An Open-Source Knowledge Graph Ecosystem for the Life Sciences

Tiffany J. Callahan MPH, PhD

The Healthcare and Life Sciences Symposium
Knowledge Graph Conference
May 2nd, 2022

COLUMBIA

COLUMBIA UNIVERSITY
DEPARTMENT OF
BIOMEDICAL INFORMATICS

@Tiff_callahan

Disclosure

I have NO financial disclosures or conflicts of interest with the material presented in this talk.

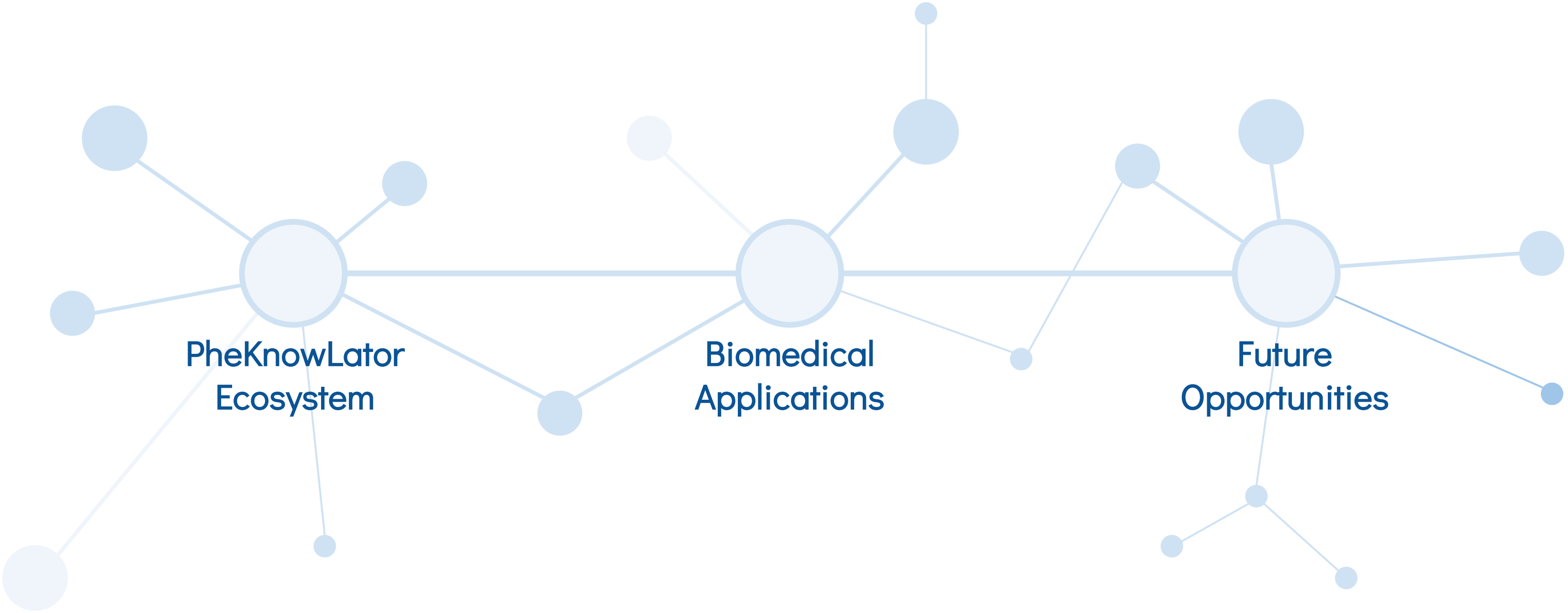
Motivation

Knowledge graphs integrate disparate data, can help decipher complex processes, and have been used to systematically interrogate the biology underlying complex systems[1]

Unsolved Challenges for Constructing Open-Source Knowledge Graphs[2,3]

1. Support only a single knowledge model
2. Standards, technical complexity, usability, and scalability
3. Biologically and clinically meaningful benchmarks

