Computational and Structural Biotechnology Journal 20 (2022) 4837–4849

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

The potential of a data centred approach & knowledge graph data representation in chemical safety and drug design

Alisa Pavel [a,b,c], Laura A. Saarimäki [a,b,c], Lena Möbus [a,b,c], Antonio Federico [a,b,c], Angela Serra [a,b,c], Dario Greco [a,b,c,d,*]

[a] Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[b] BioMediTech Institute, Tampere University, Tampere, Finland
[c] Finnish Hub for Development and Validation of Integrated Approaches (FHAIVE), Tampere, Finland
[d] Institute of Biotechnology, University of Helsinki, Helsinki, Finland

ARTICLE INFO

Article history:
Received 29 June 2022
Received in revised form 26 August 2022
Accepted 26 August 2022
Available online 5 September 2022

Keywords:
Knowledge graph
Big data
Data integration
Toxicology
Drug design
Chemical safety

ABSTRACT

Big Data pervades nearly all areas of life sciences, yet the analysis of large integrated data sets remains a major challenge. Moreover, the field of life sciences is highly fragmented and, consequently, so is its data, knowledge, and standards. This, in turn, makes integrated data analysis and knowledge gathering across sub-fields a demanding task. At the same time, the integration of various research angles and data types is crucial for modelling the complexity of organisms and biological processes in a holistic manner. This is especially valid in the context of drug development and chemical safety assessment where computational methods can provide solutions for the urgent need of fast, effective, and sustainable approaches. At the same time, such computational methods require the development of methodologies suitable for an integrated and data centred Big Data view. Here we discuss Knowledge Graphs (KG) as a solution to a data centred analysis approach for drug and chemical development and safety assessment. KGs are knowledge bases, data analysis engines, and knowledge discovery systems all in one, allowing them to be used from simple data retrieval, over meta-analysis to complex predictive and knowledge discovery systems. Therefore, KGs have immense potential to advance the data centred approach, the re-usability, and informativity of data. Furthermore, they can improve the power of analysis, and the complexity of modelled processes, all while providing knowledge in a natively human understandable network data model.

Contents

* Corresponding author at: Arvo Ylpön katu 34, 33520 Tampere, Finland.
E-mail address: dario.greco@tuni.fi (D. Greco).

https://doi.org/10.1016/j.csbj.2022.08.061

1. Introduction

The development of new drugs and chemicals is a long and expensive endeavour [1–3]. An integral part of the safety assessment process is the evaluation of the safety and efficacy of new compounds, which relies on tests that are time consuming, costly, and ethically challenging [3,4]. Therefore, a shift towards alternative methods for the traditional assessment of apical endpoints is taking place. Such efforts promote the reduction of animal experimentation and the use of integrated approaches where multiple testing strategies and research angles are combined [5]. At the same time, the efforts to reduce and replace experimental animals poses novel challenges for the evaluation of systemic effects and long-term outcomes of chemical exposures. Regardless of the test system, a comprehensive understanding, modelling, and prediction of organism-level responses requires the integration and analysis of multiple data layers. In this sense, data becomes even more central and valuable, and the computational strategies applied, grow increasingly important as large sets of data need to be analysed and integrated.

Constantly new independent data sets, relevant for chemical design and safety assessment, are generated. However, these data sets are often highly scattered, not comparable, and of varying quality. Significant efforts have been made to establish standards for data sharing and management through the establishment of the FAIR (Findable, Accessible, Interoperable and Reproducible) principles [6], however they often fall short when large amounts of data need to be integrated [7]. In addition to defining data accessibility standards, metadata reporting standards, robust data integration methodologies need to be further developed. These aspects are fundamental for combined data analysis and modelling of complex processes while also improving the optimal use of all available data. Nonetheless, the data may still not reach its full potential unless it is stored in a structure that enables straight-forward integration of various data types and layers. To this end, Knowledge Graphs (KG) present a suitable framework. A KG is a data structure that contains and conveys knowledge about the "real world under investigation", in which the data is stored in a graph based format [8–10]. This opens unprecedented possibilities also for drug and chemical design and safety assessment. KGs are an extension of a knowledge base, to which a reasoning engine is applied to generate and infer new facts about the world [8]. KGs are data collections that model structured knowledge in a graph based format and can be used 1) as a knowledge base or database (e.g. the Google search engine), 2) to analyse the data by making use of graph based metrics and methods (e.g. traffic routing systems) and 3) to infer new facts about the world (e.g. Amazon recommendation engine). The latter application is what distinguishes KGs from classical knowledge bases [8]. The underlying graph model can be directed, undirected, heterogeneous or property graph structures, that can contain edge and node labels as well as attributes [9,11].

How such a graph is stored on disk, can change between database management systems. An example of a more specific graph database model are, triple stores, which store everything as an edge, including properties, while other graph database engines store data in different manners, for example as a multigraph or adjacency list [12,13] However this review will not cover this topic in more detail, and while there are performance difference to be observed between different data management and storage options for specific use cases, this topic is of lower relevance for users wishing to use a ready-made solution, i.e. a database management system. More information on this topic can be found in these reviews [12,13].

The application of KGs and its benefits in life sciences have been extensively described [14–22]. However, while the avenues to explore by using KGs are vast and exciting, the limitations and roadblocks need to be addressed as well. This review describes available data sources that can support the drug design and chemical safety assessment process. It reviews currently available as well as possible KG applications for drug design and chemical assessment. Lastly it discusses how the use of KGs can advance the integration and analysis of the different data layers in the context of chemical and drug development, and address the challenges standing in the way of the full exploitation of these data structures from a data centric view.

2. Available data sources for chemical safety and drug design

The *in silico* drug design and chemical safety assessment process relies on knowledge from different areas of the life sciences [5,16,23,24]. For example, the structural information of compounds can be complemented with toxicogenomics data to study their mechanisms of action (MOA), defined as the underlying molecular processes happening in the biological systems under specific conditions [25–27]. Moreover, this knowledge can be integrated with clinical (trial) data to further explore the compound effects on a large population. This knowledge can be supported by organism specific information, data from systems biology and lab based experimental data.

An overview of possible data sources that can be used to address different aspects of the chemical assessment and drug design processes are listed in Table 1. KGs can easily support the integration of such heterogeneous data. In Fig. 1, we present a possible high level schema for a KG, that in its whole or on a sub-graph level can be used for chemical safety assessment and drug design.

Many of the data types listed in Table 1 are by nature link-orientated (e.g. protein–protein interactions [46,52]), drug target information [[37,78,79]) or directly produced or represented in graph structures (e.g. co-expression networks [80,81], regulation networks [54]) [82]. This natural network representation and link orientation of the data is one of the main advantages to model

Table 1
Examples of existing relevant data sources for drug design and safety assessment with possible insights these data can provide. How these data can be linked to other entity nodes is displayed in Fig. 1.

Related Node Type	Data Type	Data Source	Possible Insights
COMPOUND	Structure	PubChem [28], STITCH [29], ZINC20 [30], QSAR-DB [31]	Structural/ descriptive information of compounds
	Effects	SIDER [32], Pharos [33], DrugCombDB [34], CTD [35], OpenTargets [36], DrugBank [37], Tox21 [38,39], ECOTOX (*cfpub.epa.gov/ecotox/*), ToxCast (*epa.gov/chemical-research/toxicity-forecasting*)	Clinical/ Toxicity/ observable effect of compounds
	MOA	GEO [40], LINCS L1000 [41], CTD [35], TG-Gates [42]	MOA of compounds
GENE (Gene Product)	Function	Ensembl [43], Panther [44,45]	Gene/ Protein Family/ Function Groups Organism matching
	Interaction	HIPPIE [46], HitPredict [47,48], HuRI [49], MINT [50], IntAct [51], String [52]	Protein Interaction
	Regulation	TRRUST [53,54], TargetScan [55,56], miRTarBase [57], InnateDB [58]	Gene Regulation
PHENOTYPE	Clinical	NCBI [59] MedGen, NCBI ClinVar [60], DisGeNet [61], Human Phenotype Ontology [62], Orphanet (orpha.net), OMIM (omim.org)	Phenotype relationships, comorbidities, descriptions
	Molecular	GEO [40], GWASCatalog [63], ArrayExpress [64], CTD [35]	MOA of Phenotypes
ASSOCIATIONS	Function & Effect	GO [65,66], MSigDB [67,68], Reactome [69], Wikipathways [70], KEGG [71,72], EnrichR [73], AOP-Wiki (*aopwiki.org*)	(Functional) Groups
CELL LINE/ TISSUE/ ORGAN	(Molecular) Characteristics	Human Protein Atlas [74], GTex (*gtexportal.org*), GEO [40], ENCODE [75], CellMiner [76,77]	MOA of biological systems under different conditions

