Building Concordant Ontologies for Drug Discovery

Hande Küçük-McGinty[1], *Saurabh Mehta*[2,3], *Yu Lin*[2], *Nooshin Nabizadeh*[2], *Vasileios Stathias*[2], *Dusica Vidovic*[2], *Amar Koleti*[2], *Christopher Mader*[2], *Jianbin Duan*[1,2], *Ubbo Visser*[1], *Stephan Schürer**[2,5]

1 Department of Computer Science, University of Miami, Coral Gables, FL
2 Center for Computational Science, University of Miami, Coral Gables, FL
3 Department of Applied Chemistry, Delhi Technological University, Delhi, India
5 Department of Molecular and Cellular Pharmacology, Miller School of Medicine, University of Miami, FL

***Abstract*— In this study we demonstrate how we interconnect three different ontologies, the BioAssay Ontology (BAO), LINCS Information FramEwork ontology (LIFEo), and the Drug Target Ontology (DTO). The three ontologies are built and maintained for three different projects: BAO for the BioAssay Ontology Project, LIFEo for the Library of Integrated Network-Based Cellular Signatures (LINCS) project, and DTO for the Illuminating the Druggable Genome (IDG) project. DTO is a new ontology that aims to formally describe drug target knowledge relevant to drug discovery. LIFEo is an application ontology to describe information in the LIFE software system. BAO is a highly accessed NCBO ontology; it has been extended formally to describe several LINCS assays. The three ontologies use the same principle architecture that allows for re-use and easy integration of ontology modules and instance data. Using the formal definitions in DTO, LIFEo, and BAO and data from various resources one can quickly identify disease-relevant and tissue-specific genes, proteins, and prospective small molecules. We show a simple use case example demonstrating knowledge-based linking of life science data with the potential to empower drug discovery.**

***Keywords*— drug discovery; bioinformatics; cheminformatics**

I. INTRODUCTION

Big data are ubiquitous in business, technology and science. Life science research data are no exception. However, the nature of research data, in particular in the life sciences brings additional challenges due to broad diversity of data types and formats, the quick evolution of knowledge and advancements in technologies to generate data. Despite large investments in information systems in the pharmaceutical industry and non-profit research organizations, the difficult problem of describing, organizing, integrating, analyzing diverse, fast evolving and large scale data in the context of biological knowledge remains a critical and not fully solved challenge.

In this study we demonstrate in a simple case study how to represent and organize such data better by using Semantic Web technologies. Although this approach is not novel, we contribute by leveraging three ontologies developed in our group and that are largely aimed at addressing different aspects of drug discovery data.

The BioAssay Ontology (BAO) [3] has been developed to formally describe knowledge of chemical biology assays and screening results using Description Logic (DL) [14] and OWL (OWL2.0) [17]. The first version of BAO [9] focused on High Throughput Screening (HTS) assays and contained descriptions of many assays from PubChem. BAO since evolved to better integrate with other ontologies and better align with established upper level models and improve usability. BAO was also extended to support profiling assays such as those in LINCS [21]. The systems biology nature of LINCS data required a formal model to describe the relations of cells, disease, tissues and relevant bio-molecules, such as proteins, transcribed genes, used in different roles the various assays. The LINCS Information FramEwork (LIFE) [20] was developed to process, integrate, query, and explore this data. The LIFE application ontology (LIFEo) was developed as a knowledge model to capture the relevant relationships to facilitate this functionality. The Drug Target Ontology (DTO) is being developed as a reference framework to formalize knowledge about drug targets in the context of simple assays and more complex model systems; it is developed as part of the Illuminating the Druggable Genome (IDG) project [22]. For example DTO can readily be used in BAO or in LINCS to describe protein targets in an assay or known targets of small molecule drugs.

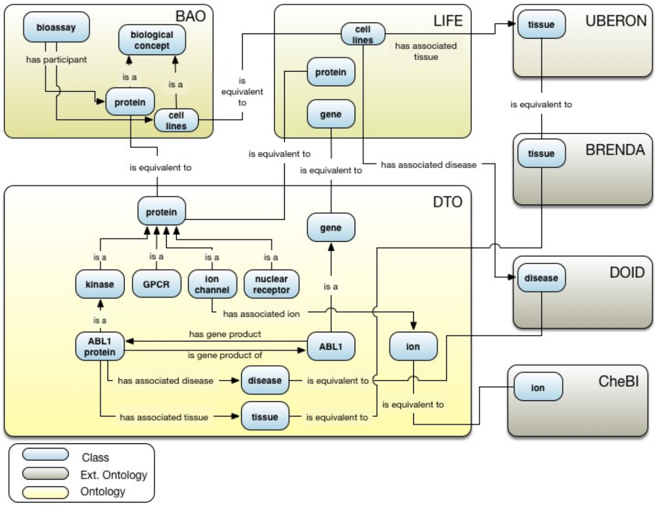

Figure 1 BAO, LIFEo, and DTO with select external ontologies

II. Methods

All three ontologies (BAO, LIFEo, and DTO) are built using the OWL language. They all use the same approach of modular architectures to facilitate maintenance and re-use [1]. For the construction of DTO we developed tools (using Java and the OWL API) to semi-automate the ontology building process; modularization in DTO further separates algorithm-generated components from expert-generated ones.

Modeling of the data requires a complex and sequential approach. BAO contains formal definitions of assay-related concepts, LIFEo contains axioms for various bio-molecules and their relationships to the assays, cells, tissues, etc, while DTO contains axioms to formalize drug target knowledge. The ontologies have been designed to complement each other and to be compatible. All ontologies make extensive use of external ontologies.

The concepts for BAO ontology are either created by our group, or extracted from external ontologies and used with their own URIs. LIFEo formally describes data generated in the LINCS project's Data and Signature Generating Centers (DSGCs). Finally, for DTO we formally describe drug target data that are the focus of the IDG Project. We further use public databases, such as UniProt [27], in an effort to cross reference and map terms.