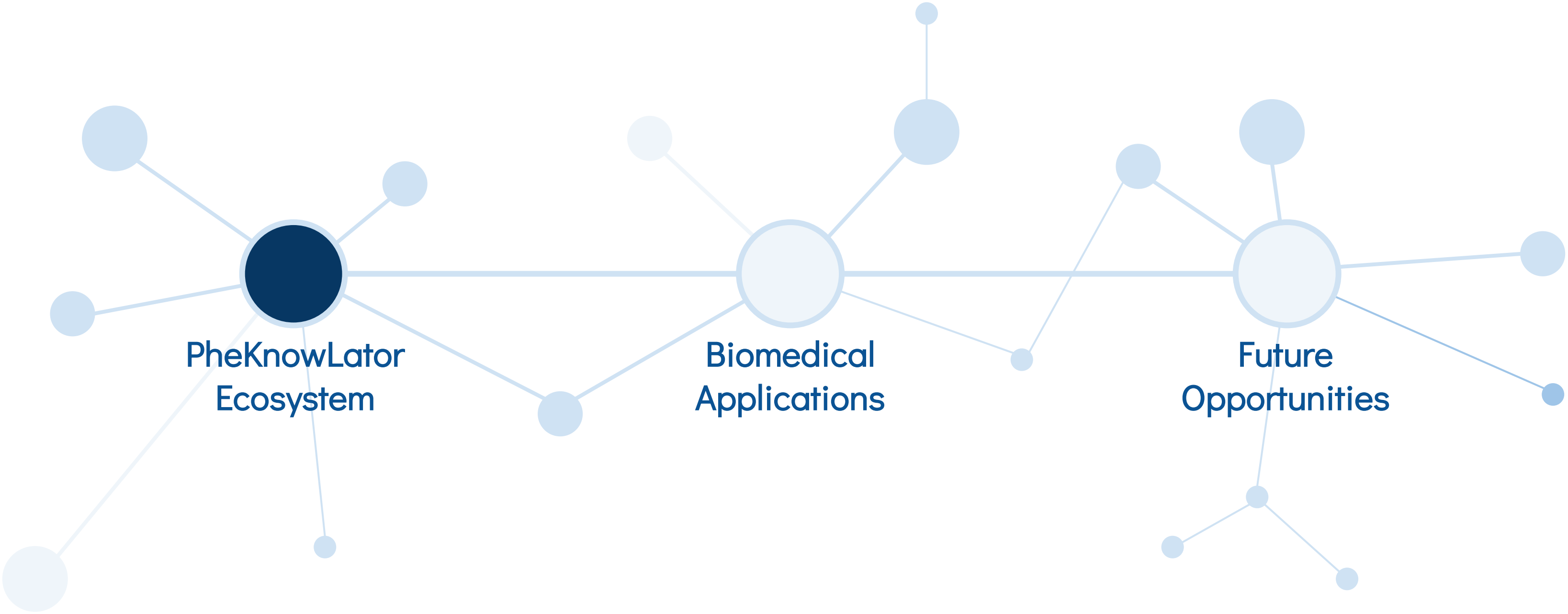
[1] PMID:21414488; [2] PMID:32637040; [3] PMID:33954284