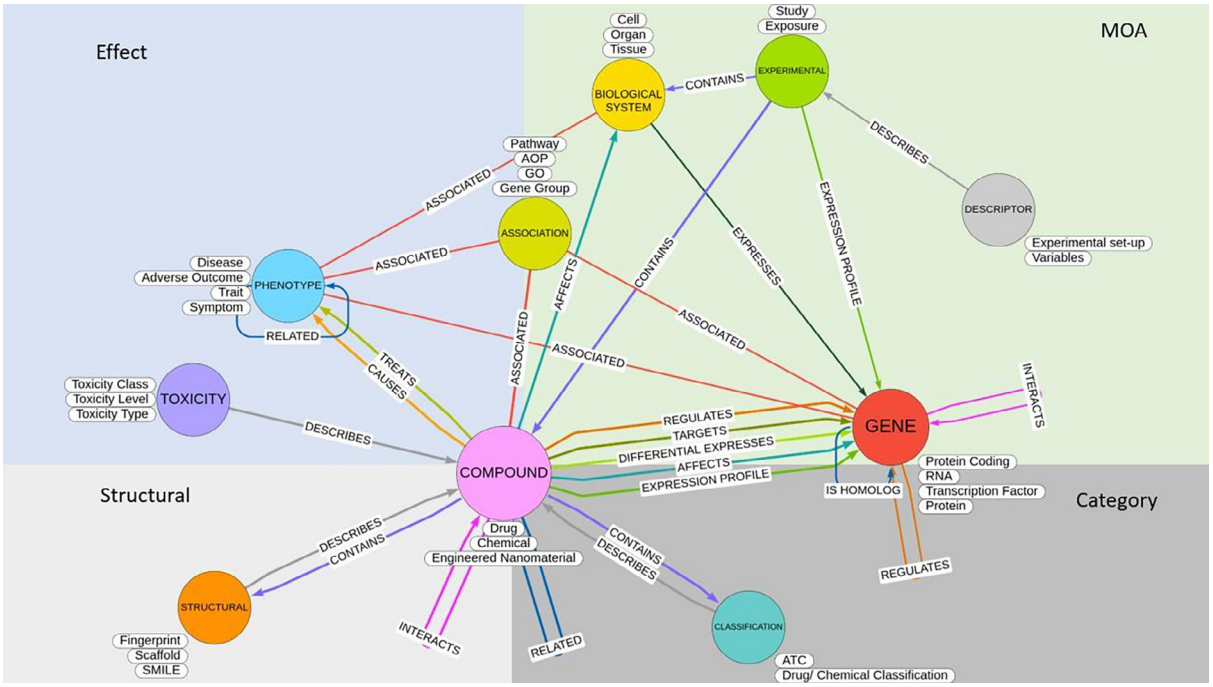


Fig. 1. Possible high-level schema of a life science KG focused on chemical safety assessment and drug development, outlining different data types & links from a compound centred perspective. Covering data describing its MOA, (observable) effect, structure and compound specific meta-data. Examples of data sources that can provide these links are listed in Table 1.

these data in an integrated network fashion. Analysing and modelling biological knowledge in a network structure, is a common methodology in systems biology [83,84], since it allows to investigate an entity with respect to all other entities in the network and to model the information flow through a network (e.g. in a protein–protein interaction network, in a regulation network or in a gene co-expression network) [20,23,80,85].

3. Advantages of data modelling with graph databases & exploration through KGs

The dual nature of KGs as knowledge base and inference engine in combination with allowing the data to be analysed form a network perspective in addition to traditional methods, can bring many advantages to studies based on drug development and compound safety assessment, which are outlined below.

3.1. KGs as a data modelling system to improve flexibility, re-usability and expandability of data rich studies

Meta-analysis is a common tool to investigate a set of studies in order to gain statistical insight into common research questions and their findings. Here we use meta-analysis as an example to showcase how large scale data management and data integration, as defined in a KG, can improve quality, re-usability and minimise the cost of such studies. Meta-analysis based studies can be cost-intensive due to the amount of manual work needed to collect and annotate research studies, perform statistical analysis, and

interpret the results. With the growing volume of research studies, this problem becomes more challenging [86]. Therefore it has been shown that optimised data representation, that allows data scaling and reuse can reduce the technical issues associated with meta-analysis based studies [87]. KGs are data structures excelling with data of varying quality, type and gaps. The time spent on the search, extraction and comparison of studies can be substantially reduced through semantic annotation. Furthermore, KGs allow flexible data representation and are easily scalable with new data and scopes. Statistical methods and analysis can directly be applied onto the KG [88], allowing not only the usage of the data stored in the KG but additional layers of information contained in the graph topology.

For example, Yang et al. [89] conducted a meta-analysis on ecological hazard data to investigate nanoplastic ecotoxicity. In their study they made use of data containing information of particle size on specific observable endpoints, such as population growth, mortality and reproduction. This data is by default link-orientated, making it easy to integrate in a KG model. On expanding the study to include more data sets, additional information or more observable endpoints can easily be retrieved from the KG in an unified format, eliminating the expensive data pre-processing step needed in most meta-analyses [87]. Wang et al. [90] combined multiple gene expression data sets covering the response to a pulmonary tuberculosis infection in order to identify possible therapeutic targets. Gene expression information in response to certain conditions can again easily be integrated into a link oriented data model, while the experimental entity can be enriched with the necessary metadata of the exposure, which can be linked to similar experiments. In addition to extracting the direct information, this also allows to identify similar studies that could be included in further studies or meta-analyses. In the pharmaceutical industry, model based meta-analyses can be used during the drug development process, which helps to leverage (prior) knowledge in order to make informed decisions about the potential of a compound [91,92]. Such data could for example contain information from previous clinical studies or information about possible competing products, which can help for example to make informed decisions about optimal dosing or to perform a risk assessment of the compound's profitability [92].

Another example how KGs can aid data centred studies by providing a unified data schema which integrates multiple layers of diverse data is showcased in Federico et al. [23]. In their study, the authors exploited multiple data types that were unified in their custom KG [20] for a drug repositioning study focused on the prioritisation of drug combinations for the treatment of human complex diseases. Since molecular targets of drugs are both soluble proteins and/or receptors, a co-expression network of the disease has been filtered by using multiple data sets (protein–protein interactions, functional relationships in biological pathways and regulatory interactions) integrated in their KG, retaining only edges (and nodes) of the co-expression network supported by these data. In this way, they were able to leverage the biological significance of the disease co-expression network, which refined the predicted drug combinations by exploiting existing molecular knowledge.

3.2. Flexibility of a graph based data model & the leveraging of hidden links

While many relevant data sources (Table 1) do already come in a network based or link oriented data format, they are still individual and independent data sources that need to be integrated into a combined data model. KGs (Graph data models) have shown the potential to be a successful framework for the integration of diverse data sets [93]. Effective integration and analysis of the comprehensive data sources could significantly increase the success rate in drug design and chemical safety assessment [92,94–96]. Zhang et al. [14] made use of the integration of drug - side effect, drug - indication and drug - target information to predict drug - adverse outcome relationships in a KG framework. Al-Saleem et al. [16] created the CAS Biomedical Knowledge Graph, which integrates multiple data sources across 11 different data layers focusing on COVID-19 relevant information in order to use the KG framework for drug repositioning studies for COVID-19.

However, while there is a general consensus that more and diverse data can provide a more complete view [97,98] on complex biological processes [99], many of the individual data sources follow different standards, were produced for different problems and therefore do not always contain the same data points or are non-complete. Graph databases and KG technology offer a good solution to this challenging integration task [93]. They are by nature schema-free and allow the integration of different types of data with different levels of quality and completeness [9] as well as allow the integration of hierarchical dependencies between data points, making the data model easily expandable and adjustable to changes over time. In comparison to relational databases, the schema-free nature implies that the database schema does not need to be defined in advance and therefore can evolve over time with the data. However this also implies that the user is responsible to keep the data "clean", i.e. to assign the same node/ edge types to data points of the same class, use the same property types for the same data properties as well as understand that while the data model allows gaps in the data, these gaps will still affect any downstream models applied. Additionally, graph structures are suitable models for biological systems, making it intuitive to understand their complex organisation. Network structures, especially their connections can be easily visualised and explored [20,80,100–104]. Moreover, the extraction and analysis of subgraphs can provide a more informative view of the process under study.

Pavel & del Giudice et al. [20] used a sub-area of the gene (product) interaction data layer in their KG infrastructure to analyse possible molecular processes associated with COVID-19. Serra et al. [105] used a network based approach to perform engineered nanomaterial (ENM) contextualisation. In their constructed network, the nodes represent four types of entities (ENMs, chemical exposures, drug treatments and human diseases), while the edges represent the similarity between entities based on their induced transcriptional alteration. The network was scanned in search of heterogeneous cliques of four nodes (one ENM, one chemical, one drug and one human disease) in order to contextualise the effect of ENM exposure with respect to the other entities. This analysis highlighted strong connections between metal oxide nanoparticles and neurodegenerative disorders. Ratajczak et al. [106] showcased that filtering KGs to only contain task relevant information can lead to significant prediction performance improvements. The authors were able to reach an improvement of up to 40 % when predicting possible drug targets via graph embedding.

The use of property graphs allows not only to add edge centred data to the KG, but to enrich nodes and edges with properties, which can be unique to a specific data point. This allows the easy integration of for example quantitative data, such as age or gene expression counts. In order to make this data comparable in a graph model, it can be assigned to classes, such as child, adolescent, adult or low, middle or high gene expression, which can be added to the graph model as their own nodes. For an example on how to classify gene expression counts, see the "Discriminant Fuzzy Pattern to Filter Differentially Expressed Genes" method [107]. By adding higher level classes of these terms as nodes, the graph topology can be used to gather further insights, while when

the accurate, individual terms/ values are needed, the properties of the datapoint can be retrieved, making this data model highly versatile in application.

In addition, KGs can efficiently be queried by specialised graph query languages, which are pattern orientated, allowing a detailed exploration of linked data (Fig. 2) and graph topologies, the latter which is information not available in other data representation formats.