We used Protégé [29] to add the manual axioms, Fact ++ [12] reasoner to reason the query view that we created and used Virtuoso [37] as our triple store.

III. Results

A. BioAssay Ontology (BAO)

BAO [3] was designed and implemented to axiomize knowledge about bioassays. As the content expanded with the addition of LINCS assays, an architectural change was implemented to the ontology so that it can maintain its core while importing external ontologies for existing information. Current version of BAO has >3300 classes, >420,000 axioms and 165 object properties.

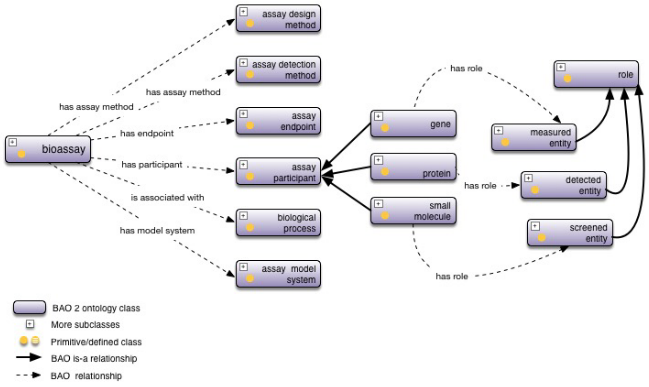

Figure 2 Concepts for Modeling of Bioassays in BAO.

Briefly, the implemented modular architecture divides the ontology into layers, starting with the vocabularies, followed by modules with BAO-native axioms, and finally, different views of the ontology can be created by combinations of modules that can contain the native as well as the external axioms. An important feature of this modularization is that it allows to create a BFO-founded version for ontology authoring and integration with other resources, but also a BAO-native version for users; since most users are not familiar with BFO terms. In addition to the ontology architecture of BAO, we aimed to standardize the assay descriptions by creating metadata and design patterns for the formal definitions. LINCS assays were axiomized in BAO using the model previously described [38] and shown in Figure 2.

B. LINCS Information FramEwork Ontology (LIFEo)

The Library of Integrated Network-Based Cellular Signatures (LINCS) project aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents. LINCS aims to use computational tools to integrate this diverse information into a comprehensive view of normal and disease states that can be applied for the development of new biomarkers and therapeutics [21]. The diverse datasets of LINCS are generated via various assays; in each assay biological molecules occur in different *roles*; formalizing this information facilitates the integration of this data and allows asking potentially novel complex queries.

To accomplish this, we have formalized LINCS assays in BAO. The systems biology nature of LINCS data required a new model we called the LIFE ontology. LIFEo is an application ontology designed to handle the different biological molecules and model systems (in particular cell lines and cells), their relationships to other concepts, such as disease and tissue and assays and their roles. LIFEo contains >49000 classes, >132,000 axioms and 62 object properties (including direct and indirect imports from BAO, DTO and other, external ontologies). By using a modular approach, LIFEo is aiming to create a useful model of how the different metadata components in LINCS align across the entire project.

The first version of LIFEo supported eight assays, namely: KINOMEscan, KiNativ, Cue Signal Response, P100, 2-Color-Apoptosis, 3-Color-Apoptosis, Cell Cycle State, and Cell Growth Assays [21].

Although LINCS assays are diverse with respect to assay technology, the detection method, model systems, and metadata entities, the main point of LIFEo is to provide an integrative model that facilitates context-specific analysis by formalizing the most important relationships.

In addition to the gene and protein modules of the LIFE ontology, we have a module for cells that are used as model systems in the various assays. They are grouped into four main categories: stem cells, primary cells, differentiated cells, and cell lines. Cell lines then are grouped by using the organs from which they were derived using UBERON.

C. Drug Target Ontology (DTO)

DTO is being created as part of the IDG project. An important goal of the IDG project is to catalyze the development of chemical probes and drug development entry points for understudied, yet relevant protein targets in the four most commonly targeted protein families (G-protein coupled receptors (GPCR), nuclear receptors, ion channels, and kinases) by integrating all available information and making it available as actionable knowledge. The current version of DTO consists of asserted class hierarchies of the ~1800 protein targets, > 13,000 classes and > 214,000 axioms.

DTO is designed to work with other ontologies, such as BAO and thus can be used to describe proteins in LINCS assays.

DTO content is being curated from various sources and the details of the development of DTO will be described elsewhere. DTO content is further annotated and linked by various ontologies. To facilitate the construction of DTO, we wrote various scripts using Java to retrieve information from databases and ontologies. These databases include UniProt and NCBI databases for ENTREZ IDs for the genes, and ChEBI [33] for ions and other small molecules. Further information from the DISEASES and TISSUES databases are incorporated [36].

We retrieved the proteins, with their tissue and disease relationships with the confidence scores that are given to the relationships. We put this data into our database and use this information while creating the ontology's axioms that refer to the probabilistic values of the relationships.

a) Knowledge Modeling of the Drug Target Ontology:

Drug Target Ontology (DTO) uses various external databases and ontologies to retrieve information. The information is retrieved via web-based applications and in-house-built scripts. The data that is used to build DTO is then housed in an internal database. To facilitate ontology development and maintenance, such as frequent updates and synchronization to other data sources, we use Java, OWL API and Jena to build modules of the ontology in a semi-automated. The details of the specific modularization architecture are shown in Figure 3.

b) Improved Modular Architecture for the Drug Target Ontology:

In contrast to BAO, which is primarily constructed manually by experts formalizing axioms, DTO integrates lots of information from different resources. We therefore separated a further module category built using only automated scripts. These are imported into modules that incorporate expert-built axioms. This way, updates from the database will not overwrite expert-modeled content.

