FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases

Maxime Delmas, Olivier Filangi, Nils Paulhe, Florence Vinson, Christophe Duperier, William Garrier, Paul-Emeric Saunier, Yoann Pitarch, Fabien Jourdan, Franck Giacomoni, et al.

Bioinformatics, 37(21), 2021, 3896–3904
doi: 10.1093/bioinformatics/btab627
Advance Access Publication Date: 3 September 2021
Original Paper

Databases and ontologies

FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases

Maxime Delmas [1], Olivier Filangi[2], Nils Paulhe [3], Florence Vinson[1], Christophe Duperier[3], William Garrier[4], Paul-Emeric Saunier [4], Yoann Pitarch[5], Fabien Jourdan [1], Franck Giacomoni [3] and Clément Frainay[1,*]

[1]Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse 31300, France, [2]IGEPP, INRAE, Institut Agro, Université de Rennes, Domaine de la Motte, Le Rheu 35653, France, [3]Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand F-63000, France, [4]ISIMA, Campus des Cézeaux, Aubière 63177, France and [5]IRIT, Université de Toulouse, Cours Rose Dieng-Kuntz, Toulouse 31400, France

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on March 2, 2021; revised on August 16, 2021; editorial decision on August 19, 2021; accepted on September 1, 2021

Abstract

Motivation: Metabolomics studies aim at reporting a metabolic signature (list of metabolites) related to a particular experimental condition. These signatures are instrumental in the identification of biomarkers or classification of individuals, however their biological and physiological interpretation remains a challenge. To support this task, we introduce FORUM: a Knowledge Graph (KG) providing a semantic representation of relations between chemicals and biomedical concepts, built from a federation of life science databases and scientific literature repositories.

Results: The use of a Semantic Web framework on biological data allows us to apply ontological-based reasoning to infer new relations between entities. We show that these new relations provide different levels of abstraction and could open the path to new hypotheses. We estimate the statistical relevance of each extracted relation, explicit or inferred, using an enrichment analysis, and instantiate them as new knowledge in the KG to support results interpretation/further inquiries.

Availability and implementation: A web interface to browse and download the extracted relations, as well as a SPARQL endpoint to directly probe the whole FORUM KG, are available at https://forum-webapp.semantic-metabolomics.fr. The code needed to reproduce the triplestore is available at https://github.com/eMetaboHUB/Forum-DiseasesChem.

Contact: clement.frainay@inrae.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The typical metabolomics data analysis workflow takes as input large collections of spectral data and returns a list of compounds, whose abundances vary significantly between experimental conditions (Giacomoni *et al.*, 2015). The subsequent step requires an extensive literature survey to decipher the relationship between these compounds and the biological condition under study. In contrast with the upstream processing of raw data, few tools are available to assist metabolomics users in processing this new high load of information. The work proposed here aims at closing this gap.

Among all omics sciences, metabolomics stands the closest to the phenotype as it captures the response of the system to environmental changes, physiological disorders or stimulus at the highest level (Fiehn, 2002). The profiling of small molecules in a tissue, a fluid or an organism using high throughput untargeted analysis is a powerful tool for the identification of biomarkers (Ludwig and Hummon, 2017). It offers valuable insights for diagnosis and patient prognosis, and opens promising opportunities for drug design and physiopathology understanding. However, the current challenge of metabolomics is to exploit these metabolic signatures to their full extent by connecting the metabolic disruptions to the observed phenotypes, giving insights into the underlying mechanisms (Johnson *et al.*, 2016). Although models of

metabolism become more accurate, yielding valuable insight from metabolic network analysis, this challenging issue requires knowledge beyond biochemical transformations alone, connecting metabolites to higher-order phenotypical information.

Scientific literature, as a primary way of reporting and sharing scientific progress, is an obvious valuable resource for the interpretation of metabolic profiles. However, the overwhelming amount of articles that need to be inspected to interpret lists of metabolites is at the core of the current metabolomics workflow limitations. This calls for computational support to deal with the information overload, which requires modeling knowledge in an appropriate form. With digitalization, annotation and indexing of scientific literature, information can be automatically extracted, paving the way for more machine-readable knowledge representation. In addition, the federation of structured data from various sources to create an exploitable centralized knowledge base could empower large-scale-assisted reasoning, and thus constitute an interesting step toward closing the gap in metabolomics.

Networks have been proven useful in various domains for storing information and modeling knowledge (Neumann and Prusak, 2007). The resulting Knowledge Graphs (KG) are especially suited for information retrieval tasks that require bridging information through indirect relationships in complex networks. A KG is a formal and structured representation of knowledge as a graph of entities and relations, commonly described in a triple formalism (Subject-Predicate-Object). In a KG, each entity, property or relation can also be described by a semantic layer, precisely defining the related concepts using structured and standardized vocabularies, usually from the appropriate ontologies (Hoehndorf *et al.*, 2015). The last decade has seen the rise of available representation of knowledge in such formalism, highlighting the potential of Semantic Web technology to aggregate and exploit these datasets (Kanza and Frey, 2019; Wu and Yamaguchi, 2014). The most important data providers in life sciences propose the content of their databases in RDF format (Fu *et al.*, 2015; Willighagen *et al.*, 2013), a standard of the Semantic Web, particularly well-suited for building KG.

One of the most used repositories of knowledge is the PubMed database that hosts more than 30 million scientific articles from life science. Although articles hold relevant information to make sense of metabolomic data, extracting it from unstructured text is a complex issue (Talib *et al.*, 2016). Fortunately, these articles are manually annotated by *National Library of Medecine* (NLM) trained experts, using MeSH descriptors (Medical Subject Headings) to describe their main topics. The MeSH Thesaurus [Ontology and Thesaurus are both structured vocabularies, but thesaurus have a lower power of reasoning than an ontology (Kim and Beck, 2006).] is a controlled vocabulary containing a broad range of biomedical concepts, hierarchically organized with parent–child relationships, from generic descriptors (e.g. *Cardiovascular Diseases*) up to representatives of specific concepts (e.g. *Coronary Occlusion*). This metadata is primarily used for information retrieval purposes, fueling the PubMed search engine. Being computationally readable, unambiguous and less noisy than text, as well as also being accessible as RDF, several resources exploit them for supporting data interpretation by performing statistical analysis of corpora's MeSH distribution (Smalheiser and Bonifield, 2016; Zhou *et al.*, 2014).

