



Databases and ontologies

# The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information

John H. Morris<sup>1</sup>, Karthik Soman<sup>2</sup>, Rabia E. Akbas<sup>2</sup>, Xiaoyuan Zhou<sup>2</sup>, Brett Smith<sup>3</sup>, Elaine C. Meng<sup>1</sup>, Conrad C. Huang<sup>1</sup>, Gabriel Cerono<sup>2</sup>, Gundolf Schenk<sup>4</sup>, Angela Rizk-Jackson<sup>4</sup>, Adil Harroud<sup>2</sup>, Lauren Sanders<sup>5</sup>, Sylvain V. Costes <sup>5</sup>, Krish Bharat<sup>2</sup>, Arjun Chakraborty<sup>2</sup>, Alexander R. Pico<sup>6</sup>, Taline Mardrossian<sup>7</sup>, Michael Keiser<sup>7</sup>, Alice Tang<sup>8</sup>, Josef Hardi<sup>9</sup>, Yongmei Shi<sup>4</sup>, Mark Musen<sup>9</sup>, Sharat Israni<sup>4</sup>, Sui Huang<sup>3</sup>, Peter W. Rose<sup>10</sup>, Charlotte A. Nelson<sup>2</sup> and Sergio E. Baranzini <sup>2,\*</sup>

<sup>1</sup>Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, San Francisco, CA 94158, USA, <sup>2</sup>Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, USA, <sup>3</sup>Institute for Systems Biology, Seattle, WA 98109, USA, <sup>4</sup>Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA 94158, USA, <sup>5</sup>Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94035, USA, <sup>6</sup>Data Science and Biotechnology, Gladstone Institutes, University of California, San Francisco, San Francisco, CA 94158, USA, <sup>7</sup>Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94143-2550, USA, <sup>8</sup>UCSF-UC Berkeley Bioengineering Graduate Program, University of California, San Francisco, San Francisco, CA 94158, USA, <sup>9</sup>Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305-5479, USA and <sup>10</sup>San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

\*To whom correspondence should be addressed.

Associate Editor: Zhiyong Lu

Received on July 26, 2022; revised on January 17, 2023; editorial decision on January 27, 2023; accepted on February 8, 2023

## Abstract

**Motivation:** Knowledge graphs (KGs) are being adopted in industry, commerce and academia. Biomedical KG presents a challenge due to the complexity, size and heterogeneity of the underlying information.

**Results:** In this work, we present the Scalable Precision Medicine Open Knowledge Engine (SPOKE), a biomedical KG connecting millions of concepts via semantically meaningful relationships. SPOKE contains 27 million nodes of 21 different types and 53 million edges of 55 types downloaded from 41 databases. The graph is built on the framework of 11 ontologies that maintain its structure, enable mappings and facilitate navigation. SPOKE is built weekly by python scripts which download each resource, check for integrity and completeness, and then create a 'parent table' of nodes and edges. Graph queries are translated by a REST API and users can submit searches directly via an API or a graphical user interface. **Conclusions/Significance:** SPOKE enables the integration of seemingly disparate information to support precision medicine efforts.

**Availability and implementation:** The SPOKE neighborhood explorer is available at <https://spoke.rbvi.ucsf.edu>.

**Contact:** sergio.baranzini@ucsf.edu

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

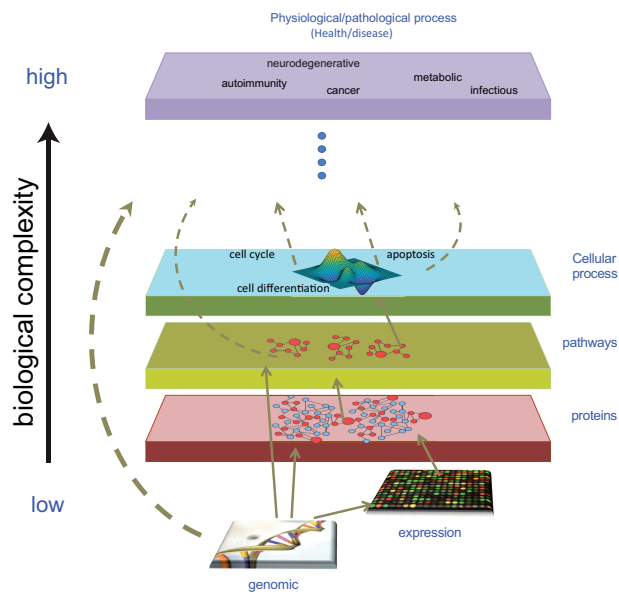


Fig. 1. Hierarchical organization of biological complexity. Biomedical information is largely compartmentalized according to disciplines. Integration of information may lead to the emergence of knowledge

## 1 Introduction

Data lead to information, and information leads to knowledge (Ackoff, 1989). Vast amounts of data are being produced at a breathtaking pace (Reinsel et al., 2018), and this explosion in the amount of generated data is causing the number and size of databases and repositories to increase exponentially. In the biomedical domain, this big data problem gets further compounded by the resulting compartmentalization of data resources according to specialty, likely driven by the enormous biological complexity underlying human physiology (Fig. 1).

Even where data and factual knowledge are stored in public repositories, their access and interpretation are still limited by physical, technical and thematic compartmentalization, making it difficult if not impossible for medical professionals to utilize this body of information and connect the dots to facilitate the emergence of knowledge.