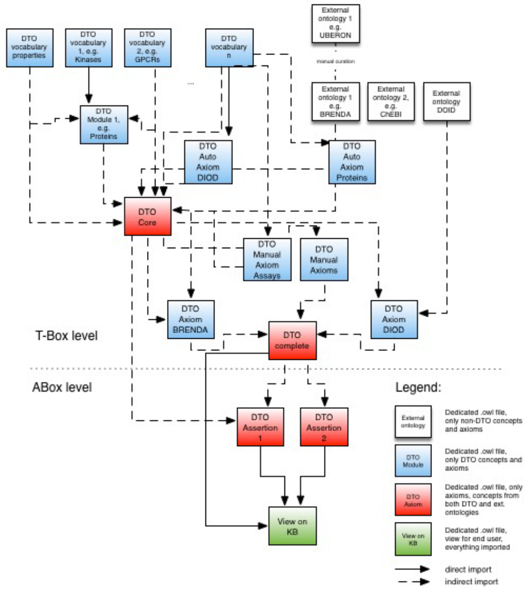
DTO vocabulary properties
DTO vocabulary 1, e.g. Kinases
DTO vocabulary 2, e.g. GPCRs
...
DTO vocabulary n
External ontology 1 e.g. UBERON
manual curation
External ontology 1 e.g. BRENDA
External ontology 2, e.g. ChEBI
External ontology DOID
DTO Module 1, e.g. Proteins
DTO Auto Axiom DIOD
DTO Auto Axiom Proteins
DTO Core
DTO Manual Axiom Assays
DTO Manual Axioms
DTO Axiom BRENDA
DTO complete
DTO Axiom DIOD
T-Box level
ABox level
DTO Assertion 1
DTO Assertion 2
View on KB
Legend:
External ontology
Dedicated .owl file, only non-DTO concepts and axioms
DTO Module
Dedicated .owl file, only DTO concepts and axioms
DTO Axiom
Dedicated .owl file, only axioms, concepts from both DTO and ext. ontologies
View on KB
Dedicated .owl file, view for end user, everything imported
direct import
indirect import

Figure 3 Modular Architecture of DTO

First, we determine the abstract horizon between TBox and ABox. Tbox contains modules, which define the conceptualization without dependencies. These modules are self-contained and well-defined with respect to the domain and they contain concepts, relations, and individuals. We can have *n* of these modules.

Second, once the *n* modules are defined, the modules with axioms that can be generated automatically are created. Those modules have interdependent axioms. At this level one could create any number of gluing modules, which import other modules without dependencies or with dependencies. It also is self-contained. This means that there is no outside term or relationship in the files.

Third, this level contains axioms created manually; however the axioms generated are independent and self-contained. The manual modules are an optional level and they inherit the axioms created automatically. A good example of axioms that may be seen in this level are axioms for protein modifications and mutations, which have been challenging modeling questions. At this level, the self-contained DTO_core is also generated with the existing modules.

Fourth, at this level we can design modules that import modules from our domain of discourse, and also from third party ontologies. Third party ontologies could be large, therefore a suitable module extraction method (e.g. Java programs using OWL API and Jena) can be used to extract only part of those ontologies (*vide supra*). We would model this in the DTO_complete level. We can have one DTO_complete file or multiple files, each may be modeled for a different purpose, e.g., tailored for various research groups. Once these ontologies are imported, the alignment takes place. The alignments are defined for concepts and relations using equivalence or subsumption DL constructs. The alignment depends on the domain experts and/or cross-references made in the ontologies. For DTO, the most significant alignment made is between UBERON [23] and BRENDA [24] ontologies for the tissue information.

Fifth, release the TBox based on the modules created from the third phase. Depending on the end-users, the modules are combined without loss of generality. With this methodology we make sure that we only send out physical files that contain our (and the absolute necessary) knowledge.

Sixth, at this level, the necessary modules ABoxes are created. ABoxes can be loaded to a triple store or to a distributed file system in a way that one could achieve pseudo-parallel reasoning.

Seventh, at this level we define *views on the knowledge base*. These are files that contain imports (both direct and indirect) from various TBoxes and ABoxes modules for the end-user. It can be seen as a *view*, using database terminology.

D. Use Case Example Query

Since BAO, LIFEo, and DTO have been constructed using a modular approach, we are able to create different *views* that would help to integrate and query relevant data, for example in the drug-discovery domain. We extracted the LINCS assays from BAO by using Jena [6], and OWL API, used the cell line module from LIFEo, and the targets from Drug Target Ontology (DTO) in order to query the following use case.

1) Query

What are the kinases used in the LINCS assays that measure protein binding and have strong evidence for being associated with cancer? Further, what are the relevant compounds targeting these kinases and are therefore relevant for the disease? What other data support the compound-disease association? The generic example cancer, is meant only for illustrative purposes.

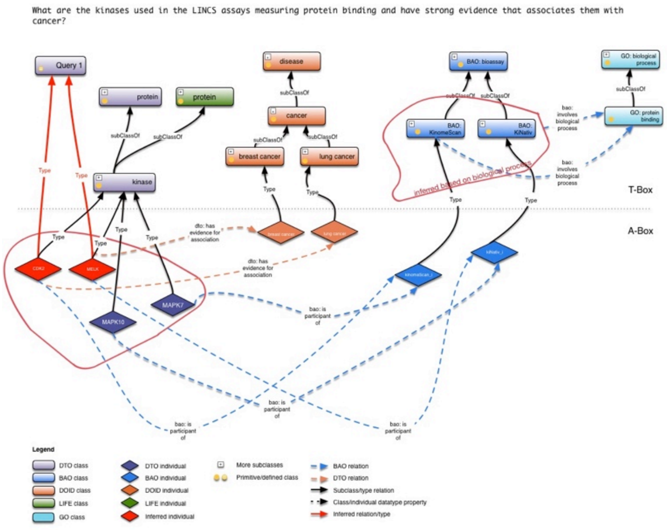

Figure 4 Illustration of the example query using the three ontologies

2) Results

This query aims to retrieve assay specific proteins based on the assays of interest. It works in two parts. In the first part we used the molecular function that is measured (i.e. protein binding) to infer the bioassays of interest. This information is formally described in BAO. We then identified the kinases used in these assays using the LINCS data axioms related with kinases in LIFEo. Finally by using DTO, we get the intersection of this subset of kinases with the kinases that have strong evidence for associations with cancer. The disease and tissue information related with different genes and proteins is formalized in DTO as described above. We further analyzed the results by using the compound data in the LIFEo. We queried the compounds used both in KINOMEscan (KS) and KiNativ (KN) assays.