PheKnowLator Ecosystem

Biomedical Applications

Future Opportunities


PheKnowLator Ecosystem
Biomedical Applications
Future Opportunities

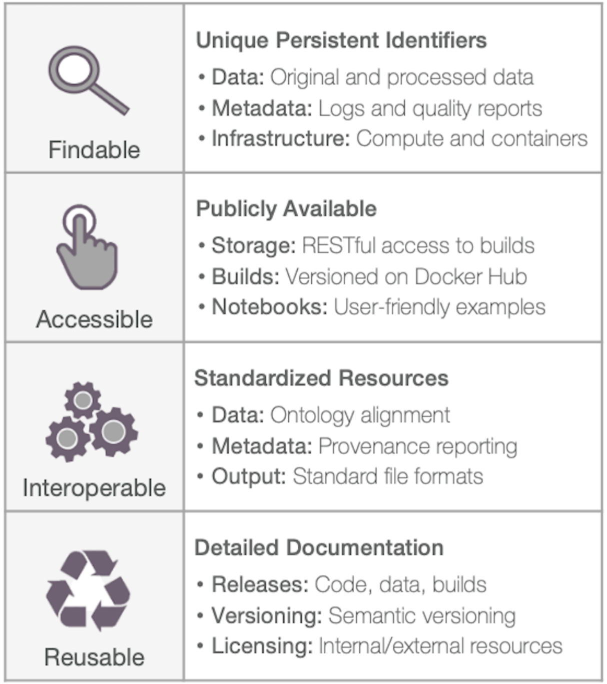
Phenotype Knowledge Translator

PheKnowLator

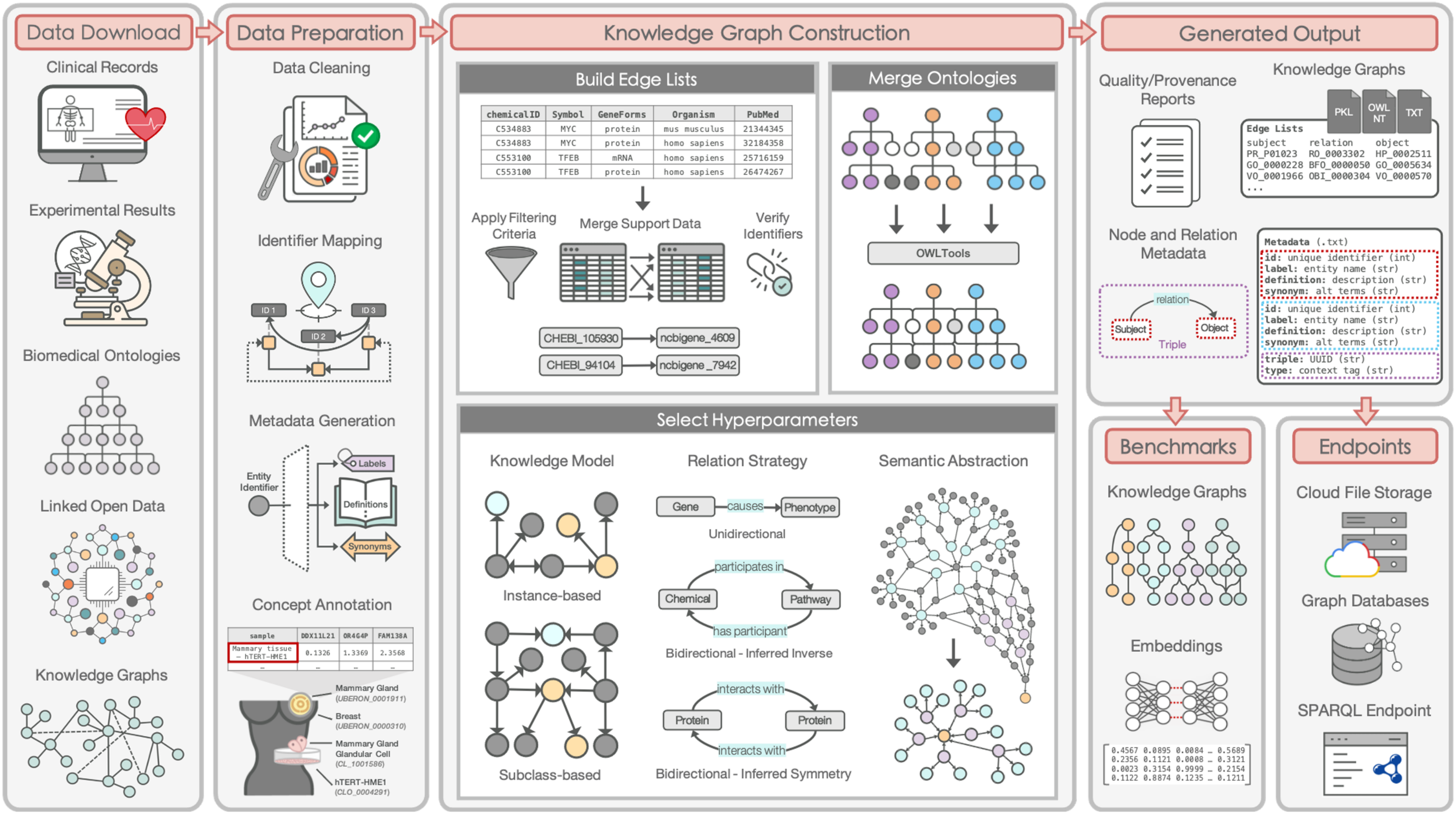
Ecosystem: Construct ontologically-grounded FAIR knowledge graphs

- *Usability*
  - Technical and laymen documentation
  - Jupyter Notebooks and interactive scripts
  - Containerization
- *Scalability*
  - System-scaled distributed execution framework
  - Flexible knowledge representation