Given the complexity of existing relationships among different biomedical fields, graph databases have recently gained popularity as a practical solution to integrate such disparate sources of information. Knowledge graphs with biomedical content have been developed using a variety of strategies, content and target applications (Fecho et al., 2021; Mattingly et al., 2006; Santos et al., 2022).

The scalable precision medicine open knowledge engine (SPOKE) is a knowledge graph that connects information from 41 specialized databases, structured as 21 different node types and 55 edge types, ranging from molecular and cellular biology to pharmacology and clinical practice. SPOKE was conceived with the philosophy that if relevant information is connected, it can result in the emergence of knowledge, and hence provide insights into the understanding of diseases, discovering of drugs and proactively improving personal health.

## 2 Materials and methods

### 2.1 Construction and enrichment of SPOKE

SPOKE currently uses 41 different data sources to construct the knowledge graph (Table 1) although new databases are being added continually. To construct SPOKE, a script downloads and processes each data source on a weekly basis. (See Supplementary material for a detailed description of databases and modeling.)

**Organisms:** Organisms in SPOKE are identified by their NCBI Taxonomy ID (Schoch et al., 2020). Species of interest are

determined by several different sources: bacterial information from KEGG (Kanehisa and Goto, 2000) and MetaCyc (Caspi et al., 2016) and pathogenic species from PathoPhenoDB (Kafkas et al., 2019).

**Proteins:** The source for all protein information in SPOKE is UniProt (Pundir et al., 2017). Both SwissProt (reviewed) and TrEMBL (unreviewed) proteins are retrieved for all of the leaf Organisms.

In addition to Protein-cleaves-to-Protein edges, we also incorporate data from several different sources to create Protein-interacts-Protein edges. For human proteins, the primary source for this information is STRING (Szklarczyk et al., 2019). In addition, all IntAct (Orchard et al., 2014) protein-protein interactions are retrieved for all proteins in SPOKE.

Finally, Protein nodes are linked to the Organism node (representing the species for that Protein) by creating Organism-encodes-Protein edges. These edges are created from the NCBI Taxonomy ID that is associated with the protein information loaded from UniProt.

**Genes:** Human gene information is imported from NCBI Gene (Maglott et al., 2011). For human genes, the gene is linked to the encoded protein using Gene-encodes-Protein edges by using the UniProt gene information described above.

**Diseases:** SPOKE uses the Human Disease Ontology (Schriml et al., 2012) as the primary identifier for Disease. The disease ontology information is read from the latest OBO file, downloaded weekly from <https://github.com/DiseaseOntology/HumanDiseaseOntology> and, in addition to creating the Disease nodes, we also create the standard ontology links Disease-isa-Disease. The DISEASES database (Pletscher-Frankild et al., 2015) is downloaded and parsed to provide Disease-associates-Gene edges, which include the sources, scores and confidence values from the DISEASES database as edge attributes. In addition to information from the DISEASES database, both OMIM (Amberger et al., 2015, 2019) and the GWAS Catalog (Buniello et al., 2019) are used to provide Disease-associates-Gene edges. Furthermore, the GWAS Catalog uses the Experimental Factor Ontology (Malone et al., 2010) to encode disease information. The GWAS lead variant  $P$ -value is added to the edge as a property.

In addition to Disease-associates-Gene edges, two more disease-related edges are included in the core: Organisms-causes-Disease and Disease-resembles-Disease. To create Organisms-causes-Disease edges, data from PathoPhenoDB (Kafkas et al., 2019) are imported, which links human pathogens to the associated disease. Disease-resembles-Disease edges are based on the co-occurrence of disease terms (based on MeSH) in PubMed. Co-occurrence is scored based on Fisher's exact test to provide both odds ratios and  $P$ -values, which are stored as edge properties along with the number papers that have both terms and the enrichment (measured as the number of papers with both terms over expected number based on a random distribution).

**Compounds:** For compound information, we chose to import ChEMBL (Mendez et al., 2019). In addition, DrugBank (Wishart et al., 2018) is used to include compounds that might not be present in ChEMBL. We also add Compound-binds-Protein edges from ChEMBL as well as BindingDB (Chen et al., 2001).

ChEMBL and DrugCentral (Avram et al., 2021; Ursu et al., 2017) both provide information about the disease targets of drugs. Disease information is stored by ChEMBL using the MeSH identifier.

Finally, we import data from the Connectivity Map project (Subramanian et al., 2017) which provides information linking perturbagens, including compounds and genes, to the regulatory effects on genes. In order to create the edges, we process the L1000 data to derive consensus signatures following the method outlined in Himmelstein et al. (2017).

As we continue to evaluate various databases that contain biological or biomedical data of interest, we integrate databases into SPOKE that augment the core with useful information but do not significantly introduce entire new ways of looking at SPOKE. Five examples of this include adding Gene Ontology (Ashburner et al., 2000) annotations for CellularComponent, MolecularFunction and BiologicalProcess; ProteinDomain and ProteinFamily from PFAM (Finn et al., 2014); the Uberon ontology (Mungall et al., 2012) for Anatomy and CellTypes from the Human Protein Atlas (Thul and

