References:

SPOKE: Mapping Biomedical Knowledge for Precision Medicine and Genomics

[1]John H Morris, [1]**Wanjun Gu**, [1]Karthik Soman, [1]Rabia E Akbas, [2]Brett Smith, [1]Gabriel Cerono, [1]Braian Peetoom, [1]Catalina Villouta-Reyes, [1]Gundolf Schenk, [1]Angela Rizk-Jackson, [3]Lauren Sanders, [3]Sylvain V Costes, [1]Yongmei Shi, [1]Sharat Israni, [2]Sui Huang, [4]Peter W Rose, [5]Charlotte A Nelson, [1]Sergio E Baranzini

[1]University of California, San Francisco **[2]Institute of Systems Biology** **[3]NASA** **[4]University of California, San Diego** **[5]Matebio.ai**

UCSF Bakar Computational Health Sciences Institute

Introduction

Vast amounts of biomedical data are being generated at an unprecedented rate, leading to rapid growth in databases and repositories. Biomedical information is fragmented across disciplines due to the complexity of human physiology, limiting accessibility and interpretation.

Knowledge graphs are gaining traction as a method to integrate disparate biomedical data sources and facilitate knowledge discovery. The biomedical field requires a structured approach to connect diverse datasets and overcome compartmentalization. Connecting relevant biomedical information could lead to new insights in disease understanding, drug discovery, and personal health.

To this end, we introduce **SPOKE** (**Scalable Precision Medicine Open Knowledge Engine**): A heterogenous biomedical knowledge graph linking molecular biology, pharmacology, and clinical practice. By connecting diverse biomedical information, SPOKE aims to facilitate knowledge emergence, enhance disease understanding, improve drug discovery, and support personalized healthcare.

Method

Source	Node Type
Cell Ontology	CellType
Disease Ontology	Disease
DrugBank, ChEMBL, KEGG	Compound
DrugCentral	PharmacologicClass
Entrez Gene	Gene
Enzyme Commission	EnzymeticActivity
FoodOn	Food
Gene Ontology	BiologicalProcess, CellularComponent, MolecularFunction
Geonames, Location databases	Location
MeSH, HPO	Symptom
MiRDB	MiRNA
NCBI Taxonomy, BV-BRC	Organism
NHANES	Blend, DietarySupplement
Pathway databases	Reaction
Pfam	ProteinDomain, ProteinFamily
SIDER	SideEffect
Uberon	Anatomy
UniProtKB	Protein
WikiPathways, Reactome, KEGG, MetaCyc	Pathway

- SPOKE Aggregates data from **51 specialized databases** across multiple biomedical domains.

- Data is structured into **33 node types** (e.g., genes, proteins, diseases, pathways) and **90 edge types** (relationships between nodes).

- **Weekly** updates for most data types, with some static sources.

Network Structure and Knowledge Graph Design

Heterogeneous Network: Nodes represent different data types; edges define known relationships.
Hierarchical Organization: Directed edges indicate **hierarchies** (e.g., disease subtypes, anatomical relationships).
Semantic and Functional Connectivity: Identifies **unknown connections** by linking previously unrelated nodes through multi-step paths.

Table 1 (left). A representative list of data sources of all nodes in SPOKE

Source	Edge Types
Bgee	Anatomy-expresses-Gene
BindingDB, DrugCentral, ChEMBL	Compound-binds-Protein
BioGRID, STRING, IntAct	Protein-interacts-Protein
ChEMBL, DrugCentral	Compound-treats-Disease
ClinVar, GWAS Catalog	Variant-associates-Phenotype
Gene Ontology, Pathway Commons	Gene-participates-Pathway
MiRDB	MiRNA-targets-Gene
OMIM, GWAS Catalog, DISEASES	Disease-associates-Gene
TFLink, TRRUST, ReMap	Protein-regulates-Gene
UniProt	Gene-encodes-Protein

Table 2 (down). A representative list of Data sources of all edges in SPOKE

Infrastructure

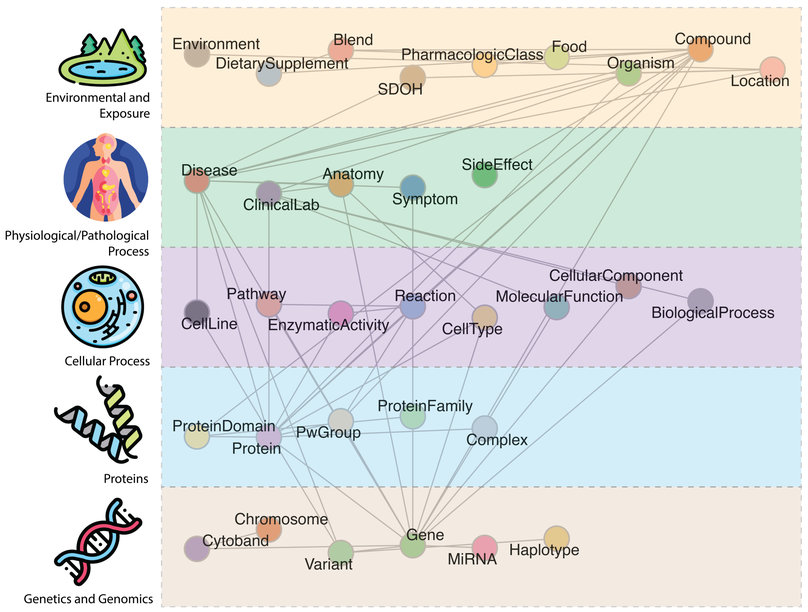


Figure 1. Schematic of SPOKE. Structure of the SPOKE knowledge graph, illustrating the interconnected biomedical entities across five hierarchical domains: **Environmental and Exposure**, **Physiological/Pathological Process**, **Cellular Process**, **Proteins**, and **Genetics/Genomics**. Nodes represent biological and medical concepts, while edges denote relationships between them, enabling integrative biomedical insights.