Benchmarks: Monthly builds of knowledge graph benchmarks



Findable	**Unique Persistent Identifiers** • **Data:** Original and processed data • **Metadata:** Logs and quality reports • **Infrastructure:** Compute and containers
Accessible	**Publicly Available** • **Storage:** RESTful access to builds • **Builds:** Versioned on Docker Hub • **Notebooks:** User-friendly examples
Interoperable	**Standardized Resources** • **Data:** Ontology alignment • **Metadata:** Provenance reporting • **Output:** Standard file formats
Reusable	**Detailed Documentation** • **Releases:** Code, data, builds • **Versioning:** Semantic versioning • **Licensing:** Internal/external resources



Data Download

Clinical Records

Experimental Results

Biomedical Ontologies

Linked Open Data

Knowledge Graphs

Data Preparation

Data Cleaning

Identifier Mapping

Metadata Generation

Concept Annotation

sample	DDX11L21	OR4G4P	FAM138A
Mammary tissue – hTERT-HME1	0.1326	1.3369	2.3568
…	…	…	…

Knowledge Graph Construction

Build Edge Lists

chemicalID	Symbol	GeneForms	Organism	PubMed
C534883	MYC	protein	mus musculus	21344345
C534883	MYC	protein	homo sapiens	32184358
C553100	TFEB	mRNA	homo sapiens	25716159
C553100	TFEB	protein	homo sapiens	26474267

Apply Filtering Criteria

Merge Support Data

Verify Identifiers

Merge Ontologies

Select Hyperparameters

Knowledge Model

Instance-based

Subclass-based

Relation Strategy

Unidirectional

Bidirectional - Inferred Inverse

Bidirectional - Inferred Symmetry

Semantic Abstraction

Generated Output

Quality/Provenance Reports

Knowledge Graphs

```
Edge Lists
subject     relation    object
PR_P01023   RO_0003302  HP_0002511
GO_0000228  BFO_0000050 GO_0005634
VO_0001966  OBI_0000304 VO_0000570
...
```

Node and Relation Metadata

```
Metadata (.txt)
id: unique identifier (int)
label: entity name (str)
definition: description (str)
synonym: alt terms (str)
id: unique identifier (int)
label: entity name (str)
definition: description (str)
synonym: alt terms (str)
triple: UUID (str)
type: context tag (str)
```

Benchmarks

Knowledge Graphs

Embeddings

```
0.4567 0.0895 0.0084 … 0.5689
0.2356 0.1121 0.0008 … 0.3121
0.0023 0.3154 0.9999 … 0.2154
0.1122 0.8874 0.1235 … 0.1211
```