Beyond the recent surge in data availability, a tremendous effort has also been put into the interoperability of life-science resources by extensive cross-references between databases and matching tools [e.g. UniChem (Chambers *et al.*, 2013), MetaNetX (Moretti *et al.*, 2016) and *E-utilities* (https://www.ncbi.nlm.nih.gov/books/NBK25500/) and BridgeDb (van Iersel *et al.*, 2010)]. Among these databases lies PubChem (Kim *et al.*, 2019), the world's largest free chemistry database, providing datasets that gather data and properties associated with chemical compounds (Fu *et al.*, 2015), including articles mentioning them in PubMed (Kim *et al.*, 2016). By navigating the cross references between resources like PubChem and PubMed, it is thus possible to link compounds to biomedical concepts. Moreover, these two entities are semantically described by ontologies or thesauri, with definitions, synonyms and other metadata, increasing the interoperability and data integration between them (Hoehndorf *et al.*, 2015). By leveraging the Semantic Web framework, relationships between concepts in the MeSH hierarchy can also be taken into account. The same can be achieved by integrating semantic annotation of compounds, using chemical classification ontologies such as ChEBI (Hastings *et al.*, 2016) and ChemOnt (Djoumbou Feunang *et al.*, 2016). These relationships enable some degree of abstraction which allows, for instance, reasoning using chemical families rather than individual compounds, raising new hypotheses through in-class generalization. Moreover, this abstraction can be implicitly performed by the KG's reasoner using the *true-path rule*. Basically, the *true-path rule* states that if an entity is annotated to a class, it implicitly also belongs to its parent classes. This rule is also used in the context of information retrieval: the Pubmed search engine uses this property to retrieve, for example, articles indexed with the MeSH descriptor 'Parkinson's disease' from a query targeting articles about the broader concept of 'Neurodegenerative Diseases' (Lu *et al.*, 2009).

Finally, all of the content of these resources and their semantic descriptions can be combined to build a KG: a network of semantically described entities connected by factual relationships, which hosts the information required to support knowledge discovery for metabolomics.

Some tools have already been developed to extract and propose gene or drug candidates related to pathological phenotypes based on literature evidence and databases (Cheung *et al.*, 2012; Hassani-Pak *et al.*, 2020; Kanza and Frey, 2019; Malas *et al.*, 2019; Sosa *et al.*, 2020). In a context closer to metabolomics, some approaches have also been developed to explore relations between chemicals and biomedical concepts based on the scientific literature, and most of them use text-mining approaches such as PolySearch (Cheng *et al.*, 2008), Alkemio (Gijón-Correas *et al.*, 2014), LimTox (Cañada *et al.*, 2017) or RELigator (Pons *et al.*, 2016). These approaches have a fair sensibility and can identify close relations between terms from co-mentions in single sentences. However, they usually offer a low precision, requiring each association be checked manually through the analysis of the sentences that support it, and usually the whole articles that contain them. The KG is a resource built to infer new hypotheses from large-scale existing associations, meaning that those constituting it must be integrated with the lowest false-positive rate, beyond current text-mining tool performances. Indeed, while a missing link would not necessarily obfuscate a latent relationship thanks to path redundancy in the KG, a spurious link between unrelated concepts could substantially alter the KG structure. The use of the NLM indexation system with the MeSH vocabulary can help to avoid these issues by taking advantage of manual labeling performed by trained experts. The association at the article topic level also grasps the full context and helps to avoid the 'noise' and ambiguities found at the sentence level.

Another tool has taken advantage of the linking of MeSH terms to PubChem compounds: Metab2MeSh (Sartor *et al.*, 2012) was a web server dedicated to the annotation of individual compounds with MeSH descriptors. It was a valuable resource that is unfortunately no longer available at the date of writing. Metab2MeSH aimed at testing the statistical significance of co-occurring concepts and compounds in literature, based on Over-Representation Analysis (ORA). ORA is a widely used approach to bring biological meaning to results, like Pathway enrichment (Xia and Wishart, 2010) or Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005), and is commonly performed on sets tested independently, setting aside relationships among them. Metab2MeSH is based on a similar methodology, effectively ignoring the MeSH thesaurus structure. Consequently, it considers 'thematic sets' of articles distinct from what a user would obtain by querying such a topic in PubMed, while missing higher-level associations that could be inferred. Nevertheless, some ORA methods account for relations between entities. For instance, GO Enrichment Analysis (Mi *et al.*, 2019) is a well-used resource to compute functional enrichments on gene sets, where those related to GO terms are determined by considering their hierarchical relations in the Gene Ontology (Ashburner *et al.*, 2000) using the *true-path rule* (Valentini, 2009). Applying this approach on the MeSH thesaurus or on chemical ontologies can thus have a significant impact on predicted relations by the ORA analysis.

Despite related end goals, none of the cited methods aim at providing a queryable KG to build upon, and they also lack open source implementation. The call for compliance with FAIR principles (*Findable*, *Accessible*, *Interoperable* and *Reusable*) has recently grown strong, as it improves the sharing and exploitation of an increasing amount of available data (Koscielny *et al.*, 2017; Williams *et al.*, 2012). Standing at the crossroads of multiple disciplines, current challenges in metabolomics need to be answered with integrative infrastructures, therefore federating knowledge, data and assertions that should be managed respecting FAIR principles (Rebholz-Schuhmann *et al.*, 2012).

In this way, we have developed FORUM, an open and fully accessible KG which allows the navigation of a network of relations between molecules and biomedical concepts by connecting PubChem, MeSH, ChEBI, ChemOnt and MetaNetX. We provide some examples to illustrate what kind of relations can be extracted from FORUM and see how their connections inside the KG can help to explore new hypotheses. Beyond its support for results interpretation, this is a proof of concept meant to illustrate the potential of open linked data in metabolomics, as FORUM could be easily extended for various purposes and grow with the endorsement of FAIR principles by the metabolomics community.

FORUM currently holds associations between 344 769 PubChem compounds, 6740 ChEBI classes and 2783 ChemOnt classes to 24 382 MeSH descriptors, implying more than 8.7 million of significant associations, with 1 096 106 related to diseases.

All associations are queryable through the web portal at https://forum-webapp.semantic-metabolomics.fr. Data are accessible by connecting to the sftp server at sftp://ftp.semantic-metabolomics.org (see information on web portal) and ready-to-mine using the Endpoint at https://forum.semantic-metabolomics.fr, also allowing to exploit the connectivity between entities in custom requests to explore more complex questions.

2 Materials and methods

2.1 Integrated data resources

The main goal of FORUM, our developed KG, is to provide links between chemical compounds and biomedical concepts supported by literature. It is composed of several inter-connected sub-graphs, each describing particular entities and relations (Fig. 1A). Each entity is identified using a specific and persistent *URI* (Uniform Resource Identifier) and relations between them are described by predicates in RDF triples format. An example of this formalism is presented in Figure 1B, showing some relations between compounds, literature and chemical classes that can be extended from the Glucose record (compound:CID5793) in the KG.

Chemical compounds are described in the PubChem RDF compound graph. It contains more than 103 million unique chemical structures extracted from PubChem Substance. Their properties, such as mass, IUPAC name and InChiKey identifiers, are exposed in the PubChem Descriptor and PubChem InchiKey graphs (triples in blue on Fig. 1B). To provide a semantic representation of chemicals, which makes it possible to infer new relationships (Hoehndorf *et al.*, 2015), we use the ChEBI (ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/chebi.owl) ontology classification available for 116 068 pubchem compounds provided in the PubChem store. We also integrated the ChemOnt (http://classyfire.wishartlab.com/downloads) ontology for 319 189 compounds using the web services of ClassyFire (Djoumbou Feunang *et al.*, 2016) and the InChiKey annotation of compounds. The ChemOnt ontology file, originally formatted in OBO (Open Biomedical Ontologies), was converted to RDF syntax using Protégé (https://protege.stanford.edu/) (v.5.50) to be imported in the KG. In ontologies, hierarchical dependencies between classes are defined using the predicate *rdfs:subClassOf*, as illustrated in Figure 1B, where the Chemont chemical class of *Hexoses* is defined as a sub-class of *Monosaccharides*. PubChem also provides a representation of the Pubmed literature (Fu *et al.*, 2015) associated with PubChem Compounds, Substances or Assays, in the PubChem Reference graph, including the MeSH indexation of each publication (triples in green on Fig. 1B). All data extracted from PubChem RDF can be found on the NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/pubchem/RDF/).

