Towards sustainable pathway identification from biowaste resources to chemicals using knowledge graphs

Adarsh Arun[a, c], **Alexei Lapkin**[a, b, c]

[a] *Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, UK* aa2133@cam.ac.uk

[b] *Chemical Data Intelligence (CDI) Pte Ltd, Robinson Road, #02-00, 068898 Singapore*

[c] *Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd. 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore*

1. Introduction

In recent years there has been great interest in transitioning the chemical industry toward a circular and sustainable model by integrating potentially net-zero biowaste resources as alternatives to traditional fossil feedstocks [1, 2]. These biowaste sources can range from agricultural residues to food waste fractions, and contain highly functionalized, complex feedstocks such as biopolymers that need to be extracted via pretreatment and further processed into value-added chemicals.

At every stage there exists a complex network of options and decisions. Ultimately, the mix of process steps and technologies from raw material to value-added chemicals must remain sustainable [3]. This leads us to a critical research question: What pathways from biowaste to value-added chemicals are the most sustainable, and how can we systematically identify them?

Answering this question requires a comprehensive understanding of data requirements within the biowaste to chemicals domain, encompassing both the biowaste to feedstocks and feedstocks to value-added chemical domains. Additionally, it involves understanding data requirements for sustainability assessment, and, most importantly, ways to contextualize and store all this knowledge to enable automated, early-stage decision-making.

Given the interconnected and complex nature of the domain, this work hypothesizes that Knowledge Graphs (KGs) [4], directed labelled graphs with different entities and relationships, are especially suited to the task. Thus, the aim of this work is to generate a proof-of-concept Knowledge Graph (KG) that can represent the biowaste-to-chemicals domain effectively and allow for sustainable pathway identification.

2. Methodology

Various KGs have been proposed in literature covering scientific domains, notably materials science [5] and biomedical [6]. These KGs often leverage existing structured datasets compiled in prior studies or extract information from research abstracts.

However, for the biowaste-to-chemicals domain, no such efforts exist. No KG has been developed, and there remains a significant data gap especially in the biowaste-to-feedstocks domain, notably specific subdomains that influence the chemical composition of biowaste sources. These include but are not restricted to: citations, locations (and potentials), biowaste sources, taxonomy (e.g. species), compositions, feedstocks, and processes (pretreatment and preprocessing) for feedstock extraction. Most of this information is contained within the unstructured text of the main body of literature papers.

Therefore, in this work, a novel KG generation workflow is proposed for the biowaste-to-feedstocks domain covering the following steps:

1. Knowledge acquisition – Assembly of a literature corpus via keyword searches and keyword co-occurrence analysis.

2. Knowledge ingestion – Schema development covering subdomains of interest, structured information extraction from text using few-shot prompted Large Language Models (LLMs) and automated generation of KG write queries.

3. Knowledge storage – Execution of KG write queries into Memgraph [7], an in-memory graph database based on the Labeled Property Graph (LPG) model.

For the feedstocks-to-chemicals domain, large reaction databases such as Reaxys [8] can be employed and ingested into the final KG via steps 2 and 3.

3. Results

The final generated KG encompasses 2.76 million nodes and 7.36 million relationships, covering all the identified subdomains. Fig. 1 illustrates the underlying schema which covers important entities and relationships. Fig. 2 provides a snapshot of the instantiated KG for the biowaste to feedstocks domain, with notable clusters and entities labeled. Fig. 3 indicates a snapshot of the KG for the feedstock to chemicals domain, with notable value-added chemicals labeled.

The utility of the KG can be demonstrated through a sequence of analytical tasks. By leveraging graph traversal algorithms, such as Depth-First search (DFS), it is possible to conduct efficient knowledge retrieval, missing property assessment, and identification of sustainable pathways. Only the latter case is shown here for illustrative purposes in Fig. 4.

Towards sustainable pathway identification from biowaste resources to chemicals using knowledge graphs

Adarsh Arun[a, c], **Alexei Lapkin**[a, b, c]

[a] *Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, UK* aa2133@cam.ac.uk

[b] *Chemical Data Intelligence (CDI) Pte Ltd, Robinson Road, #02-00, 068898 Singapore*

[c] *Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd. 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore*

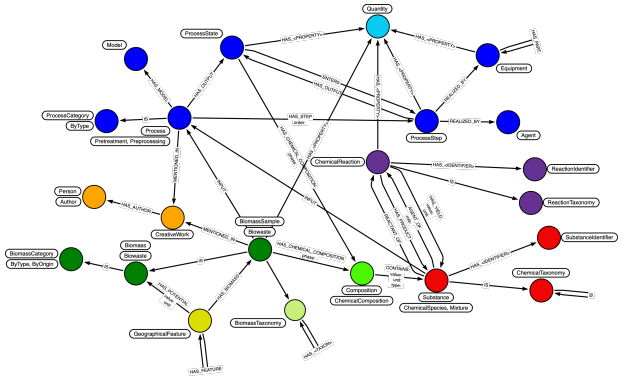

Fig. 1: High-level illustration of the KG schema covering relevant subdomains for the biowaste-to-chemicals domain.