Table 1 Results for the Query

Assays	**Results for Kinases**		
	Protein Name	***Gene***	***Disease***
KS& KN	cyclin-dependent kinase 2	CDK2	cancer
KS& KN	cyclin-dependent kinase 16	CDK16	cancer
KS	death-associated protein kinase 3	DAPK3	cancer
KS& KN	insulin-like growth factor 1 receptor	IGF1R	cancer
KS& KN	interleukin-1 receptor-associated kinase 1	IRAK1	cancer
KS	maternal embryonic leucine zipper kinase	MELK	cancer

Table 2 Small Molecules and Proteins that are used in KS and KN Assays

Small Molecule Name	Small Molecule LINCS ID	Proteins
OTSSP167	LSM-6340	CDK2, DAPK3, IGF1R, IRAK1, MELK
5z-7-oxozeaenol	LSM-43344	CDK2, CDK16, IRAK1,MELK
XMD16-144	LSM-43287	CDK2, IRAK1
Sorafenib	LSM-1008	CDK2, IRAK1,CDK16
GW-5074	LSM-1029	CDK2, IRAK1
SB590885	LSM-1046	CDK2, IRAK1
PLX-4720	LSM-1049	CDK2, IRAK1
AZ-628	LSM-1050	CDK2, IRAK1
PLX4032	LSM-1068	CDK2, IRAK1
NPK76-II-72-1	LSM-1070	CDK2, IRAK1
Torin1	LSM-1079	CDK2, IRAK1
Torin2	LSM-1080	CDK2, IRAK1
XMD11-50	LSM-1086	CDK2, IRAK1
JWE-035	LSM-1092	CDK2, IRAK1
XMD8-85	LSM-1093	CDK2, IRAK1
XMD8-92	LSM-1094	CDK2, IRAK1
XMD-12	LSM-1106	CDK2, IRAK1
Ibrutinib	LSM-1129	CDK2
XMD11-85h	LSM-5577	CDK2, IRAK1
QL-X-138	LSM-5803	CDK2, IRAK1
WZ3105	LSM-5970	CDK2, CDK16, IRAK1
HG-6-64-01	LSM-43248	CDK2, IRAK1

We combined the resulting kinases of Query1 with the 22 compounds. Table 2 shows the specific kinases that were targets of the same assays as the 22 compounds used both in KINOMEscan and KiNativ assays.

In summary, assays with their molecular functions of interest are axiomized in BAO. Kinases have assay related axioms in LIFEo, which we retrieve as the second step in the query. We then explore more about the proteins by using the axioms related with their associated disease information from DOID [8] encoded in the DTO. As cell lines are linked to diseases, compounds can further be identified based on the growth inhibition assays.

Our results showed us that with the three ontologies, BAO, LIFEo, and DTO, we were able to connect different data types and content related to drug-discovery data. The uniform architecture along with the complex and sequential modeling templates we use for the diverse types of data, allows us to combine different modules and create different views in order to reach the components of interest faster.

IV. Discussion

Here we presented three ontologies built for three related, yet different projects, and how they can work together in queries crossing several concepts important for drug discovery. This is facilitated by the similar modular architectures of the ontologies, which enable their integration of diverse information into a triple store.

BAO has been developed to formalize complex chemical biology assays, such as HTS assays, which are one of the primary methods to identify novel entry points for drug discovery projects. BAO facilitates re-use of this data. LIFEo provide a simple model to address the systems biology aspects, specifically relations of disease model systems, tissues, protein targets, small molecules and assays. DTO describes drug targets formally and integrates information from many sources. All ontologies utilize external ontologies, which serve as an integration point, such as disease and tissue.

BAO was used in the BioAssay Research Database (BARD) software system [19] and it is used in several projects and organizations [36] after we had initially demonstrated its use in the semantic software application BAOSearch (http://baosearch.ccs.miami.edu/). We have also used BAO to describe omics profiling assays in the LINCS program via the LINCS Information Framework (LIFE) (http://life.ccs.miami.edu/).

DTO provides a formal classification of four protein families based on function and phylogenetic and describes their clinical classifications and relations to diseases and tissue expression. DTO is already used in the IDG main Portal Pharos (https://pharos.nih.gov/) and the TinX software application (http://newdrugtargets.org/) to prioritize drugs by novelty and importance. DTO is publicly available at http://drugtargetontology.org/, where it can be visualized and searched.

We have illustrated how DTO, LIFEo, and BAO and included external ontologies are used to describe, integrate, and query drug discovery related data. We are also in the process of integrating these knowledge models with the recently released LINCS Data Portal (http://lincsportal.ccs.miami.edu/). For the purpose of this paper, we have integrated only a part of the available LINCS data in a local triple store to demonstrate the basic concept of our approach of integration. Much more work is required to fully integrate and model all LINCS data. As we expand the LINCS and DTO knowledge models, we can construct more complex queries. A particular goal is to enable the context-sensitive integration and querying of data. We will also integrate further ontologies for example the Cell Line Ontology (CLO) to formalize LINCS cell lines.

We continue to develop BAO and DTO to maximize their utility for the research community. We are constructing a more advanced LINCS MetaData Ontology towards the goal of a comprehensive systems-based model of LINCS signature and drug discovery data.