Endpoints

Cloud File Storage

Graph Databases

SPARQL Endpoint

https://github.com/callahantiff/PheKnowLator

Benchmarks

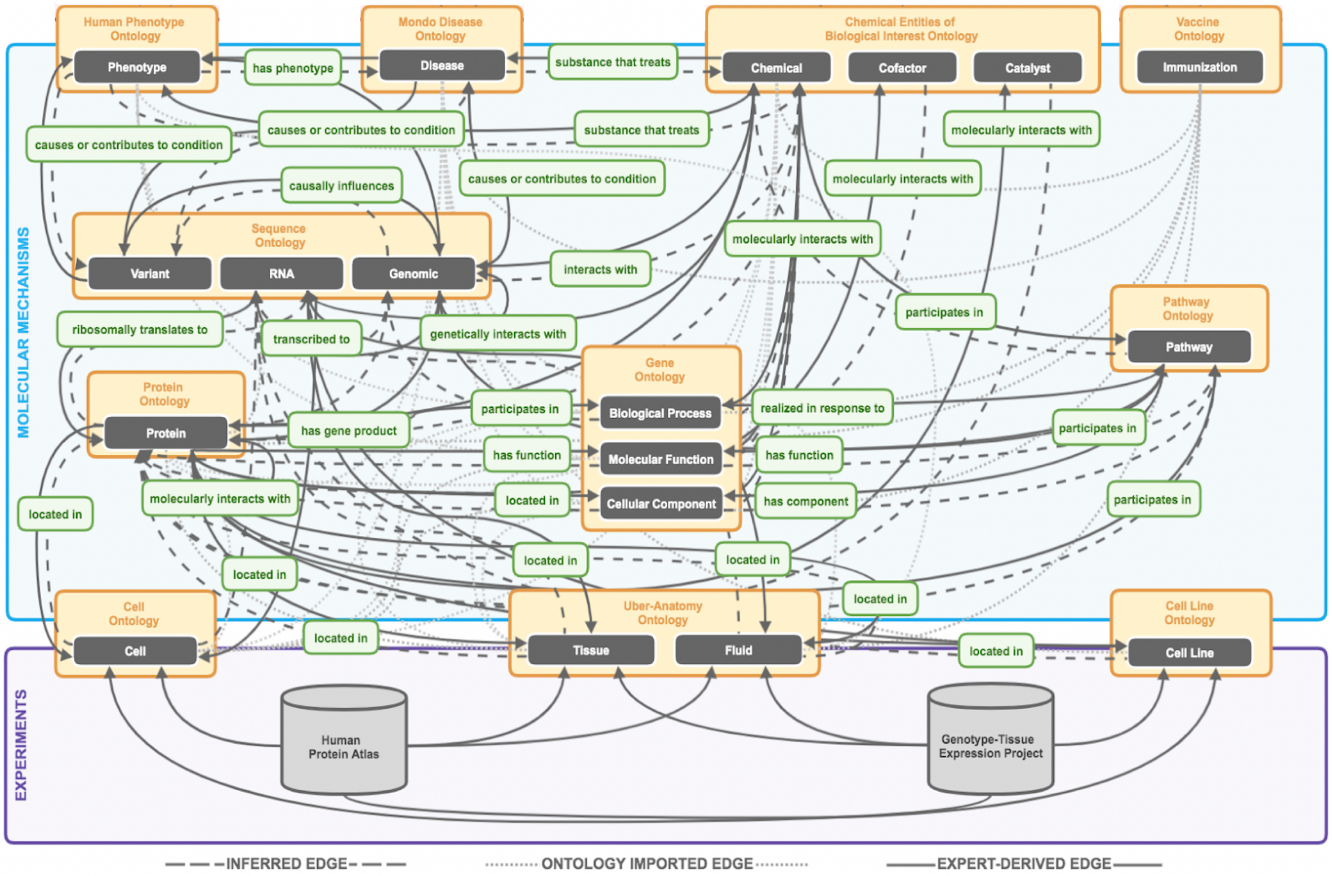
Human Disease Mechanisms

Anchor Ontologies
- 12 ontologies

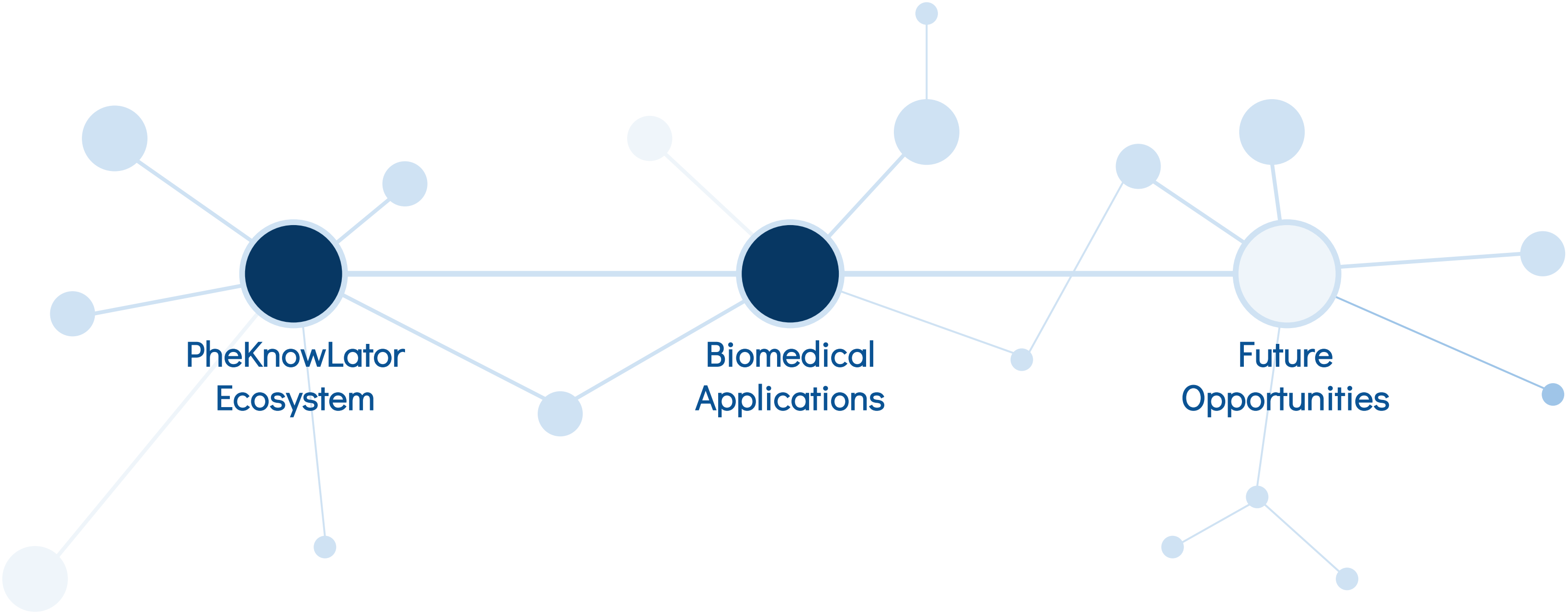
Edge Data
- 22 public datasets
- 2 genome-wide analyses

