organized by annotation, the SemPhysKB can provide an integrated, gene-to-organism perspective on physiological processes and their physical entity participants. We have previously described this type of view as a PhysioMap (Cook, et al., in press), a directed graph linking physiological processes and their participants that represents the qualitative connections between biophysical phenomena based on the mathematical formulations parsed from models. We contend that the SemPhysKB, by providing a collection of linked PhysioMaps, will allow a better understanding of physiological functions. For example, a cardiac physiologist interested in understanding the role of calcium in heart contraction could use this view to investigate all the processes in the cardiac myocyte that involve calcium. For easy access, we plan to package the SemPhysKB with our SemGen software and implement a query-generator tool for accessing its contents. Rather that constructing queries in Manchester OWL syntax as done here, a simple user-friendly interface requiring a minimum of ontological expertise will be provided. Furthermore, as composite annotations accrue in the SemPhysKB, they will be made available for reuse to facilitate subsequent model annotation tasks.

Currently, our knowledge base contains 429 classes that represent composite annotations - about five classes per model. This number was lower than expected, given that the CellML and MML models we annotated usually contained more than five composite annotations. We found that the SBML model annotations for chemical species and reactions sometimes lacked sufficient precision, due, in part, for lack of exact terms in biomedical ontologies. For example, BioModels curators often use the "isVersionOf" relation rather than the more exact "is" relation in their annotations. SemSim models require precise semantics; thus, during SBML-to-SemSim conversion, we only include ontology annotations for chemical species and reactions if the "is" relation is used. Otherwise, we instantiate a custom species or reaction that does not include an ontology-based definition. If a composite annotation includes a custom term, we do not convert it into a singular form when creating the SemPhysKB. This omission of the less precise annotations in SBML models is one reason the number of classes representing composite annotations was lower than expected.

A limitation of our work is that the SemSim format is currently restricted to representing algebraic and ordinary differential equation models. In the future we hope to extend SemSim to accommodate spatially resolved models. We also acknowledge that the vast majority of models used are chemical network models where we could exploit the substantial annotation efforts of the BioModels team. As the number of annotated models in the CellML and NSR repositories grows, we plan to include them in the SemPhysKB.

In this first exercise, we have mined the code of biosimulation models as exclusive inputs to our SemPhysKB. However, there is no reason that the SemPhysKB cannot be augmented by inputting hypotheses by hand or by text-mining information from the literature. In practice, we envision such purpose-built SemPhysKBs that could be aggregated to nucleate a true multiscale, multidomain, semantic model of the human phyiome. Given that the SemSim architecture preserves the computational as well as semantic aspects of models, the aggregation of SemSim models into physiome-level resources offers an architecture that solves many of the challenges faced by efforts such as the Virtual Physiological Human and Virtual Physiological Rat projects in developing holistic, computational representations of physiology.

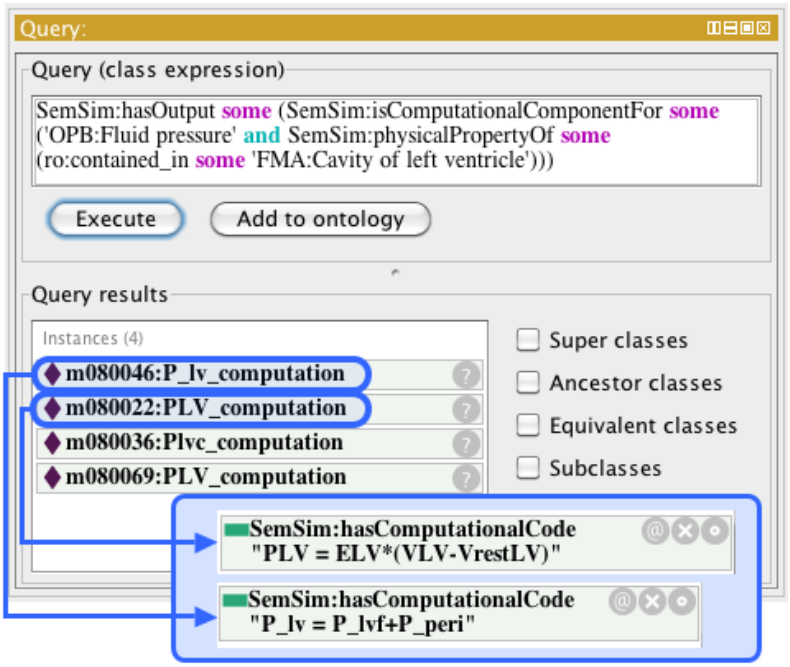


**Fig. 5.** Query expression and results for retrieving the different formulations for left ventricular blood pressure available in the SemPhysKB. In one model the codeword representing this pressure (P_lv) depends on pericardial pressure (P_peri), the codeword in the other model (PLV) does not.