3.3. Leveraging the topology of a KG

One main advantage of the integration of many data layers in a single KG is the possibility to retrieve the so-called "hidden links", which are relationships, associations and correlations that are contained indirectly in the data but not visible in the raw data without the additional topological information (Fig. 2). These hidden links can easily be spotted in a graph based format, but are difficult to investigate in a relational data format. For example, by integrating knowledge about gene product interactions (e.g. protein–protein interaction networks), with drug - target information as well as gene product - phenotype information drug - phenotype links can directly be retrieved from the network even though this information is not directly contained in any of the integrated datasets. Fig. 3 shows how such links can be explored. In addition, classical topological network metrics can be used to evaluate entities. Such metrics are for example degree centrality, closeness centrality or edge betweenness centrality [88,108].

Pavel, del Giudice et al. [20] leveraged shortest paths to identify genes that link known gene sets associated with COVID-19, in order to identify possible genes associated with the disease but are neither direct interactors of the virus or measurable in differential expression analysis. With the help of the applied topological exploration, the authors were able to identify a set of intermediate genes and link them to relevant biological processes, such as vascular processes. Through the additional integration of drug - gene target information in their KG model, the authors were able to suggest possible drug repositioning candidates based on the identified gene sets. Zhu et al. [109] constructed a drug KG, which they used to explore possible drug repositioning candidates. Next to an embedding based approach they also explored paths that connected diseases with drugs in order to extract the connectivity information between a drug - disease pair. In their study, investi-

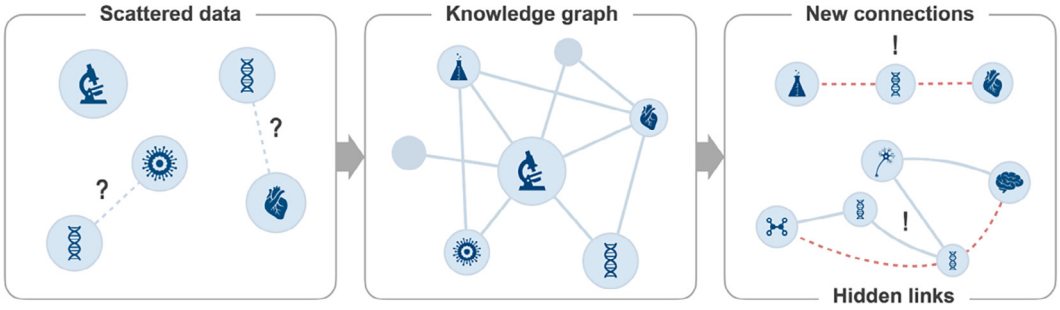
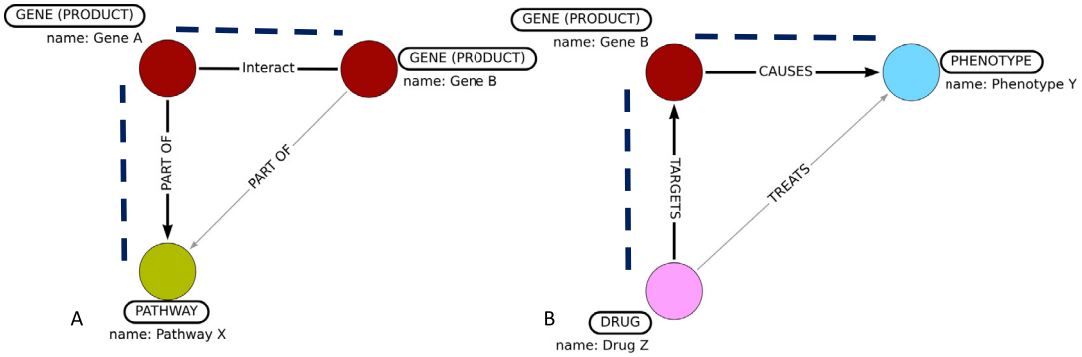


Fig. 2. Diverse data sources can be integrated into a unified data model, such as a KG. Through data integration, hidden links from the individual data sources can be made visible. In addition, the KG can be used to generate/ infer new knowledge (links) based on existing data.



Cypher Query

MATCH (n1:PATHWAY {name: "Pathway X"})<-[:PART OF]-(n2:`GENE (PRODUCT)`)-[:Interact]-(n3:`GENE (PRODUCT)`)
RETURN n3

MATCH (n1:PHENOTYPE {name: "Phenotype Y"})<-[:CAUSES]-(n2:GENE)<-[:TARGETS]-(n3:DRUG)
RETURN n3

Explanation

Find a GENE that interacts with a GENE belonging to PATHWAY X.
This gene could also be a part of PATHWAY X & therefore suggest a new "PART OF" connection.

Find a DRUG, which targets a GENE known to cause PHENOTYPE Y.
This drug could be a possible medication for condition PHENOTYPE Y and suggest a new "TREATS" connection.

Fig. 3. Example of graph exploration with respective Cypher (Neo4j query language) commands. The figure shows an example of a subset of a KG, each with 3 nodes and 2 edges. The grey lines are links that could be inferred from the existing data, via exploration of the dashed lines. A) Gene (products) possibly belonging to a specific pathway are inferred, through one step neighbours of known gene (products) belonging to this specific pathway. B) A possible drug to treat a certain phenotype via the knowledge of a gene (product) causing this phenotype as well as a drug - gene (product) relationship is sought. Below the figure, examples of cypher (Neo4j query language) queries are shown, which show how the graph can be explored and missing links can be inferred, in a very simplistic manner. If the graph would contain multiple genes or drugs that would fit the criteria outlined in the queriers, multiple results would be returned.

gating the mechanism of action of engineered nanomaterials in *in vivo* and *in vitro*, Kinaret et al. [110] showed that by exploring the expression profiles via gene co-expression networks [85] and functional groups contained in them (communities) [108] the in vivo *&* *in vitro* functional responses converged, which was not observable when comparing the differential expressed genes directly. Madi et al. [111] built an antigen-antigen correlation network from antigen microarray data and by extracting their minimum spanning tree they were able to create immune trees in order to compare these between mothers and their newborns. In Pavel et al. [85], they compared the mechanism of action of dasatinib and mitoxantrone via topological properties of gene co-expression networks. In Federico et al. [23], the authors prioritised potentially relevant drugs by considering the MOA of drugs, their structure and topological properties of the disease network. Drug combinations are prioritised based on having "long" shortest paths between their targets on the created cancer co-expression network, so as to target non-overlapping areas. In addition, drugs that target central genes in terms of degree centrality in the cancer co-expression network are prioritised. The criterion behind this assumption is, that by targeting genes that are central in the network it is possible to indirectly expand the effect of the drug to the widest area of the network. This means that the selected drug combinations target genes that show high connectivity in the cancer network, covering, in this way, the widest area of the network, so as to maximise the therapeutic effect of the combination, and minimising the functional overlap of the drugs.

Topological information of the graph can be used on a local level to assess the quality of knowledge of individual entities or whole subgraphs. For example, similar entities can be compared based on their connectivity profile to evaluate the quality of individual relationships [112] or individual relationships can be scored based on their likelihood to be true based on topologically close entities as well as the connection to similar node entities in the graph. The same principle could be applied to assess the correctness of node or edge labels or to add possible correct labels [113]. The underlying assumption is that similar entities should be connected to similar other entities.

This idea is explored in for example network matching algorithms [114] as well as in node embedding algorithms, such as node2vec [115], which leverages random walks to translate the graph space into a vector space, where close/ similar connected nodes are translated to be near in space.

3.4. Making the graph space interpretable by classical machine learning algorithms through node embedding

A lot of current approaches that use KGs to gain new insight into biological processes are based on node embedding methodologies [14,17,21,22,116], such as node2vec [115], in combination with a classification algorithm, such as logistic regression-based classifiers, to solve the link prediction problem present in a KG (a new fact about the world under investigation translates into a new edge in the KG, which reduces most prediction problems on a KG to a link prediction problem). Embedding based methods have the advantage that they translate the graph into a vector space, making it suitable for the application of existing prediction/ classification models.

Zhang et al. [14] made use of a KG and its custom node embedding, based on the word2vec algorithm [117], to link drugs with their potential adverse drug reactions, based on a logistic regression classifier applied to the vectorized node embeddings. Karim et al. [22] propose a framework leveraging KG embedding methodologies to predict possible interactions between drugs, Myklebust et al. [118] assessed the ecotoxicological effect of chemicals via KG embedding and Mohamed et al. [119] predicted possible drug targets via KG embeddings.

4. Example applications of KGs in drug development and safety assessment

While KGs represent a valuable instrument that facilitate the integration of multi-source and heterogeneous data, they provide an unprecedented opportunity to gain new knowledge to guide *de novo* drug design. However, disentangling and understanding data of such high complexity and diversity is perhaps the biggest challenge of big data exploitation. A schematic representation of how KGs can be applied to aid clinical trials during compound development and risk assessment is displayed in Fig. 4 while multiple examples of KGs and the knowledge gained from them in different areas of toxicology and chemical/ drug development are outlined in this section.