**Table 1.** SPOKE nodes

Node	Label	Description	Count	Source
1	Anatomy	Tissue (from UBERON)	15 239	<a href="http://obophenotype.github.io/uberon/">http://obophenotype.github.io/uberon/</a>
2	AnatomyCellType	Intermediate node built by combining cell type and anatomy	102	N/A
3	BiologicalProcess	From Gene Ontology	13 343	<a href="http://geneontology.org">http://geneontology.org</a>
4	CellType	From Gene Ontology	54	<a href="https://www.ebi.ac.uk/ols/ontologies/cl">https://www.ebi.ac.uk/ols/ontologies/cl</a>
5	CellularComponent	From Gene Ontology	1722	<a href="http://geneontology.org">http://geneontology.org</a>
6	Compound	pharmacological or metabolic compound	2 112 091	<a href="https://www.ebi.ac.uk/chembl/">https://www.ebi.ac.uk/chembl/</a>
7	Disease	Disease	10 932	<a href="https://disease-ontology.org">https://disease-ontology.org</a>
8	EC	Enzymatic activity	8 287	<a href="https://iubmb.qmul.ac.uk/enzyme/">https://iubmb.qmul.ac.uk/enzyme/</a>
9	Food	Food	992	<a href="https://foodon.org">https://foodon.org</a>
10	Gene	Gene (Entrez)	20 086	<a href="https://www.ncbi.nlm.nih.gov/gene">https://www.ncbi.nlm.nih.gov/gene</a>
11	MolecularFunction	FROM gene ontology	3488	<a href="http://geneontology.org">http://geneontology.org</a>
12	Nutrient	Nutrient	39	<a href="https://www.ebi.ac.uk/chembl/">https://www.ebi.ac.uk/chembl/</a>
13	Organism	Organism (NCBI taxonomy)	10 030	<a href="https://www.ncbi.nlm.nih.gov/taxonomy">https://www.ncbi.nlm.nih.gov/taxonomy</a>
14	Pathway	Biological pathway	3454	<a href="https://reactome.org">https://reactome.org</a>
15	PharmacologicClass	Pharmacological class	577	<a href="https://www.ebi.ac.uk/chembl/">https://www.ebi.ac.uk/chembl/</a>
16	Protein	Protein (UniProt)	24 805 918	<a href="https://www.uniprot.org">https://www.uniprot.org</a>
17	ProteinDomain	Protein domain (Pfam)	19 178	<a href="https://pfam.xfam.org">https://pfam.xfam.org</a>
18	ProteinFamily	Protein family (Pfam)	645	<a href="https://pfam.xfam.org">https://pfam.xfam.org</a>
19	Reaction	Metabolic reaction (KEGG or Metacyc)	22 370	<a href="https://www.kegg.jp">https://www.kegg.jp</a>     <a href="https://metacyc.org">https://metacyc.org</a>
20	SideEffect	Compound side effect (SIDER)	6061	<a href="http://sideeffects.embl.de">http://sideeffects.embl.de</a>
21	Symptom	Symptom (MeSH)	1759	<a href="https://www.ncbi.nlm.nih.gov/mesh/">https://www.ncbi.nlm.nih.gov/mesh/</a>
	Total		27 056 367	

Lindskog, 2018); PharmacologicClass of Compounds from DrugCentral (Avram *et al.*, 2021); and Symptom from MeSH terms.

The InterPro database (Blum *et al.*, 2021) is used to provide ProteinDomain-partof-Protein edges, which provides the linkage between ProteinDomain and the SPOKE core. The ProtCID database (Xu and Dunbrack, 2020) provides information about known interactions between protein domains as well as between protein domains and compounds.

Finally, the Bgee (Bastian *et al.*, 2021) resource is used to determine differential expression of genes across tissues. This information is used to encode Anatomy-upregulates-Gene, and Anatomy-downregulates-Gene edges.

**Pathways:** Initially, we imported human pathway information from WikiPathways (Martens *et al.*, 2021) and Pathway Commons (Cerami *et al.*, 2011). These resources were used to add a Pathway node type, which is connected to Gene with Gene-participates-Pathway edges.

To import metabolic pathways, we read data from KEGG (Kanehisa and Goto, 2000), MetaCyc (Caspi *et al.*, 2016) and PATRIC (Wattam *et al.*, 2014). We use a reaction-centric model, adding a Reaction node that links to the metabolites with Reaction-consumes-Compound and Reaction-produces-Compound edges. A key part of the model is the addition of an EC node that links to the Reaction through an EC-catalyzes-Reaction edge. The EC node also links to the Proteins that have that EC using Protein-has-EC edges.

**Food:** The current version of SPOKE contains two food databases: FoodDB (Scalbert *et al.*, 2011) and the Australian Food Composition Database from Food Standards Australia New Zealand. Two edges are derived from the databases, Food-contains-Compound and Food-contains-Nutrient. We are currently integrating the FoodOn (Dooley *et al.*, 2018), an ontology of foods that we will use to map foods from the various databases into a consistent ontology.

## 2.2 REST API

All of the nodes and edges discussed above are accessible through the SPOKE REST API. The API was designed primarily to support

the Neighborhood Explorer graphical user interface (Fig. 3) but also provides reasonable access to the SPOKE database for other potential uses. The API can be roughly divided into three different parts: calls that return meta-information, calls that return information about nodes and calls that return networks. All API calls begin with the prefix: <https://spoke.rbvi.ucsf.edu/api/v1/>. The API is documented more fully at <https://spoke.rbvi.ucsf.edu/swagger/>. The metagraph call returns a cytoscape.js (Franz *et al.*, 2016) formatted JSON file that reflects the current SPOKE metagraph. The SPOKE call for getting information about nodes is the full-text search call search. The search call takes two arguments: a node type and a query term. This call uses the Neo4j full-text capability to quickly return a set of matching nodes of the indicated type that match that query, where the query is a lucene-formatted (Bialecki *et al.*, 2012) query.