To build an explicit network of compound to article relations, we used the NCBI-Eutils Elink utility (https://dataguide.nlm.nih.gov/eutilities/utilities.html), allowing the extraction of crosslinks between entities from different Entrez databases to define these links. Relations between PubChem compounds and Pubmed articles are provided from complementary sources: by external data contributors [e.g. IBM Almaden Research Center (https://www.research.ibm.com/labs/almaden/index.shtml)], by journal publishers or by matching PubChem chemical names on MeSH annotated to articles (Kim *et al.*, 2016). These links are then instantiated in a RDF graph, named *PMID-CID* (Fig. 1A), with the *cito:discusses* predicate (shown in orange in Fig. 1B), the CiTO ontology being particularly suited to describe bibliographic resources (Peroni and Shotton, 2012). Contributors of each association are also referenced in an alternative graph, named *PMID-CID-endpoint*. The connections between compounds and articles then allow us to extract compounds to MeSH links from the chain of relationships embedded in our federation of resources. Those relations are symmetric, thus allowing to retrieve compounds from a MeSH descriptor as well.

The RDF graph of MeSH Thesaurus (ftp://ftp.nlm.nih.gov/online/mesh/rdf/) (Bushman *et al.*, 2015) provides a machine-readable representation of MeSH concepts, organized as a set of trees, where the *treeNumber* of a MeSH indicates its tree location. Then, the hierarchical relationships between descriptors in the MeSH tree can be accessed through the *meshv:parentTreeNumber* predicate (Fig. 1B). Finally, additional ontologies were also integrated to define the nature of each entity or property used in the KG. For example, CiTO and FaBiO (Peroni and Shotton, 2012) describe bibliographic resources, Cheminf (Hastings *et al.*, 2011) defines chemical properties such as *is_stereoisomer_of*, dcterms (Weibel and Koch, 2000) references metadata, SIO (Dumontier *et al.*, 2014) annotates additional properties and SKOS (Miles and Bechhofer, 2009) is used to instantiate extracted relations.

The KG was built on a Virtuoso triplestore (version 7.20.3229) packaged in a Docker image (https://hub.docker.com/r/tenforce/virtuoso), providing access to data through the SPARQL endpoint. The SPARQL request language supports complex queries and allows to exploit the semantics defined in ontologies to extract stated relations and infer new ones from the KG. It also ensures the accessibility and the use of data and metadata available in the KG, following *FAIR* principles. The total number of articles associated with each compound, chemical class and MeSH descriptor, as well as the co-occurrences between them, was then determined using specific SPARQL requests against the endpoint. For associations with chemical classes, we only considered those with less than one thousand associated compounds, as classes like *Carboxylic acids and derivatives* may be too broad to bring meaningful results.

2.2 Computation of associations

By performing co-occurrence counts using the Virtuoso reasoner, associations between chemical entities or MeSH descriptors and PubMed articles are implicitly propagated to their ancestors in their respective ontologies/thesaurus, through the *subClassOf*/*parentTreeNumber* properties, according to the *true-path rule*. For example, this implies that a publication which is related to both *Glucose* (compound:CID5793) and *Diabetes Mellitus Type 2* (mesh:D003924) also implicitly supports the association between the chemical class of *Monosaccharides* (chemont:0001540) and the disease family of *Glucose Metabolism Disorders* (mesh:D044882), even if this relation was not explicit in the KG (Fig. 1B). Associations were only computed on a selection of MeSH categories relevant in the studied context (Sartor *et al.*, 2012): Diseases, Anatomy, Chemicals and Drugs, Phenomena and Processes, Organisms, Psychiatry and Psychology, Anthropology, Education, Sociology and Social Phenomena, Technology, Industry and Agriculture.

The over-representation of each association in the scientific literature was then tested using a right-tailed fisher exact test,

