COMMUTE
Quarterly Bulletin
March 2025

Prof. Dr. Martin Hofmann-Apitius, Coordinator and Colleague

Introducing Work Package 2 – the Data & Knowledge graph analysis

Investigating the Link Between COVID-19 and Neurodegenerative Diseases in COMMUTE
COMMUTE is a collaborative research initiative bringing together experts across Europe to explore how SARS-CoV-2 and the COVID-19 pandemic impact neurodegenerative diseases.

Work Package 2 (WP2): Data Management and Knowledge Graphs play a central role in the project.
WP2 focuses on integrating existing data and knowledge to support two major tasks in COMMUTE:

1. identifying shared causal mechanisms among COVID-19, Alzheimer's, and Parkinson's diseases based on existing knowledge.
2. supporting the interpretation of patterns and predictive features identified in the data-driven approaches taken in the Data Science and AI work package (WP3) and in experimental approaches using cellular assays, organoid systems, and clinical samples in Work Package 4.

Fundamental to the work of WP2 is the development and implementation of data and knowledge management strategies, serving as the central hub for data harmonization, curation, and annotation. Given the diverse datasets in COMMUTE used and generated by the different work packages – including electronic health records (EHRs), organoid data from the wet lab, real-world environmental data, and knowledge graphs in various formats derived from the scientific literature – WP2 plays a critical role in providing shared, formal semantics for COVID-19 and neurodegenerative diseases to ensure interoperability. Another key objective is maintaining and enhancing already existing COVID-19 and neurodegenerative disease knowledge graphs, such as PDmap and NeuroMMSig, up to date. These graphs represent essential knowledge about disease etiology and pathophysiology mechanisms. They are made accessible in a unified, harmonized fashion through a dedicated database. AI/NLP-driven literature analysis, provided by partner Kairntech, is used to expand their insights.

Additionally, WP2 focuses on the improvement of the algorithmic utility of knowledge graphs for various purposes:

- the mapping between heterogeneous data types and e.g., common data models,
- the embedding of knowledge and semantic concepts for AI-based modeling,
- for in-silico simulation.

WP2 partners Fraunhofer and the University of Luxembourg are using WP2 resources to identify shared disease mechanisms. Systematic literature mining and systematic "confrontation" of knowledge graphs and relevant data sets enable in-silico validation of comorbidity hypotheses.

As a major outcome of WP2, the partners collaborating in this work package will design and implement the **COMMUTE Evidence Base**, a platform that provides testable hypotheses in the form of computable cause-and-effect graphs, AI-accessible clinical guidelines, information on candidate biomarkers, and ML/AI models of patient-level data designed for the discovery of co-morbidity mechanisms. The Evidence Base is the central hub to access shareable models, shared knowledge (graphs), and comorbidity hypotheses in computable form. It is the anchor point for fostering collaboration and innovation in the project.

Interoperability of data and knowledge

The COMMUTE WP2 Semantic framework plays a crucial role in creating unified and shared semantics for the description ("annotation") of data and knowledge. Shared semantics is a term coined for methods and tools that ensure that when two people use the same term, they mean the same. Ontologies and terminologies are instrumental in making the meaning of terms used in scientific communication explicit and in providing definitions for each term. They therefore constitute fundamental tools for the harmonization of annotations of objects and processes across heterogeneous information sources. The Gene Ontology (GO) is a prime example of how a widely adopted ontology harmonizes the use of terms in science and standardizes the meaning linked to the term.

WP2 partners have already developed ontologies for Alzheimer's disease, Parkinson's disease, and COVID-19 and are currently actively updating these ontologies to enhance interoperability between data and knowledge by controlling object names and their meaning. This process is pivotal for achieving semantic harmonization across the project's workflows and resources.

An entire semantic framework assembled for COMMUTE not only includes the ontologies developed by WP2 but also incorporates other publicly available ontologies and terminologies that cover all domains of data and knowledge relevant to our project.

A critical component of this framework involves comparing the data sources used in the Data Science and AI work package (WP3) with existing common data models (which are essentially unified variable spaces), already available for Alzheimer's disease (AD) and Parkinson's disease (PD). They have been developed by Fraunhofer during the last two years. Furthermore, we adopt the OMOP data model from OHDSI, which makes us interoperable with a wide spectrum of data science approaches at a global scale. Building on these resources, we aim to create a new common data model (CDM) tailored specifically for the COMMUTE project, harmonizing the feature set in our data with the standardized ontologies and terminologies available. By doing so, we ensure a comprehensive and cohesive approach to managing heterogeneous datasets.