The SPOKE network calls are more complicated to allow more complicated filters and cutoffs. The three network calls all return cytoscape.js JSON networks. The sea call takes a SMILES (Weininger, 1988) string or a ZINC (Irwin and Shoichet, 2005) identifier as an argument and returns the SEA (Keiser *et al.*, 2007) network. The neighborhood call is similar to the node call and takes node\_type, attribute and value arguments. See <https://spoke.rbvi.ucsf.edu/swagger/> for more information. The final network call is the expand call, which takes as its input a node type and an internal node ID to expand along with a list of existing node identifiers.

## 3 Results

SPOKE is a knowledge graph connecting information from 41 biomedical databases. The current release contains more than 27 056 367 nodes of 21 different types (Table 1) and 53 264 489 edges of 55 types (Supplementary Table S1). SPOKE uses 11 different ontologies as a framework to organize and connect data in a semantically meaningful manner.

SPOKE strategically collects content from a range of biomedical data sources (i.e. providers of facts or established knowledge). In order to enhance its relevance to human health, SPOKE focuses on

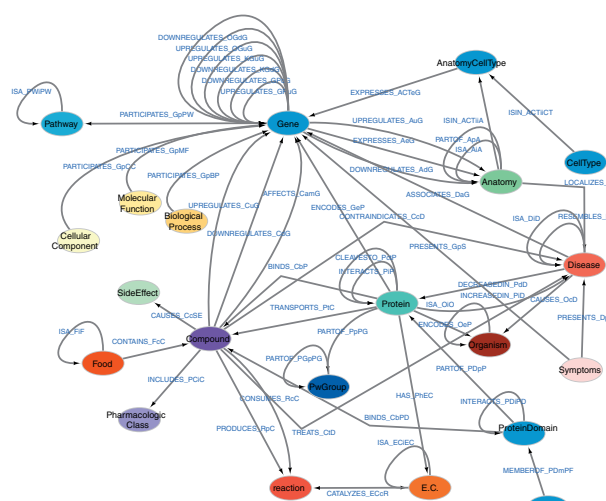


Fig. 2. SPOKE Metagraph. Nodes denote biomedical concepts and links show how data is related and connected in the graph. For details on the node and edge nomenclature, please refer to Table 1 and Supplementary Table S1

experimentally determined information. Thus, computational predictions and text mining from the literature are not currently prioritized. SPOKE is implemented as a Neo4j Community instance and built weekly from scratch by a series of custom python scripts which download each resource, check for integrity and completeness, and then create a ‘root table’ of nodes and edges. Finally, a Cypher script is used to upload the root table into Neo4j (Supplementary Fig. S1). Graph queries are translated by a REST API and users can submit searches directly via the API or via the graphical user interface (Neighborhood Explorer).

The SPOKE metagraph (Fig. 2) shows all node types connected by biologically meaningful, semantic relationships. Both nodes and edges retain source properties that are exposed to the user and include provenance, context, descriptions, etc. If available, additional details are encoded as edge properties, such as association *P*-value and odds ratio (or Beta value) for an associated genetic variant, etc.

### 3.1 Ontologies

Ontologies are used to provide hierarchical structure to the graph, which enables anchoring of additional concepts and facilitates logical navigation. SPOKE also uses ontologies to mark up the datasets coming into the knowledge graph so that the data can be linked consistently across all other datasets. Whenever practical, SPOKE also adheres to the Biolink model. While not strictly an ontology, the Biolink model aims at standardizing the types and relational structures present in biomedical knowledge graphs.

### 3.2 Identifiers

For each type of node in SPOKE, a unique identifier must be chosen. While several different identifiers can be found for the same concept, one identification is selected as primary (SPOKE uses Ensembl). To enable cross-referencing, additional identifiers available for a given concept are kept as node properties.

### 3.3 Modeling

To preserve and make optimal use of available information, SPOKE considers genes and proteins as separate concepts (genes and transcripts remain unified). This distinction is particularly useful to describe protein isoforms, to properly map disease associations to genes, to accurately describe gene–gene regulations, and to distinguish drug–protein interactions from drug–gene (transcript) regulation. In most cases, data are downloaded and integrated ‘as is’, thus no modification to the source data is introduced.