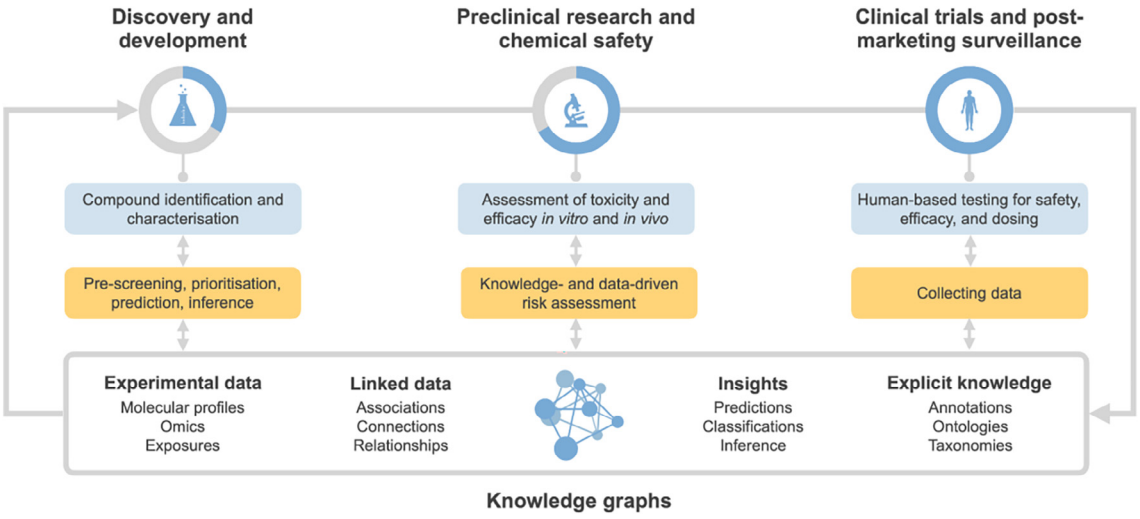


Fig. 4. Schematic representation for compound development and risk assessment. In a data-driven pipeline only compounds that pass the knowledge-based risk assessment, for example via a KG, are allowed to continue into experimental based evaluations. This reduces development costs, increases safety and improves development speed since only compounds with a high probability of success are allowed to continue. New data generated can constantly be re-fed into the KG and used to re-evaluate the compounds for the next step in the pipeline. All information gained during the process is added to the KG and can be used for other compounds in the future.

4.1. Drug adverse outcome & drug target predictions

Prediction models to link chemicals or drugs with their possible phenotypic outcomes, such as possible side effects/ adverse reactions, have been developed in a KG framework [14,15]. Often the prediction of adverse drug reactions is carried out by considering one data layer at a time, such as the chemical structure, ADME (absorption, distribution, metabolism, excretion), or its molecular targets. KGs give the opportunity to investigate a drug or a set of drugs over multiple data layers at the same time, in a combined data model and analysis framework. These approaches aid the drawing of connections among drugs, relying on more robust predictions that are based on a bigger number of characteristics with respect to the past. Zhang et al. [14] constructed a KG comprising drug, indication, target and side effect (adverse outcome) nodes, and three relationship types between them (has side effect, has target and has indication) [14]. Through node embedding and the application of a classifier they tried to link drugs with possible adverse outcomes. They tested their model on a dataset of 862 FDA approved drugs, containing information if the drug has a risk or no risk of inducing liver injury. While they were not able to infer the severity of the risk of liver injury, their model was able to discriminate between those not inducing liver injury and those that can induce liver injury.

One of the crucial steps in drug development is the identification of drug targets. Predicting the ability of a certain compound to interact with a molecular target and the effect that the compound has on it is a challenging task, which can be simplified and sped up through the application of KGs and thorough integration of large scale and diverse data layers. For example, Thafar et al. [120], developed a computational method, called DTiGEMS+, that can predict drug-target interactions by combining graph embedding, by means of the node2vec algorithm [115], and machine learning classifiers such as artificial multilayer perceptron [121], random forests [122] and adaptive boosting [123].

4.2. Predicting drug-drug interactions

Predicting chemical-chemical or drug-drug interactions *in silico* can reduce development costs significantly as well as improve their safety [22]. Different KG based frameworks to predict drug-drug interactions have been proposed recently [21,22,124]. Both Wang et al. [21] and Karim et al. [22] propose a framework leveraging KG embedding methodologies to predict possible interactions between drugs, while Abdelaziz et al. [124] make use of similarity metrics computed on drug information and KG structure applied to a logistic regression model to identify potential drug-drug interactions.

4.3. Drug repositioning

One of the fastest and cost-efficient methods to treat existing or new diseases is through drug repositioning, where already approved/ existing drugs are applied to other conditions and therefore only require a fraction of the assessment and approval step than novel compounds would. However, testing multiple compounds for a set of conditions in an in vivo/ clinical setting is not feasible. Therefore narrowing down compounds to likely successful candidates is necessary. Data integration and the application of KGs can provide a feasible computational infrastructure for large scale drug repositioning candidate detection [16–18]. Such KG based frameworks have gained a lot of attention recently in their possible application for new occurring diseases where a rapid response is required, such as COVID - 19 [16,20,125], drug target prediction applications [116] as well as closing the genotype-phenotype gap [126].

4.4. Chemical risk assessment

Assessing a compound's toxicity, being it onto the environment or organisms *in silico*, can significantly reduce costs and time needed to invest in *in vivo* or *in vitro* studies. In addition, possible toxic compounds can be discarded, if needed, already during the early development process instead of during later state testing. KGs are a framework that allow the fast screening and assessment of compounds with respect to their possible toxic effects on the environment or organisms as well as can provide necessary background information of specific compounds [118,127,128]. Myklebust et al. [127] created the TERA KG to assess chemical toxicity via node embedding. TERA combines chemical information from 3 data sources, toxicity information from ECOTOX (https://cfpub.epa.gov/ecotox/index.cfm) with taxonomy data from 2 data sources. To cirumvere the entity mapping challenge (s. next section), Myklebust et al. [127] used the Wikidata mapping engine (wikidata.org). They evaluated 9 different node embedding models on TERA to show the improvement node embedding can have on the prediction accuracy of neural networks. Zheng et al. [128] showcased the usage of KGs as an integrated data source, where data from unstructured documents were collected through a deep learning based entity recognition system, with the goal to create a unified system, containing information about the effective risk management of hazardous chemicals.

4.5. Biological drugs

Biological drugs or biologics are products of living organisms, or contain parts of living organisms, such as recombinant proteins, mRNA-based vaccines, blood components, cells, antibodies, etc. The development of biological drugs has substantially increased in the last years, since they offer many advantages compared to small molecules, especially with regards to their high target specificity [129,130]. To date, the KG framework has been only marginally exploited in the R&D of biological drugs, there is no specific technical challenge preventing biological drug properties data to be integrated into a KG. Interactions between biologics (e.g. peptides, antibodies or viral nanoparticles) and other already discussed compounds [131], such as cells or gene products, can be modelled in a KG natively due to their relationship focused data. The same applies for associations of these biologics to phenotypes, their ability to bind certain (chemical) compounds, when used as carriers [132], as well as attributes describing their 3D/ 2D structure or makeup. Such a KG can be used to design biologics with desired binding capabilities, with respect to both their target destination and/or their binding compound.

4.6. Possible KG application for the toxicological definition of point of departure

While, to our knowledge, there have been no efforts to date to investigate the possibility of KG models for time and/or dose-dependent predictions, such as the identification of safe doses for novel compounds, KG and big data models in combination with experimental data could be promising. Under the assumption that compounds with similar chemical characteristics would exert the same effects, KGs can be exploited to predict effective doses of a new compound for specific experimental conditions. This could be achieved by applying a read-across based approach, where knowledge from structurally similar compounds in the KG is used to infer possible behaviour for an unknown compound. This can be useful to speed up the initial phases of chemical development and increase the success rate of the process (Fig. 4). On the other hand, when dose-dependent modelling of transcriptomic experiments is performed [133–135], a list of dose-dependent genes with effective

doses is identified. KGs can be used to further enrich functional information about these genes, and their interaction with specific chemical structures or target information. Moreover, subgraphs contained in the KG could be used to identify or compare compounds with similar dose-dependent alteration profiles in order to categorise and characterise their effectiveness.

4.7. Clinical trials

During clinical trials, large amounts of data are gathered that need to be processed and ideally managed in a way that makes them available for future studies (being lab based or *in silico* based). By exploring the data model and database side of KGs, these data can be integrated into a KG for easy access and use as well as to link findings to other data, which can be of the same type (e.g., to access frequency or quality of the results) or of a different type (to link it to other types of knowledge). Chen et al. [136] proposed the Clinical Trials KG to combine information about different clinical trials, such as drugs and conditions studied, and evaluated its suitability for drug repositioning (via node embedding) and the identification of similar medical entities (e.g. to find a similar study of a specific study). By combining the Clinical Trials KG with some of the previously mentioned KGs, for example containing drug (structural) information, phenotypic information or information about a compound's MOA, we believe that for example the linkage of chemical sub-structures to clinical trial outcomes could be possible. By analysing successful trials of similar compounds/ phenotypes, suitable clinical trial set-ups could be suggested by the KG, in addition to leveraging the knowledge gathered during analysis and evaluation of the study. Such a possible workflow of data gathering, creating and analysing during all steps of a compound's development is represented in Fig. 3.

5. Challenges associated with KG in drug development and chemical safety assessment

The previously mentioned examples showcase the effectiveness of KGs in chemical risk assessment and drug development, however they may not have yet achieved their full potential. Many of these introduced KGs are constructed from a limited amount of data sources, data layers (Table 2) and are problem specific. While context specific KGs are easier to construct and leverage, they are limited in their re-usability to other problem domains. The limitation of data sources and data types further introduces context/ data specific biases into the KG and in result into its analysis. This section outlines multiple challenges associated with KGs, especially for KGs associated with the drug development and chemical safety assessment domain, limiting the potential and growth of current KG systems.

5.1. Lack of standards on data management and reporting

Successful large scale data integration highly depends on standardizations of individual data sets of the same data type, detailed

Table 2
Examples of KGs, their size and integrated data layers.