This standardization effort allows us to bridge the gap between real-world data and the diverse knowledge graphs in the project (e.g., the DiseaseMaps provided by partner University of Luxembourg). These activities aim to link patterns in data to knowledge represented in graphs. Ultimately, this is a key step toward bringing together real-world data and contextual knowledge, enabling unprecedented functional interpretation of signals and patterns identified in patient-level data.

COMMUTE Knowledge Graphs – Modelling and Analysis

A major task in WP2 is the development and analysis of disease-specific knowledge graphs (KGs) that model comorbidity (shared pathophysiology) mechanisms between COVID-19 and neurodegenerative diseases (NDDs), with particular emphasis on Alzheimer's (AD) and Parkinson's diseases (PD). This work is a joint effort between the Applied Semantics team at Fraunhofer SCAI and the BioCore team at LCSB, University of Luxembourg, who bring complementary expertise in building, curating, and analyzing large-scale biological data and knowledge. Fraunhofer SCAI contributes manually curated, high-quality KGs encoded in Biological Expression Language (BEL), capturing key molecular interactions and cause-and-effect relationships in COVID-19, AD, PD, and other NDD-related disease pathophysiology. Additionally, Fraunhofer provides a comorbidity KG built from the literature corpus provided by COMMUTE WP4 (clinical expert) team, focusing on publications that are considered highly relevant by our clinical partners. The University of Luxembourg enhances these efforts by developing disease maps for AD, PD, and COVID-19, available in SBML/SBGN formats. These specialized diagrams represent manually curated molecular mechanisms specific to each condition, providing in-depth insights into their underlying biology. Together with the BEL graphs, these disease maps are being integrated into a unified graph database, which will serve as a centralized KG at the core of the COMMUTE Evidence Base. Once the integration is complete, the WP2 team will employ advanced graph analytics and multimodal modeling techniques to find hypotheses on the molecular mechanisms driving COVID-19-induced neurodegeneration, which will subsequently be tested by experimental partners in WP4.

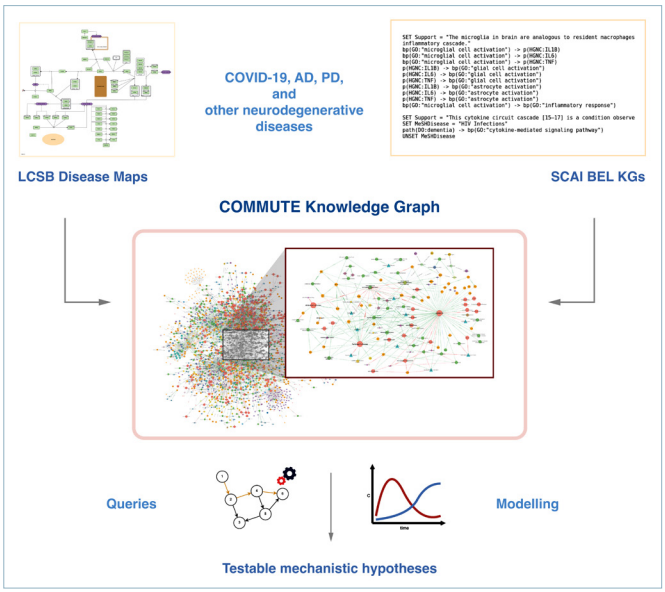

Fig. 1: Illustration of the knowledge-driven approach taken in COMMUTE. Disease maps and BEL KGs (related to COVID-19 and neurodegenerative diseases, focusing on AD and PD) are integrated into a unique NEO4J database, which serves as a centralized knowledge graph for the project. The knowledge graph is then queried and analyzed to find hypotheses on the molecular mechanisms driving COVID-19-induced neurodegeneration, which will be tested by WP4.

COMMUTE Architecture & Automatic Knowledge Extraction

To build a comprehensive knowledge graph representing the comorbidity between COVID-19 and neurodegenerative diseases (NDDs), we use advanced text mining and natural language processing (NLP) techniques to extract key insights from biomedical literature. Our process starts by identifying

relevant research articles and abstracts from databases like PubMed, using specific keywords related to COVID-19 and NDDs. We then apply the Sherpa platform (developed by partner Kairntech), which uses NLP techniques such as named entity recognition (NER) and relation extraction (RE) to identify and link biomedical entities, including diseases, genes, proteins, pathways, cell types, and other important concepts.