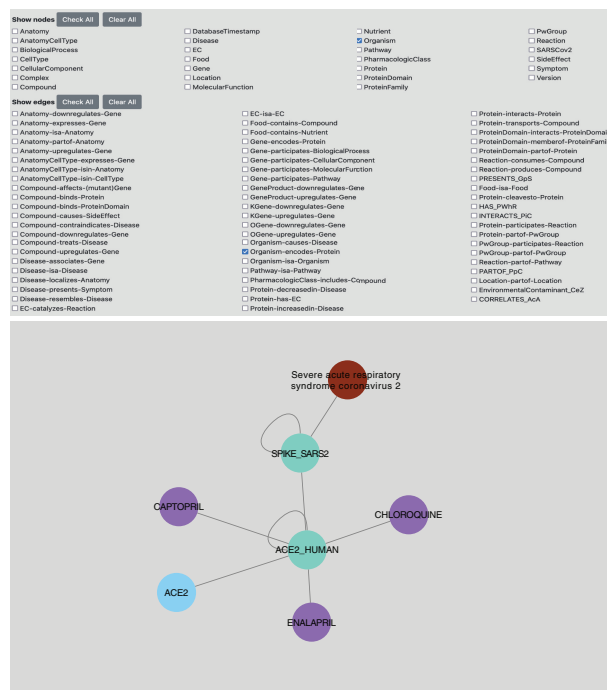


Fig. 3. A view of the SPOKE Neighborhood explorer. The top panel shows the controls that allow a user to select nodes/edges for expansion as well as other key parameters. The bottom panel shows an example of the graph neighbors of the SARS-CoV-2 Spike protein (light blue), which includes three human genes (blue) and the proteins they encode (green). One such protein (ACE2\_HUMAN) acts as the virus receptor in humans and has edges connecting it to three compounds (two of them FDA-approved and one -ORE-100- in experimental phase)

### 3.4 Database download and update scripts

SPOKE is supported by a collection of Python scripts that identify the URL for the resource, downloads data tables, matches identifiers and creates nodes and edges between corresponding concepts.

### 3.5 Graphical user interface: the SPOKE Neighborhood Explorer

SPOKE can be accessed via the Neighborhood Explorer (Fig. 3). The SPOKE Neighborhood Explorer is a simple web interface (<https://SPOKE.rbvi.ucsf.edu>) that allows a researcher to query a given drug, disease, gene or protein and returns its neighbors in graph space—with precise controls (i.e. options) over the kind of nodes and edges that will be retrieved to the user, and a mouse-over function that displays the node/edge metadata (including its provenance). To preserve integrity of the original databases and prevent redistribution of content, SPOKE is not available as a bulk download.

### 3.6 Uses for SPOKE

**Drug discovery capabilities:** Compounds with therapeutic evidence (FDA-approved) or under experimentation, can be directly searched via their ChEMBL identifier or by typing in free text. Relationships to diseases (ChEMBL and DrugCentral), protein binding (ChEMBL and bindingDB), side effects (SIDER) or gene regulation (LINCS L1000) are available for selection (Fig. 4). Predicted binding to human proteins [pre-computed by the SEA algorithm (Keiser et al., 2007)] can be retrieved by entering the compound’s SMILES ID. Starting from a SPOKE search, advanced graph analytic and machine learning approaches can be employed to use multi-node drug neighborhoods as a ‘functional fingerprint’ to complement its molecular profile for drug discovery or repurposing approaches.

**Anatomy-driven searches:** The class hierarchy view among anatomical terms can be explored by expanding any term using the subsumption relationships (Anatomy-isa-Anatomy, UBERON).

- Compound-affects-(mutant)Gene
- Compound-binds-Protein
- Compound-binds-ProteinDomain
- Compound-causes-SideEffect
- Compound-contraindicates-Disease
- Compound-downregulates-Gene
- Compound-treats-Disease
- Compound-upregulates-Gene

Fig. 4. Available relationships for compounds. Example of user options to select different relationships for compounds. Compound binds protein (BINDS\_CbP) reflects molecular interactions obtained from ChEMBL and BINDINGdb. Compound causes side effect (CAUSES\_CcSE) relationships are retrieved from SIDER. Compound contraindicates disease (CONTRAINS\_CcD) and Compound treats disease relationships are obtained from ChEMBL. Compound downregulates gene (DOWNREGULATES\_CdG) and Compound upregulates gene (UPREGULATES\_CuG) are obtained from LINCS1000

The edges Anatomy-partof-Anatomy describe relationships between Anatomy nodes (also from Uberon) indicating physical inclusion, for example, ‘brain’ is a part of ‘central nervous system’.

Nominal or enriched gene expression information by each anatomy can be retrieved by Anatomy-expresses-gene or anatomy-upregulates-gene edges (Bgee). Cell types are connected to anatomies via intermediate AnatomyCellType nodes and AnatomyCellType-isin-Celltype edges. This modeling strategy was implemented to disambiguate cases in which the same cell type is found in different organs but they express different genes in each case [e.g. squamous epithelial cells can be found in several anatomies and their expressed genes/protein profiles can be different (Fig. 5)].

**Food-driven searches:** With the incorporation of FooDB and Australian Food Composition Database, thousands of edges connect chemicals to common foods. When available, a numeric quantity describes the amount as an edge property. This is useful when connecting foods with metabolic reactions or components of the gut microbiota. Indeed, a SPOKE search can be initiated with any available foods and use a combination of Extend and Options to display a complete picture of the role of its neighborhood. For example, a user can start a search with the term ‘(arabica) coffee’ and bring the compound caffeine as one of its components (Fig. 6). An unrestricted extension of caffeine brings nodes of different types, including proteins (Adenosine receptors, acetylcholinesterase and monoamine oxidases) known to bind this compound. As some of them are enzymes (MAO-A and MAO-B), a connection to the corresponding E.C. (Monoamineoxidase) can be retrieved. In addition, protein domains (light blue) from each protein can be retrieved. Caffeine is also connected to the gene TLR4 (by an edge Compound\_upregulates\_gene), as reported by LINCS L1000. Additional information is available for caffeine, such as its pharmacological class (xantines), associated side effects (e.g. feeling jittery) and disease contraindications (e.g. epilepsy). In addition, caffeine is linked to a series of metabolic reactions (red nodes), some of which are endogenous (monoxygenase and Cytochrome P450) and some are bacterial (e.g. a demethylase and a dehydrogenase) corresponding to *Pseudomonas putida* (Yu et al., 2009). Thus, SPOKE was able to reconstruct a large body of knowledge by linking information deposited in multiple databases (Fig. 6).

