Article | Open access | Published: 11 April 2024

An open source knowledge graph ecosystem for the life sciences

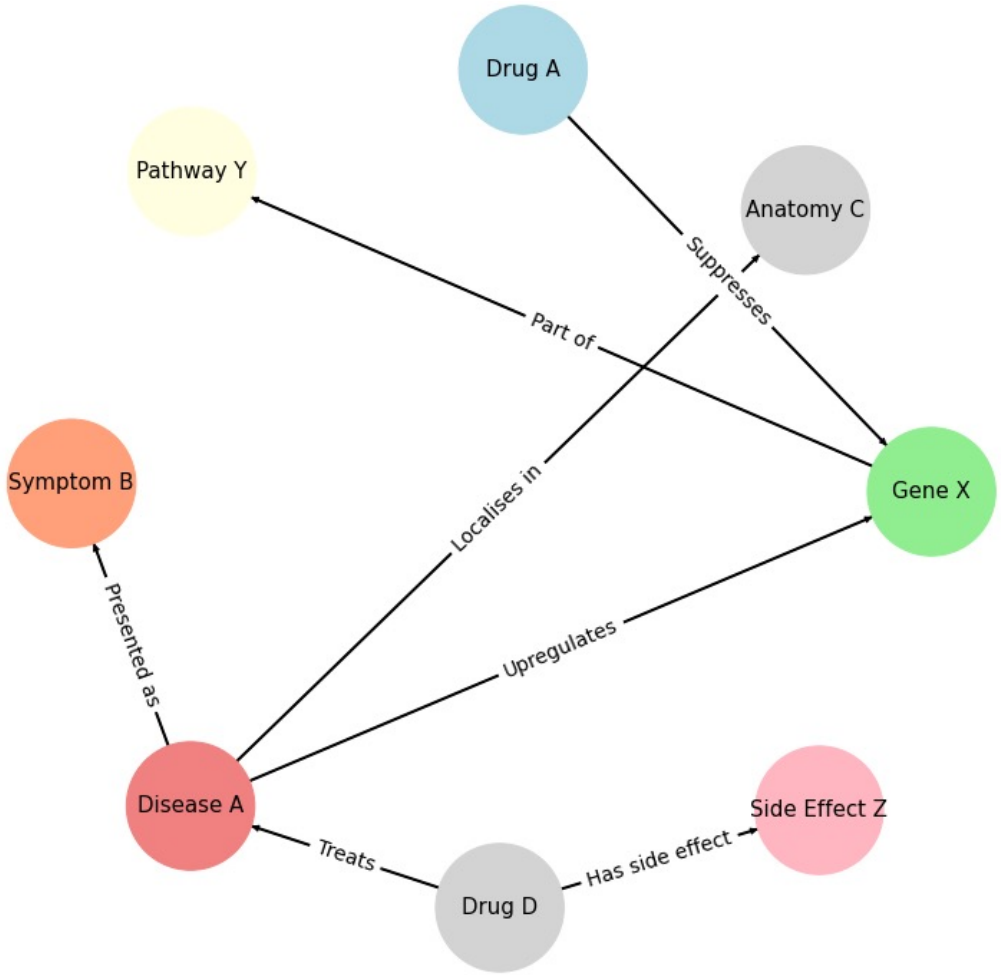
Tiffany J. Callahan, Ignacio J. Tripodi, Adrianne L. Stefanski, Luca Cappelletti, Sanya B. Taneja, Jordan M. Wyrwa, Elena Casiraghi, Nicolas A. Matentzoglu, Justin Reese, Jonathan C. Silverstein, Charles Tapley Hoyt, Richard D. Boyce, Scott A. Malec, Deepak R. Unni, Marcin P. Joachimiak, Peter N. Robinson, Christopher J. Mungall, Emanuele Cavalleri, Tommaso Fontana, Giorgio Valentini, Marco Mesiti, Lucas A. Gillenwater, Brook Santangelo, Nicole A. Vasilevsky, … Lawrence E. Hunter

+ Show authors

Scientific Data **11**, Article number: 363 (2024) | Cite this article

Yusuf Abdulle

Introduction – A Knowledge Graph?