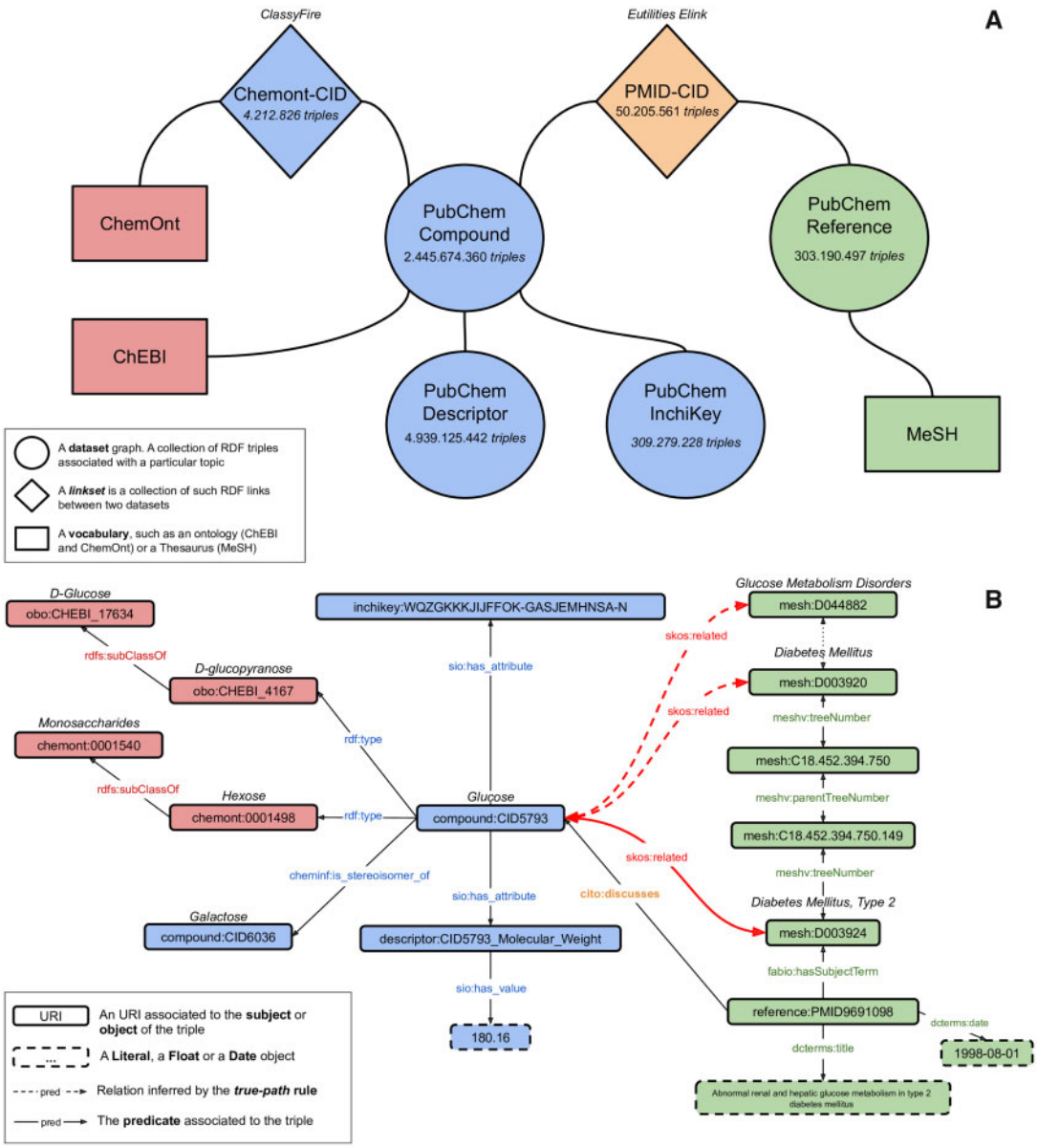

Fig. 1. A general view of the KG architecture (**A**) and an example of interrelations between compounds, literature and chemical classes, that can be extracted from the KG, centered around Glucose (**B**). In (A), only graphs containing the basic relations between entities are indicated while graphs containing inferred relations from statistical analysis (*skos:related* relations) are hidden. Blue boxes represent triples exposing properties associated with PubChem compounds, green triples described publications, orange triples established links between molecules and publications, and red are related to chemical classification of molecules. In (B), gray arrows represent explicit relations that are instantiated in the original dataset and bold red arrows represent added relations, inferred from the enrichment analysis. Among those relations, dashed ones represent relations inferred using the true-path rule, associating broader concepts by transitivity. The predicate *rdf:type* is used to state that a PubChem Compound belongs to a given chemical class (ChEBI or ChemOnt); *rdfs:subClassOf* defines the hierarchy between chemical classes; *cheminf:is_stereoisomer_of* connects two PubChem compound that are stereoisomers; *sio:has_attribute* and *sio:has_value* refer to additional properties of chemical compounds; *cito:discusses* indicates that a PubMed article discusses statements, ideas or conclusions related to a PubChem compound; *skos:related* describes an associative link; *dcterms:title* and *dcterms:date* indicate respectively the title and the publication date of an article; *fabio:hasSubjectTerm* refers to MeSH annotations of the article; *meshv:treeNumber* and *meshv:parentTreeNumber* are used together to express hierarchical relations in the MeSH Thesaurus

reporting the *odds-ratio* and *P*-value, which was adjusted for multiple comparisons using the Benjamini–Hochberg procedure, providing the *q*-value. Associations can be ranked by effect-size using odds-ratios: the higher the odd-ratio, the more the two entities are frequently mentioned together in the literature, suggesting their interdependence. The χ^2 statistics is also indicated to provide a ranking between associations using the *q*-value, even if it was not precisely computable and was approximated to 0 (Sartor *et al.*, 2012). Associations with a *q*-value $<$ 1*e*-6 are considered significant and are instantiated in specific assertion sub-graphs as follows: *chemical_entity skos:related mesh:descriptor*, the *skos:related* predicate being a symmetric and non-transitive property used to assert associative links. An example is provided in Figure 1B, the instantiated relation between *Glucose* and *Diabetes Mellitus, Type 2* is assumed to be supported by other publications and is statistically significant. According to the *true-path* rule, the relation between Glucose and ancestors of *Diabetes Mellitus, Type 2* can also be inferred and instantiated. For clarity reasons, only these relations are indicated on the Figure, but in the same way, a *skos:related* relation could also exists between *Glucose* related chemical classes (e.g. *Monosaccharides*) and *Diabetes Mellitus, Type 2*, as well as its parent descriptors.

The use of a threshold on the *q*-value is a convenient and widely used statistical approach to select relevant results from hypothesis

tests, but is not sufficient for interpreting association strength (Solla *et al.*, 2018). Thus, we added the absolute number of supporting articles, as well as the creation of a complementary index named *fragility index*. We define the *fragility index* as the minimal number of articles, included in Jeffrey's proportion confidence interval (DasGupta *et al.*, 2001), that, when removed from the association corpus, increase the *P*-value over the significance threshold (1*e*-6). Details of this approach are presented in Supplementary Materials (Supplementary Section S3.2.1).

Code and SPARQL requests required to re-build all the Virtuoso triplestore and reproduce results are available at https://github.com/eMetaboHUB/Forum-DiseasesChem.

3 Results

3.1 Advantages of the semantic level on association extraction

The true-path rule affects corpora sizes of chemical classes and MeSH descriptors by propagating annotated literature to broader concepts, and thus also influences counts used for statistical testing of independence. For instance, a broad descriptor like *Neoplasm* (mesh:D009369) goes from 96 373 explicitly annotated articles to 955 848 (~10-fold) using literature related to sub-concepts. For more information, see Supplementary Materials (Supplementary Section S3.1). Beyond changes in counts, we analyzed the effect of the true-path rule on extracted relations. A Venn diagram of significant associations between MeSH descriptors and PubChem compounds detected with and without the use of the true-path rule is presented in Figure 2. The number of significant associations detected using propagation is almost twice the number of initially detected associations. More than half of these new associations link compounds to MeSH terms that are never directly co-annotated in the same article, but inferred from semantic relationships. Furthermore, 6% of initially detected associations were discarded by performing the analysis with the semantic elevation, as they no longer met our inclusion criteria (q-value < 1*e*-6).

To illustrate what kind of extracted associations can be affected by the method, we selected some typical examples and analyzed the MeSH annotation frequency of their supporting corpus, in order to illustrate the context surrounding the association. All MeSH descriptors explicitly annotated to publications supporting the co-occurrence between both entities were extracted and we selected the top 20 most important descriptors, using a score analogous to TF-IDF with MeSH terms in metadata (see details in Supplementary Section S3.2.2). The score increases with the frequency of the MeSH appearing in the subset of articles that describe the relation, and is adjusted by the MeSH's general frequency in the literature to filter out unspecific terms.

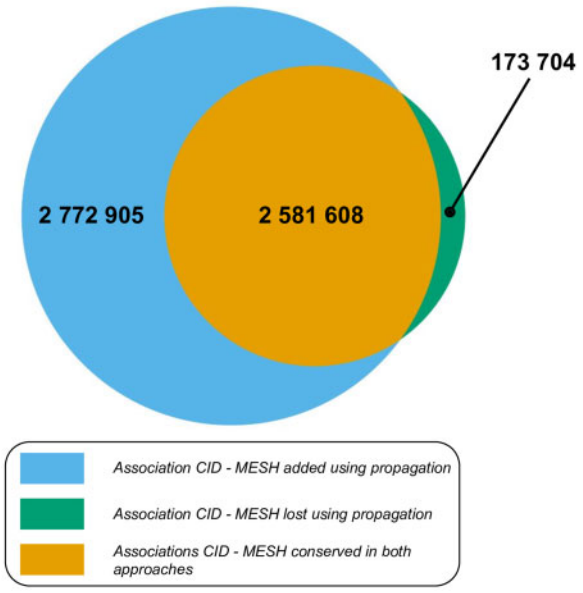

Fig. 2. Overlap between sets of statistically significant associations (q-value < 1*e*-6) detected with or without the true-path rule. 58% (1 614 120/2 772 905) of newly detected associations are completely novel as there are no direct links between these entities in the KG without using propagation