Publication	Problem	Number of Data Layers	Data Layers	KG Size
Zhang et al. [14]	Prediction of Adverse Drug Reactions	3	Drug - Side Effect Drug - Target Drug - Indication	12,473 nodes 154,239 relationships
Al-Saleem et al. [16]	Drug Repositioning for Covid-19	11	Gene - Gene Gene - Virus Gene - Disease Gene - Biological Process Gene - Pathway Gene - Molecular Function Gene - Small Molecule Small Molecule - Side Effect Small Molecule - Clinical Trial Clinical Trial - Virus Clinical Trial - Disease	> 6 M nodes > 18 M edges
Pavel & del Giudice et al. [20]	Identification of Genes Associated with Covid-19	2	Gene - Gene Gene - Drug	27,892 nodes 5,964,612 edges
Wang et al. [21]	Prediction of Drug - Drug Interactions	5	Drug - Gene (3 relationship types) Gene - Pathway Pathway - Phenotype	NA
Thafar et al. [120]	Prediction of Drug - Target Interactions	1	Usage of multiple benchmarking data sets [137] containing Drug - Gene relationships	
Mohamed et al. [116]	Prediction of Drug - Target Interactions	1	Usage of multiple benchmarking data sets [137,138] & a KEGG [72] based one; containing Drug - Gene relationships	
Abdelaziz et al. [124]	Prediction of Drug - Drug Interactions	At least 6	Drug - Gene Drug - Disease Gene - Gene Gene - Disease Chemical - Pathway Gene - Function	NA
Zhang et al. [125]	Drug repositioning for Covid-19	At least 15	Based on subset of SemMedDB [139] & CORD-19 [140]	331,427 nodes 20,017,236 edges
Chen et al. [136]	Collection of Clinical Trial data	21	Meta data & results of the clinical trials but not linked to additional information outside of this data	

metadata reporting as well as the accessibility of the data through computational means (e.g. APIs, computational processable reporting formats). Many of the available datasets have been generated independently and for different purposes and therefore vary greatly with respect to their quality, data points, data identifiers and metadata reported, making it challenging to compare and integrate these data sets. While FAIR is a start in introducing standardisation and re-usability of produced data, it has recently been criticised for lacking in quality standardisation [7]. In addition, it is mainly aimed at individual data sets and not towards large scale integration of multiple data sets, which would require in addition guidelines for naming and identification standards, especially across sub-disciplines.

5.2. Diversity of standards and ID systems

The biological research field is by tradition a highly fractured field, where a major difference in naming standards, processes and protocols can be found between sub-disciplines [98,141], making large scale data integration of multiple data sources, especially coming from different sub-disciplines, highly challenging. The basis to solving this problem is not an algorithmic challenge but a semantic one: common naming standards and ID systems need to be generated across the different sub-disciplines as well as extensive computational mapping systems should be provided publicly. In the context of drug development and chemical assessment many different data layers are affected by this same problem of which some examples are outlined in Table 3. These data reporting and identification related challenges are one of the main underlying issues, why large scale and problem unspecific KGs have not been developed yet, yielding mostly low data-layer, low data-source and application specific KGs, not aimed at re-use, as shown in Table 2. This suggests that a lack of semantic definitions and agreement between agencies (e.g. NCBI vs Ensembl) and sub-disciplines has a long lasting impact on the Big Data leverage possibilities of the life sciences. However, while all life science research fields would ideally agree on the same semantic database to be used, this likely is an unrealistic world view. How can you for example globally unify language dependent differences, appoint a single authority that makes decisions for every-one (across sub-disciplines) as well as ensure that such a consortium has unlimited funding and the necessary authority to enforce such a semantic database. Through data integration strategies and (manual) data mapping, it is possible to identify the *most* shared entities across data sets and to create links between knowledge from different data domains. However, the emphasis is on *most,* indicating that researchers need to accept that while data integration will provide more data and knowledge it is possible that parts of individual data sets become "unusable" (e.g., through not being able to be mapped to other data source identifiers) or that through automatic entity mapping systems errors will occur.

While the Natural Language Processing (NLP) field works fiercely on developing methods to extract information from academic (or free) texts as well as to provide methods to map between terms [142], they are often struggling with the specificity of biological terms and often require manual adjustments. For example, the

Table 3
Data integration related challenges for different data types possibly needed in a drug and chemical centred KG.

Data Type	Common Identifiers & Ontologies	Associated Challenges for the Data Integration Task
Chemicals/Drugs/ Compounds	SMILE Canonical SMILES Fingerprints Molecular descriptors inchKEy	While canonical SMILES are defined, they are not always used in reporting but instead their parent identifiers of simple SMILES are used, which change based on where in the compound structure they are started. Therefore multiple SMILES for the same compound can be created.Depending on the features used to compute chemical fingerprints or molecular descriptors, the same fingerprint/ descriptor can be computed for compounds varying in their 3D structure (e.g. through bond rotation)
	Brand or company	.
	Active principle NameThe Drug Ontology (https://purl.obolibrary.org/obo/dron.owl)	Drug names are often brand and language dependent, yielding therefore different names for the same compound.
Genes/ Gene products	Entrez Ensembl Gene symbols Location proteinID Probe ID	Between different identification systems there is not always a 1-1 mapping available. In addition, different platforms have different algorithms underneath to detect possible genes, making them vary in location, identification and even in what is considered a gene.
Gene Sets	Pathways Disease Associations AOPs GO	Even though for example pathways are defined on a conceptual level, pathways are not 1–1 mappable between platforms. Pathways/ GO terms that are considered the same, may not always have the same gene sets associated with them. Key Events within an AOP are manually created, inducing human error, such as duplicated Key Events due to differences in describing/ naming the underlying event.
Clinical Data/ Phenotypes	Name Description Ontology of Adverse Events [144] ICD UMLS OMIM MESH Orphanet Rare Disease Ontology [145] LOINC OMOP	Medical terms are often language dependent, making an international mapping challenging. In addition many different "unified" standards have been proposed, which use different terms and classes, indicating that a 1-1 mapping does not always exist, in addition to the challenge that every user will have their own preferences to which naming system to use. For medical professionals, the patient is at the centre and not the re-use of reporting of insights in a computational readable and processable format. Even if computational/ electronic health records are used, their standards vary across disciplines, borders and institutes. In addition their main purpose is to record a patient's health (or specified study) and not large scale, integratable research data.
Celllines / Tissues	Name Cell Ontology [146] Cell Line Ontology [147] The BRENDA Tissue Ontology [148,149]	There is no agreed standard on how to report cell-line or tissue names and especially for commercial cell-lines the names may be producer dependent.

meaning of a term can change with a single word, such as "not, upregulated, downregulated, increase, decrease", which will yield a high matching score in the algorithm, even though the terms may actually describe opposite events. The usage of different terms to describe the same "thing", or the usage of abbreviations [143], also proves challenging for NLP algorithms and often requires them to be provided with a pre-defined dictionary, which needs to be created (mostly) manually [143].

5.3. Concept mapping and data linking challenges

Going hand in hand with the challenges outlined in 5.1 and 5.2 another difficulty to overcome in order to make Big Data and KGs suitable for chemical safety assessment and drug development is that data right now is not at the forefront in all sub-disciplines of the life sciences. From a clinical point of view, the patient is at the centre and the re-use of such data in the best case is an afterthought and may result in a case report next to free-text entries in their medical record [150]. While during clinical studies, disease progressions, treatment responses or comorbidities may be outlined and reported, this is often done in writing, which traditionally is challenging to process computationally in combination with the previous outlined challenges. The same can be said for the academic research field though, where experimental outcomes are again often only reported in a publication, and if the data is provided, as outlined in 5.2., a lot of details are getting lost in translation. While the NLP field is working on methods to extract valuable medical information from text [142,150], there are still multiple draw-backs and challenges associated with it and until now no consensus has been reached on what method may work the most reliable [150]. This puts at the current moment in time the responsibility back towards the data generators, which requires every sub-discipline to realises the value of data, to understand that humans cannot process the amount of data available as well as that data coming from different sub-disciplines only in combination will provide insight into the bigger picture. While data provision and reporting becomes more common, it still needs to become more wide-spread together with a general understanding of computation methods, and data management by every researcher in the field, in order for.

individual researchers to make informed decisions on how and what to report. However, we expect this to automatically change over time, with computers playing a large role in the daily lives of current and next generation researchers together with an increase in computational methods taught during their education.

5.4. Unavailability of negative data

In order to learn the most from available data, not only positive results should be reported but negative ones as well and integrated into the knowledge base. Commonly, such negative results are not reported and therefore are not available to the wider community, resulting in the loss of valuable information. Therefore researchers should adapt to a more data centred approach [82], with the goal of reporting everything - from metadata to failed approaches. This allows on the one hand to learn negative samples from the data as well as allows other researchers to not waste valuable resources on the same or similar experiments. Many of the previously described KG applications relied on supervised classification tasks [14,22,127]. However the life science domain often struggles with the availability of true negative relationships, since from an experimental point of view they are not worth testing or reporting. For example Zhang et al. [14], used in their adverse outcome prediction problem drug indication pairs as negative data points for their classifier. However from a biological point of view a drug's indication and adverse outcome are closely related and may even be dose or situation dependent. This suggests that drug indication pairs and drug adverse outcome pairs are not significantly different from a biological point of view, making them highly unsuitable as substitutes for true unrelated drug phenotype relationships. But without the existence of true negative data points the training and validation of such classifiers stays difficult.