Validation
- Domain expert review
- Wet lab validation[1]

Monthly Builds
- 12 knowledge graphs
- 15M nodes and 47M edges



[1] PMID: 32387679



PheKnowLator Ecosystem

Biomedical Applications

Future Opportunities

Biomedical Application

- Programs like All of Us[1] and the National COVID Cohort Collaboration (N3C)[2] have made a ton of observational data available for research, but most do not yet integrate molecular data
- For rare diseases like Sickle Cell, prevention and treatment differs based on genotype[3]
- Observational data (alone) is often insufficient to determine genotype in the absence of newborn screening[4]

Objective: Can PheKnowLator enable the genotype of pediatric Sickle Cell Disease patients to be inferred from an independent population of pediatric genotyped patients?

[1] https://allofus.nih.gov/; [1] https://ncats.nih.gov/n3c; [3] PMID:24991875; [4] PMID:29202133

Methods

Clinical Data: Children's Hospital of Colorado (CHCO)

- 2,646 rare disease patients (≥10 visits)

External Genotyped Data: Gene Expression Omnibus (GEO)

- Whole blood gene expression data[1]

Node Embeddings: Walking RDF and OWL[2]

- **CHCO:** conditions, medications, measurements
- **GEO:** gene expression signature-adjusted embeddings

Evaluation

- K-Means clustering
- Patient similarity-based analyses

CHCO Rare Disease Patients

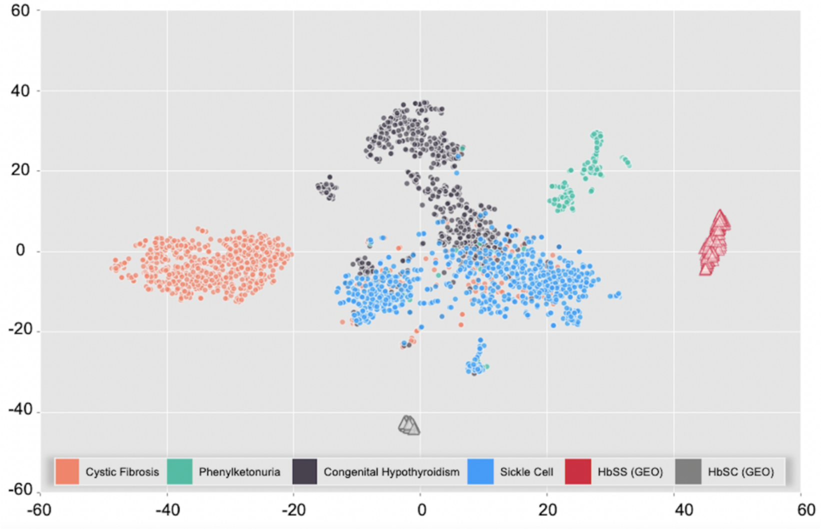
PKU: Phenylketonuria (n=235)
CH: Congenital Hypothyroidism (n=760)
SCD: Sickle Cell Disease (n=816)
CF: Cystic Fibrosis (n=835)