3.1.1 Associations discarded by corpus size adjustment

The case of *Water* and *Eukaryote* describes the removal of associations due to the true-path rule. *Water* (CID 962) has a very large corpus of 156 845 annotated articles and an initial co-occurrence with *Eukaryote* of 330 publications. However, this appears highly significant (*P*-value ≈ 2.9*e*-49) as *Eukaryote* was explicitly annotated to only 7276 articles. Using propagation to broader descriptors, *Eukaryote* is found to be related to 7 700 235 articles (~87% of all publications in the KG), which seems more realistic according to the expected representation of articles related to Human and model organisms in PubMed. Although the co-occurrence increases by 82 293 articles, the association is not supported by the test (*P*-value = 1). The MeSH descriptors annotating the articles bearing the co-occurrences (Supplementary Fig. S1) consist of mostly unrelated sets of terms, from *Plant Roots* to *Time Factors*, which also have low importance scores indicating that they are not specifically associated with this relation, but widely distributed in the literature.

3.1.2 New associations inferred from semantic relatedness

An example of an association revealed using propagation is *Oxyfedrine*, a vasodilator and beta-agonist agent, and the disease family of *myocardial ischemia* (a disorder of cardiac function caused by insufficient blood flow to the muscle tissue of the heart). Originally, no publication directly links Oxyfedrine to myocardial ischemia: it is only linked to specific diseases of this class: *Angina Pectoris*, *Coronary Disease* or *Myocardial Infarction*. By propagating annotation to broader terms, the union of articles associated with these diseases supports the relation with *myocardial ischemia*, as well as the higher concepts of *Cardiovascular Diseases* and *Heart Diseases* which are also found to be highly associated with Oxyfedrine. By looking at the other annotations of the supporting corpus (Supplementary Fig. S2), we see that unlike the association between Water and Eukaryota, it bears a specific set of topics, with high importance scores, related to Heart Diseases (*Blood Pressure*, *Coronary Circulation*, etc.).

However, propagation through broader descriptors up to the highest hierarchical concepts, have sometimes yielded reasonable associations of average interest. An example is the association between *Levocabastine*, an antihistaminic agent used in treatment for conjunctivitis allergies, and the *Eukaryota* kingdom. Their relatedness is completely inferred from semantic relationships as there are no publications explicitly annotated with both. The supporting literature is mainly focused on concepts such as *ophthalmic solutions*, *anti-allergic agents* and the organisms *Human*, *Rats* and *Mice* (Supplementary Fig. S3), which yield the association of Levocabastine with the whole kingdom.

3.2 Relevance of extracted associations

To evaluate our approach, we used previously published test-cases from the Metab2MeSH article to determine a reference set of diseases related to Cyclic AMP, and compounds related to Phenylketonuria using literature evidence [data were, respectively, retrieved from the Supplementary Figs S2 and S3 of the original Metab2MeSH article (Sartor *et al.*, 2012)].

3.2.1 Diseases related to cyclic AMP

Most of the 13 disease-related MeSH associated with Cyclic AMP (cAMP) are malignant neoplasms, such as neuroblastoma or glioma. The top 20 of MeSH descriptors reported by FORUM is compared to the reference set (see Supplementary Table S1). FORUM retrieves

all diseases annotated in the reference set as over-represented in the cAMP corpus. Also, by taking advantage of the true-path rule and the semantic relations between MeSH terms, it enriches results with new associations involving broader descriptors which represent the disease families in which the previously identified disorders belong. For instance, '*Neoplasms, Neuroepithelial*' is a parent in the MeSH Thesaurus of '*Glioma*' and '*Neuroblastoma*', and '*Parathyroid Diseases*' is the direct parent of '*Hypoparathyroidism*' and '*Hyperparathyroidism*'. This higher level of description allows the direct identification of the type of disorders that could be related to a given compound, in addition to reporting the list of specific diseases.

3.2.2 Compounds related to Phenylketonuria

In the second test case, we compare results for compounds associated with Phenylketonurias (Supplementary Tables S2 and S3), a group of metabolic diseases induced by a defect in the production of phenylalanine hydroxylase (Blau *et al.*, 2014). Some compounds (12/25) associated with this disease in the reference set are not present in our compound list (in red), as there are no or few articles associated with them in our KG. This could have stemmed from compound disambiguation by PubChem links providers, since some of them were rarely occurring D-configuration amino acids, while L-forms are still found. Differences in the approach used to link literature to compounds, as well as changes in the PubChem database since 2012, make comparison with previous methods difficult. The majority of the compounds newly associated with Phenylketonurias in our list are pterin derivatives. Supplementary Tables S4 and S5 describe results obtained using ChEBI and ClassyFire ontologies providing a semantic description of molecules. This representation highlights molecule families associated with Phenylketonuria rather than simple compounds, like the family of *aromatic amino acids, biopterins and derivatives* or *pterins* whose members were also found to be significantly associated with the disease in the previous analysis.

3.2.3 Deeper association extraction

We used FORUM to extract significant associations between chemical entities and biomedical concepts and explicitly connect them in the KG. This new layer of relations enriches the KG, allowing us to explore and study more complex questions.

We illustrate the potential of our query engine coupled with the extended KG using a previously reported association extracted from a complex chain of relationships. Literature-based discovery was first introduced by Swanson (1986) and aimed to extract new hypotheses from initially separated pieces of knowledge developed in scientific articles. In his first example, he proposed as a new hypothesis that fish oil might be interesting in the treatment of Raynaud's syndrome, by connecting the knowledge that a fish oil diet is rich in Eicosapentaenoic acid (EPA), which has antiplatelet actions and increases production of PGI3. Second, PGI3 is an analog of PGI2, a vasodilatator and antiplatelet agent, used in the treatment of Raynaud's syndrome. All building blocks for reconstructing Swanson's reasoning can be found in our KG, using the created *skos:related* relations between entities from the subgraph shown on Figure 3. The relation between therapeutic MeSH descriptors and the MeSH descriptor representing Raynaud's syndrome have been extracted using the same approach as described in Section 2, similarly to (Smalheiser and Bonifield, 2016). The SPARQL requests that can be used to retrieve these connections are provided in Supplementary Materials (Supplementary Section S3.3).

4 Discussion

Federating knowledge is an essential process in successfully inferring novel hypotheses from metabolomics. For this purpose, we gathered and connected complementary resources in a KG to provide a semantic path between chemical and biomedical concepts through the scientific literature. By performing an enrichment analysis on their related corpora, we estimated the effect-size and relevance of more than 120 million assertions and selected 8.7 million of them to be instantiated, using a threshold on the resulting q-value at 1e-6. According to this threshold, we statistically expect less than nine falsely created relations in the KG, providing a robust resource to explore connections between chemical entities and biomedical concepts.