**Disease-driven searches:** Diseases can be explored by entering a DOID or text and selecting any of the available Options, which include relationships to genes, symptoms, indications, similarity and anatomy (in addition to exploring the disease ontology). For example, it is possible to search for Alzheimer’s disease (AD, DOID: 10652), and retrieve just its symptoms (PubMed) and all sub-types

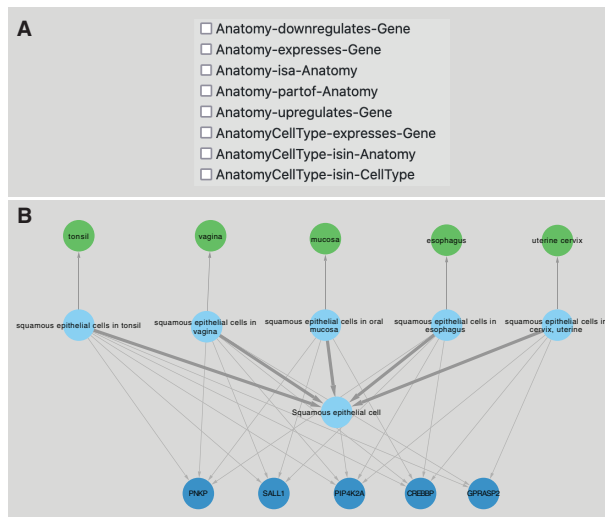


Fig. 5. Anatomy-cellType-gene (protein) relationships. (A) An example of user options to select and control visualization of anatomy nodes in SPOKE. (B) An example of how the cell type squamous epithelial cell (blue—from cell ontology-), can be present in multiple organs/tissues (green—from UBERON). A hybrid node (AnatomyCellType) was created to represent the conditional statement that a given cell type can be found in multiple different tissues/anatomies. This allowed for the representation of immunostaining of histological specimens (from The Protein Atlas). For example, squamous epithelial cells in Tonsil express SALL1, but squamous epithelial cells in mucosa do not

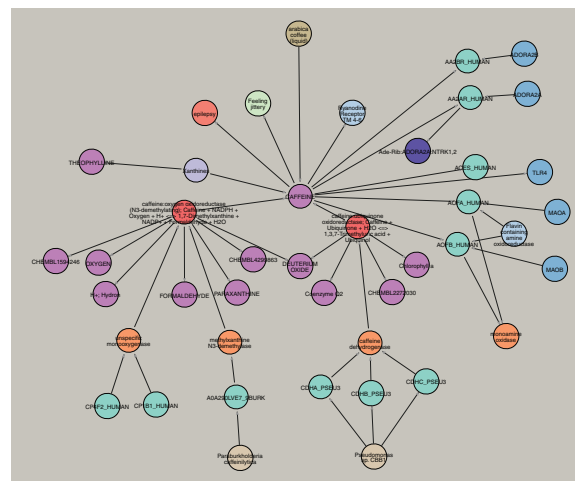


Fig. 6. A search for Coffee reveals molecular, pharmacological and metabolic pathways of caffeine. A multi-step search for coffee can provide a deep understanding of its relationship to human metabolism. In this example, Arabica coffee (food) contains caffeine (compound), which, together with theophylline, is a xanthine (pharmacological class). In addition, caffeine binds two adenosine receptors (AA2AR and AA2BR), encoded by the genes ADORA2A and ADORA2B, acetylcholinesterase (ACES) and monoamine oxidase A (ACFA) and B (ACFB). Caffeine also binds the protein domain Ryanodine receptor, upregulates the gene TLR4, causes a feeling jittery side effect and is contraindicated in epilepsy. In the left-hand side of the figure, two metabolic reactions that consume caffeine are depicted. A mono-oxygenase catalytic activity is denoted for cytochrome P450 complex in humans, and a methylxanthine demethylase activity in *Pseudomonas putida*. A dehydrogenase activity is carried out by enzymes in *Pseudomonas* sp. CBB1

of the disease described in the DO (AD1, AD2, etc.) (Fig. 7). An extension to this search can be performed to bring genes associated with each disease subtype (GWAS Catalog, OMIM and DISEASES), the proteins these genes encode (NCBI Gene) and their domains and families (PFAM). Entire classes of diseases can be explored at once,

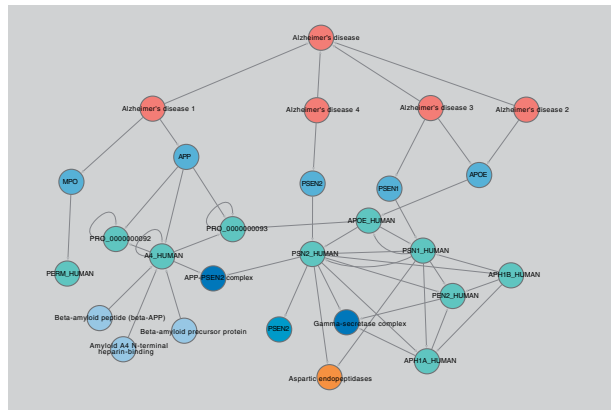


Fig. 7. Main types of Alzheimer's disease and their relationships to symptoms, genes, proteins, domains and families. Four subtypes of Alzheimer disease are depicted, each with its corresponding genetic association. Type 1 is related to variation/mutation in MPO and APP, Type 2 is related to APOE, Type 3 is related to APOE and PSEN1, and Type 4 is related to PSEN2 (blue). The corresponding proteins encoded by those genes are also depicted (teal). The enzymatic proteolysis of APP into the different amyloid peptides by the secretase complex (encoded by presenilins 1 and 2) is depicted at the bottom of the figure

by leveraging the disease ontology. For example, all Mendelian or metabolic diseases can be retrieved in a single query. For a given metabolic disease, it is possible to explore relationships to gene, protein, enzymatic activity, all the way down to the metabolic reaction affected by the gene defect. This strategy is particularly useful when searching for compounds that can reverse the damage either by reducing degradation or by increasing production of the affected metabolite.