Acknowledgment

This work was supported by grants U54CA189205 (Illuminating the Druggable Genome Knowledge Management Center, IDG-KMC) and U54HL127624 (BD2K LINCS Data Coordination and Integration Center, DCIC). The IDG-KMC is a component of the Illuminating the

Druggable Genome (IDG) project (https://commonfund.nih.gov/idg) awarded by the NCI. The BD2K LINC DCIC is awarded by the National Heart, Lung, and Blood Institute through funds provided by the trans-NIH Library of Integrated Network-based Cellular Signatures (LINCS) Program (http://www.lincsproject.org/) and the trans-NIH Big Data to Knowledge (BD2K) initiative (http://www.bd2k.nih.gov). Both IDG and LINCS are NIH Common Fund projects.

References

[1] Abeyruwan, Saminda, et al. "Evolving BioAssay Ontology (BAO): modularization, integration and applications." *Journal of biomedical semantics* 5.1 (2014): 1.

[2] Oprea, Tudor I., et al. "Systems chemical biology." *Nature chemical biology*3.8 (2007): 447-450.

[3] Visser Ubbo; Abeyruwan Saminda; Vempati Uma; Smith Robin P; Lemmon Vance; Schurer Stephan C., "BioAssay. Ontology, a semantic description of bioassays and high-throughput screening results " *BMC bioinformatics* 12 (2011): 257.

[4] Harmar, Anthony J., et al. "IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels." *Nucleic acids research* 37.suppl 1 (2009): D680-D685.

[5] Stevens, Robert, Carole A. Goble, and Sean Bechhofer. "Ontology-based knowledge representation for bioinformatics." *Briefings in bioinformatics* 1.4 (2000): 398-414.

[6] Carroll, Jeremy J., et al. "Jena: implementing the semantic web recommendations." *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, 2004.

[7] Hille, Bertil. *Ion channels of excitable membranes*. Vol. 507. Sunderland, MA: Sinauer, 2001.

[8] Kibbe, Warren A., et al. "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data." *Nucleic acids research* 43.D1 (2015): D1071-D1078.

[9] Schürer, Stephan C., et al. "BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets." *Journal of biomolecular screening* 16.4 (2011): 415-426.

[10] Gene Ontology Consortium. "Gene ontology consortium: going forward."*Nucleic acids research* 43.D1 (2015): D1049-D1056.

[11] Brinkman, Ryan R., et al. "Modeling biomedical experimental processes with OBI." *Journal of biomedical semantics* 1.1 (2010): 1.

[12] Tsarkov, Dmitry, and Ian Horrocks. "FaCT++ description logic reasoner: System description." *Automated reasoning*. Springer Berlin Heidelberg, 2006. 292-297.

[13] Dumontier, Michel, and Melanie Courtot. "Proceedings of the 8th International Workshop on OWL: Experiences and Directions." (2011).

[14] Baader, Franz. *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[15] Hitzler, Pascal, Markus Krotzsch, and Sebastian Rudolph. *Foundations of semantic web technologies*. CRC Press, 2009.

[16] Gruber, Thomas R., Nicola Guarino, and Roberto Poli. "Formal ontology in conceptual analysis and knowledge representation." *Chapter" Towards principles for the design of ontologies used for knowledge sharing" in Conceptual Analysis and Knowledge Representation* (1993).

[17] W3C Owl Working Group. "{OWL} 2 Web Ontology Language Document Overview." (2009).

[18] Xiang, Zuoshuang, et al. "OntoFox: web-based support for ontology reuse."*BMC research notes* 3.1 (2010): 175.

[19] Howe, E. A., et al. "BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types."*Nucleic acids research* (2014): gku1244.

[20] Library of Integrated Network-based Cellular Signatures (LINCS) Information FramEwork, http://life.ccs.miami.edu/life/, Last visited on 05/05/2016

[21] Library of Integrated Network-Based Cellular Signatures (NIH LINCS) program, http://www.lincsproject.org/, Last visited on 05/05/2016

[22] Illuminating the Druggable Genome | NIH Common Fund, https://commonfund.nih.gov/idg/index, Last visited on 05/05/2016

[23] Mungall, Christopher J., et al. "Uberon, an integrative multi-species anatomy ontology." *Genome Biol* 13.1 (2012): R5.

[24] Maglott, Donna, et al. "Entrez Gene: gene-centered information at NCBI."*Nucleic acids research* 33.suppl 1 (2005): D54-D58.

[25] Gremse, Marion, et al. "The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources." *Nucleic acids research* 39.suppl 1 (2011): D507-D513.

[26] Gaulton, Anna, et al. "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic acids research* 40.D1 (2012): D1100-D1107.

[27] UniProt Consortium. "UniProt: a hub for protein information." *Nucleic acids research* (2014): gku989.

[28] Callahan, Alison, et al. "RegenBase: a knowledge base of spinal cord injury biology for translational research." *Database* 2016 (2016): baw040.

[29] Protégé, http://protege.stanford.edu/, Last visited on 06/10/2015

[30] Smith, Barry, Anand Kumar, and Thomas Bittner. "Basic formal ontology for bioinformatics." *Journal of Information Systems* (2005): 1-16.

[31] Dumontier, Michel, and Robert Hoehndorf. "Realism for scientific ontologies."*FOIS*. 2010.

[32] Gaulton, Anna, et al. "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic acids research* 40.D1 (2012): D1100-D1107.

[33] Degtyarenko, Kirill, et al. "ChEBI: a database and ontology for chemical entities of biological interest." *Nucleic acids research* 36.suppl 1 (2008): D344-D350.

[34] Wache, Holger, et al. "Ontology-based integration of information-a survey of existing approaches." *IJCAI-01 workshop: ontologies and information sharing*. Vol. 2001. 2001.

[35] Clark AM, Bunin BA, Litterman NK, Schürer SC, Visser U. (2014) Fast and accurate semantic annotation of bioassays exploiting a hybrid of machine learning and user confirmation. *PeerJ.*,**14**(2), e524.