## ACKNOWLEDGEMENTS

We thank Robert Hoehndorf, Michal Galdzicki, Bryan Bartley and Evren Sirin for their suggestions on improving the manuscript. Work partially funded by NIH grant NIH-NIGMS GM094503.

## REFERENCES

Bassingthwaighte, J.B. (2000) Strategies for the physiome project, *Annals of Biomedical Engineering*, 28, 1043-1058.

Beard, D.A., Neal, M.L., Tabesh-Saleki, N., et al. (2012) Multiscale modeling and data integration in the Virtual Physiological Rat Project, *Annals of Biomedical Engineering*, 40, 2365-2378.

Cook, D.L., Bookstein, F.L. and Gennari, J.H. (2011) Physical Properties of Biological Entities: An Introduction to the Ontology of Physics for Biology, *PLoS ONE*, 6, e28708.

Cook, D.L., Neal, M.L., Hoehndorf, R., et al. (2013) Representing physiological processes and their participants with PhysioMaps, *Journal of Biomedical Semantics*, 4(Suppl 1):S2.

Gennari, J.H., Neal, M.L., Carlson, B.E. and Cook, D.L. (2008) Integration of multi-scale biosimulation models via light-weight semantics, *Proceedings of the Pacific Symposium on Biocomputing*, 13, 414-425.

Gennari, J.H., Neal, M.L., Galdzicki, M. and Cook, D.L. (2011) Multiple ontologies in action: Composite annotations for biosimulation models, *Journal of Biomedical Informatics*, 44, 146-154.

Gennari, J.H., Neal, M.L., Hoehndorf, R., et al. (2012) Discovering model-model connections in biological model repositories, *Proceedings of the Virtual Physiological Human conference, Sept 18-20, London, UK.*

Hoehndorf, R., Dumontier, M., Gennari, J.H., et al. (2011) Integrating systems biology models and biomedical ontologies, *BMC Systems Biology*, 5.

Hucka, M., Finney, A., Sauro, H.M., et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models, *Bioinformatics*, 19, 524-531.

Hunter, P.J. and Borg, T.K. (2003) Integration from proteins to organs: the Physiome Project, *Nature Reviews Molecular Cell Biology*, 4, 237-243.

Le Novere, N., Bornstein, B., Broicher, A., et al. (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems, *Nucleic Acids Research*, 34, D689-D691.

Lloyd, C.M., Lawson, J.R., Hunter, P.J. and Nielsen, P.F. (2008) The CellML model repository, *Bioinformatics*, 24, 2122-2123.

Neal, M.L. (2010) Modular, semantics-based composition of biosimulation models. Ph.D. dissertation. University of Washington.

Neal, M.L., Gennari, J.H., Arts, T. and Cook, D.L. (2009) Advances in semantic representation for multiscale biosimulation: A case study in merging models, *Proceedings of the Pacific Symposium on Biocomputing*, 14, 304-315.

Rosse, C. and Mejino, J.L.V. (2003) A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy, *Journal of Biomedical Informatics*, 36, 478-500.

Ruch, P., Gobeill, J., Lovis, C. and Geissbuhler, A. (2008) Automatic medical encoding with SNOMED categories, *BMC medical informatics and decision making*, 8, S6.

Sirin, E., Parsia, B., Grau, B.C., et al. (2007) Pellet: A Practical OWL-DL reasoner, *Web Semantics: science, services and agents on the World Wide Web*, 5, 51-53.

Smith, B., Ceusters, W., Klagges, B., et al. (2005) Relations in Biomedical Ontologies, *Genome Biology*, 6, R46.