AI-powered Framework to Predict Environmental Impacts of Organic Chemicals via Retrosynthesis

Shaohan Chen, Tim Langhorst, Benedikt Winter, Johannes Schilling, André Bardow\*

ETH Zurich, 8092 Zurich, Switzerland

abardow@ethz.ch

Abstract

The transition toward a sustainable chemical industry relies on the ability to quantify the environmental impacts of chemical production. Life cycle assessment (LCA) is a holistic method that quantifies the environmental impacts of chemical productions across multiple impact categories. However, conducting a thorough LCA study is challenging in the early process design stages due to limited LCA data availability. This work estimates LCA data of organic chemicals by designing a pathway-resolved framework using machine-learning-based retrosynthesis. Our method automatically predicts the life cycle inventories (LCI) and the corresponding environmental impacts solely using SMILES codes of target chemicals as the input. We verify this framework with a benchmark dataset of 136 organic chemicals, including industrially validated LCIs. The results show that our framework can accurately predict LCIs and the environmental impact of all impact categories. Our framework thus allows for filling data gaps in LCA databases for early-stage process design and accelerates the transition toward a sustainable chemical industry.

**Keywords**: Environmental impacts, predictive LCA, retrosynthesis, machine learning.

* 1. Introduction

Integrating environmental impacts at the early stages of process design is essential for fostering a sustainable chemical industry. Life cycle assessment (LCA) is a standardized method that comprehensively quantifies potential environmental impacts across various impact categories, such as climate change and toxicity (ISO 14040). However, current LCA databases only cover a limited number of chemicals, necessitating tedious LCA work to assess any other chemicals. These LCAs require extended life cycle inventory (LCI) data that quantify the material and energy requirements needed to produce the target chemical. This data is challenging to acquire during the early stages of process design (Chebaeva et al., 2021).

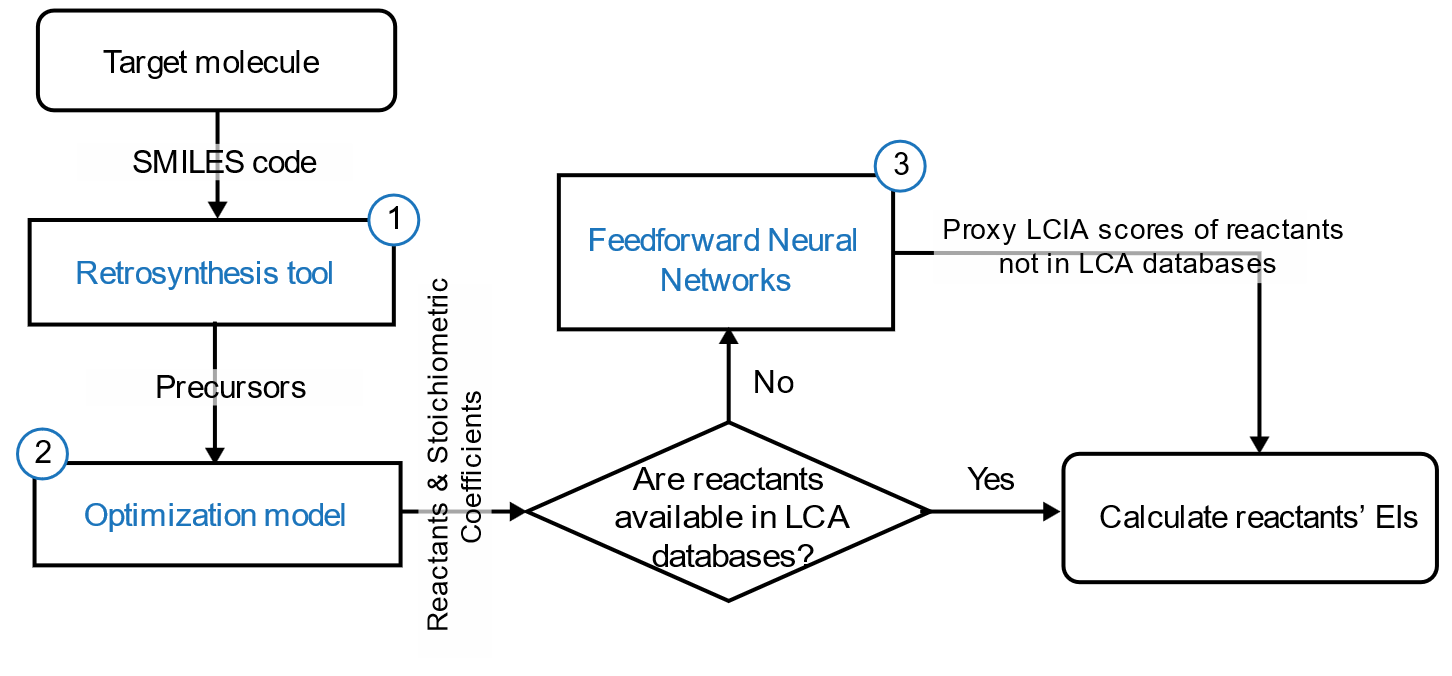
To mitigate the lack of LCA data, two types of methods have been developed to predict the environmental impacts of chemical production: 1) machine-learning-based (ML-based) methods and 2) stoichiometry-based methods. ML-based methods, for example, Kleinekorte et al. (2023), usually use molecular descriptors as the input. The resulting accuracy is limited because of data scarcity for training, and the input ignores the actual conversion pathway. Stoichiometry-based methods, for example, Langhorst et al. (2023), integrate chemical synthesis details from literature but are, therefore, also time-consuming and labor-intensive. In particular, the impacts of the reactants need to be known so that the data problem is only shifted to them.