6. Summary and outlook

This review provided an overview of the advantages of data modelling and explorations by means of graph databases and KGs in the context of chemical safety assessment and drug design. These processes rely on vast and diverse data sets from many different areas in the life sciences. KGs can significantly improve data integration, data re-use, data access and data quality of such diverse data sets. In this review, examples of successful KG applications for different tasks were provided, such as drug repositioning, drug target prediction, drug-drug interaction, its application in clinical trials and for chemical risk assessment. Finally current challenges, which are suggested to hinder KGs to reach their full potential in drug development and chemical safety assessment were outlined. This review suggests a shift in mentality across the multiple sub-fields in the life sciences, towards a data centred approach, where semantic standards, data creation and availability methods and data re-use are at its centre.

Additionally, more research into large scale KGs need to be performed, especially for their application into the life sciences. KGs have found widespread use in the technical industry. However the data included in these KGs is often less diverse, has lower variance in quality and the KG usages are of less variance than when KGs are applied in the life sciences. Therefore it is necessary that more research into the applicability domain of KGs, especially for the life sciences, has to be conducted.

In conclusion, KGs are emerging as a successful tool for drug & chemical development and their safety assessment. This review suggests that the use of data-driven approaches on top of a KG infrastructure, in combination with a data centred view, can accelerate these processes significantly and solve multiple challenges associated with the compound development process and its safety assessment.

Funding

This study was supported by the Academy of Finland [322761], EU H2020 NanoSolveIT project [814572] and European Research Council (ERC) programme, Consolidator project "ARCHIMEDES" [101043848].

CRediT authorship contribution statement

Alisa Pavel: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Laura A. Saarimäki:** Writing – review & editing, Visualization. **Lena Möbus:** Writing – review & editing, Supervision. **Antonio Federico:** Writing – review & editing, Supervision. **Angela Serra:** Conceptualization, Writing – review & editing, Project administration, Supervision. **Dario Greco:** Conceptualization, Writing – review & editing, Project administration, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] DiMasi JA, Feldman L, Seckler A, Wilson A. Trends in risks associated with new drug development: success rates for investigational drugs. Clin Pharmacol Ther 2010;87:272–7. https://doi.org/10.1038/clpt.2009.295.

[2] DiMasi JA, Reichert JM, Feldman L, Malins A. Clinical approval success rates for investigational cancer drugs. Clin Pharmacol Ther 2013;94:329–35. https://doi.org/10.1038/clpt.2013.117.

[3] Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. Health Policy 2011;100:4–17. https://doi.org/10.1016/j.healthpol.2010.12.002.

[4] Freires IA, Sardi J de CO, de Castro RD, Rosalen PL. Alternative animal and non-animal models for drug discovery and development: bonus or burden? Pharm Res 2017;34:681–6. https://doi.org/10.1007/s11095-016-2069-z.

[5] Serra A, Fratello M, Federico A, Ojha R, Provenzani R, Tasnadi E, et al. Computationally prioritized drugs inhibit SARS-CoV-2 infection and syncytia formation. Brief. Bioinformatics 2022;23. https://doi.org/10.1093/bib/bbab507.

[6] Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:. https://doi.org/10.1038/sdata.2016.18160018.

[7] Saarimäki LA, Melagraki G, Afantitis A, Lynch I, Greco D. Prospects and challenges for FAIR toxicogenomics data. Nat Nanotechnol 2022;17:17–8. https://doi.org/10.1038/s41565-021-01049-1.

[8] Ehrlinger L, Wöß W. Towards a Definition of Knowledge Graphs. Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16), vol. 1695, Leipzig, Germany: CEUR-WS; 2016.

[9] Hogan A, Blomqvist E, Cochez M, D'amato C, Melo GD, Gutierrez C, et al. Knowledge graphs. ACM Comput Surv 2021;54:1–37. https://doi.org/10.1145/3447772.

[10] Sheth A, Padhee S, Gyrard A, Sheth A. Knowledge graphs and knowledge networks: the story in brief. IEEE Internet Comput 2019;23:67–75. https://doi.org/10.1109/MIC.2019.2928449.

[11] Ji S, Pan S, Cambria E, Marttinen P, Yu PS. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst 2021;PP.. https://doi.org/10.1109/TNNLS.2021.3070843.

[12] Ali W, Saleem M, Yao B, Hogan A, Ngomo A-C-N. A survey of RDF stores & SPARQL engines for querying knowledge graphs. VLDB J 2022;31:1–26. https://doi.org/10.1007/s00778-021-00711-3.

[13] Paul S, Mitra A, Koner C. A Review on Graph Database and its representation. 2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC), IEEE; 2019, p. 1–5. https://doi.org/10.1109/ICRAECC43874.2019.8995006.

[14] Zhang F, Sun B, Diao X, Zhao W, Shu T. Prediction of adverse drug reactions based on knowledge graph embedding. BMC Med Inform Decis Mak 2021;21:38. https://doi.org/10.1186/s12911-021-01402-3.

[15] Nováček V, Mohamed SK. Predicting polypharmacy side-effects using knowledge graph embeddings. AMIA Jt Summits Transl Sci Proc 2020;2020:449–58.

[16] Al-Saleem J, Granet R, Ramakrishnan S, Ciancetta NA, Saveson C, Gessner C, et al. Knowledge graph-based approaches to drug repurposing for COVID-19. J Chem Inf Model 2021;61:4058–67. https://doi.org/10.1021/acs.jcim.1c00642.

[17] Xiong Z, Huang F, Wang Z, Liu S, Zhang W. A multimodal framework for improving in silico drug repositioning with the prior knowledge from knowledge graphs. IEEE/ACM Trans Comput Biol Bioinform 2021;PP.. https://doi.org/10.1109/TCBB.2021.3103595.

[18] Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics 2019;35:5191–8. https://doi.org/10.1093/bioinformatics/btz418.

[19] Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J 2020;18:1414–28. https://doi.org/10.1016/j.csbj.2020.05.017.

[20] Pavel A, del Giudice G, Federico A, Di Lieto A, Kinaret PAS, Serra A, et al. Integrated network analysis reveals new genes suggesting COVID-19 chronic effects and treatment. Brief Bioinformatics 2021. https://doi.org/10.1093/bib/bbaa417.

[21] Wang M, Wang H, Liu X, Ma X, Wang B. Drug-Drug Interaction Predictions via Knowledge Graph and Text Embedding: Instrument Validation Study. JMIR Med Inform 2021;9:. https://doi.org/10.2196/28277e28277.

[22] Karim MdR, Cochez M, Jares JB, Uddin M, Beyan O, Decker S. Drug-Drug Interaction Prediction Based on Knowledge Graph Embeddings and Convolutional-LSTM Network. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics - BCB '19, New York, New York, USA: ACM Press; 2019, p. 113–23. doi: 10.1145/3307339.3342161.

[23] Federico A, Fratello M, Scala G, Möbus L, Pavel A, Del Giudice G, et al. Integrated Network Pharmacology Approach for Drug Combination Discovery: A Multi-Cancer Case Study. Cancers (Basel) 2022;14. doi: 10.3390/cancers14082043.

[24] Serra A, Önlü S, Coretto P, Greco D. An integrated quantitative structure and mechanism of action-activity relationship model of human serum albumin binding. J Cheminform 2019;11:38. https://doi.org/10.1186/s13321-019-0359-2.

[25] Kinaret PAS, Serra A, Federico A, Kohonen P, Nymark P, Liampa I, et al. Transcriptomics in toxicogenomics, part I: experimental design, technologies, publicly available data, and regulatory aspects. Nanomaterials (Basel) 2020;10. doi: 10.3390/nano10040750.

[26] Federico A, Serra A, Ha MK, Kohonen P, Choi J-S, Liampa I, et al. Transcriptomics in toxicogenomics, part II: preprocessing and differential expression analysis for high quality data. Nanomaterials (Basel) 2020;10. doi: 10.3390/nano10050903.

[27] Serra A, Fratello M, Cattelani L, Liampa I, Melagraki G, Kohonen P, et al. Transcriptomics in toxicogenomics, part III: data modelling for risk assessment. Nanomaterials (Basel) 2020;10. doi: 10.3390/nano10040708.

[28] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47:D1102–9. https://doi.org/10.1093/nar/gky1033.

[29] Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res 2016;44:D380–4. https://doi.org/10.1093/nar/gkv1277.

[30] Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. J Chem Inf Model 2020;60:6065–73. https://doi.org/10.1021/acs.jcim.0c00675.

[31] Ruusmann V, Sild S, Maran U. QSAR DataBank repository: open and linked qualitative and quantitative structure-activity relationship models. J Cheminform 2015;7:32. https://doi.org/10.1186/s13321-015-0082-6.

[32] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016;44:D1075–9. https://doi.org/10.1093/nar/gkv1075.

[33] Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen D-T, Bologa CG, et al. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Res 2021;49:D1334–46. https://doi.org/10.1093/nar/gkaa993.

[34] Liu H, Zhang W, Zou B, Wang J, Deng Y, Deng L. DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. Nucleic Acids Res 2020;48:D871–81. https://doi.org/10.1093/nar/gkz1007.

[35] Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Res 2021;49:D1138–43. https://doi.org/10.1093/nar/gkaa891.

[36] Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: supporting systematic drug-target identification and prioritisation. Nucleic Acids Res 2021;49:D1302–10. https://doi.org/10.1093/nar/gkaa1027.

[37] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006;34:D668–72. https://doi.org/10.1093/nar/gkj067.

[38] Thomas RS, Paules RS, Simeonov A, Fitzpatrick SC, Crofton KM, Casey WM, et al. The US Federal Tox21 Program: a strategic and operational plan for continued leadership. ALTEX 2018;35:163–8. https://doi.org/10.14573/altex.1803011.