[36] Pletscher-Frankild, Sune, et al. "DISEASES: Text mining and data integration of disease–gene associations." *Methods* 74 (2015): 83-89.

[37] Erling, Orri, and Ivan Mikhailov. "RDF Support in the Virtuoso DBMS."*Networked Knowledge-Networked Media*. Springer Berlin Heidelberg, 2009. 7-24.

[38] Vempati, Uma D., et al. "Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the BioAssay Ontology (BAO)." *PloS one* 7.11 (2012): e49198.

Based on all the reviewers' request, we added pages with larger figures.
We couldn't see a way to add supplementary materials.

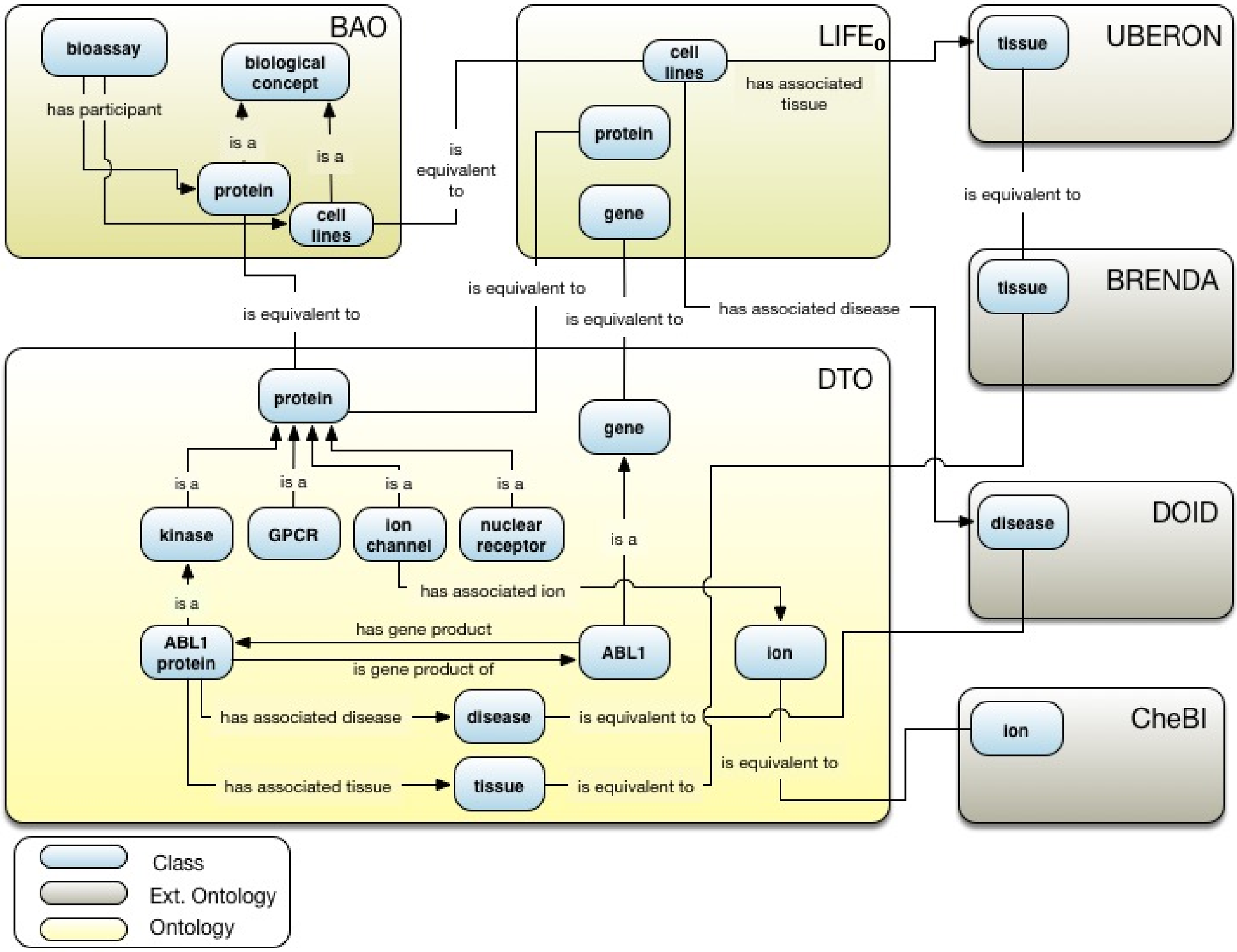

Figure 1

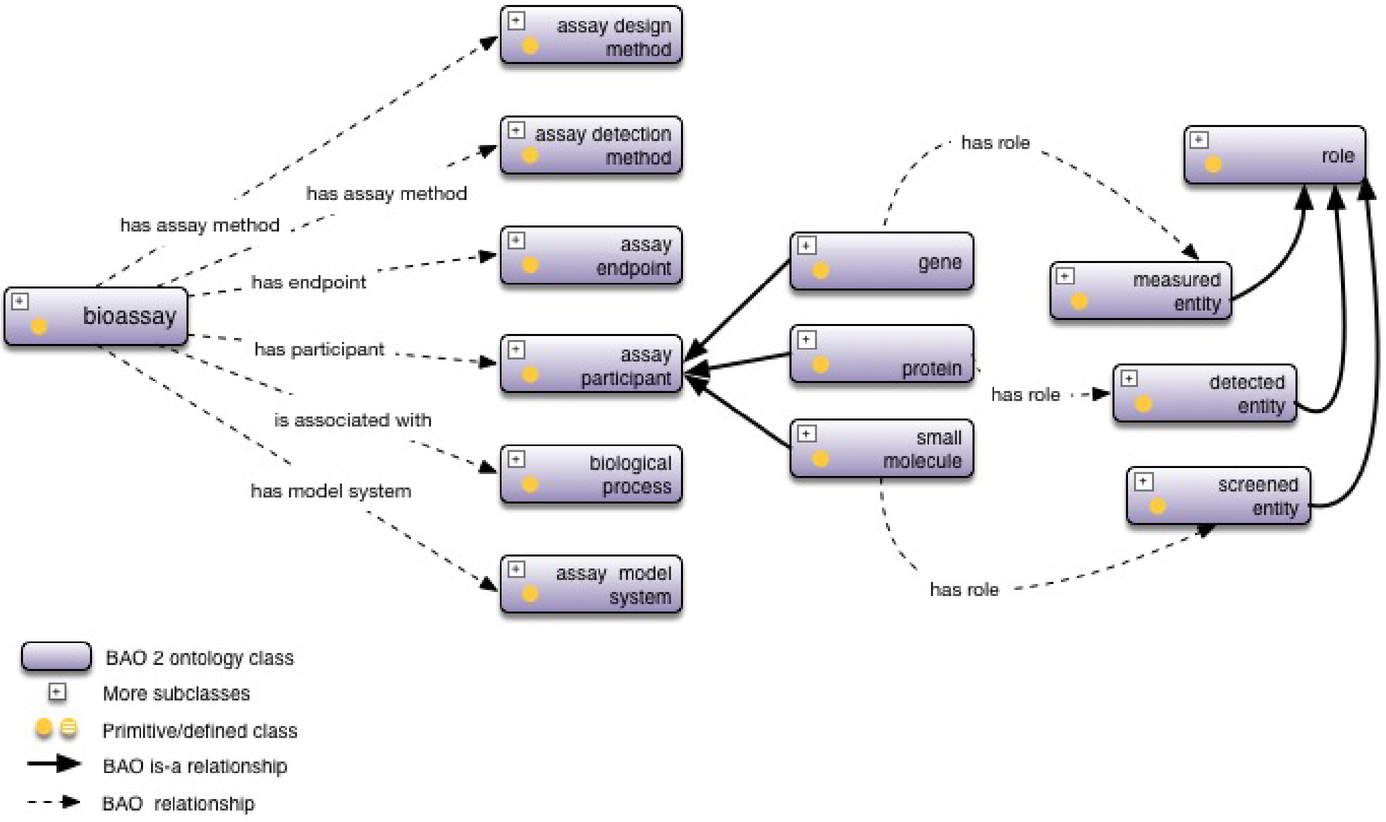
assay design method
assay detection method
has assay method
has assay method
assay endpoint
has endpoint
bioassay
has participant
assay participant
is associated with
biological process
has model system
assay model system
gene
protein
small molecule
has role
has role
has role
role
measured entity
detected entity