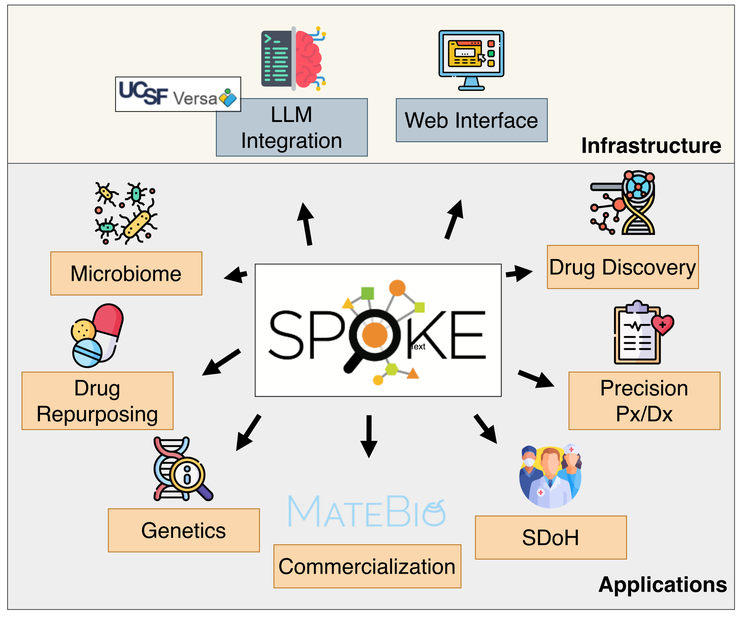


Figure 2. Overview of SPOKE's infrastructure and applications. The infrastructure integrates UCSF Versa with large language models (LLMs) and a web interface. Key applications include microbiome research, drug repurposing, genetics, social determinants of health (SDoH), precision medicine (Px/Dx), and drug discovery.

Applications

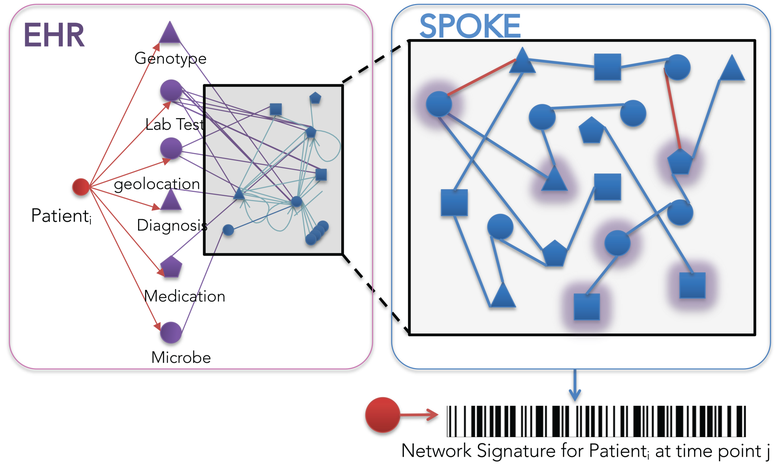


Figure 3. Integrating electronic health record (EHR) data into SPOKE for precision medicine. Multi-modal patient data (e.g., genotype, lab tests, diagnoses, medications, microbes) is embedded into the SPOKE knowledge graph, enabling network-based patient representations. These embeddings facilitate personalized disease progression prediction, subtyping, and diagnostics, supporting data-driven precision medicine.

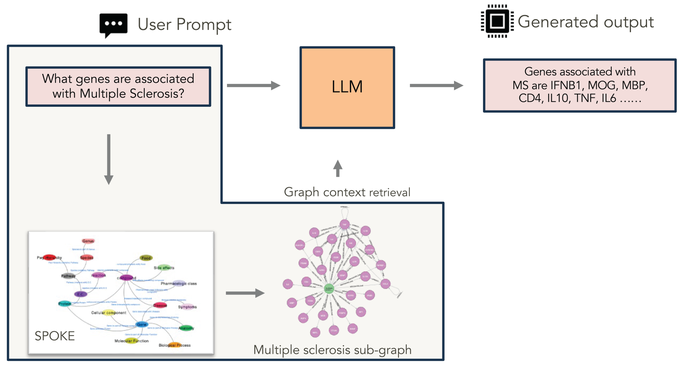
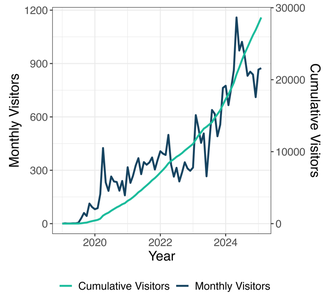


Figure 4. Integration of SPOKE with a large language model (LLM) using a retrieval-augmented generation (RAG) system. When a user queries a biomedical concept (e.g., "What genes are associated with Multiple Sclerosis?"), the system retrieves relevant subgraphs from SPOKE, providing structured contextual information to the LLM. The LLM then synthesizes an informed response based on graph-derived knowledge, enabling accurate and interpretable answers to biological and medical inquiries.



Other Applications: A subnetwork of SPOKE has been adapted into a pathway analysis tool, enabling the interpretation of omics data (transcriptomics, metabolomics, proteomics) to uncover molecular biology insights.

Figure 5. Growth in SPOKE's popularity over time. The dark blue line represents the number of monthly visitors, showing an increasing trend with fluctuations. The green line represents cumulative visitors, indicating a steady rise in SPOKE's user engagement and adoption since its inception.