## 4 Discussion

Knowledge can be considered an emergent property of the interconnected web of trusted information and known facts. The space of the 'unknown knowns' is growing fast and remains vastly underexplored. Concretely, in order to effectively mine them, we must 'connect the dots' from several information sources. We argue that when heterogeneous networks are connected at a massive scale, new knowledge can be extracted as an emergent property of the network. In this article, we present SPOKE, a large biomedical knowledge graph that amalgamates data and information from a large spectrum of databases ranging from molecular to physiological processes.

SPOKE has been used for a variety of biomedical applications including drug repurposing (Himmelstein and Baranzini, 2015), disease prediction and interpretation of transcriptomic data (Himmelstein and Baranzini, 2015), among others. More recently, we developed an algorithm to embed electronic health records onto SPOKE, which, when combined with machine learning techniques, enables a wide range of applications relevant to precision medicine (Nelson et al., 2019, 2022). This approach uses an original embedding method based on the Page rank algorithm that enables the creation of concept-specific vectors (PSEV) trained in millions of de-identified electronic health records. These vectors describe cohorts of patients that share one specific concept (e.g. patients treated with the drug Metformin or patients with tremor as a symptom). Each of these embeddings represents the importance of each node in SPOKE for that cohort, based on the training data, and can later be combined to represent the status of a given patient at a particular point in time. For details, see Nelson et al. (2019). This approach has been successfully implemented to predict a diagnosis of multiple sclerosis with up to 83% accuracy 3 years before the first disease code was found in the EHR (Nelson et al., 2022). A similar approach is now being used to predict diagnosis of other chronic diseases, such as Parkinson's and Alzheimer.

A number of biomedical knowledge graphs exist, but without clear standards for their representation and modeling, a wide variety of strategies have been implemented. Naturally, such knowledge graphs have been difficult to create, as they require deep expertise in

Table 2. Comparison of biomedical knowledge graphs

	SPOKE	CGK	ROBOKOP	ARAX	KeyGEN	CTD
User friendly	+++	+	++	+	+	+++
Experimental info rich	+++	+	+	+	n/a	+++
Literature rich	+	+++	++	+++	ONLY	++
Ontologies	+++	+	++	++	++	+
Food info	YES	YES	NO	YES	n/a	NO
Metabolic info	YES	NO	NO	NO	n/a	NO
Microbiome info	YES	NO	NO	NO	n/a	NO
Full analytics workbench	NO	YES	NO	YES	NO	NO
Automatically generated	NO	NO	NO	NO	YES	NO
Installation needed	NO	YES	YES	NO	NO	NO

a variety of domains. In particular, biomedicine has been slow to adopt this potentially transformative approach, in part due to the complexity of the underlying information. While some focus on experimentally determined information, others include primary data, literature mining and predicted relationships. In addition, these resources can be implemented as property graphs or using RDF (triples) representation (DataCommons <https://www.datacommons.org/>), which largely determines the range of applications they can be used for. Finally, some biomedical graphs are built using semi-automated methods (Rossanez et al., 2020; Santos et al., 2022), and others like SPOKE, Robokop (Fecho et al., 2021) and the comparative toxicogenomics database (Mattingly et al., 2006), CTD, require extensive manual curation (Table 2 illustrates key features of some of the most relevant biomedical graphs available).

The Biomedical Data Translator project (Translator, for short) (<https://ncats.nih.gov/translator>) is a novel and ambitious undertaking by the National Institutes of Health's National Center for Advancing Translational Sciences involving a large and collaborative cadre of scientists from a variety of scientific domains including semantic representation, computer science and biomedical experts. The Translator project aims at developing a comprehensive, relational, N-dimensional Biomedical Data Translator that integrates multiple types of existing data sources, including objective signs and symptoms of disease, drug effects and intervening types of biological data relevant to understanding pathophysiology. SPOKE is one of the knowledge providers of the Translator project.

The National Science Foundation's Convergence Accelerator Program catapulted the development of SPOKE and other open knowledge graphs in the content of track A, which started in 2019. The program prompted a 10× growth in SPOKE in terms of number of nodes, edges and types of information incorporated. Current applications in development include graph traversal, embeddings and drug repurposing efforts, among others.

Machine and deep learning models such as neural networks have traditionally been considered 'black boxes', capable of delivering predictions, but in and of themselves, no new knowledge. This perceived limitation has slowed down their adoption in a range of chemical and biological contexts, under the sensible argument that a technique a scientist, clinician or engineer cannot understand will in turn provide no guarantee of correctness in a true discovery context. Similarly, biomedicine, and human health in general, has been a 'black box' field for predictions and prognoses. In this context, SPOKE can be used to predict biomedical outcomes in a biologically meaningful manner thus representing 'clear box' (i.e. explainable) models. With SPOKE, the paradigm of knowledge graphs—amply proven in Search—is ready to be tested and ultimately applied in biomedicine.