We have shown that FORUM results are consistent with reported test cases and enriched with new relevant associations, derived from initial outcomes. Metab2MeSH matched chemical compound names to MeSH concepts to determine links to articles. However, these links are a subset of all links that can be extracted using Elink, which also takes advantage of external providers and

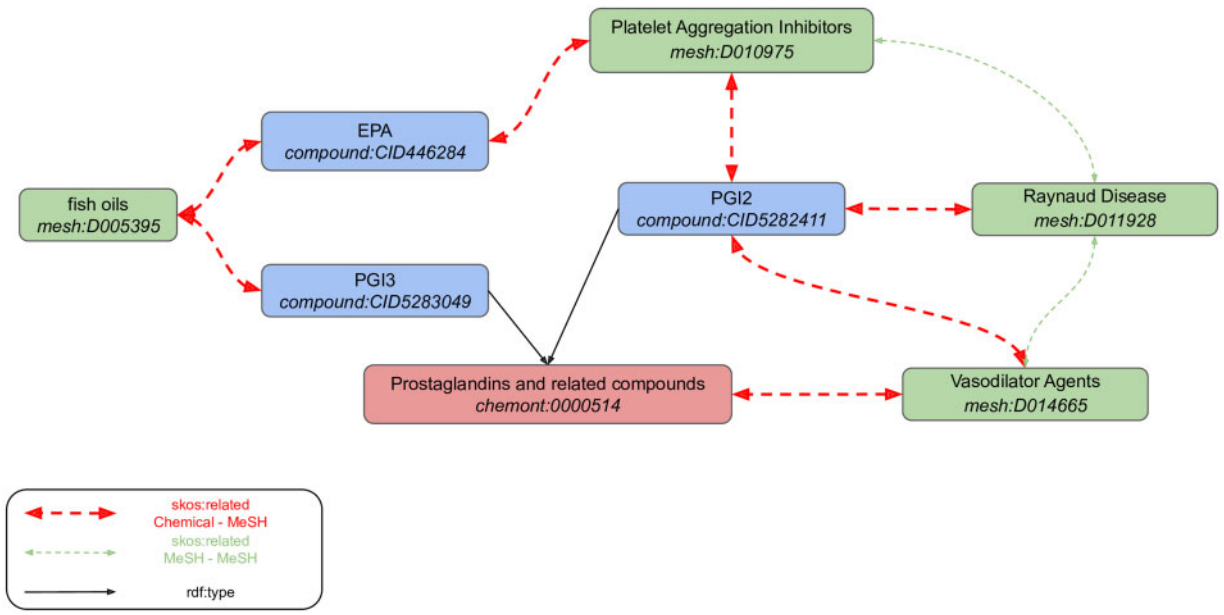

Fig. 3. Retrieving the Swanson's hypotheses fragments in FORUM KG. All skos:related associations have been built from the processing of underlying chains of links found in the federation of linked datasets. There are also *skos:related* links directly linking the chemical class of *Prostaglandins and related compounds* to *fish oils*, *Platelet Aggregation Inhibitors*, *Vasodilator Agents* and *Raynaud Disease* in the KG, but those are not shown for clarity reasons

publishers. Moreover, the semantic description of data, combined with the use of the *true-path* rule, allows the propagation of annotations through the hierarchical relations defined between entities. This enables the computation of associations with several degrees of abstraction from both chemical and biomedical perspectives, making it possible to consider entire disease or chemical families instead of simple compounds or diseases, which can be useful for providing a broader vision on results and guiding interpretation. Thus, we have been able to relate such specific disease families like *Neoplasms, Neuroepithelia* and *Parathyroid diseases* to cyclic AMP and some chemical classes like *pterins* and *aromatic amino acids* to Phenylketonuria (Blau *et al.*, 2014). Here, we mainly focused our examples on disease cases, but the other MeSH trees are also meaningful and present similar behavior regarding the impact of the *true-path* rule.

The approach used in FORUM allows us to double the number of associations found by exploiting the inference through the hierarchical relationships in the MeSH Thesaurus, compared with the approach that omits term relationships. As MeSH descriptors are organized in trees, the corpora extension is null for leaves and becomes stronger as we get closer to the roots, which gather all articles referring to the main topics. Newly extracted relations are thus associated with MeSH descriptors describing increasingly broader concepts, with larger corpora, enabling the creation of new hypotheses. For example, we were able to significantly relate the Oxyfedrine to the *Myocardial Infarction* disease family. Furthermore, given that associated descriptors highlight a drug related context (vasodilator agent and Adrenergic beta agonist), this may suggest hypotheses for the use of Oxyfedrine for other diseases, such as alternative forms of myocardial ischemia. Since some precise concepts may not always be extracted from relations, abstraction to chemical classes or biomedical broader concepts can still help find relevant information at a higher level. In the same way, we also know that most chemical compounds are not well described in the scientific literature because, for instance, they cannot be correctly identified with current experimental methods. However, for these compounds, concepts linked to the closest chemical class using articles discussing other members of that class can provide relevant prior knowledge. The Chemont structural ontology is particularly suited to this purpose as the classification of any molecule can be determined using structural information such as SMILES or InchiKey identifiers. Compounds potentially related to understudied phenotypes could also be estimated by looking at a broader disease family. It can also help to discard irrelevant relations for broader descriptors affected by the propagation. Indeed, the co-occurrence between some chemicals and MeSH descriptors may originally appear significant, but mostly because the related MeSH corpus can be underestimated without considering the literature related to subconcepts (e.g. *Eukaryota*). Although the extracted relations can thus be used to identify diseases related to one particular compound for instance, these also represent knowledge fragments that can be assembled to highlight new hypotheses. Indeed, like Swanson's deduction extracted from our KG, by following paths between relations to infer new connections between concepts and/or chemicals, FORUM could also support knowledge discovery in metabolomics. Finally, the MeSH annotations of compounds in FORUM can be used to create literature-based descriptors of those compounds, which could fuel other computational approaches. Indeed, various molecular representations, from structure to bioactivity data (Bento *et al.*, 2014; Berman, 2000; Kim *et al.*, 2019; Mattingly *et al.*, 2006; Wang *et al.*, 2012; Wishart, 2006), have already been used to draw chemical similarity to support drug discovery (Duran-Frigola *et al.*, 2020; Katsila *et al.*, 2016) or computational toxicology (Nigsch *et al.*, 2009).

4.1 Limitations

Since FORUM knowledge extraction is based on literature mining, it is evident that this strongly depends on the quality of the considered corpora. While it can be expected that products of scientific misconduct and malpractice are marginal in the PubMed corpus, it can have a substantial impact on associations based on small corpora. For example, one particular harmful practice consisting of multiplying articles from single experiments, coined 'salami-slicing' (Byrne *et al.*, 2019; Errami and Garner, 2008; Spielmans *et al.*, 2010), may pass peer-review and evade retraction, while still inducing bias in the contingency table and thus potentially yield spurious associations. Beyond the articles' content, the metadata quality, namely MeSH indexing and PubChem annotation (Kim *et al.*, 2016), also have a critical impact on the results. To support the identification of associations that could be yielded by such practices, we propose the *fragility index*, which estimates the minimum number of spurious articles that, if removed, would trigger the exclusion of the association in the KG.