To bridge the gap between the two types of pioneering methods, we design a pathway-resolved machine-learning framework to automatically estimate LCIs and the corresponding environmental impacts of organic chemical production. For this purpose, we integrate an ML-based retrosynthesis tool (Schwaller et al., 2020).

* 1. Method

Our pathway-resolved ML-based framework predicts the LCI data required to produce a specific chemical and its reactants. The LCI data is used subsequently to calculate the environmental impacts. The framework consists of three modules to ensure reliable environmental impact predictions for chemical production (see Figure 1):

1. First, an ML-based retrosynthesis tool (Schwaller et al., 2020) provides possible synthesis pathways and potential precursors to produce the target chemicals.
2. Second, this synthesis pathway data is analyzed by an optimization model to determine the stoichiometric coefficients of the reactants and possible byproducts. Thereby, we know the stoichiometries along the full predicted synthesis pathway.
3. Lastly, the reaction equations along the predicted synthesis pathway are used for stoichiometric estimation methods to predict the LCIs of producing the target chemical. To calculate the environmental impacts, the LCIA scores of reactants are taken from available LCA databases. For reactants not available in the considered LCA databases, we use feedforward neural networks with a similar structure as the one proposed by Wernet et al. (2008) to predict proxy LCIA scores of the reactants.



**Figure 1.** Flow chart of the proposed pathway-resolved ML-based framework to predict reactants' environmental impacts (EIs).