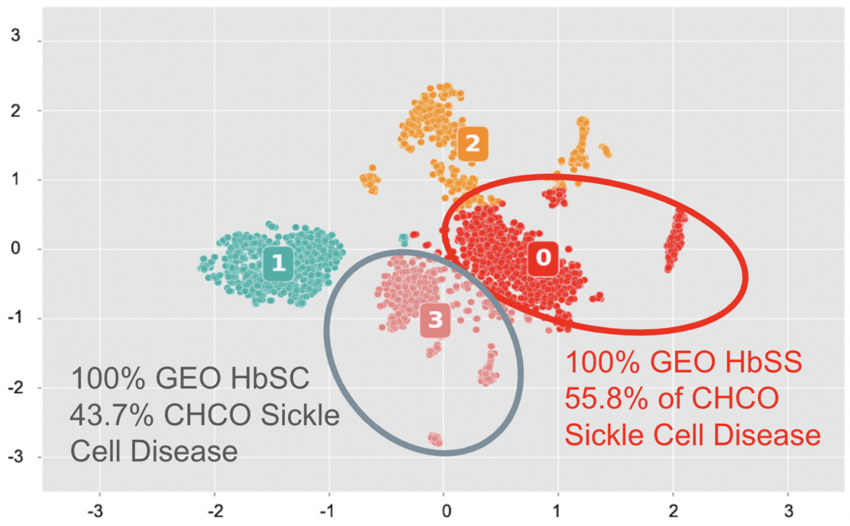
GEO SCD Patients

HbSS: Homozygous Hemoglobin S (n=147)
HbSC: Homozygous Hemoglobin C (n=51)
Control: (n=61)