## Acknowledgements

S.E.B. holds the Heidrich Family and Friends Endowed Chair of Neurology at UCSF. S.E.B. holds the Distinguished Professorship in Neurology I at UCSF.

## Funding

The development of SPOKE and its applications are being funded by grants from the National Science Foundation [NSF\_2033569], NIH/NCATS [NIH\_NOA\_1OT2TR003450] and the Marcus Program in Precision Medicine Innovation.

*Conflict of Interest:* S.E.B. is a co-founder of MATE Bioservices.

## Data availability

To preserve integrity of the original databases and prevent redistribution of content under multiple licenses, SPOKE is not available as a bulk download.

## References

- (2020) *Australian Food Composition Database*.
- Ackoff,R.L. (1989) From data to wisdom. *J. Appl. Syst. Anal.*, **16**, 3–9.
- Amberger,J.S. *et al.* (2015) OMIM.org: online Mendelian inheritance in man (OMIM<sup>®</sup>), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–798.
- Amberger,J.S. *et al.* (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.*, **47**, D1038–D1043.
- Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
- Avram,S. *et al.* (2021) DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.*, **49**, D1160–D1169.
- Bastian,F.B. *et al.* (2021) The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res.*, **49**, D831–D847.
- Bialecki,A. *et al.* (2012) Apache lucene 4. In: *SIGIR 2012 Workshop on Open Source Information Retrieval, Portland, OR, USA*. p. 17.
- Blum,M. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
- Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
- Caspi,R. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
- Cerami,E.G. *et al.* (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
- Chen,X. *et al.* (2001) BindingDB: a web-accessible molecular recognition database. *Comb. Chem. High Throughput Screen.*, **4**, 719–725.
- Dooley,D.M. *et al.* (2018) FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Sci. Food*, **2**, 23.
- Fecho,K. *et al.* (2021) A biomedical knowledge graph system to propose mechanistic hypotheses for real-world environmental health observations: cohort study and informatics application. *JMIR Med. Inform.*, **9**, e26714.
- Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–230.
- Franz,M. *et al.* (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.
- Himmelstein,D.S. and Baranzini,S.E. (2015) Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. *PLoS Comput. Biol.*, **11**, e1004259.
- Himmelstein,D.S. *et al.* (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, **6**:e26726. <https://doi.org/10.7554/eLife.26726>.
- Irwin,J.J. and Shoichet,B.K. (2005) ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, **45**, 177–182.
- Kafkas,Ş. *et al.* (2019) PathoPhenoDB, linking human pathogens to their phenotypes in support of infectious disease research. *Sci. Data*, **6**, 79.
- Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
- Keiser,M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.
- Maglott,D. *et al.* (2011) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
- Malone,J. *et al.* (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
- Martens,M. *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
- Mattingly,C.J. *et al.* (2006) The comparative toxicogenomics database (CTD): a resource for comparative toxicological studies. *J. Exp. Zool. A Comp. Exp. Biol.*, **305**, 689–692.
- Mendez,D. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
- Mungall,C.J. *et al.* (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
- Nelson,C.A. *et al.* (2022) Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *J. Am. Med. Inform. Assoc.*, **29**, 424–434.
- Nelson,C.A. *et al.* (2019) Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nat Commun.*, **10**, 3045.
- Orchard,S. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–363.
- Pletscher-Frankild,S. *et al.* (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods*, **74**, 83–89.
- Pundir,S. *et al.* (2017) UniProt protein knowledgebase. *Methods Mol. Biol.*, **1558**, 41–55.
- Reinsel,D. *et al.* (2018) The digitization of the world from edge to core. *IDC White Paper*.
- Rossanez,A. *et al.* (2020) KGen: a knowledge graph generator from biomedical scientific literature. *BMC Med. Inform. Decis. Mak.*, **20**, 314.
- Santos,A. *et al.* (2022) A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol.*, **40**, 692–702.
- Scalbert,A. *et al.* (2011) Databases on food phytochemicals and their health-promoting effects. *J. Agric. Food Chem.*, **59**, 4331–4348.
- Schoch,C.L. *et al.* (2020) NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**.
- Schriml,L.M. *et al.* (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.
- Subramanian,A. *et al.* (2017) A next generation connectivity map: 1 1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.e1417.
- Szklarczyk,D. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
- Thul,P.J. and Lindskog,C. (2018) The human protein atlas: a spatial map of the human proteome. *Protein Sci.*, **27**, 233–244.
- Unni,D.R. *et al.*; Biomedical Data Translator Consortium. (2022) Biolink model: a universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clin. Transl. Sci.*, **15**, 1848–1855.
- Ursu,O. *et al.* (2017) DrugCentral: online drug compendium. *Nucleic Acids Res.*, **45**, D932–D939.
- Wattam,A.R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.
- Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
- Wishart,D.S. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
- Xu,Q. and Dunbrack,R.L. Jr. (2020) ProtCID: a data resource for structural information on protein interactions. *Nat. Commun.*, **11**, 711.
- Yu,C.L. *et al.* (2009) Two distinct pathways for metabolism of theophylline and caffeine are coexpressed in *Pseudomonas putida* CBB5. *J. Bacteriol.*, **191**, 4624–4632.