Furthermore, as with all literature-based analysis, FORUM associations are also affected by topic imbalance, meaning that the topics that attract most funding and follow publication trends will be overrepresented (Cheung *et al.*, 2013). However, while the ubiquity of these associations can impair their relevance, their validity remains.

The use of a threshold on the q-value is a convenient and widely used statistical approach to select relevant results from hypothesis testing, but this can also lead to misinterpretation, which should be treated with care (Tanha *et al.*, 2017). Although statistically relevant associations could be used to derive new hypotheses, the absence of associations should not be used to this end (Open-world assumption). Indeed, because the KG is incomplete by nature and by design, since path-search is at the core of its exploitation, irrelevant shortcuts should be avoided as much as possible. We thus focused on conservative methods to retain the 'strongest' associations. Other values that help to gauge the effect size and the volatility are provided in complement to the q-value.

By increasing the number of associations, including somewhat redundant information with close parent concepts, we also tend to favor information overload. Indeed, we aim at gathering a large collection of strong associations for computational analysis, consequently, the output can result in a large set of machine-readable associations that can be difficult to navigate manually. Some associations shared by closely related MeSH descriptors may increase the number of triplets to read without adding much information. Also, associations with overly broad concepts can be highly statistically significant but may not be useful or relevant depending on the context. However, some methods (Barriot *et al.*, 2007) exist to filter such information and have been already applied in similar fields (e.g. Gene Ontology), which is a perspective of improvement for FORUM.

The relations drawn between compounds and medical concepts are solely based on the non-independence expectation, meaning that a scientific article mentioning the former is more likely to also mention the latter. This doesn't guarantee any direct relationship, and if one exists, its nature is unknown. Thus, without further inquisition, one should not assume that an association between a compound and a disease imply, for example, its use as a treatment of said disease. Beyond therapeutic uses, the associations can encompass, among others: implication in physiopathology, use as a biomarker for diagnosis, a reagent in related analytical tools, a metabolic product of drug side-effects and so on. Some publications may also report inconclusive results from the test of a drug or a metabolic biomarker, which still count as co-occurrences of a condition and a compound, and thus contribute towards their association. While these articles are not expected to be in a sufficient proportion to lead to an association in our knowledge base, the user should also check this aspect before inferring hypotheses. Some text-mining methods could be applied to abstracts of articles supporting significant associations to unveil the nature of the relationship, pushing the analysis of FORUM results even further. We have shown that even simple MeSH frequency analysis of association supporting corpus can also provide some insights (see also Imidocarb example in Supplementary Fig. S4).

We have also identified some subtleties related to the propagation process, as associations inferred to broader concepts are not necessarily relevant for all sub-classes. In the example of

Levocabastine and Eukaryota, we see that the supporting corpus is mainly related to ophthalmic solutions for human use, but end up relating the compound to a class of organism where the presence of conjunctiva is an exception. These types of relations can nonetheless generate new hypotheses but this requires further analyses, and generally associations with broader concepts should be interpreted as 'related to SOME x' and not 'related to ALL x'. Further developments need to be performed to correctly identify generalization opportunities from inferred relations to broader concepts or chemical classes. Finally, associations with all MeSH categories are included for the sake of exhaustivity and their relevance left at the judgement of the user. We nonetheless believe that the *organisms* category should be interpreted with care, as many large Phylums display associations driven by very few overrepresented model organisms. For example, 81% of articles related to Eukaryota are derived from Human, Rats or Mice studies.

4.2 Conclusion

The KG of FORUM provides a resource to connect chemical entities to biomedical concepts through millions of scientific articles, fully accessible and searchable, offering a useful tool to support the interpretation of results in metabolomics and yield new hypotheses. FORUM could allow a user to directly discover the main MeSH descriptors (e.g. diseases) related to its observed metabolites, and then shape the following bibliographic search with more targeted queries. Potentially shared concepts, or families of concepts, related to discriminant metabolites can also be identified and guide the researcher in the interpretation of the metabolic fingerprint. Symmetrically, FORUM could also be used to suggest compounds related to a particular condition, in order to design a targeted metabolomics analysis. FORUM is particularly suited for clinical and biomedical metabolomic data analysis, but may also be applied to other domains as long as the relevant literature is available in PubMed, with the right level of descriptiveness in the MeSH thesaurus. Built with an in-depth annotation of the literature, and collected using approaches from data mining to manual curation, relations are extracted from large corpora analysis. Therefore, they are less prone to errors than text-mining approaches which require a second examination, as a price for higher descriptiveness. Extracted relations also create higher-level relations whose connections can be explored and easily expanded through the semantic framework, by linking to larger knowledge collections such as Wikipathway (Slenter *et al.*, 2018) to infer new hypotheses. Indeed, the more data are accessible and linked between various resources, the more we open the path to infer new connections and reveal hidden relations. Moreover, while provided associations only involve one chemical entity regarding a precise biomedical concept, FORUM can also be used to give insights on custom questions. Indeed, when the issue can be formulated as an enrichment problem using a reference set of articles that can be semantically described in the KG, the SPARQL endpoint can be used to extract required data by combining several MeSH descriptors or chemical classes in a request for instance. Interfacing SPARQL requests using some dedicated frameworks such as AskOmics (Bettembourg *et al.*, 2015) could also improve the accessibility and manipulation of the KG.

By gathering knowledge from various resources and inferred assertions inside a FAIR infrastructure, organized around a linked data model, we think that FORUM can be an interesting tool, both to support the interpretation of metabolomics data in several contexts, and to serve as a sand-box to give clues on various questions. As the availability of linked data in life science continues to grow, we hope that FORUM could also be considered as a proof of concept demonstrating the efficiency of this type of approach in information extraction.

Acknowledgements

The authors express their gratitude to the PubChem, NCBI Eutils and MetaNetX support teams for helping us in the exploitation of their datasets. We thank the Askelys team who provided insight and expertise on the semantic web technology, as well as our colleagues from the EMPREINTE CATI group for their technical support. We are also very grateful to Juliette Cooke for proofreading the manuscript.

Funding

This project has received funding from the INRA SDN (Schéma Directeur du Numérique) and the European Union's Horizon 2020 research and innovation program under grant agreement GOLIATH No. 825489. This work was supported by the French Ministry of Research and National Research Agency as part of the French MetaboHUB infrastructure [Grant ANR-INBS-0010].

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Barriot,R. *et al.* (2007) How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. *BMC Bioinformatics*, **8**, 332.

Bento,A.P. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bettembourg,C. *et al.* (2015) AskOmics: intégration et interrogation de réseaux de régulation génomique et post-génomique. In: *In OVIVE (INtégration de Sources/Masses de Données Hétérogènes et Ontologies, Dans le Domaine Des Sciences du VIVant et de L'Environnement)*, Rennes, France. Institut National de Recherche en Informatique et en Automatique (INRIA). Rennes, FRA, p. 7.

Blau,N. *et al.* (2014) Molecular genetics and diagnosis of phenylketonuria: state of the art. *Expert Rev. Mol. Diagn.*, **14**, 655–671.

Bushman,B. *et al.* (2015) Transforming the medical subject headings into linked data: creating the authorized version of MeSH in RDF. *J. Library Metadata*, **15**, 157–176.