It is noteworthy that the representation of published knowledge allows us to test whether published knowledge actually reflects patterns in data. This is important because we face a reproducibility crisis in translational clinical research. The degree of non-reproducible results in biomedicine is very high, and we simply cannot trust published knowledge without a critical assessment of each and every triple in our graphs.

Our advanced NLP methods allow us to capture various relationships, such as co-occurrences, causality, and molecular interactions, and identify shared processes between the two diseases. The identified relationships are represented as BEL (Biological Expression Language) triples, which describe cause-and-effect links between entities (see Fig. 2, part 1).

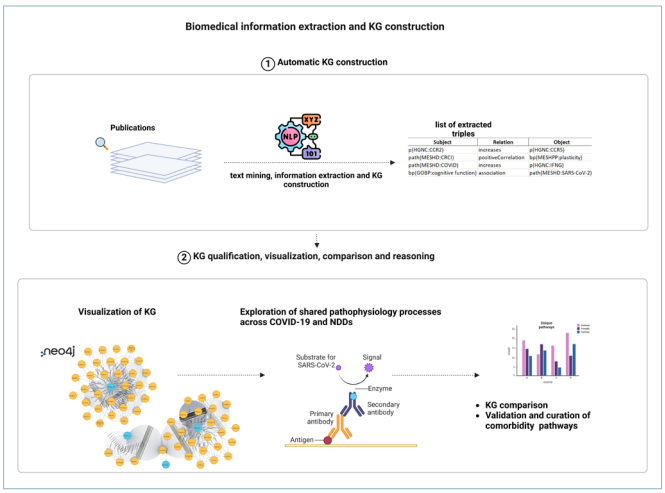

Fig. 2: The process of converting publications into triples using the Sherpa workflow (part 1) and combining the resulting triples into a knowledge graph for identifying comorbidity pathways (part 2).

To enhance the visualization and exploration of the knowledge graph, we use the NEO4J graph platform, enabling interactive examination of the constructed graph by browsing and querying for direct connections and paths in the graph. We further analyze the graph using various graph algorithms, searching for shared mechanisms and phenotypes to identify pathways representing comorbidity between COVID-19 and NDDs (see Fig. 2, part 2). To query the knowledge graph modern chatbot-based approaches, such as GraphRAG (Graph based retrieval augmented generation), can be used.

GraphRAG works like this:

1. **Gathering Information:** Imagine you have a giant library with all relevant publications, experiments, and simulation models in the Evidence Base. GraphRAG collects information from all the various sources.
2. **Creating Connections:** Instead of treating each piece of information separately, it builds a "map" showing how different pieces of information are related to each other, like connecting dots in a web.
3. **Finding Answers:** When you ask a question, GraphRAG looks at this map to find the most relevant pieces of information. It uses a dedicated knowledge graph query language (e.g., Cypher). It doesn't just pull one answer; it considers how different pieces fit together.
4. **Generating a Response:** Finally, it uses the connected information to create a clear and accurate answer, just like a knowledgeable friend would do by combining what they know from different sources.

In short, GraphRAG helps computers find and understand information better by looking at how everything is linked, making it smarter in answering questions (cf. Figure 3).

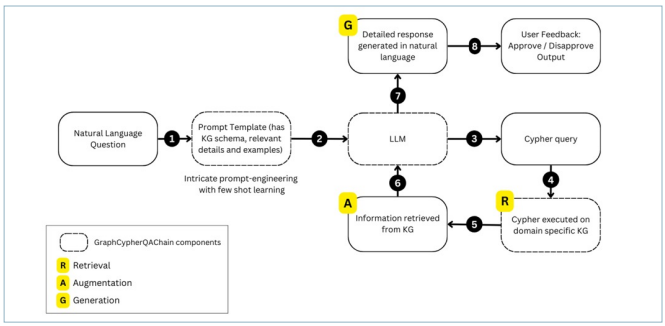

Fig. 3: Cypher Generating Expert System Workflow (GraphRAG): A natural language question is processed through the GraphCypherQAChain. The system converts the question into a Cypher query, retrieves relevant information from the KG, and uses a LLM to generate a detailed natural language response.

Contact

Fraunhofer Institute for Algorithms
and Scientific Computing SCAI
Schloss Birlinghoven 1
53757 Sankt Augustin

Prof. Dr. Martin Hofmann-Apitius
Phone +49 2241 14-4103
martin.hofmann-apitius@
scai.fraunhofer.de

www.commute-project.eu