* 1. Results

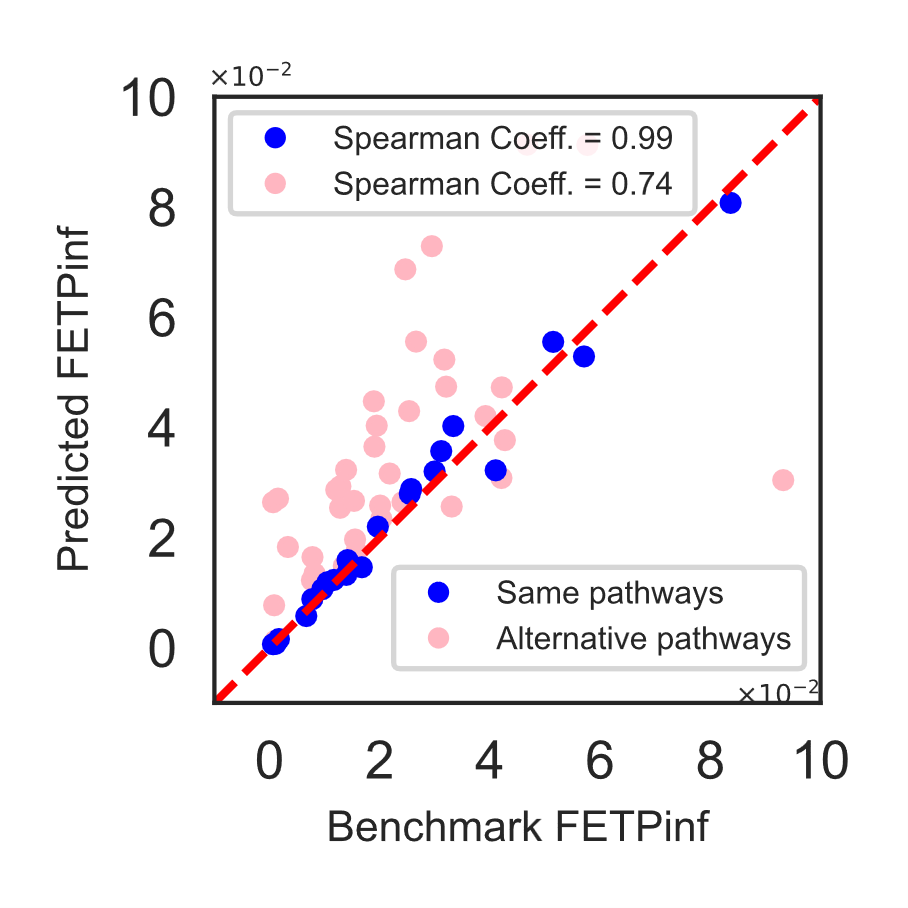
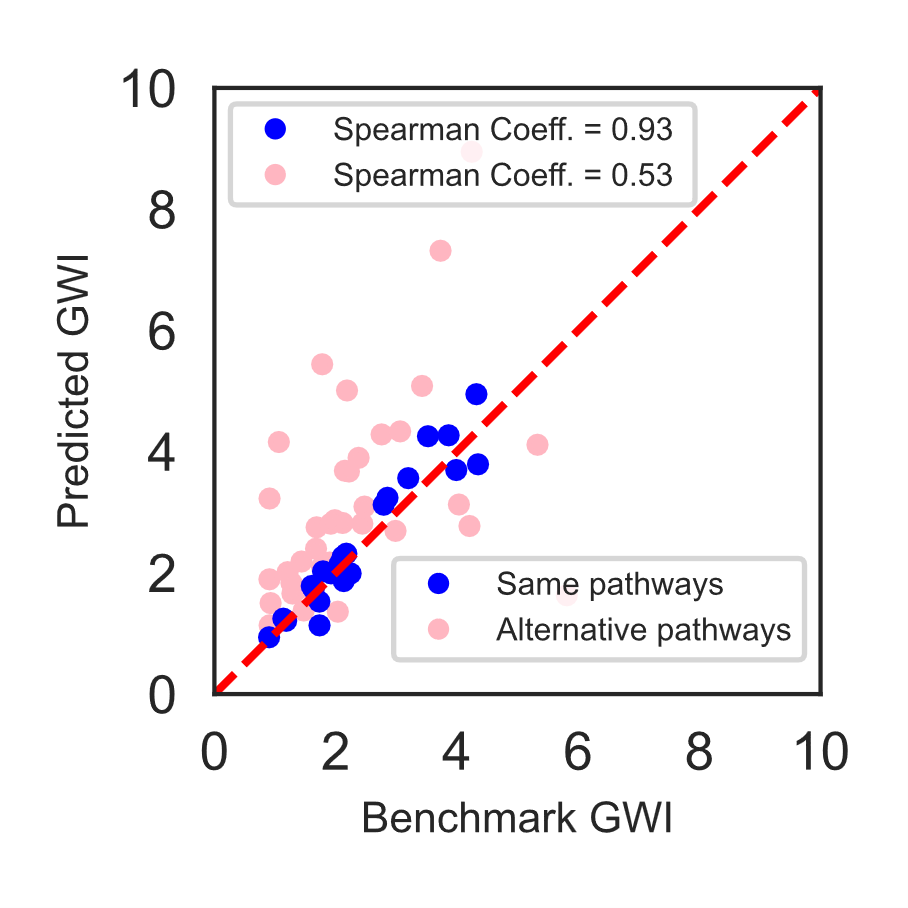
Our pathway-resolved ML-based framework is verified by comparing the predicted LCIs and the estimated environmental impacts with a benchmark dataset of 136 organic chemicals (Langhorst et al., 2023). This dataset includes industrially validated LCIs for each chemical. As an underlying LCA database, we use Ecoinvent version 3.5 (Wernet et al., 2016) and the ReCiPe (H) V1.13 method (Huijbregts et al., 2017). Since the retrosynthesis tool mainly identifies the reactions along the synthesis pathways, we focus our analysis on the predicted reactants' environmental impacts.

In our dataset, our framework identifies 60 chemicals with predicted reactants sourced entirely from the Ecoinvent database. For the remaining 76 chemicals, for which the reactants are not fully listed in the Ecoinvent database, the framework calculates proxy LCIA scores via neural networks.

For the 60 chemicals with reactants in Ecoinvent, Figure 2 exemplifies the predicted reactants' environmental impacts on global warming impact (GWI) and freshwater ecotoxicity potential (FETPinf). This analysis distinguishes between two groups of chemicals: those for which our framework predicts the same synthesis pathways as the one in the benchmark dataset and those for which an alternative pathway is predicted.