screened entity
BAO 2 ontology class
More subclasses
Primitive/defined class
BAO is-a relationship
BAO relationship

Figure 2

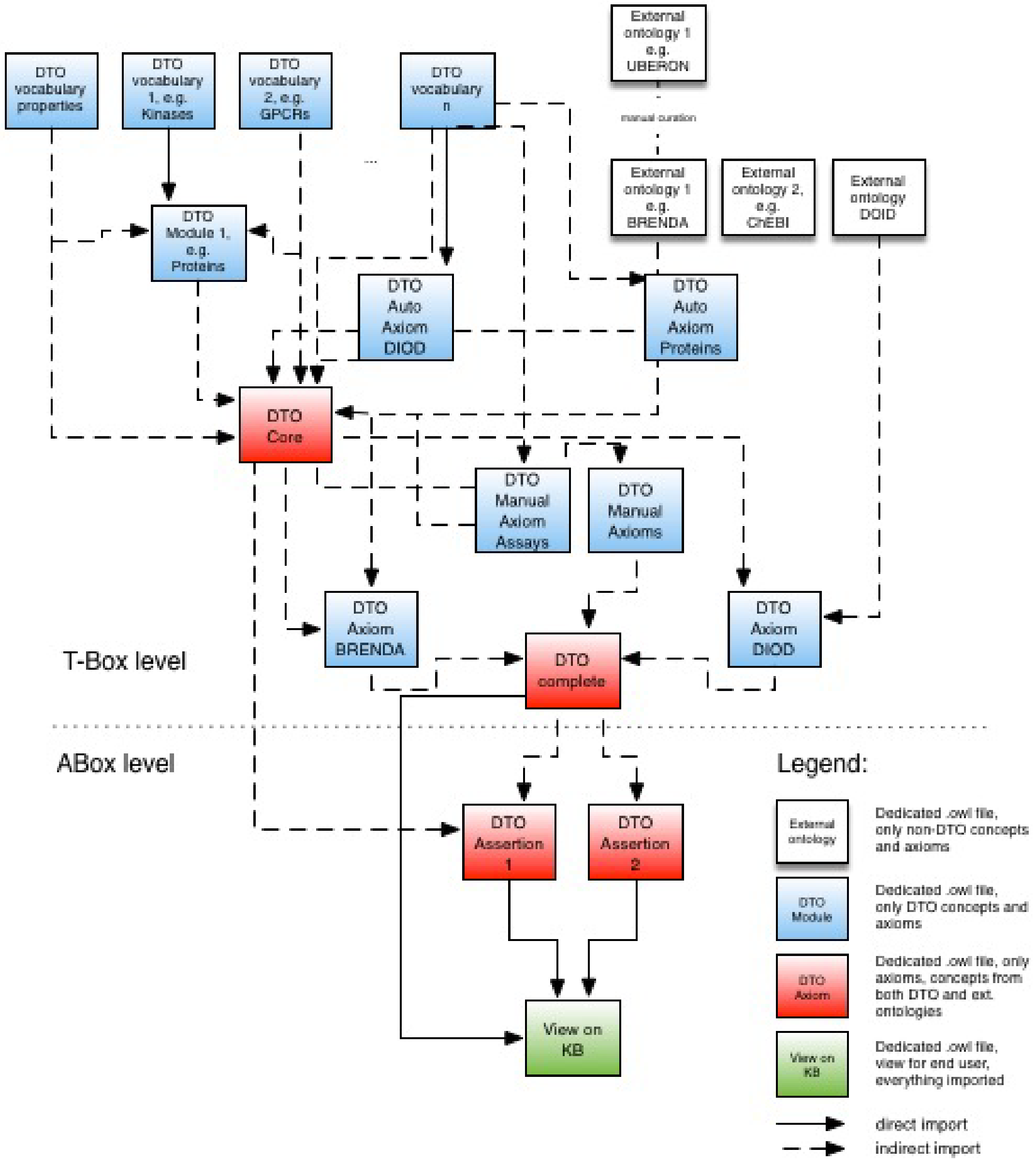
DTO vocabulary properties
DTO vocabulary 1, e.g. Kinases
DTO vocabulary 2, e.g. GPCRs
...
DTO vocabulary n
External ontology 1 e.g. UBERON
manual curation
External ontology 1 e.g. BRENDA
External ontology 2, e.g. ChEBI
External ontology DOID
DTO Module 1, e.g. Proteins
DTO Auto Axiom DIOD
DTO Auto Axiom Proteins
DTO Core
DTO Manual Axiom Assays
DTO Manual Axioms
DTO Axiom BRENDA
DTO complete
DTO Axiom DIOD
T-Box level
ABox level
DTO Assertion 1
DTO Assertion 2
View on KB
Legend:
External ontology
Dedicated .owl file, only non-DTO concepts and axioms
DTO Module
Dedicated .owl file, only DTO concepts and axioms
DTO Axiom
Dedicated .owl file, only axioms, concepts from both DTO and ext. ontologies
View on KB
Dedicated .owl file, view for end user, everything imported
direct import
indirect import