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35007; [2] PMID:28449114

Results



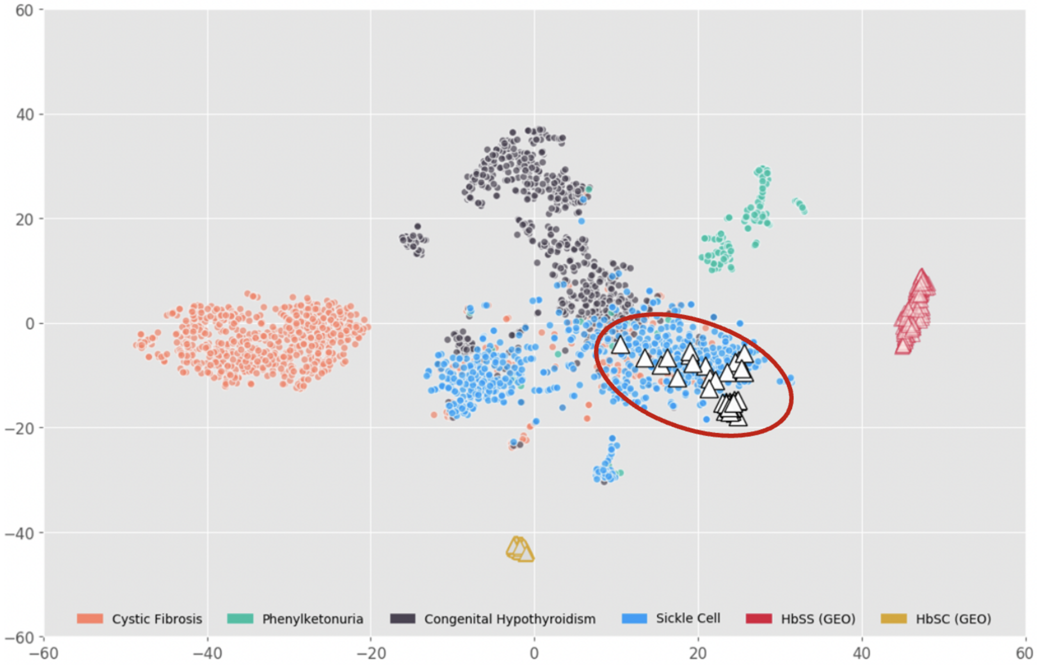


100% GEO HbSC
43.7% CHCO Sickle
Cell Disease

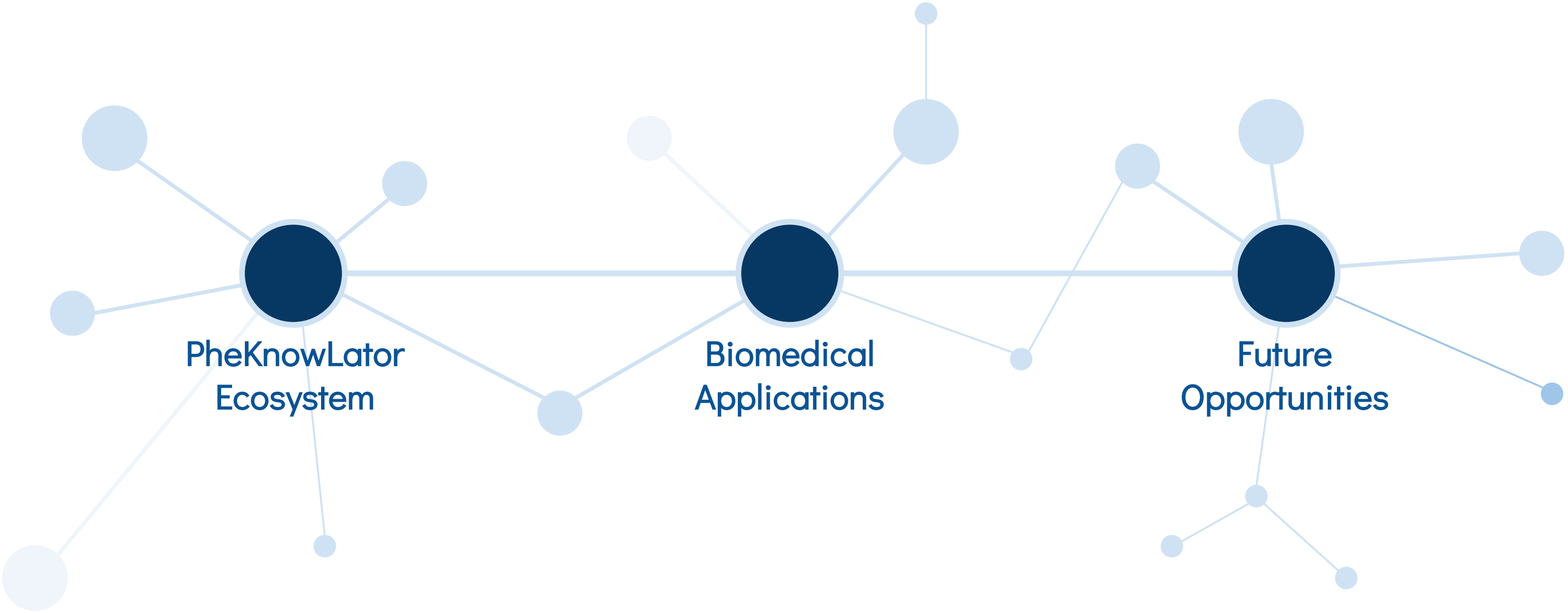
100% GEO HbSS
55.8% of CHCO
Sickle Cell Disease

Geo HbSS (p<0.001) and HbSC (p<0.001) patients were significantly more similar to CHCO Sickle Cell Disease patients than the other rare disease patients

Results



43 CHCO Sickle Cell Disease patients in the HbSS GEO cluster had at least 1 HbSC and HbSS diagnosis



PheKnowLator Ecosystem

Biomedical Applications

Future Opportunities

Future Opportunities

Solve real-world problems within healthcare, life sciences, and public health

1. Phenotype development and evaluation
2. Treatment justification
3. Causal inference

Phenotype Development and Evaluation

Development

- Recommendation systems like PHOEBE[1], rely heavily on the metadata and mappings provided by an ontology or vocabulary
- PheKnowLator can leveraging knowledge of the biological mechanism(s) underlying disease

Evaluation

- CohortDiagnostics[2] helps users determine if a phenotype is of sufficient quality by providing a detailed characterization of the underlying data
- PheKnowLator can enable a more targeted adjudication process by providing knowledge-driven filtering

[1] https://data.ohdsi.org/PHOEBE/; [2] https://ohdsi.github.io/CohortDiagnostics/

Treatment Justification

Electronic health records do not explicitly connect a treatment and disease

PheKnowLator can explain why a particular drug was prescribed for a given indication

- Classify drugs by indication to help identify patients taking alternative treatments and 'off-label' medication use
- Improve data quality by helping determine whether treatments without justification are due to missingness or malpractice

Causal Inference

Causal inference requires expert knowledge to formulate and answer scientific questions

- Identification and adjustment for confounding variables
- Knowledge of and adjustment for features not present in the data

PheKnowLator can help combine what is learned from observational data with what is known

Design	Interpretation
Identify confounders and negative controls	Explain associations between inputs and outputs
Generate causal subgraphs	Assess biological plausibility

Questions?

William A. Baumgartner, Jr.
University of Colorado
Anschutz Medical Campus

Patrick B. Ryan
Columbia University
Janssen R&D

Lawrence E. Hunter
University of Colorado
Anschutz Medical Campus

George Hripcsak
Columbia University

PheKnowLator