Byrne,J.A. *et al.* (2019) The possibility of systematic research fraud targeting under-studied human genes: causes, consequences, and potential solutions. *Biomarker Insights*, **14**, 1177271919829162.

Cañada,A. *et al.* (2017) LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. *Nucleic Acids Res.*, **45**, W484–W489.

Chambers,J. *et al.* (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminf.*, **5**, 3.

Cheng,D. *et al.* (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36**, W399–W405.

Cheung,W.A. *et al.* (2012) Quantitative biomedical annotation using medical subject heading over-representation profiles (MeSHOPs). *BMC Bioinformatics*, **13**, 249.

Cheung,W.A. *et al.* (2013) Compensating for literature annotation bias when predicting novel drug-disease relationships through Medical Subject Heading Over-representation Profile (MeSHOP) similarity. *BMC Med. Genomics*, **6**, S3.

DasGupta,A. *et al.* (2001) Interval estimation for a binomial proportion. *Stat. Sci.*, **16**, 101–133.

Djoumbou Feunang,Y. *et al.* (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminf.*, **8**, 61.

Dumontier,M. *et al.* (2014) The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J. Biomed. Seman.*, **5**, 14.

Duran-Frigola,M. *et al.* (2020) Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nat. Biotechnol.*, **38**, 1087–1096.

Errami,M. and Garner,H. (2008) A tale of two citations. *Nature*, **451**, 397–399.

Fiehn,O. (2002) Metabolomics—the link between genotypes and phenotypes. In: Town,C. (ed.) *Functional Genomics*. Springer Netherlands, Dordrecht, pp. 155–171.

Fu,G. *et al.* (2015) PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. *J. Cheminf.*, **7**, 34.

Giacomoni,F. *et al.* (2015) Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics*, **31**, 1493–1495.

Gijón-Correas,J.A. *et al.* (2014) Alkemio: association of chemicals with biomedical topics by text and data mining. *Nucleic Acids Res.*, **42**, W422–W429.
Hassani-Pak,K. *et al.* (2021) KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. Plant Biotechnol. J., **19**, 1670–1678.
Hastings,J. *et al.* (2011) The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS ONE*, **6**, e25513.
Hastings,J. *et al.* (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.
Hoehndorf,R. *et al.* (2015) The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinf.*, **16**, 1069–1080.
Johnson,C.H. *et al.* (2016) Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.*, **17**, 451–459.
Kanza,S. and Frey,J.G. (2019) A new wave of innovation in Semantic web tools for drug discovery. *Exp. Opin. Drug Discov.*, **14**, 433–444.
Katsila,T. *et al.* (2016) Computational approaches in target identification and drug discovery. *Comput. Struct. Biotechnol. J.*, **14**, 177–184.
Kim,S. and Beck,H.W. (2006) A practical comparison between thesaurus and ontology techniques as a basis for search improvement. *J. Agric. Food Inf.*, **7**, 23–42.
Kim,S. *et al.* (2016) Literature information in PubChem: associations between PubChem records and scientific articles. *J. Cheminf.*, **8**, 32.
Kim,S. *et al.* (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.
Koscielny,G. *et al.* (2017) Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.*, **45**, D985–D994.
Lu,Z. *et al.* (2009) Evaluation of query expansion using MeSH in PubMed. *Inf. Retrieval*, **12**, 69–80.
Ludwig,K.R. and Hummon,A.B. (2017) Mass spectrometry for the discovery of biomarkers of sepsis. *Mol. BioSyst.*, **13**, 648–664.
Malas,T.B. *et al.* (2019) Drug prioritization using the semantic properties of a knowledge graph. *Sci. Rep.*, **9**, 6281.
Mattingly,C. *et al.* (2006) The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J. Exp. Zool. A Comparative Exp. Biol.*, **305**, 689–692.
Mi,H. *et al.* (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.*, **47**, D419–D426.
Miles,A. and Bechhofer,S. (2009) *SKOS Simple Knowledge Organization System Reference. W3C Recommendation.* World Wide Web Consortium.
Moretti,S. *et al.* (2016) MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.*, **44**, D523–D526.
Neumann,E. and Prusak,L. (2007) Knowledge networks in the age of the Semantic Web. *Brief. Bioinf.*, **8**, 141–149.
Nigsch,F. *et al.* (2009) Computational toxicology: an overview of the sources of data and of modelling methods. *Exp. Opin. Drug Metab. Toxicol.*, **5**, 1–14.
Peroni,S. and Shotton,D. (2012) FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *J. Web Semantics*, **17**, 33–43.
Pons,E. *et al.* (2016) Extraction of chemical-induced diseases using prior knowledge and textual information. *Database*, **2016**, baw046.
Rebholz-Schuhmann,D. *et al.* (2012) Text-mining solutions for biomedical research: enabling integrative biology. *Nat. Rev. Genet.*, **13**, 829–839.
Sartor,M.A. *et al.* (2012) Metab2MeSH: annotating compounds with medical subject headings. *Bioinformatics*, **28**, 1408–1410.
Slenter,D.N. *et al.* (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
Smalheiser,N. and Bonifield,G. (2016) Two similarity metrics for Medical Subject Headings (MeSH): an aid to biomedical text mining and author name disambiguation. *J. Biomed. Discov. Collab.*, **7**, e1.
Solla,F. *et al.* (2018) Why a P-value is not enough. *Clin. Spine Surg.*, **31**, 385–388.
Sosa,D.N. *et al.* (2020) A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. *Pac. Symp. Biocomput.*, **25**, 463–474.
Spielmans,G.I. *et al.* (2010) A case study of salami slicing: pooled analyses of duloxetine for depression. *Psychother. Psychosomatics*, **79**, 97–106.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.
Swanson,D.R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.*, **30**, 7–18.
Talib,R. *et al.* (2016) Text mining: techniques, applications and issues. *Int. J. Adv. Comput. Sci. Appl.*, **7**, 2016.
Tanha,K. *et al.* (2017) P-value: what is and what is not. *Med. J. Islamic Republic Iran*, **31**, 65–378.
Valentini,G. (2009) True path rule hierarchical ensembles. In: Benediktsson,J.A. *et al.* (eds.) *Multiple Classifier Systems. Lecture Notes in Computer Science*, Vol. **5519**. Springer, Berlin, Heidelberg. pp. 232–241.
van Iersel,M.P. *et al.* (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, **11**, 5.
Wang,Y. *et al.* (2012) PubChem's BioAssay database. *Nucleic Acids Res.*, **40**, D400–D412.
Weibel,S.L. and Koch,T. (2000) The Dublin core metadata initiative. *D-lib Mag.*, **6**, 1082–9873.
Williams,A.J. *et al.* (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today*, **17**, 1188–1198.
Willighagen,E.L. *et al.* (2013) The ChEMBL database as linked open data. *J. Cheminf.*, **5**, 23.
Wishart,D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
Wu,H. and Yamaguchi,A. (2014) Semantic Web technologies for the big data in life sciences. *Biosci. Trends*, **8**, 192–201.
Xia,J. and Wishart,D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.*, **38**, W71–W77.
Zhou,X. *et al.* (2014) Human symptoms–disease network. *Nat. Commun.*, **5**, 4212.