Introduction – Knowledge Graph Construction Types

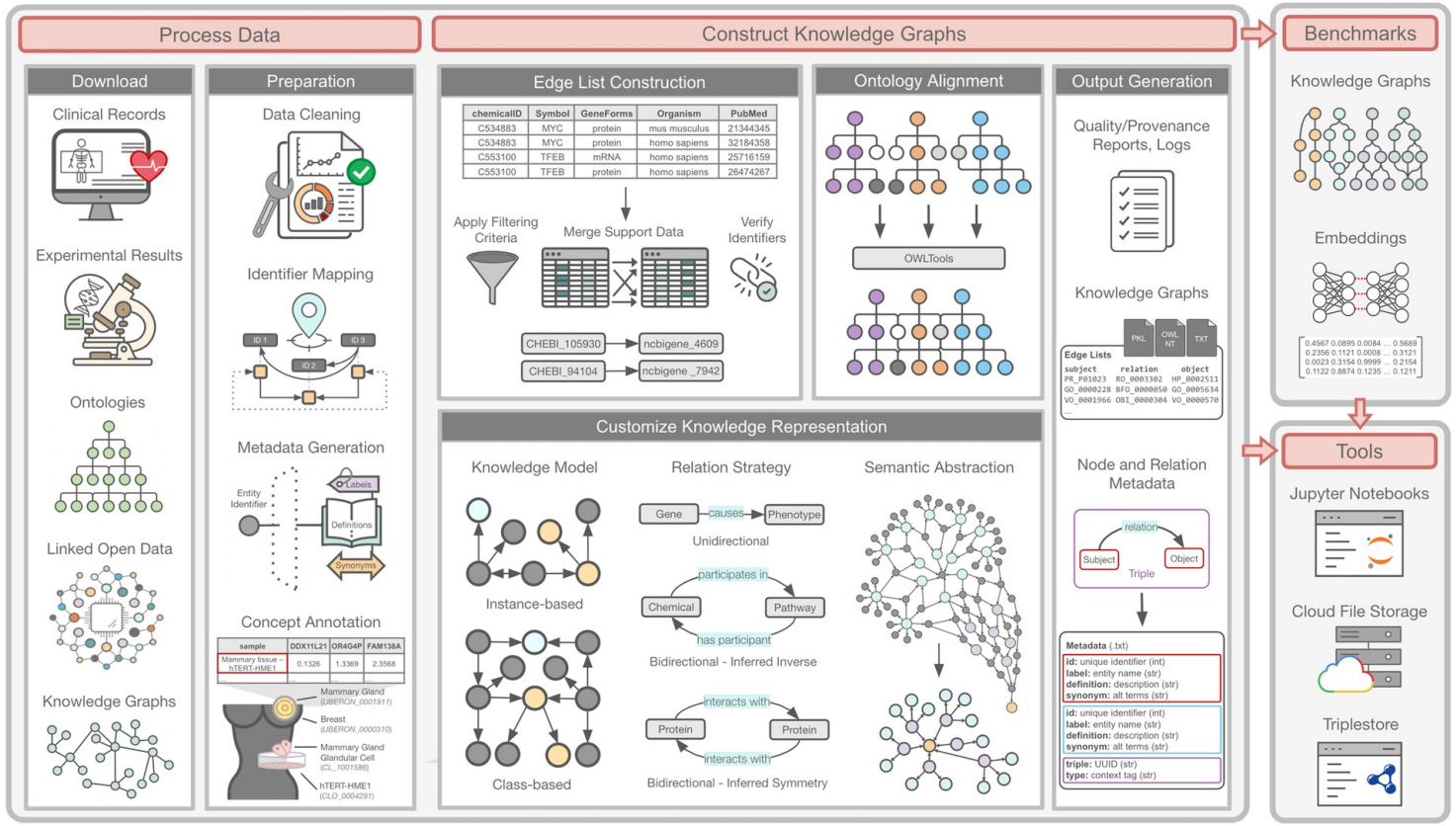
- Construction of Knowledge Graph has 3 different approaches:
  - Simple: Joining nodes as a series of edges
  - Hybrid: Join nodes as a series of edges to existing ontologies
  - Complex: Use of formal semantics with/without formal/existing ontologies

Introduction – What's the problem and the solution?

- **Problem:** Integrating data across different scales (genes, proteins, diseases) is crucial for biomedical research, but the ever-growing data available makes it difficult.

- **PheKnowLator:** A new framework that allows researchers to build KGs with customisable knowledge representation (simple, hybrid, complex). It uses FAIR principles (Findable, Accessible, Interoperable, Reusable) for KG construction.