Figure 3

What are the kinases used in the LINCS assays measuring protein binding and have strong evidence that associates them with cancer?

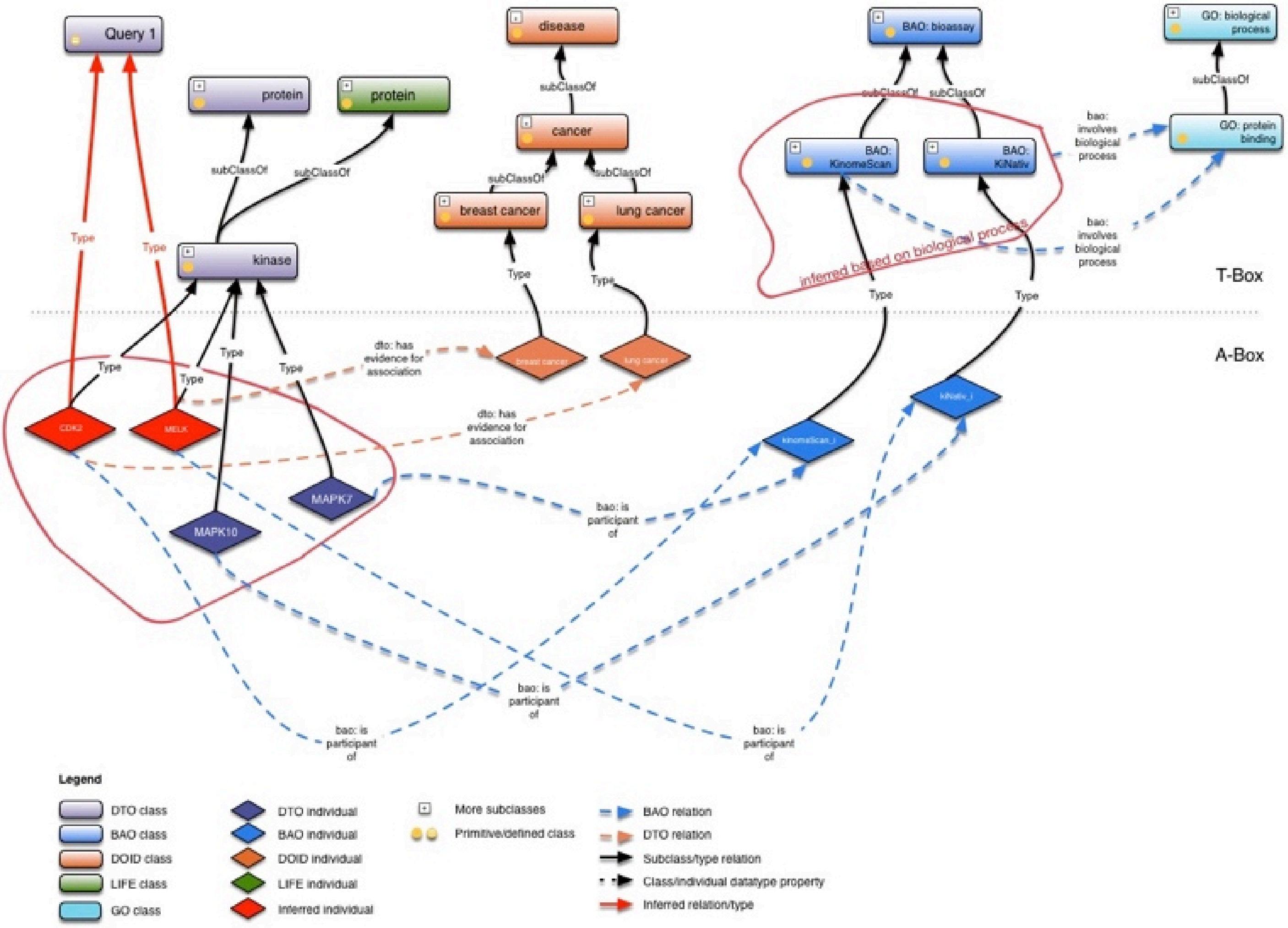

Figure 5