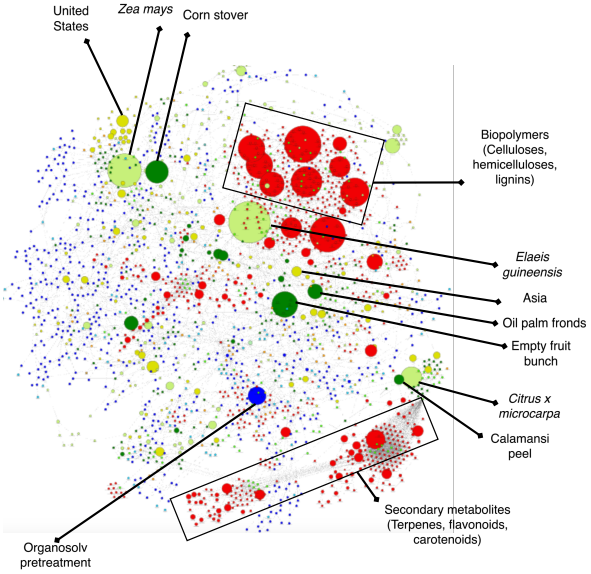

Fig. 2: Snapshot of the instantiated KG for the biowaste-to-feedstocks domain, with relevant entities labeled.

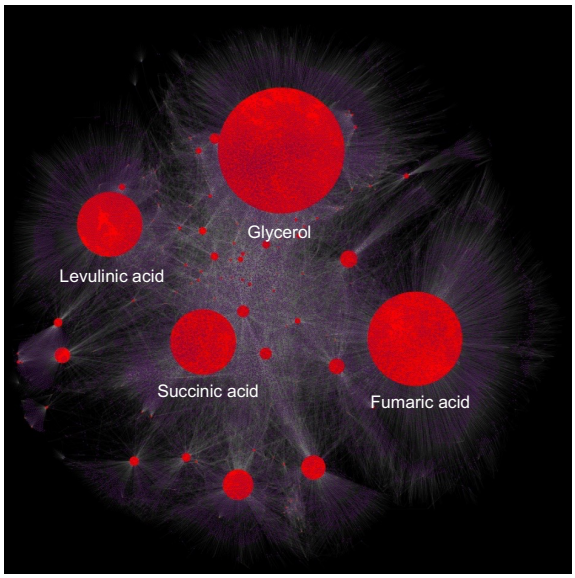

Fig. 3: Snapshot of the instantiated KG for the feedstocks-to-chemicals domain with relevant value-added chemicals labeled. The size of each node corresponds to how well-connected it is to the rest of the KG via chemical reactions (in-degree and out-degree).

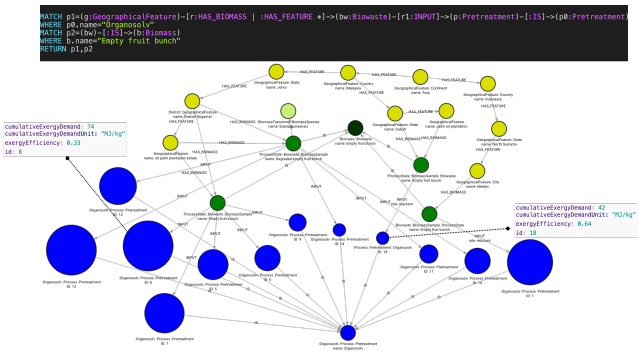

Fig. 4: Simplified cypher query DFS and subgraph of the most sustainable organic solvent pretreatment processes (blue nodes) for conversion of Empty Fruit Bunch (EFB) into biopolymers lignin, cellulose and hemicellulose. The size of the nodes correspond to cumulative exergy demand (smaller is better).

Fig. 4 provides summarized results of a case study starting from a common agricultural waste source available in the Southeast Asia region, Empty Fruit Bunch (EFB). Using a Depth-First Search (DFS) query, relevant organic solvent pretreatment processes acting on EFB can be identified and retrieved from the KG (blue nodes). The size of each node corresponds to the cumulative exergy demand of the process, which was evaluated through automated process simulations based on process data stored in the KG. Cumulative exergy demand is one possible sustainability metric that reflects the total energy demand of a process accounting for any thermodynamic inefficiencies. In Fig. 4, the identified most sustainable pretreatment process (smallest node) has an exergy demand of 42 MJ/kg input biowaste at an exergy efficiency of 64 %.

4. Future Work

The developed Knowledge Graph (KG) needs to be scaled up further and additional case studies need to be conducted using a broader variety of sustainability metrics. In the long run, it is hoped that the KG will serve as a key data resource in the community that facilitates early-stage sustainable pathway identification and development for a circular chemical economy integrating bioresources, while also supporting machine learning and analytical workflows.

Towards sustainable pathway identification from biowaste resources to chemicals using knowledge graphs

Adarsh Arun[a, c], **Alexei Lapkin**[a, b, c]

[a] *Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, UK* aa2133@cam.ac.uk

[b] *Chemical Data Intelligence (CDI) Pte Ltd, Robinson Road, #02-00, 068898 Singapore*

[c] *Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd. 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore*

References

[1] Guo, Z., Yan, N. & Lapkin, A. A. Towards circular economy: integration of bio-waste into chemical supply chain. *Curr Opin Chem Eng* **26**, 148–156 (2019).

[2] Tuck, C. O., Pérez, E., Horváth, I. T., Sheldon, R. A. & Poliakoff, M. Valorization of biomass: deriving more value from Waste. *Science (1979)* **337**, 695–699 (2012).

[3] Clark, J. H. *et al.* Green chemistry and the biorefinery: A partnership for a sustainable future. *Green Chemistry* **8**, 853–860 (2006).

[4] Singhal, A. Introducing the Knowledge Graph: things, not strings. *Google* (2012).

[5] Venugopal, V. & Olivetti, E. MatKG: An autonomously generated knowledge graph in Material Science. *Sci Data* **11**, 217 (2024).

[6] Santos, A. *et al.* A knowledge graph to interpret clinical proteomics data. *Nat Biotechnol* **40**, 692–702 (2022).

[7] Memgraph. Memgraph. https://memgraph.com/ (2024).

[8] Reaxys - An expert-curated chemistry database. https://www.elsevier.com/solutions/reaxys.