[39] Richard AM, Huang R, Waidyanatha S, Shinn P, Collins BJ, Thillainadarajah I, et al. The tox21 10K compound library: collaborative chemistry advancing toxicology. Chem Res Toxicol 2021;34:189–216. https://doi.org/10.1021/acs.chemrestox.0c00264.

[40] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10. https://doi.org/10.1093/nar/30.1.207.

[41] Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell 2017;171:1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

[42] Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res 2015;43:D921–7. https://doi.org/10.1093/nar/gku955.

[43] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res 2021;49:D884–91. https://doi.org/10.1093/nar/gkaa942.

[44] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129–41. https://doi.org/10.1101/gr.772403.

[45] Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res 2021;49:D394–403. https://doi.org/10.1093/nar/gkaa1106.

[46] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res 2017;45:D408–14. https://doi.org/10.1093/nar/gkw985.

[47] Patil A, Nakai K, Nakamura H. HitPredict: a database of quality assessed protein-protein interactions in nine species. Nucleic Acids Res 2011;39:D744–9. https://doi.org/10.1093/nar/gkq897.

[48] López Y, Nakai K, Patil A. HitPredict version 4: comprehensive reliability scoring of physical protein-protein interactions from more than 100 species. Database (Oxford) 2015;2015. doi: 10.1093/database/bav117.

[49] Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature 2020;580:402–8. https://doi.org/10.1038/s41586-020-2188-x.

[50] Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT: the Molecular INTeraction database. Nucleic Acids Res 2007;35:D572–4. https://doi.org/10.1093/nar/gkl950.
[51] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2014;42:D358–63. https://doi.org/10.1093/nar/gkt1115.
[52] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47:D607–13. https://doi.org/10.1093/nar/gky1131.
[53] Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, et al. TRRUST: a reference database of human transcriptional regulatory interactions. Sci Rep 2015;5:11432. https://doi.org/10.1038/srep11432.
[54] Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 2018;46:D380–6. https://doi.org/10.1093/nar/gkx1013.
[55] Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. ELife 2015;4. https://doi.org/10.7554/eLife.05005.
[56] McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, et al. The biochemical basis of microRNA targeting efficacy. Science 2019;366. https://doi.org/10.1126/science.aav1741.
[57] Huang H-Y, Lin Y-C-D, Li J, Huang K-Y, Shrestha S, Hong H-C, et al. updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res 2020;2020(48):D148–54. https://doi.org/10.1093/nar/gkz896.
[58] Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. Nucleic Acids Res 2013;41:D1228–33. https://doi.org/10.1093/nar/gks1147.
[59] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2022;50:D20–6. https://doi.org/10.1093/nar/gkab1112.
[60] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46:D1062–7. https://doi.org/10.1093/nar/gkx1153.
[61] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48:D845–55. https://doi.org/10.1093/nar/gkz1021.
[62] Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res 2021;49:D1207–17. https://doi.org/10.1093/nar/gkaa1043.
[63] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019;47:D1005–12. https://doi.org/10.1093/nar/gky1120.
[64] Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update – from bulk to single-cell expression data. Nucleic Acids Res 2019;47:D711–5. https://doi.org/10.1093/nar/gky964.
[65] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9. https://doi.org/10.1038/75556.
[66] The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 2021;49:D325–34. https://doi.org/10.1093/nar/gkaa1113.
[67] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
[68] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27:1739–40. https://doi.org/10.1093/bioinformatics/btr260.
[69] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2020;48:D498–503. https://doi.org/10.1093/nar/gkz1031.
[70] Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res 2021;49:D613–21. https://doi.org/10.1093/nar/gkaa1024.
[71] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61. https://doi.org/10.1093/nar/gkw1092.
[72] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.
[73] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44:W90–7. https://doi.org/10.1093/nar/gkw377.
[74] Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science 2015;347:1260419. https://doi.org/10.1126/science.1260419.
[75] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74. https://doi.org/10.1038/nature11247.
[76] Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. Cell Miner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res 2012;72:3499–511. https://doi.org/10.1158/0008-5472.CAN-12-1370.
[77] Shankavaram UT, Varma S, Kane D, Sunshine M, Chary KK, Reinhold WC, et al. Cell Miner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genomics 2009;10:277. https://doi.org/10.1186/1471-2164-10-277.
[78] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82. https://doi.org/10.1093/nar/gkx1037.
[79] Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res 2017;45:D985–94. https://doi.org/10.1093/nar/gkw1055.
[80] Marwah VS, Kinaret PAS, Serra A, Scala G, Lauerma A, Fortino V, et al. Inform: inference of network response modules. Bioinformatics 2018;34:2136–8. https://doi.org/10.1093/bioinformatics/bty063.
[81] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinf 2006;7(Suppl 1):S7. https://doi.org/10.1186/1471-2105-7-S1-S7.
[82] Tolani P, Gupta S, Yadav K, Aggarwal S, Yadav AK. Big data, integrative omics and network biology. Adv Protein Chem Struct Biol 2021;127:127–60. https://doi.org/10.1016/bs.apcsb.2021.03.006.
[83] Albert R. Network inference, analysis, and modeling in systems biology. Plant Cell 2007;19:3327–38. https://doi.org/10.1105/tpc.107.054700.
[84] Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinformatics 2018;19:1370–81. https://doi.org/10.1093/bib/bbx066.
[85] Pavel A, Federico A, Del Giudice G, Serra A, Greco D. Volta: adVanced mOLecular neTwork Analysis. Bioinformatics 2021. https://doi.org/10.1093/bioinformatics/btab642.
[86] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7:. https://doi.org/10.1136/bmjopen-2016-012545e012545.
[87] Tiddi I, Balliet D, ten Teije A. Fostering Scientific Meta-analyses with Knowledge Graphs: A Case-Study. In: Harth A, Kirrane S, Ngonga Ngomo A-C, Paulheim H, Rula A, Gentile AL, et al., editors. The semantic web: 17th international conference, ESWC 2020, heraklion, crete, greece, may 31–june 4, 2020, proceedings, vol. 12123, Cham: Springer International Publishing; 2020, p. 287–303. https://doi.org/10.1007/978-3-030-49461-2_17.
[88] Badkas A, De Landtsheer S, Sauter T. Topological network measures for drug repositioning. Brief. Bioinformatics 2021;22. https://doi.org/10.1093/bib/bbaa357.
[89] Yang T, Nowack B. A meta-analysis of ecotoxicological hazard data for nanoplastics in marine and freshwater systems. Environ Toxicol Chem 2020;39:2588–98. https://doi.org/10.1002/etc.4887.
[90] Wang Z, Arat S, Magid-Slav M, Brown JR. Meta-analysis of human gene expression in response to Mycobacterium tuberculosis infection reveals potential therapeutic targets. BMC Syst Biol 2018;12:3. https://doi.org/10.1186/s12918-017-0524-z.
[91] Upreti VV, Venkatakrishnan K. Model-based meta-analysis: optimizing research, development, and utilization of therapeutics using the totality of evidence. Clin Pharmacol Ther 2019;106:981–92. https://doi.org/10.1002/cpt.1462.
[92] Chan P, Peskov K, Song X. Applications of model-based meta-analysis in drug development. Pharm Res 2022. https://doi.org/10.1007/s11095-022-03201-5.
[93] Yan J, Wang C, Cheng W, Gao M, Zhou A. A retrospective of knowledge graphs. Front Comput Sci 2016;12:1–20. https://doi.org/10.1007/s11704-016-5228-9.
[94] Mallon A-M, Häring DA, Dahlke F, Aarden P, Afyouni S, Delbarre D, et al. Advancing data science in drug development through an innovative computational framework for data sharing and statistical analysis. BMC Med Res Methodol 2021;21:250. https://doi.org/10.1186/s12874-021-01409-4.
[95] Qian T, Zhu S, Hoshida Y. Use of big data in drug development for precision medicine: an update. Expert Rev Precis Med Drug Dev 2019;4:189–200. https://doi.org/10.1080/23808993.2019.1617632.
[96] Kiriiri GK, Njogu PM, Mwangi AN. Exploring different approaches to improve the success of drug discovery and development projects: a review. Future J Pharmaceutical Sci 2020;6:27. https://doi.org/10.1186/s43094-020-00047-9.
[97] Gupta S, Modgil S, Gunasekaran A. Big data in lean six sigma: a review and further research directions. Int J Prod Res 2019:1–23. https://doi.org/10.1080/00207543.2019.1598599.
[98] Leonelli S. What difference does quantity make? on the epistemology of big data in biology. Big Data & Society 2014;1. https://doi.org/10.1177/2053951714534395.
[99] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights 2020;14:1177932219899051. doi: 10.1177/1177932219899051.
[100] Recanatini M, Cabrelle C. Drug research meets network science: where are we? J Med Chem 2020;63:8653–66. https://doi.org/10.1021/acs.jmedchem.9b01989.
[101] Arrell DK, Terzic A. Network systems biology for drug discovery. Clin Pharmacol Ther 2010;88:120–5. https://doi.org/10.1038/clpt.2010.91.
[102] Knapen D, Angrish MM, Fortin MC, Katsiadaki I, Leonard M, Margiotta-Casaluci L, et al. Adverse outcome pathway networks I: Development and

