# Challenges and opportunities in Life Cycle Inventory data generation for the production of chemicals

**Bianca Köck, Anton Friedl, Sebastián Serna Loaiza, Walter Wukovits and Bettina Mihalyi-Schneider**

*Institute of Chemical Engineering, TU Wien, Getreidemarkt 9, 1060 Vienna, Austria;*
*Correspondence: bianca.koeck@tuwien.ac.at; Tel.: +43-664-8635967*

***Introduction and study objectives***

Globally, the production and use of chemicals are a significant contributor to environmental degradation and human health impacts. Therefore, the shift to green chemicals and sustainable processes have become the most important goal for researchers, regulators and industry. Life cycle assessment (LCA) is considered as a central method to estimate the environmental impact of products, processes, and services over their entire life cycle, from raw material extraction to disposal or recycling. (cradle to grave or cradle to cradle). Also in the new safe-and-sustainable-by-design (SSbD) concept developed by the joint research center (JRC) that takes a systems approach by integrating safety, sustainability, and functionality throughout a product's the life cycle.

However, the availability of reliable and complete data sets for life cycle inventory is a challenge, especially when the technology/chemical under investigation is still at an early stage of development. There are several options to deal with data gaps for the evaluation of chemical manufacturing processes, such as estimation, omitting modules, calculating with similarities, but these lead to a high variance of the results, which in turn can cause poor decisions. Furthermore, the complexity and diversity of chemical structures, properties, and application make it difficult to develop a uniform procedure for data collection and assessment methodologies.

Even if the principles of green chemistry were taken into account at an early stage of development, the entire manufacturing process and its integration into the overall process chain must still be evaluated. Similarly, for new substances and materials, comprehensive human and environmental toxicity results are often not yet available for sufficient environmental assessment. In order to identify new solutions here, the aim of this study was to collect methods for generating reliable data for LCA of chemicals and processes in the literature and to review and evaluate their applicability, focusing on those that are not yet included in LCA databases.

***Methodology***

To address the aim of this study, a systematic literature review was performed. A comprehensive search was conducted in relevant scientific databases such as Scopus and Web of Science to identify relevant publications. The investigation was limited to papers published from 2008 onwards. The search terms were selected to identify papers that focused on the reduction of data gaps in chemical process inputs and outputs. Specific criteria were established to select publications that

address the objective of this study by evaluating or supplementing empirical data, or where simulation-based approaches are used to close the data gaps in LCA studies.

***Results and discussion***

There are various methods to generate data for life cycle inventory (LCI), depending on the specific objectives of the study. These methods include knowledge engineering and data mining, process simulation, predictive life cycle assessment using machine learning and multivariate statistics, and computer-aided molecular design. Knowledge engineering and data mining involve extracting data and knowledge from existing databases, literature, and expert knowledge to identify relevant data sources and fill data gaps. In process simulation, the production process or process step is modeled using special software, and the detailed information calculated as a result provides the corresponding input and output data for the respective parameter settings. By varying the parameters, the operating areas with the lowest environmental impact can be identified. Predictive LCA uses multivariate statistical techniques and machine learning to assess the environmental impact of chemicals based on their properties and characteristics, one example is a structure-based model. Computer Aided Molecular Design involves using computational methods to design and identify chemicals with desired properties and characteristics, potentially reducing environmental impacts.

These methods are not mutually exclusive, and can be used in combination to improve the accuracy and reliability of LCA data generation. The methods can be applied at different Technology Readiness Levels (TRLs), as, for example Computer Aided Molecular Design and Predictive LCA, despite their greater uncertainties, provide useful information when limited process information or general data on the chemical is available at an early stage of development. However, their application can also be meaningful at higher TRLs, as poorly characterized chemicals may still be present in products on the market and not considering these chemicals could lead to an incomplete or distorted assessment of their environmental impacts.

The use of the discussed techniques can introduce uncertainties in the data generation process. For instance, predictive LCA and Computer Aided Molecular Design may have high uncertainties due to the lack of reliable and representative input data, while knowledge engineering and data mining methods may also introduce uncertainties due to the quality and representativeness of the data sources.

***Conclusions***

Various methodologies, such as knowledge engineering and data mining, process simulation, predictive LCA using machine learning and multivariate statistics, and Computer Aided Molecular Design were evaluated in terms of the uncertainties associated with these methods. The outcomes of this review will contribute to the development of best practices guidelines in order to generate reliable and comprehensive data for LCA, supporting informed decision-making towards sustainable production of chemicals and materials. Moreover, open science practices, such as sharing programming codes and data sources, could enhance the transparency and reproducibility of the results and facilitate the identification and resolution of potential sources of uncertainties.