Methods - Overview



- The **PheKnowLator** ecosystem aims to provide more comprehensive resource to aid with the construction of KGs in the Life Sciences into three components:

    1. **Component 1: KG Construction Resources**
    2. **Component 2: Benchmarking KGs**
    3. **Component 3: KG Tools**

- The **PheKnowLator** ecosystem was also assessed against functionality, availability, usability, maturity and reproducibility

Methods – Component 1: Knowledge Graph Construction Resources

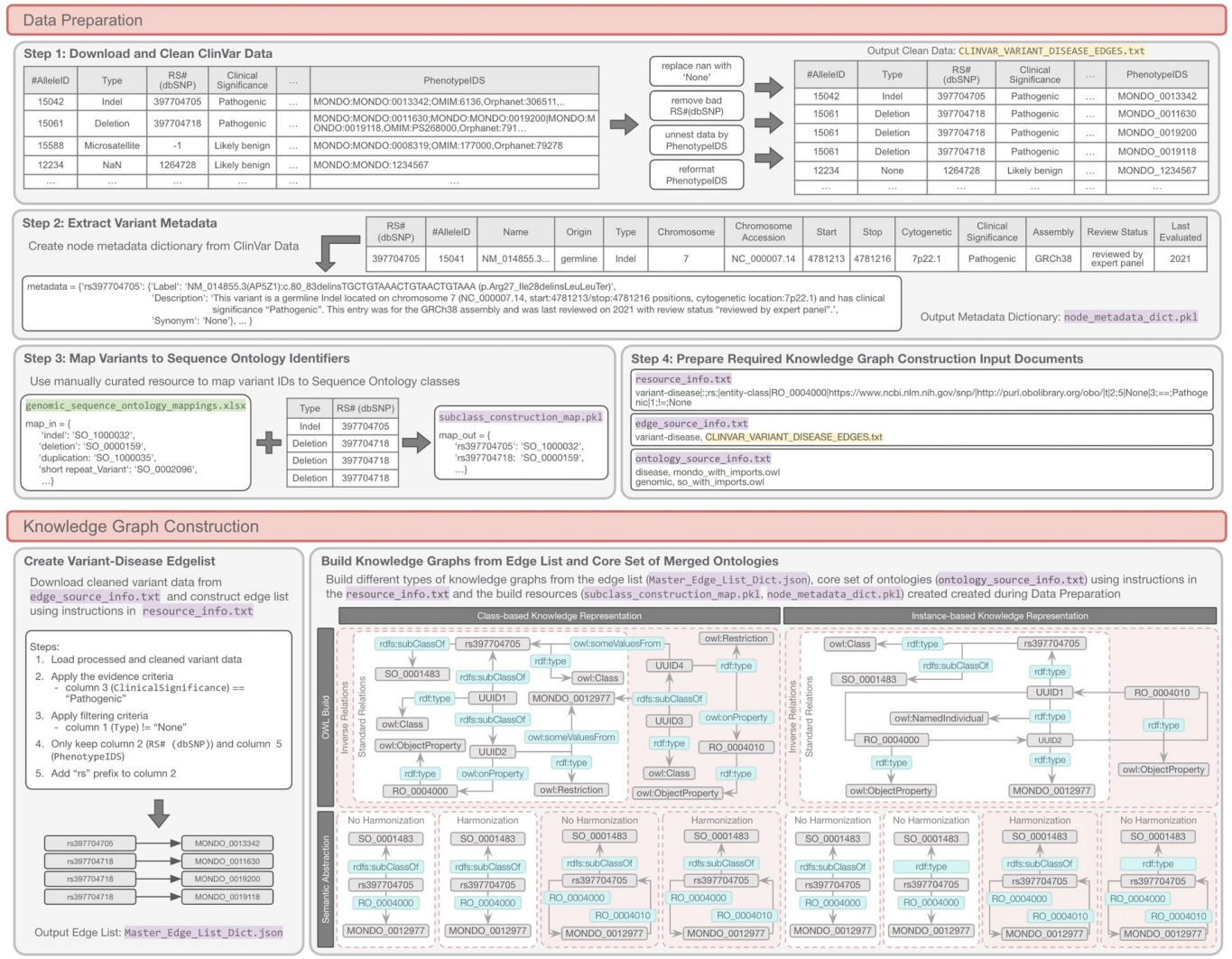
- This component is the foundation of PheKnowLator; uses various resources and algorithms to construct heterogenous KGs and has 3 key features:

1. **Processing Data:**
   - Download Data
   - Preparation of Data

2. **Constructing KGs**
   - Edge List Construction
   - Ontology Alignment
   - Customisable Knowledge Representation

3. **Output Generation**

Methods – Component 2: Knowledge Graph "Benchmarks"



- This component focuses on creating and using pre-built knowledge graphs (KGs) for benchmarking purposes
- With PheKnowLator, there can be generation of different **variations** of a KG based on <u>user-selected parameters</u>, which focus on 3 aspects of the KG:
  - Knowledge Model
  - Relation Strategy
  - Semantic Abstraction
- Component also provides embeddings
- Benefits of "benchmarking" include comparing KG structures and evaluation of downstream tasks

Methods – Component 2: Knowledge Graph "Benchmarks"

Knowledge Model	Relation Strategy	Semantic Abstraction	Variation
Class-based	Standard	OWL-NETS Only	Variation 1
Class-based	Standard	OWL-NETS + Harmonization	Variation 2
Class-based	Inverse	OWL-NETS Only	Variation 3
Class-based	Inverse	OWL-NETS + Harmonization	Variation 4
Instance-based	Standard	OWL-NETS Only	Variation 5
Instance-based	Standard	OWL-NETS + Harmonization	Variation 6
Instance-based	Inverse	OWL-NETS Only	Variation 7
Instance-based	Inverse	OWL-NETS + Harmonization	Variation 8
(Repeat above options without Semantic Abstraction)			Variation 9-12

- This component focuses on creating and using pre-built knowledge graphs (KGs) for benchmarking purposes
- With PheKnowLator, there can be generation of different **variations** of a KG based on user-selected parameters, which focus on 3 aspects of the KG:
  - Knowledge Model
  - Relation Strategy
  - Semantic Abstraction
- Component also provides embeddings
- Benefits of "benchmarking" include comparing KG structures and evaluation of downstream tasks

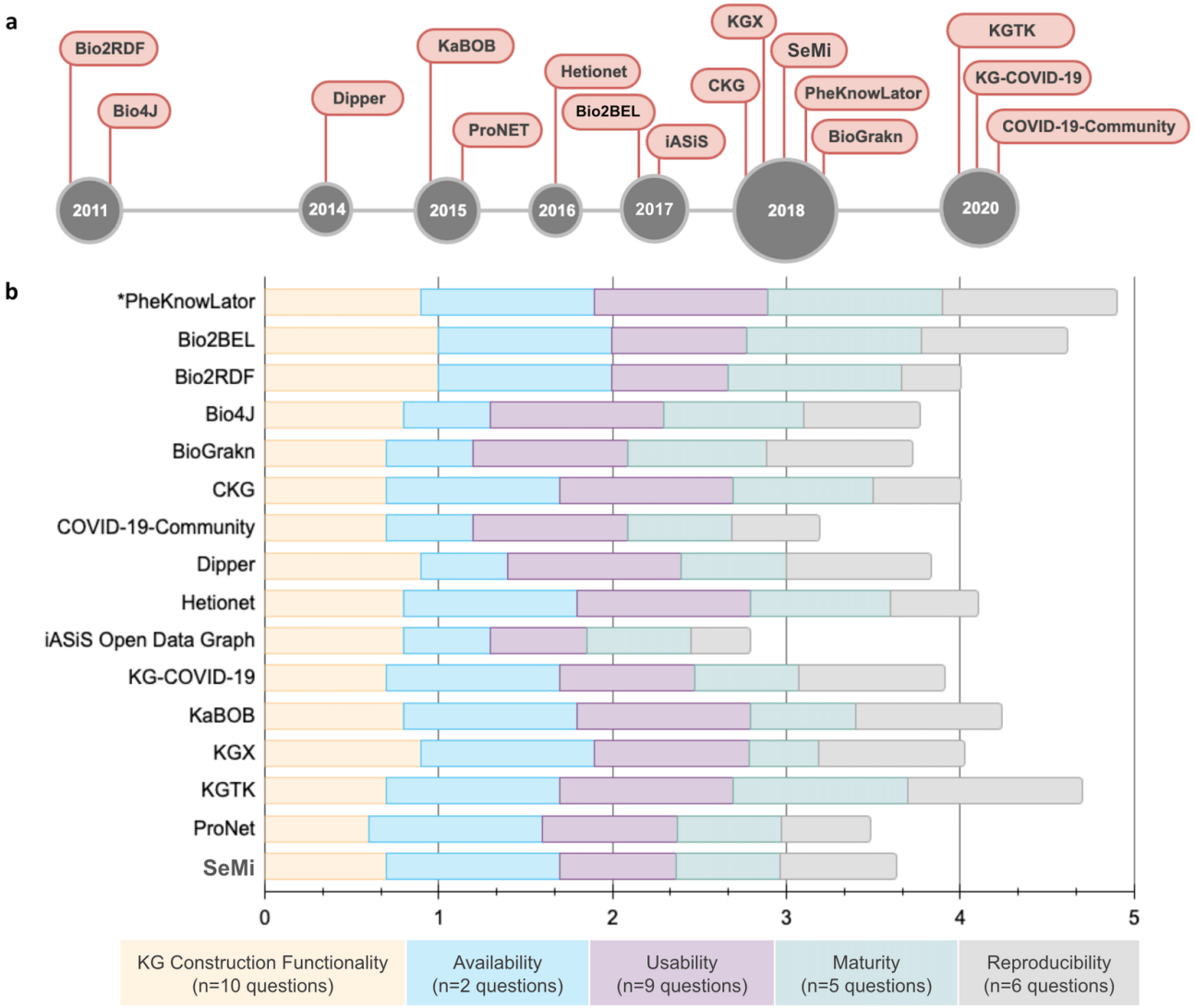
Methods – Component 3: Knowledge Graph Tools