applications. Environ Toxicol Chem 2018;37:1723–33. https://doi.org/10.1002/etc.4125.
[103] Blucher AS, McWeeney SK, Stein L, Wu G. Visualization of drug target interactions in the contexts of pathways and networks with ReactomeFIViz. [version 1; peer review: 3 approved]. F1000Res 2019;8:908. doi: 10.12688/f1000research.19592.1.
[104] Tanoli Z, Alam Z, Ianevski A, Wennerberg K, Vähä-Koskela M, Aittokallio T. Interactive visual analysis of drug-target interaction networks using Drug Target Profiler, with applications to precision medicine and drug repurposing. Brief Bioinformatics 2018. https://doi.org/10.1093/bib/bby119.
[105] Serra A, Letunic I, Fortino V, Handy RD, Fadeel B, Tagliaferri R, et al. INSIdE NANO: a systems biology framework to contextualize the mechanism-of-action of engineered nanomaterials. Sci Rep 2019;9:179. https://doi.org/10.1038/s41598-018-37411-y.
[106] Ratajczak F, Joblin M, Ringsquandl M, Hildebrandt M. Task-driven knowledge graph filtering improves prioritizing drugs for repurposing. BMC Bioinf 2022;23:84. https://doi.org/10.1186/s12859-022-04608-y.
[107] Glez-Peña D, Alvarez R, Díaz F, Fdez-Riverola F. DFP: a Bioconductor package for fuzzy profile identification and gene reduction of microarray data. BMC Bioinf 2009;10:37. https://doi.org/10.1186/1471-2105-10-37.
[108] Pavel A, Serra A, Cattelani L, Federico A, Greco D. Network analysis of microarray data. Methods Mol Biol 2022;2401:161–86. https://doi.org/10.1007/978-1-0716-1839-4_11.
[109] Zhu Y, Che C, Jin B, Zhang N, Su C, Wang F. Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. Health Informatics J 2020;26:2737–50. https://doi.org/10.1177/1460458220937101.
[110] Kinaret P, Marwah V, Fortino V, Ilves M, Wolff H, Ruokolainen L, et al. Network analysis reveals similar transcriptomic responses to intrinsic properties of carbon nanomaterials in vitro and in vivo. ACS Nano 2017;11:3786–96. https://doi.org/10.1021/acsnano.6b08650.
[111] Madi A, Kenett DY, Bransburg-Zabary S, Merbl Y, Quintana FJ, Tauber AI, et al. Network theory analysis of antibody-antigen reactivity data: the immune trees at birth and adulthood. PLoS ONE 2011;6:. https://doi.org/10.1371/journal.pone.0017445e17445.
[112] Lao N, Mitchell T, Cohen W. Random walk inference and learning in a large scale knowledge base. In: Proceedings of the 2011 conference on empirical methods in natural language processing. p. 529–39.
[113] Steenwinckel B, Vandewiele G, Weyns M, Agozzino T, Turck FD, Ongenae F. INK: knowledge graph embeddings for node classification. Data Min Knowl Discov 2022;36:620–67. https://doi.org/10.1007/s10618-021-00806-z.
[114] Berlingerio M, Koutra D, Eliassi-Rad T, Faloutsos C. Netsimile: A scalable approach to size-independent network similarity. ArXiv Preprint ArXiv:12092684 2012.
[115] Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. KDD 2016;2016:855–64. doi: 10.1145/2939672.2939754.
[116] Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. Bioinformatics 2020;36:603–10. https://doi.org/10.1093/bioinformatics/btz600.
[117] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space 2013.
[118] Myklebust EB, Jimenez-Ruiz E, Chen J, Wolf R, Tollefsen KE. Knowledge graph embedding for ecotoxicological effect prediction. In: Ghidini C, Hartig O, Maleshkova M, Svátek V, Cruz I, Hogan A, et al., editors. The semantic web – ISWC 2019: 18th international semantic web conference, auckland, new zealand, october 26–30, 2019, proceedings, part II, vol. 11779, Cham: Springer International Publishing; 2019, p. 490–506. doi: 10.1007/978-3-030-30796-7_30.
[119] Mohamed SK, Nounu A, Nováček V. Drug target discovery using knowledge graph embeddings. Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing – SAC '19, New York, New York, USA: ACM Press; 2019, p. 11–8. https://doi.org/10.1145/3297280.3297282.
[120] Thafar MA, Olayan RS, Ashoor H, Albaradei S, Bajic VB, Gao X, et al. DTiGEMS +: drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. J Cheminform 2020;12:44. https://doi.org/10.1186/s13321-020-00447-2.
[121] Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal 2000;22:717–27. https://doi.org/10.1016/S0731-7085(99)00272-1.
[122] Breiman L. Random Forests. Springer Science and Business Media LLC 2001. doi: 10.1023/a:1010933404324.
[123] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55:119–39. https://doi.org/10.1006/jcss.1997.1504.
[124] Abdelaziz I, Fokoue A, Hassanzadeh O, Zhang P, Sadoghi M. Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. Web Semantics: Science, Services and Agents on the World Wide Web 2017;0. doi: 10.1016/j.websem.2017.06.002.
[125] Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H. Drug repurposing for COVID-19 via knowledge graph completion. J Biomed Inform 2021;115:. https://doi.org/10.1016/j.jbi.2021.103696103696.
[126] Hu J, Lepore R, Dobson RJB, Al-Chalabi A, Bean DM, Iacoangeli A. DGLinker: flexible knowledge-graph prediction of disease-gene associations. Nucleic Acids Res 2021;49:W153–61. https://doi.org/10.1093/nar/gkab449.
[127] Myklebust EB, Jimenez-Ruiz E, Chen J, Wolf R, Tollefsen KE. Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings. Semantic Web – Interoperability, Usability, Applicability n.d.
[128] Zheng X, Wang B, Zhao Y, Mao S, Tang Y. A knowledge graph method for hazardous chemical management: ontology design and entity identification. Neurocomputing 2021;430:104–11. https://doi.org/10.1016/j.neucom.2020.10.095.
[129] Leader B, Baca QJ, Golan DE. Protein therapeutics: a summary and pharmacological classification. Nat Rev Drug Discov 2008;7:21–39. https://doi.org/10.1038/nrd2399.
[130] Muralidhara BK, Wong M. Critical considerations in the formulation development of parenteral biologic drugs. Drug Discov Today 2020;25:574–81. https://doi.org/10.1016/j.drudis.2019.12.011.
[131] Oren EE, Tamerler C, Sahin D, Hnilova M, Seker UOS, Sarikaya M, et al. A novel knowledge-based approach to design inorganic-binding peptides. Bioinformatics 2007;23:2816–22. https://doi.org/10.1093/bioinformatics/btm436.
[132] Chung YH, Cai H, Steinmetz NF. Viral nanoparticles for drug delivery, imaging, immunotherapy, and theranostic applications. Adv Drug Deliv Rev 2020;156:214–35. https://doi.org/10.1016/j.addr.2020.06.024.
[133] Saarimäki LA, Kinaret PAS, Scala G, del Giudice G, Federico A, Serra A, et al. Toxicogenomics analysis of dynamic dose-response in macrophages highlights molecular alterations relevant for multi-walled carbon nanotube-induced lung fibrosis. NanoImpact 2020;100274. https://doi.org/10.1016/j.impact.2020.100274.
[134] Serra A, Saarimäki LA, Fratello M, Marwah VS, Greco D. BMDx: a graphical Shiny application to perform Benchmark Dose analysis for transcriptomics data. Bioinformatics 2020;36:2932–3. https://doi.org/10.1093/bioinformatics/btaa030.
[135] Serra A, Fratello M, Del Giudice G, Saarimäki LA, Paci M, Federico A, et al. TinderMIX: time-dose integrated modelling of toxicogenomics data. GigaScience 2020;9. https://doi.org/10.1093/gigascience/giaa055.
[136] Chen Z, Peng B, Ioannidis VN, Li M, Karypis G, Ning X. A knowledge graph of clinical trials ([Formula: see text]). Sci Rep 2022;12:4724. https://doi.org/10.1038/s41598-022-08454-z.
[137] Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 2008;24:i232–40. https://doi.org/10.1093/bioinformatics/btn162.
[138] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008;36:D901–6. https://doi.org/10.1093/nar/gkm958.
[139] Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 2012;28:3158–60. https://doi.org/10.1093/bioinformatics/bts591.
[140] Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset. ArXiv 2020.
[141] Marx V. Biology: The big challenges of big data. Nature 2013;498:255–60. https://doi.org/10.1038/498255a.
[142] Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017;73:14–29. https://doi.org/10.1016/j.jbi.2017.07.012.
[143] Grossman Liu L, Grossman RH, Mitchell EG, Weng C, Natarajan K, Hripcsak G, et al. A deep database of medical abbreviations and acronyms for natural language processing. Sci Data 2021;8:149. https://doi.org/10.1038/s41597-021-00929-4.
[144] He Y, Sarntivijai S, Lin Y, Xiang Z, Guo A, Zhang S, et al. OAE: the ontology of adverse events. J Biomed Semantics 2014;5:29. https://doi.org/10.1186/2041-1480-5-29.
[145] Maiella S, Rath A, Angin C, Mousson F, Kremp O. Orphanet et son réseau : où trouver une information validée sur les maladies rares. Rev Neurol (Paris) 2013;169:S3–8. https://doi.org/10.1016/S0035-3787(13)70052-3.
[146] Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semantics 2016;7:44. https://doi.org/10.1186/s13326-016-0088-7.
[147] Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, et al. CLO: The cell line ontology. J Biomed Semantics 2014;5:37. https://doi.org/10.1186/2041-1480-5-37.
[148] Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res 2011;39:D507–13. https://doi.org/10.1093/nar/gkq968.
[149] Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 2021;49:D498–508. https://doi.org/10.1093/nar/gkaa1025.
[150] Kersloot MG, van Putten FJP, Abu-Hanna A, Cornet R, Arts DL. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. J Biomed Semantics 2020;11:14. https://doi.org/10.1186/s13326-020-00231-z.