Our framework performs remarkably regarding Spearman’s rank correlation coefficient when the predicted pathways are the same as in the benchmark, 0.93 for the GWI and 0.99 for the FETPinf, respectively. The corresponding mean absolute percentage error (MAPE) is 0.11 for the GWI and 0.12 for the FETPinf, respectively. The MAPE is between 0.11 and 0.16 for all 18 impact categories.

The validation is not possible in cases where an alternative synthesis pathway is predicted since this difference does not imply an incorrect prediction. Manual analysis of the predicted pathways using SciFinder shows that 89% of them exactly match the ones in the existing literature. The presented framework is transparent by providing the details on the assumed pathways.

  
**Figure 2.** Predicted reactants' environmental impacts compared to the benchmark reactants' environmental impacts. Left: global warming impact (GWI), right: freshwater ecotoxicity potential (FETPinf). Blue dots indicate chemicals for which our framework predicts the same synthesis pathways as considered in the benchmark dataset. Pink (light) dots represent chemicals for which our framework predicts synthesis pathways different from those in the benchmarks.

These results demonstrate that our framework can reliably predict the LCIs for organic chemical production, accurately estimate the environmental impacts when the predicted synthesis pathways match the benchmark, and predict reasonable alternative pathways in a transparent manner for the LCA practitioner to check and verify.

* 1. Conclusions

This work proposed a pathway-resolved framework using ML-based retrosynthesis to predict the environmental impacts of organic chemical production by using the SMILES code of the target chemical as the only input. Our framework provides transparent information on synthesis pathways, stoichiometric coefficients of the predicted chemical reactions, and the estimated byproducts.

The results show that our framework can accurately predict LCIs and the environmental impact of all impact categories. Due to the transparent modelling of the synthesis pathways and LCIs of chemical productions, the framework can be adapted for specific LCA databases, LCIA methods, and system models. Our framework thus allows for filling data gaps in LCA databases for early-stage process design and accelerates the transition toward a sustainable chemical industry.

**Acknowledgments**

This work is part of the Usorb-DAC Project supported by a grant from The Grantham Foundation for the Protection of the Environment to RMI’s climate tech accelerator program, Third Derivative. It is also partly funded by the LIFE Klimastiftung and the ETH Zurich Foundation.

References

[1] ISO 14040, *Environmental management - Life cycle assessment-Principles and framework*; International Organization for Standardization, 2006.

[2] N. Chebaeva, M. Lettner, J. Wenger, J. P. Schöggl, F. Hesser, D. Holzer, and T. Stern (2021). Dealing with the eco-design paradox in research and development projects: The concept of sustainability assessment levels. *Journal of Cleaner Production*, *281*, 125232.

[3] J. Kleinekorte, J. Kleppich, L. Fleitmann, V. Beckert, L. Blodau, and A. Bardow (2023). APPROPRIATE Life Cycle Assessment: A PROcess-Specific, PRedictive Impact AssessmenT Method for Emerging Chemical Processes. *ACS Sustainable Chemistry & Engineering*.

[4] T. Langhorst, B. Winter, D. Roskosch, and A. Bardow (2023). Stoichiometry-Based Estimation of Climate Impacts of Emerging Chemical Processes: Method Benchmarking and Recommendations. *ACS Sustainable Chemistry & Engineering*, *11*(17), 6600-6609.

[5] P. Schwaller, R. Petraglia, V. Zullo, V.H. Nair, R.A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, and T. Laino (2020). Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, *11*(12), 3316-3325.

[6] G. Wernet, S. Hellweg, U. Fischer, S. Papadokonstantakis, K. Hungerbühler (2008).  Molecular-structure-based models of chemical inventories using neural networks. *Environmental science & technology, 42(17),* 6717-6722.

[7] G. Wernet, C. Bauer, B. Steubing, J. Reinhard, E. Moreno-Ruiz, and B. Weidema (2016). The ecoinvent database version 3 (part I): overview and methodology. *The International Journal of Life Cycle Assessment*, *21*, 1218-1230.

[8] M. A. J. Huijbregts, Z. J. N. Steinmann, P. M. F. Elshout, G. Stam, F. Verones, M. D. Vieira, A. Hollander, M. Zijp, and R. van Zelm (2017). ReCiPe 2016 v1.1: a harmonized life cycle impact assessment method at midpoint and endpoint level report I: characterization [RIVM Report 2016 -0104a]. *National Institute for Public Health and the Environment RIVM*, 2016-0104.

[9] SciFinder. Chemical Abstracts Service, n.d. <https://scifinder.cas.org> (accessed November 8, 2023).