- Component 3 provides the tools needed to analyse the KGs generated by the PheKnowLator ecosystem
- There are 5 main tools offered:
  - Jupyter Notebooks
  - Cloud-based storage
  - APIs
  - Triplestores
  - SPARQL Endpoint

Methods – FAIR Data Principles & Evaluation

- The ecosystem is built on the FAIR data principles
  - **Findability**
  - **Accessibility**
  - **Interoperability**
  - **Reusability**

Results: Systematic comparison of open-source KG construction methods



- A survey assessed 15 open-source methods on: **functionalities, availability, usability, maturity and reproducitbility** – **PheKnowLator** performed competitively across all these criteria

- PheKnowLator differentiated from these other methods in the following 4 areas:
  - Tools for assessing ontology quality
  - Logging / documentation of metadata for KG construction
  - Customisable knowledge representation for advancing reasoning techniques
  - Ability to generate multiple KG versions for benchmarking purposes

Results: Human disease knowledge graph benchmark comparison and construction performance

- 12 different KG variations were constructed for a large human disease KG

- 4 key findings
  1. Size varied depending on the chosen parameters
  2. All the benchmark KGs were **highly sparse**
  3. Semantic abstraction with instance-based models resulted in larger KGs with higher average degree
  4. Knowledge model harmonisation had minimal impact on class-based models w/ semantic abstraction

Discussion

1. PheKnowLator aims to create a unique system for creating biomedical / life science KGs

2. Strengths of this work:
    1. Customisation
    2. Advanced Analysis
    3. Real-World Applications

3. Limitations of this work:
    1. KG Method Comparison Needs Improvement
    2. Performance Metrics Incomplete
    3. Integration with New Standards
    4. Outdated Tools
    5. Validating Large KGs Challenging
    6. User-friendliness Enhancements Needed

Discussion – cont.

1. Application of this work:
    1. Identifying causal factors in diseases
    2. Predicting potential drug interactions
    3. Exploring connections between microbes and disease
    4. Developing new tools for biomedical research

Usefulness to our work?

- Identification of new research pathways

- LLMs: KnowMedLM
  - We can use this tool to build customised KGs to train LLMs for **specific biomedical tasks**

  - Better reasoning / inference to enable better prediction capabilities, i.e. of potential drug interactions based on known relationships

Thank you

PheKnowLator Tutorial - Entity Linking

https://github.com/callahantiff/PheKnowLator/blob/master/notebooks/tutorials/entity_search/Entity_Search.ipynb

PheKnowLator Paper

https://www.nature.com/articles/s41597-024-03171-w

PheKnowLator GitHub

https://github.com/callahantiff/PheKnowLator/tree/master